Noname manuscript No.
(will be inserted by the editor)

Human migration: the Big Data perspective

Alina Sîrbu · Gennady Andrienko · Natalia Andrienko · Chiara Boldrini · Marco Conti · Fosca Giannotti · Riccardo Guidotti · Simone Bertoli · Jisu Kim · Cristina Ioana Muntean · Luca Pappalardo · Andrea Passarella · Dino Pedreschi · Laura Pollacci · Francesca Pratesi · Rajesh Sharma

Received: date / Accepted: date

Abstract How can Big Data help to understand the migration phenomenon? In this paper we try to answer this question through an analysis of various phases of migration, comparing traditional and novel data sources and models at each phase. We concentrate on three phases of migration, at each phase describing the state of the art and recent developments and ideas. The first phase includes *the journey*, and we study migration flows and stocks, providing examples where big data can have an impact. The second phase discusses *the stay*, i.e. migrant integration in the destination country. We explore various datasets and models that can be used to quantify and understand migrant integration, with the final aim of providing the basis for the construction of a novel multi-level integration index. The last phase is related to the effects of migration on the source countries and *the return* of migrants.

Keywords human migration · big data · migration flows · migration stocks · integration · return of migrants

A. Sîrbu, D. Pedreschi, L. Pollacci, F. Pratesi
Department of Computer Science, University of Pisa, Pisa, Italy.
E-mail: alina.sirbu@unipi.it, dino.pedreschi@unipi.it, laura.pollacci@di.unipi.it, francesca.pratesi@isti.cnr.it

G. Andrienko, N. Andrienko
Fraunhofer Institute IAIS, Sankt Augustin, Germany & City, University of London, UK.
E-mail: gennady.andrienko@iais.fraunhofer.de, natalia.andrienko@iais.fraunhofer.de

C. Boldrini , M. Conti, A. Passarella
IIT - CNR, Pisa, Italy.
E-mail: chiara.boldrini@iit.cnr.it, marco.conti@iit.cnr.it, andrea.passarella@iit.cnr.it

R. Guidotti, F. Giannotti, C.I. Muntean, L. Pappalardo
ISTI-CNR, Pisa, Italy. E-mail: fosca.giannotti@isti.cnr.it, riccardo.guidotti@isti.cnr.it, cristina.muntean@isti.cnr.it, luca.pappalardo@isti.cnr.it

S. Bertoli
Université Clermont Auvergne, CNRS, IRD, CERDI, Clermont-Ferrand, France.
E-mail: simone.bertoli@uca.fr

J. Kim
Scuola Normale Superiore, Pisa, Italy. E-mail: jisu.kim@sns.it

R. Sharma
University of Tartu, Estonia. E-mail: rajesh.sharma@ut.ee

1 Introduction

The phenomenon of human migration has been a constant of human history, from the earliest ages until now. As such, the study of migration spans various research fields, including anthropology, sociology, economics, statistics and more recently physics and computer science. We are at a moment where various types of data not typically used to study migration are becoming increasingly available. These include so called social big data: digital traces of humans generated by using mobile phones, online services, online social networks (OSNs), devices within the internet of things. At the same time, new technologies are able to extract valuable information from these large datasets. Both traditional and novel models and data are currently being employed to understand different questions on migration, including monitoring migration flows, the economic and cultural effects on the migrants and also on the source and destination communities. In this paper we provide a survey of existing approaches, both traditional and data-rich andwe propose new methods and datasets that could contribute significantly to the study of human migration. We concentrate on three different phases of migration: the journey - analysing migration

flows and stocks; the stay - studying migrant integration and changes in the communities involved; the return - the study of migrants returning to the origin country.

The journey. At the moment, information about migration flows and stocks comes from official statistics obtained either from national censuses or from the population registries. Given that migration intrinsically involves various nations, data is often inconsistent across databases, and offers poor time resolution. With the availability of social Big Data, we believe it should be possible to estimate flows and stocks from available data in real time, by building models that map observed measures extracted from these unconventional data sources to official data, i.e. nowcasting stocks and flows. We also look at migration phenomena within smaller communities, such as scientific migration, where even prediction of migration events can be possible. An important step in understanding migration flows is suitable visualisation, which we also explore.

The stay. Migration might generate cultural changes with both long- and short-term effects on the local and incoming population. Migrant integration is generally measured through indicators related to the labour market, economic status or social ties. Again, these statistics are available with low resolution and not for all countries. A new direction is that of observing integration and perception on migration through Big Data. For instance, OSN sentiment analysis specific to immigration topics can allow us to evaluate perception of immigration. Analysis of retail data can enable us to understand if immigrants are integrated economically but also if they change their habits during their stay. Scientific data can help us understand how migration benefits both the host countries and the migrants themselves. Through these data we can derive novel integration indices that take into account the traces of human activity observed.

The return. Besides effects on the receiving communities, the source communities may also see effects of migration. In fact, migrants can maintain a strong attachment to their home countries and eventually return there. This can bring multiple benefits: economic growth, new skills, entrepreneurship, better healthcare, different participation in governance issues and many others. We discuss various approaches to analysing these cases based on existing data.

Both traditional and new methods to analyse migration depend highly on the availability of data. Hence, infrastructures that can catalogue the various datasets and make them available to the community, ensuring privacy and ethical use, are very useful. At the same time, with new methods being developed, means of facilitating their use by the research community are necessary. An example of framework that aims to achieve these requirements is the SoBigData infrastructure [79] (`www.sobigdata.eu`). This includes a catalogue of methods, datasets and training material, grouped in so-called *exploratories*. Virtual research environments allow users to use some of the data and methods directly in the SoBigData engine. The exploratory on migration studies includes many of the methods and datasets presented below.

The rest of the paper is organised as follows. The study of migration flows and stocks is discussed in Section 2. This compares traditional data (Section 2.1) with social big data (Section 2.2) including scientific migration (Section 2.2.1), providing also a review of tools for visualisation of migration data (Section 2.3). Section 3 concentrates on migrant integration and perception of migration. We start by looking at approaches based on traditional data sources (Section 3.1) and move on to social big data including retail data (Section 3.2.1), mobile data (Section 3.2.2), language and sentiment in OSNs (Sections 3.2.3 and 3.2.4), ego-networks (Section 3.2.5). The return of migrants is discussed in Section 4, while Section 5 concludes the paper with a summary and a discussion on ethical issues.

2 The journey: migration flows and stocks

In this section we discuss various means of analysing migration flows and stocks. We start with traditional approaches and data types, and then move to new datasets that can be employed for the task, underlining advantages and disadvantages of each approach.

2.1 Traditional data sources and challenges

Tracking international migrants' flows and stocks is an important task but also challenging. At the moment, many researchers and policy makers rely on traditional data sources to study the journey of migrants. Such data sources come from either official statistics or from administrative data. Studying the journey of migrants with these traditional data sources, however, come with various limitations as migration intrinsically involves various nations. For instance, the data are often inconsistent across databases as different countries employ various definitions of a *migrant*. A lot of efforts have been made so far from both researchers and international organizations to improve quality and harmonize traditional data sources [51,149,172]. Interna-

tional organizations such as the United Nations provide also guidelines and suggestions[1] which countries should employ when dealing with migration statistics. In this section, each type of data source is described in detail and evaluated.

Census data and surveys are official statistics collected by institutions. They provide socio-demographic information of the population, including immigrants. However, the two types of data have different focus. The census data are collected once in five years or once in ten years, depending on the country. For example, the most recent data available in the United States is the 2010 census data, while in Europe the last census was performed in 2011. By the recommendation given by the United Nations[2], countries should collect the data every year that ends with zero in order to establish a consistency across different migration datasets. But as the process of collecting data is expensive and time consuming, some developing countries do not collect the data as it is recommended, creating inconsistency across different countries' databases. The high cost is due to the fact that the majority of countries carry out door-to-door or phone interviews to a randomly selected sample of population to collect the data. For instance, the Chinese population is almost 1.4 billion[3], so about 6 million enumerators are needed to conduct all the interviews. On the other hand, most European countries retrieve the data from administrative registries which makes the procedure faster[150, 63].

In the census data, migration related information collected is the following; citizenship, country of birth, last place of residence as well as length of stay. However, depending on the countries' characteristics of immigrants and the immigration system[4], they do not use the same information to count the number of immigrants. In Europe for example, the focus is also given on different migrant groups depending on whether they are from the European Member States or third country[5]. On the other hand, the United States counts everyone born outside of their territory as immigrants. Yet, the recommendation of the United Nations defines an international migrant as 'a person who moves to a country other than that of his or her usual residence for a period of at least a year'. The difference in the definition of immigrants creates incomparability across different migration data. Furthermore, information about returning migrants is not well captured through the census data. This is due to the fact that returning migrants are not obliged to declare their departure. In the leaving country's data, they would simply exit from the data, meaning that information about these migrants is difficult to track.

Census data is usually published in aggregated form by the authorities that organised the census. Typically, immigration rates are made available at country or at most regional level. For instance, historical immigration data can be found on the websites of Eurostat [64], the WorldBank [166], OECD [165] and other local authorities and research institutions [98, 96, 97, 68, 62]. However, in certain situations having data with higher spatial resolution can be useful. Recently, the Joint Research Centre of the European Union published a data challenge[6] where they make available for research high resolution immigration data from the 2011 census, for selected European countries. However, similar data is more difficult to obtain for other regions.

Surveys also collect information about the flows and stocks of immigrants and they are retrieved more often than the census data. Unlike the census data, they are generally conducted to collect information on households, labour market or community, depending on their main purpose. As a result, there are very few questions related to migration. For instance, in the employment survey in France, there are two questions which are about country of origin and date of arrival. With these two details, it is difficult to infer the immigrants' journey since a clear definition of immigrants cannot be established. As a consequence, it has low accuracy level in capturing immigrants' flows and stocks and real-time observation cannot be done. In addition, information retrieved from surveys refers to a small subset of the entire population.

Administrative data are retrieved from registries. It can be from health insurance, residence permits, labour permits or border statistics, which gather also information about immigrants. Registry data can provide more detail and are less costly than official statistics as the information is intrinsically and directly given by the individuals. For instance, data collected from the residence permits include details about intention and length of stay. They also require specific details on place of origin and address in the country of stay. The same applies to labour permit data. Nevertheless, in Europe where the freedom of movement and work is established, it is difficult to know flows and stocks of EU immigrants using these administrative data unless all the individuals

[1] Recommendations on Statistics of International Migration, Revision1(p.113). United Nations, 1998.

[2] Idem.

[3] "The Statistics Portal." Statista. Retrieved from www.statista.com

[4] "Sources and comparability of migration statistics". OECD, Retrieved from https://www.oecd.org/migration/mig/43180015.pdf.

[5] Those born outside of Europe

[6] Data Challenge on Integration of Migrants in Cities (D4I), `https://bluehub.jrc.ec.europa.eu/datachallenge/`

are registered. An alternative is to use health insurance data. With these, it is possible to infer the stocks more accurately, provided the immigrants register for health insurance. In addition, registries can also collect information about asylum seekers[7] and refugees[8]. However, this information is not always present in all migration data. In some countries like France, Italy, United Kingdom and so on, asylum seekers residing at least 12 months in a country are included in the data. In other countries like Belgium, Sweden and Finland, they are excluded [63]. Again, an application of different definitions makes it difficult to compare data across different countries. When studying the journey with administrative data, caution should be used when inferring the immigrants' journey as it is difficult to identify the true movements of immigrants.

The use of traditional data in studying the journey of immigrants is definitely useful. These can be used for building models of migration [145] and understanding the determinants of migration. But for the reasons discussed above, several drawbacks have to be taken into account. To improve data quality, institutions provide estimates to impute the gaps between years, or use *the double-entry matrix*[9] firstly introduced by UNECE[10] to establish comparability across different nations' data (see for instance [143,144,51]). Nevertheless despite of the efforts, the data still appear inconsistent and unreliable. With the availability of social big data sources, researchers hope not only to overcome the limitations of traditional data, but also to be able to conduct real-time analyses at a higher accuracy level.

2.2 Alternative data sources - is nowcasting possible?

In recent studies, the use of social big data in the study of immigrants' journey is increasing. A variety of data types can fall under this category. They can be data from social media, internet services, mobile phones, supermarket transaction data and more. These datasets contain detailed information about their users. Furthermore, they cover larger sets of population than some of the traditional data sources which are limited in terms of sample size. Yet, the literature points out that the data may be biased because of users' characteristics in the sample. For instance with Twitter data, it is known that the majority of the users are young and that it cannot represent the whole population. Nevertheless, various of studies state that the observed estimates of immigrants' flows and stocks extracted from these unconventional data sources can still improve the understandings of migration patterns (see for instance, [184, 88,127]).

Big Data allows researchers to study immigrants' movements in real-time. Twitter data for instance, provide geo-located timestamped messages. Geo-located messages are often the key variable in estimating the flows and stocks but not the only one. In the work of [184], the authors infer migration patterns from Twitter data by looking at where the tweets were posted. Other studies like [127] assume origins of immigrants from language used in tweets, whether the local language was used or not. These studies conclude that Twitter data allow researchers to localize the flows and stocks of immigrants and to observe recent trends even before the official statistics are published. The results of these studies are validated by matching the big data results to official data.

In one of our recent works, we have also analysed geolocalised Twitter data, with the aim of quantifying diversity in communities, by computing a superdiversity index [140] (see also Section 3). This index correlates very well with migration stocks, hence we believe it can become an important feature in a now-casting model. A different line of work we are pursuing is that of estimating user nationality from Twitter data. As seen above, language can be important in understanding nationality, however we believe that this can be refined by employing also the connections among users. The model can be validated with data collected through monitoring frameworks such as that presented in [21]. Once users are assigned a nationality, we can use these for a now-casting model of migration stocks. Additionally, we can define communities on Twitted based on nationality, and study the flow of ideas among communities, and the role of migrants in the spreading of information. Furthermore, these data could enable analysis of ego-networks of migrants (see Section 3.2.5 bellow).

Skype Ego networks data can also be used to explain international migration patterns [102]. In this case, the IP addresses that appear when users login to their account can be used to infer the place of residence. More precisely, they look at how often the users login to their IP address, which allows them to label the location as the users' place of residence. The users' place of residence then can be used to observe whether migration took place or not.

[7] Asylum seekers are individuals who seek to obtain refugee status

[8] Individuals with subsidiary protection are also referred as refugees

[9] It compares statistics of both immigrants and emigrants between a set of country. The degree of underestimation of number of emigrants can be inferred by doing so.

[10] United Nations Economic Commission for Europe

Big data can also be used to study movements of individuals in the time of crisis. For instance, [30] propose to use mobile phone data to trace individuals' movements in the occurrence of earthquake in Haiti. With these data, the authors are able to trace users as the phone towers provide information about their locations. They conclude that Big Data can be used to observe movements in real-time, which cannot be done through traditional data.

Another limitations in using traditional data source is that it is difficult to anticipate immigrants' movement. In the work of [36], they study whether the GTI[11] can now-cast the immigrants' journey. However, as authors point out, not every search means that searchers have intention to migrate. To address this issue, they compare Gallup World Poll data[12] with the results obtained with GTI data. The Gallup data is a survey done on more then 160 countries and it contains questions on whether the individuals are planning to move to another country and if so, whether the plan will take place within 12 months and lastly, whether they have made any action to do so, i.e., visa applications or research for information. The comparison validates that the GTI data can indeed now-cast the "genuine migration intention".

Unconventional Big Data has its limitations like traditional data. Nevertheless, new big data methods are developing in order to address the newly arising issues. In addition, Big Data covers worldwide users with very fine granularity of information on immigrants' journey. The hope is that by merging knowledge from both traditional and novel datasets we will be able to overcome some of the issues and build accurate models for now-casting immigrant journeys and immigration rates.

2.2.1 Scientific migration

Given its importance to scientific productivity and education, the study of scientific migration has attracted a growing interest in the last years, fostered by the availability of massive data describing the publications and the careers of scientists in several disciplines [139, 48, 130, 156]. Understanding the mechanisms driving scientists' decision to relocate can help institutions and governments manage scientific mobility, implement policies to attract the best scientists or prevent their departure, hence improving the quality of research. At the same time, predictive models explaining when, and where, scientists migrate can facilitate the design of job recommender systems for scientists based on their profile [157], or help search committees seek successful candidates for their research jobs.

The studies proposed in the literature on scientific migration can be grouped into three main strands of research. A first group of studies focus on country-level movements or on movements between universities [20, 132, 126]. Relying on a large-scale survey, Appelt et al. find that geographic distance, as well as socio-economic disparities and scientific proximity, negatively correlate with the mobility of scientists between two countries [14]. By investigating the professional and personal determinants of the decision to relocate to a new institution, Azoulay et al. [22] find that scientists are more likely to move when they are highly productive and their local collaborators are fewer and less accomplished than their distant collaborators, while they find it costly to disrupt the social networks of their children. Gargiulo and Carletti [69] investigate the movements of scientists between universities and find that, starting from a lower rank institution lowers the probability of reaching a top rank academy and makes higher the probability to remain in a low rank one and, on the contrary, starting from a high ranked university strongly lowers the probability of ending in a low rank one.

A second strand of research focuses on understanding the impact of a scientist's relocation to their scientific impact. In this context, it has been discovered that while moves from elite to lower-rank institutions lead to a moderate decrease in scientific performance, moves to elite institutions does not necessarily result in subsequent performance gain [53]. Sugimoto [162] analyses the migration traces of scientists extracted from Web of Science and reveals that, regardless the nation of origin, scientists who relocate are more highly cited than their non-moving counterparts.

In the context of studying labor mobility, the availability of massive datasets of individuals' career path fostered works on predicting individuals' next jobs (outside the academia) [157]. Paparrizos et al. [136] build a system to recommend new jobs to people who are seeking a job, using all their past job transitions as well as their employees data. They train a predictive model to show that job transitions can be accurately predicted, significantly improving over a baseline that always predicts the most frequent institution in the data. Recently, Li et al. [112] propose a system to predict next career moves based on profile context matching and career path mining from a real-world LinkedIn dataset. They show that their system can predict future career moves, revealing interesting insights in micro-level labor mobility.

Our recent work, conducted within the SoBigData projects, is placed on the line of conjunction of the

[11] Google Trend Index, https://trends.google.com/trends/
[12] http://gallup.com

aforementioned strands of research. In particular, we investigate how a scientist's scientific profile influences the decision to move, based on a massive dataset consisting of all the publications in journals of the American Physical Society (APS) from 1950 to 2009 – 360,000 publications, 3,500 institutions and 60,000 scientists [99]. We approach the problem by constructing a two-stage predictive model. We first predict, using data mining, which scientist will change institution in the next year. We describe a scientist's profile as a multidimensional vector of variables describing three aspects: the recent scientific career, the quality of scientific environment and the structure of the scientific collaboration network. From the constructed predictive model, we identify the main factors influencing scientific migration. Secondly, for those scientists who are predicted to move, we predict which institution they will choose using the *performance-social-gravity model*, an adaptation of the gravity model of human mobility to include the above mentioned factors.

A different recent line of work in the SoBigData project is to understand, by using ORCID data, what was the effect of the Brexit referendum on scientific migration in and out of the United Kingdom. Preliminary results (still unpublished) show an increase in UK researchers moving from the EU to the UK and an increase of EU researchers moving out of the UK.

2.3 Visual Analytics of migration data

The phenomenon of migration is strongly associated with human movement. Analysis of movement data is one active topics in Visual Analytics research. The monograph [11] systematically considers a variety of possible representations of movement data. Frequently used representations are trajectories (sequences of time-stamped positions of individuals), time series (e.g. counts of departing, arriving or transit visitors over time), and events (e.g. movement with abnormal speed or unusual concentration of moving objects). A special case of trajectories is a trajectory consisting of only two time-stamped positions, origin and destination of a trip. This representation is frequently used in migration studies, since more detailed information is often not available.

The following three main classes of techniques are applied for visualization of origin-destination (OD) flows: OD matrix [85], OD flow map [168], and a hybrid of a matrix and a map called OD map [183]. In an OD matrix, the rows and columns correspond to locations and the cells contain flow magnitudes represented by color shades. The rows and columns can be automatically or interactively reordered for uncovering connectivity patterns. Disadvantages of the matrix display are the lack of spatial context and the limited number of different locations that can be represented. In OD flow maps, links between locations are represented by straight or curved linear symbols analogously to node-link diagrams. Various possible representations of directed links are discussed and evaluated by Holten et al. [93]. Flow magnitudes are shown by proportional line widths or by color shades. OD maps [183] use a map-like grid layout with embedded maps that represent movement from/to selected locations to/from all other locations that correspond to remaining maps.

A straightforward approach to showing time-variant flows is to use multiple displays (e.g. OD matrices, OD flow maps, or OD maps) arranged either temporally in animation or spatially in a small multiple display. Map animation is not effective [171] because the user cannot memorize and mentally compare multiple spatial situations. In small multiples, a limited number of spatial situations can be shown simultaneously; hence, this approach is not suitable for long time series. Clustering of spatial situations [13] can be used to reduce the number of distinct situations that need to be shown. A completely different approach is to show the time series of flow magnitudes separately from maps, for instance, as it is done in FlowStrates [38]; however, the spatial situations and their changes over time cannot be seen.

The paper [12] defines a workflow for analysis of long-term origin-destination data. The approach starts with aggregation of flows by origin or destination regions, directions and distances of move, and time intervals. Next, time intervals are clustered according to feature vectors composed from descriptors of all origins representing magnitudes of flows in all considered directions and distances. The proposed system enables exploration and continuous refinement of clustering results. The process is supported by space- (flow maps, diagram maps) and time-based (calendar showing temporal dynamics of situations by colors of dates) visualizations.

The techniques described in this section have been successfully applied or are potentially applicable to analysis of long-term migration data, for detecting patterns and changes of migration.

3 The stay: effects on communities, immigrant integration

The study of the effects of migration on the communities involved includes various traditional lines of study. Immigrant integration is a complex process that can reflect a progressive adoption of the norms that prevail in the destination country or a return to the habits of the home countries. Integration has been analysed from

multiple viewpoints. Here, we outline some of these lines of work, with some recent examples, and we provide a few directions for development using big data. However, this section is not intended to be a complete survey of methods, since the complexity of the issue would require much more than a few pages to describe. For more comprehensive reviews on migrant integration please see e.g. [42].

3.1 Current practices

In general, immigrant integration and cultural changes have been traditionally analysed using census data, administrative registries and surveys. In this section we describe the different criteria used for analysis. We start with a discussion of research studying social integration (social network, mixed marriages), then we move on to labour market integration, and language adoption of immigrants. We conclude the section with a discussion of the effects of immigration on educational attitudes, on economic prosperity and on political attitudes.

The effect of the social network on migration was analysed by [119] using survey data on Mexican migrants to the US. The richness of the social networks is shown to promote migration of low-skill migrants, while for communities where the social networks are weak, high-skill migrants are present. In terms of migrant integration, social networks in schools were analysed in European countries by [159]. They show that homophilic attitudes develop differently for immigrants and natives, with the former being positively influenced by multi-ethnicity in class. Ego networks of Turks and Moroccans in the Netherlands are studied in [173], using survey data. The authors show that in general closest friends come from the same ethnic group. The effect is stronger for women and those that are culturally more dissimilar to the natives.

In terms of marriage relationships, in the United States, marriage with whites is analysed for different ethnicities and education levels [147,146]. Divorce rates are shown to be higher for mixed than for non-mixed couples in the Netherland, particularly for couples coming from very distant cultures [158]. The relation between mixed marriages and the immigration rate in Italian communities was studied by [4]. The authors show that there are differences between large cities and smaller municipalities, and they argue, based on probabilistic interaction models, that this is due to the structure of the social network, which is disconnected in large communities. The presence of female immigrants was found to increase the risk of separation of native couples in Italy, using survey data and official statistics [178].

Integration in the labour market has been analysed for various western and non-western countries by [160]. They show that general patterns of integration and factors affecting it are very similar between western and non-western countries. Factors that affect the probability to find a job are language exposure, cultural distances, economic advancement of the origin country. Recent work shows that language training has an important effect on labour market integration of immigrants in France [113]. The effect of education on employment is analysed in [133] for Mexican immigrants in the USA. Integration in the labour market can also depend on the location where immigrants settle. In some cases, such as refugee situations, locations are assigned centrally. Recent work[24] has used data on past employment success to provide better matches between locations and refugees, showing that the probability of being employed can be increased by 40 to 70%.

Both mixed marriages and labour market integration was analysed using official data from Spain by [26]. They show using insights from statistical physics that while mixed marriages seem to be driven by peer interaction, this is missing when it comes to labour market integration. The same approach can be used to forecast integration from the two points of view [50].

Language adoption is a very important factor contributing to the success of an immigrant in the host country, since it provides opportunities for education, employment, social interaction. Integration in the US was analysed by [6], looking at the language spoken at home by third-generation immigrants. The study shows that while Asians and European adopted the language at a similar pace, Spanish speaking families were still preserving some of their mother language. A different study [174] looks at the dynamics of language adoption in the US, and show that education is an important factor positively influencing speed of adoption, while group size provides negative influence. A related issue is that of naming children [1]. A recent study of early US census data shows that people coming from families where children were given foreign names were less successful in terms of education and earnings, and were more likely to marry foreign spouses. The bilingual settings was studied in [175], i.e. language adoption of immigrants in Belgium. The study shows that immigrants adopt faster the more international language.

The above mentioned works study integration by looking mostly at the immigrant population. However, effects on the local population due to integration of migrants exist too. For instance, educational expectations of middle school children were shown to change in children both from native and immigrant communities, in Italy, based on survey data [124,123]. Immigrant chil-

dren increased their expectations in the presence of native children with high expectations. Native children studying in multiethnic classes seemed more prone to high expectations. The effect of school class composition and ethnic attitudes was analysed in [39], showing that a balanced composition is beneficial for all ethnic groups involved.

A different effect that can be studied is related to economic prosperity of the target society. Diversity of birthplace was shown to increase economic prosperity [7], especially in the case of high-skill migrants moving to rich countries. The cultural diversity of the origin country was also analysed, showing that there is an optimal cultural distance for immigrants to maximise the beneficial economic effects. At the same time, however, [25] show that competition in the labour market and public services, together with cultural differences, generates a shift in political inclination. For instance, a shift of votes towards the left-wing parties was observed in Italy. Similar changes were observed in Austria, where one factor was the concern about the quality of the neighbourhood [87].

3.2 Towards a novel integration index using alternative data sources

While the type of studies exemplified in the previous section have been instrumental in understanding the effects of migration, the fact that they are based on traditional data makes them inherit the disadvantages of these data. Big Data can help to analyse the issues above, and others, with the advantage of producing real-time results, and enabling analysis at higher spatial resolution. For instance, retail data can help understand how immigrants adopt habits and values of the new community they live in. Mobile call data records (CDR) can be used to describe social interaction and mobility patterns of immigrants, and understand segregation. OSN data can help study various topics, such as social integration, language adoption, changes in the local language, sentiment towards immigrants, etc. All these data types can be also combined to build a novel multi-level integration index than takes into account all of these criteria. In the following we will exemplify some of these topics, including existing results from our project and new directions to pursue.

3.2.1 Retail data: Tell me what you eat, I will tell you who you are.

The measures for immigrant integration discussed in Section 3.1 capture choices that can be easily observed and potentially exposed to social sanctions. Moreover, they are usually measured at one point in time, while integration is a dynamic phenomenon. The analysis of retail data from a supermarket chain can enable us to understand if immigrants are converging to or diverging from the norms and habits of the destination country. By observing immigrants' food consumption baskets we can estimate the *degree of integration* and how this varies in time. This behaviour is less prone to social sanction, since the food basket is not generally known to people outside a family. Furthermore, we can identify which are the most relevant factors for the integration. The degree of integration can be considered both with respect to economic aspects but also based on how immigrant customers change their habits during their stay in terms of purchased products.

Market basket analysis and the study of food consumption have been widely used in the literature for different purposes, such as defining individual indicators of customer predictability [80], studying GDP trends [81], analyzing customers with respect to their temporal purchasing patterns [83], and classifying them as residents or tourists according to their shopping profile [82]. Exploiting retail data to study the migration phenomenon from an individual and collective point of view that is not exposed to social sanctions and with multiple observations in time can bring to the light novel results useful for better understanding the migration phenomenon and also for developing well-being policies.

Our project owns a key data source for these analyses, composed of scanner data from a large Italian retail market chain, that are available since January 2007 for more than 1.1 million customers holding a fidelity card. The dataset includes the price, quantity, promotional sales (if any) and the name of the good purchased out of a set of around 600,000 products. Besides this information, for each customer the country of birth is available and the date on which the fidelity card was obtained. About 7% of the customers are foreign-born, when the immigration rate in Italy is currently around 8.5%. On average, a foreign customer is observed 5 times per month, with a mean monthly food expenditure of about EUR150. In Figure 1 we report the cumulative number of customers joining the fidelity club for Albania, Romania and France. We observe how the trend is stable for Albania, while the number of customers with fidelity card is growing for Romania and decreasing for France. These indices are in line with the immigration trends from European official statistics, indicating that these data could be representative of the migrant population. In the following we discuss research directions that our project is pursuing.

To understand whether there is a convergence in food consumption choices of immigrants (by country of

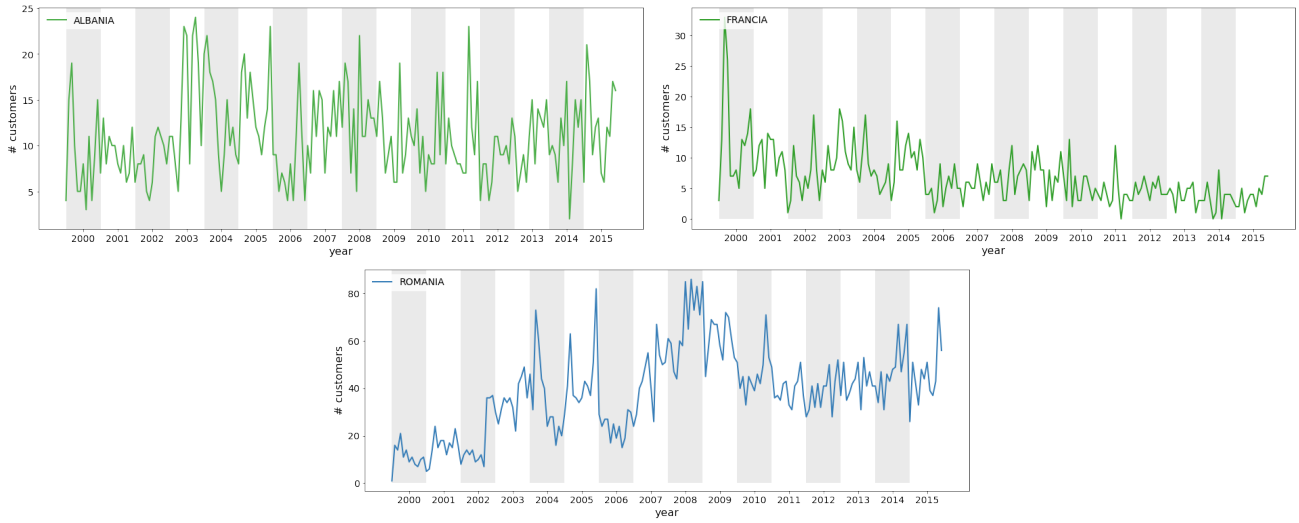


Fig. 1 Association to Italian Supermarket Chain. Trends of the number of customers with fidelity card for Albania, France and Romania.

birth), two orthogonal approaches can be followed. A *top down* approach aims at analyzing aggregated variables among the various items purchased that take into account for each foreign-born customer the difference between the normalized amount spent on a specific period and the mean spent in that period by Italian customers. In this way for each foreign-born customer we can obtain a time series indicating if that customer is converging or diverging from the Italian norms. Hence, we can find foreign countries having customers with homogeneous behaviors but also countries with different integration behaviors.

A weakness of the top down approach is that it is not easy to understand which are the products leading to the convergence/divergence. A *bottom up* approach analyzing the basket composition can provide this kind of information. In particular, our idea is to extract for different periods for each customer their *individual representative baskets* using the algorithm defined in [84]. Then re-cluster for each country the representative basket of the customers and develop *national collective representative baskets*. This can allow, through a set-based distance measure, to develop an indicator of shopping divergence/convergence with respect to the Italians typical baskets.

Finally, we underline that 14 percent of the foreign-born customers disappear from the dataset after some activity. The purchases of these customers could also be used for studying *the return* to the origin country.

3.2.2 Call data records

A large amount of work has been done using call data records (CDR) in understanding individual [76,71,138, 182] as well as group mobility [90,137,115,169]. These range from empirical analyses of large CDR datasets [76, 138,137,71,182] to proposal of theoretical mobility models [155]. Initiatives to motivate researchers to analyse CDR data have also appeared, through data challenges such as the Data for Development (D4D) challenge in Senegal [35] or the recent Data for Refugees (D4R) challenge in Turkey [170,153]. Readers can refer to [34] for a survey of works related to using CDR data for individual mobility studies and models.

A recent example is the study of the flocking and mobility behavior of the population after the Haiti earthquake using CDR data [114]. Researchers found that mobility patterns of the population after natural calamities is predictable. People tend to move to destinations where they have been making more calls before the disaster. In another natural calamity study done in NewZealand with respect to Christchurch earthquake which happened in Feb, 2011 [2], the researchers found that people either moved to Big cities like Auckland or to the small towns. However, no correlation between the mobile phone calls before and after the disaster has been reported. In all cases, this is an important outcome, as it can help in timely and effective infrastructural decisions in the time of emergencies or natural disaster [52].

In a different dimension, mobility patterns have also been studied with respect to socioeconomic development [137]. Authors found strong correlation between human mobility patterns with socio-economic indicators. It has also been shown that mobility patterns can be used for creating detailed maps of population distribution which are more accurate and recent. This approach is in particular useful for poor countries. This in

turn can help in creating proper socio-economic policies for the population [52].

However, while mobility analyses are abundant, not much work has been done to analyse the international migration phenomenon using CDR. This due to several reasons. First, CDR datasets typically span only one nation. Secondly, in general, due to privacy reasons, no information on the nationality of the customer is provided. Without these pieces of information, studying migration with these data is difficult. One exception is the above mentioned D4R challenge, where refugee status of customers is made available. Our project has participated in this challenge, together with several other teams, concentrating on five different aspects: health, integration, unemployment, safety and security, and education. For details on result obtained by other teams please see the published collection of articles [152]. Our objective was to analyse integration and combine the Turktelekom data with other datasets [31]. We observed that integration seemed to increase in time for refugees, and also that the presence of refugees influenced the house market in Turkey, decreasing housing prices.

Another recent example where CDR data was used to analyse transnational mobility is [5], using CDR data that includes mobile roaming events. Transnational population mobility can be defined as living and working in two or more countries. Understanding this phenomenon with traditional statistics and register-based data is impossible. The authors show that roaming data can enable the analysis of travel behavior and social profile of visitors. They can differentiate between tourists, cross-border commuters, foreign workers, and transnationals.

3.2.3 Language in online social networks

Language allows us to express needs, feelings and achieve our communication goals. Society changes and grows more complex over time, thus language must evolve and adapt itself to the new needs of its population. As a consequence, this evolution leads to changes, creation, and vanishing of expressions, dialects and even whole languages [75]. Over the past two decades, globalization has driven social, cultural and linguistic changes panorama in societies all over the world. The earlier multiculturalism, since the 1990s, intended as the ethnic minorities paradigm, turned in what Vertovec [177] calls *Superdiversity*. The concept aims to acquire the increasingly complex and less predictable set of relationships between ethnicity, citizenship, residence, origin, and language. Thanks to the influence of pioneering works of linguistic anthropologists, mixing, mobility patterns and historical framework became key issues in the study of the languages and of the language groups [33]. Over time, linguists and sociologists analysed variation and changes in both oral [106] and written [29] language by exploiting surveys, corpora, and records [75]. In the last decade, the pervasive use of online social networking and micro-blogging services led to the availability of freely-made contents never seen before. This unprecedented wealth of written data allowing us to recover a detailed picture of language evolution both from the geographical and the time points of view [131].

The literature regarding the language in social networks applied to migration studies is wide and involves several research fields, including but not limited to mobility patterns, migrations stocks and flows, Well-Being and Sentiment Analysis. Even though some works focused more on metadata instead of the real data contents, the text bears a wealth of information, starting from the language in which is written [108]. For instance, Kulkarni et al. [105] have proposed a novel method allowing to detect English linguistic variation and quantify its significance among geographic regions; Ibrahim et al. [95] have combined different data to present a sentiment analysis system for standard Arabic and Egyptian dialectal Arabic; The language has been also investigated in the spatial distribution as well as the spatial extension of dialects. In [117], geolocated tweets are exploited to identify localized patterns in language usage and to analyse the language diversity over different countries; Mocanu et al. [125] have characterized the worldwide linguistic geography by aggregating multi-scale OSN data; Jurdak et al. [100] have compared Twitter mobility patterns with patterns observed through other technologies, eg. CDRs., by using individuals' spatial orbit as the measure of how far they move; Gonçalves et al. [75] have found two global super-dialects in the modern-day Spanish; and Doyle [55] have proposed a Bayesian method to build conditional probability distributions of the spatial extension of English dialects.

Within the SoBigData project, we have analysed the concept of Superdiversity theorized by Vertovec (2007), and proposed a measure to quantify it [140]. We focus on the conjunct analysis of both language and geographic dimensions starting from a Twitter dataset. Our ground hypothesis is built on the idea that different cultures use the language in different ways and, in consequence, the emotional value associated with words changes depending on the culture of the person that writes a tweet. We introduced a *Superdiversity Index* (SI), that is based on the diversity of the emotional content expressed in texts of different communities. Specifically, we extract the emotional valences of words used by a community from Twitter data produced by that community. We compare the obtained valences with a

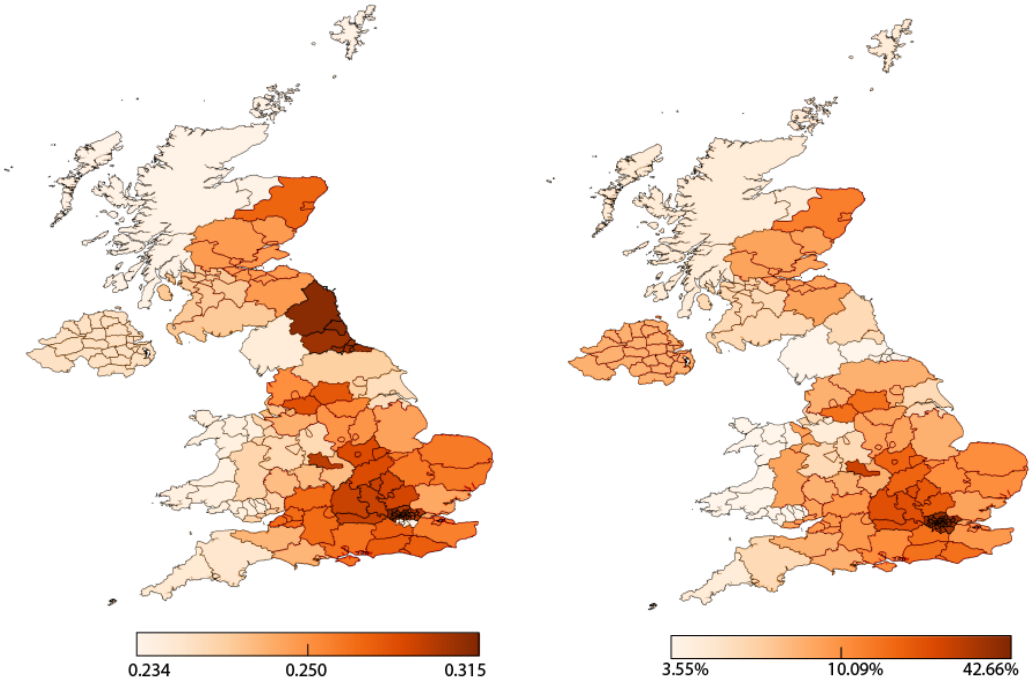


Fig. 2 Superdiversity index (left) and immigration levels (right) across UK regions at NUTS2 level [140]

standard dictionary tagged with sentiment. The distance between the community and the standard valences is a measure of superdiversity for the community. This SI measure is computed at different geographical scales based on the Classification of Territorial Units for Statistics (NUTS) for two different nations: Italy and United Kingdom, and validated with data from the above mentioned D4I challenge (see Section 2). We observe a very high correlation with immigration rates at all geographical levels. Figure 2 shows the case of the United Kingdom, where we observe that the geographical distribution of the SI proposed matches very well that of official immigration rates. Thus, we believe that, besides quantifying the cultural changes that migrants instill on the community, our SI can also become a key measure in a now-casting model for migration stocks.

3.2.4 Migration and sentiment

One way of studying migrant integration is by analysing the opinions of the locals related to migration topics and different migrant groups. While performing targeted surveys is one way of collecting such opinions, using Online Social Networks (OSNs) is a novel direction that can overcome some limitations of survey data. Using Twitter for *opinion mining and to study sentiment and user polarization* is a vast subject [135]. The existence of polarization in Social Media was first studied by Adamic et al. [3] who identified a clear separation in the hyperlink structure of political blogs. Conover et al. [49] studied afterwards the same phenomenon on Twitter, evaluating the polarization based on the retweets. Most of the studies on polarization are still based on sentiment analysis of the content. The *sentiment analysis* methods proposed are numerous and they are mainly based on dictionaries and on learning techniques through unsupervised [134] and supervised methods (lexicon-based method [164]) and combinations [104]. Opinion mining techniques are widely used in particular in the political context [3] and in particular on Twitter [46]. Recently new approaches based on polarization, controversy and topic tracking in time have been proposed [70,47]. The idea of these approaches is to divide users of a social network in groups based on their opinion on a particular topic and tracking their behavior over time. These approaches are based on network measures and clustering [70] or hashtag classification through probabilistic models [47] with no use of dictionary-based techniques.

Regarding the migration topic, in Coletto et al. [45] we propose an analytical framework aimed at investigating different views of the discussions regarding polarized topics which occur in OSNs. The framework supports the analysis along multiple dimensions, *i.e., time, space and sentiment* of the opposite views about a controversial topic emerging in an OSN, and is applied to the perception of the refugee crisis in Europe and Brexit. The sentiment analysis method adopted is efficient in tracking polarization over Twitter compared to

other methods. Concerning other approaches for studying social phenomena, we do not base our analyses on the change of location of Twitter users to measure the flow of individuals through space, but rather we aim at understanding the impact on the EU citizens perception of migrants' movements and their resulting decision to vote for Brexit.

The framework, initially presented in [44], allows to monitor in a scalable way the raw stream of relevant tweets and to automatically enrich them with location information (user and mentioned locations), and sentiment polarity (positive vs. negative). The analyses we conducted show how the framework captures the differences in positive and negative user sentiment over time and space. The resulting knowledge supports the understanding of complex dynamics by identifying variations in the perception of specific events and locations.

We used the Twitter Streaming API under the *Gardenhose* agreement (granting access to 10% of all tweets) to collect the English tweets posted in two periods: from mid August to mid Sept 2015 for the refugees dataset, and from mid June to the beginning of July 2016 for the Brexit dataset, respectively. We filtered out the tweets not related to the specific events analysed. The first dataset refers to the *Refugees crisis* and contains about 1.2 M tweets, while the second one refers to the *Brexit referendum* and contains about 4.3 M tweets. The datasets[13] are available for use through Transnational Access in the SoBigData project infrastructure.

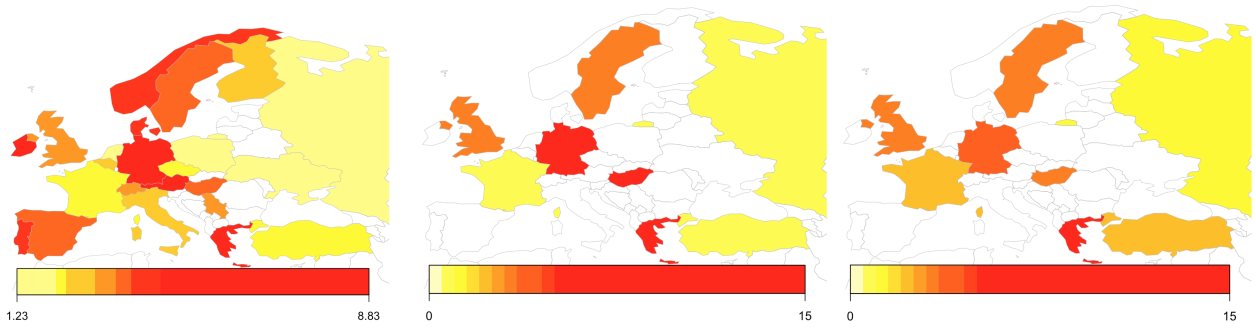
In our study we try to answer the following analytical questions: What is the evolution of the discussions about refugees migration in Twitter? What is the sentiment of users across Europe in relation to the refugee crisis? What is the evolution of the perception in the countries affected by the phenomenon? Are users more polarized in the countries that are most impacted by the migration flow? Is the polarization of the users about refugees and the Brexit referendum somehow correlated? For this purpose, we analyse the ratio between pro- and against-refugee users across Europe. For example, Figure 3 shows the geographical distribution of this ratio considering all users residing in a country, but also internal and external perception (perception of the users residing inside/outside a country C related to the refugees in C). We observe that Eastern countries in general are less positive than Western countries. Also, we note that for internal perception Russia, France and Turkey have a really low sentiment. We conjecture that the sentiment of a person, when the problem involves directly his/her own country, could be more negative since we are generally more critical when issues are closer to ourselves. External perception is generally higher in countries most affected by the refugee crisis, such as France, Russia and Turkey, with the exception of Germany where the decision to open borders seems to have produced positive internal sentiment.

[13] `https://sobigdata.d4science.org/group/resourcecatalogue/data-catalogue?path=/dataset/hpc_twitter_dumps`

3.2.5 Ego-networks and their effect on migration

Personal networks of migrants have been shown to play a strategic role in the destination country chosen by the migrant, in the well-being of the migrant (once settled in), and in the professional outcome [66,181,73,10,176]. For this reason, studying the properties of migrants' personal network is a particularly promising avenue of research in digital demography, in order to characterise both the journey and the stay. In this section, we review the basic concepts of ego networks and some existing applications, and we argue that studying ego-networks from OSN platforms can be a powerful tool in the analysis of migration.

It is a well-established result from sociology that personal networks, i.e., the ensemble of social relationships that an individual entertains with other people, have a significant influence on the quality of life of the individual in terms of, e.g., job opportunities [77, 78], social support [101], power and influence in organization/communities [154,122,129,109]. Personal networks are also closely related to the concept of social capital, i.e., the network of connections, loyalties, and mutual obligations [73] that translates into favors and preferential treatment. In this perspective, studying the evolution of personal networks over time is the ideal approach to characterise the modification of migrants' social structures (or lack thereof), due to the migration process. This is related to one of the main subjects of study in this area, i.e., the characterisation of integration of migrants. Integration is typically measured in terms of *assimilation* and *transnationalism*. Assimilation is defined as the gradual adoption of customs and traditions from the receiving country by the migrant, and can be full [8,9], partial [72] or segmented [142]. As a consequence of assimilation, the composition of a migrant's personal network is expected to change significantly over time. At the opposite side of assimilation, there is the phenomenon of transnationalism, whereby migrants continue to participate in the political, economic and cultural life of origin societies and of fellow migrants from the same country [141]. Many researchers have postulated that the widespread availability of Internet connectivity and OSNs has made easier to keep alive these transnational links with the ori-



(a) Global perception (b) Internal perception (c) External perception

Fig. 3 Sentiment related to the refugee crisis across European countries (from [45]: red corresponds to a higher predominance of positive sentiment, yellow indicates lower positive sentiment. (a) Refers to the whole dataset. (b) Is limited to users when mentioning locations in the their own country. (c) Is limited to users otherwise.

gin country [110]. Again, this should be reflected into the personal network of migrants, in terms of number and relationship strength of links towards migrants and non-migrants from the same origin area. These changes can be studied using traditional data coming from targeted surveys, but also from OSN data that can fill some of the gaps present in survey data.

While most migration studies of personal networks are qualitative, quantitative studies are available in the literature on generic social networks. Quantitative studies often explore the graph-theoretical concept of *ego networks*. An ego network is the graph-based abstraction that models the personal network of an individual (called *ego)*. Beside the ego, the nodes in the ego network correspond to the people the ego entertains social relationships with. These people are referred to as *alters*. The ego and each alter are connected by an edge, whose weight corresponds to the strength of their social relations (often referred to as *emotional closeness*). Depending on the ego network model used, ties between alters can also be included [65]. More rarely, only the alter-alter ties are considered for extracting ego network properties [118]. Several structural properties of ego networks can be derived [86].

Ego network models have been used in the literature to characterise human cognitive constraints and their impact on the social processes. In particular, evolutionary anthropology has studied the structure of ego networks (as a representation of human personal networks) in terms of the cognitive investment required from the ego to actively maintain it. Dunbar [56] has found that the humans' neocortex size places an upper limit on the number of *meaningful* relationships that can be maintained. Specifically, the group size predicted by the human neocortex size is around 150 alters and it has been validated studying tribal, traditional, and modern societies [59,91]. This limit on the size of the ego network determined by the cognitive effort required to maintain active social relationships is known as the *social brain hypothesis* [58]. Additional investigations of this cognitive constraint have shown that the alters in the ego networks are organised into concentric circles around the ego, where the emotional closeness decreases and the number of alters increases as we move from the ego outwards [91,187]. When looking at the size of the circles, a typical scaling ratio around 3 between the size of consecutive circles has been observed [187], with the size of individual circles concentrating around the values of 5, 15, 50, 150, respectively.

Quite interestingly, ego networks formed through many interaction means, including face-to-face contacts [58], letters [91,187], phone calls [116], co-authorships [17], and, remarkably, also OSN, are well aligned with the above model. Specifically, very similar properties have been found also in Facebook and Twitter ego networks [57, 19]. In this sense, OSN become one of the outlets that is taking up the brain capacity of humans, and thus are subject to the same limitations that have been measured for more traditional social interactions, and are not capable of "breaking" the limits imposed by cognitive constraints to our social capacity [60]. Tie strengths and how they determine ego network structures have been the subject of several additional works. For example, in [74] authors provide one of the first evidences of the existence of an ego network size comparable to the Dunbar's number in Twitter. The relationship between ego network structures and the role of users in Twitter was analysed in [148]. In general, ego network structures are also known to impact significantly on the way information spreads in OSN, and the diversity of information that can be acquired by users [15]. More in general, many traits of human social behaviour (resource sharing, collaboration, diffusion of information) are chiefly determined by the structural properties of

ego networks [163]. Less studied (typically due to the lack of data) but equally important are the dynamic properties of ego networks, which characterise the evolution of personal networks over time. Arnaboldi et al. [16,18] found that, unexpectedly, the strongest social relations in Twitter change frequently for the majority of generic users and also for the special class of politicians. This is a marked difference with respect to offline networks, where high-frequency relationships correspond to stable and intimate ties [91].

While data from OSNs have been recently used for migration studies, as detailed in previous sections, the graph-theoretical perspective has been rarely taken into account. The only exceptions are [89], [92], and [108]. In [89], community-centric metrics are used to study cultural assimilation as a function of the number of social ties between migrant communities and local people using the set of friendship links extracted from Facebook. The graph in this case is unweighted, i.e., the effect of different emotional closeness between node pairs is not taken into account. Lamanna et al. [108] again focus on cultural assimilation but from the spatial segregation standpoint. In this case, they use a bipartite graph structure, connecting tweet languages and cities. In [92], Facebook is used to study the network of teenagers in the Netherlands, concentrating on ethnicity and gender. The analysis shows that ethnicity plays a stronger role in link formation. However, the extended Facebook networks are less segregated, in general, compared to core ego-networks.

To the best of our knowledge, ego networks of migrants built from OSN data have never been investigated in the related literature. This is quite surprising, as it is well-known that many facets of the human behaviour chiefly depend on the ego network structure. This includes features intrinsically related to migration and integration, such as willingness to cooperate with alters, resilience to problems and possibility of seeking for assistance from trusted alters [161]. As discussed before, migrants' ego networks have been studied previously in the sociology literature, but only traditional data sources had been considered, and the approach to the analysis is typically more qualitative than quantitative. Here we advocate, along the lines of digital demography, that it is crucial to integrate traditional and innovative data sources to provide a timely and deeper understanding of personal networks and their impact on the migratory phenomena. For non-migrant users, the integration of OSN data has already proven successful and has highlighted properties that would have been impossible to extract from offline data alone [57]. Given the role played by personal networks on migration flows and integration, we believe it is crucial to fill this gap. OSN are particularly appealing for accomplishing this task. In fact, they allow to reach scales far beyond what can be obtained from traditional data sources and they can also allow researcher to easily analyse temporal variations in the ego networks, ultimately allowing forms of nowcasting of the migration phenomena.

Two research questions are particularly pressing: understanding and quantifying the relationship between the migrant's online ego network and their migration choices, as well as measuring cultural assimilation and transnationalism through the evolution of online ego networks over time. With respect to the first question, it would be important to study the influence that alters in the different layers of migrants' ego networks exert on the ego's migration choices, distinguishing between the role played by weak and strong ties. These results can then be used to attempt predictions of the future migration choices of people, similarly to what is discussed in [99] for scientists. With respect to the second question, online ego networks can be a strategic asset for studying cultural assimilation, as they are typically easy to monitor for a prolonged amount of time, going beyond the single snapshot problem mentioned in [151]. As the migrant "moves" into the receiving society, we expect to observe a turnover in the ego network layers, reflecting the changes in his/her social relationships. This turnover can be measured in terms of similarity between layers across different temporal snapshots and observing the jumps that alters perform in the ego's network (similar to what [18,37] do for the ego networks of politicians and journalists on Twitter). Special attention should be reserved to the movements, inside the ego network, of co-nationals vs natives of the receiving country. Cultural assimilation predicts that the first class of ties should weaken progressively, while the latter should thrive. As a result, we expect to observe outward movements for co-nationals and inwards movements for natives inside the ego network. If this is not the case, we can postulate poor or imperfect assimilation and/or strong transnational ties linking migrants to their origin country.

4 The return: migrants returning to the country of origin

Migration is commonly seen as a permanent change in residence habits. However, when considered as a temporary phenomenon, several implications arise. Return migration is increasing in several countries, i.e. Mexico [40], China [186], Jamaica [167], Tunisia [121,120], and Mali [43], with several effects observed. The most recent literature almost completely agrees in underlining the

benefits led by returning migrants. These advantages concern a very wide range of fields and include the rise of business activity, and the wages increase [179,180], the improvement of educational attainment and health conditions, the increase of electoral participation [43], and the decrease in violence [41].

The origin country can benefit economically from temporary migration in at least two ways [121,120]. The authors show, taking the example of Tunisia, that money transfers from abroad to the migrant families are a sizeable income. Secondly, new skills learned and savings can enable return migrants to start their own business in the origin country. The SoBigData project also performed research in this field, with an approach based on data journalism that resulted in a documentary on return migration in Senegal: "Demal Te Niew" [23]. Zhao [186] has analysed the determinants of return migration and the economic behaviour of return migrants in China. Its findings result partially in mild contrast with those already discussed. The author found that out-migration is still dominant, while the return migration led by both push and pull factors is limited in scale. However, inspecting the employment-related field, the results show that return migrants invest more in productive farm activity. However, they do not show higher tendencies to engage in local non-farm activities than natives and migrants. As well as most of the literature, Zhao findings testify the return migrants key role in the modernisation process of developing or less rich countries.

A lot of research has been focused on the "brain gain" provided by the return of high-skilled individuals, i.e., scientists returning in the country of birth. Scholars found that even if migration leads to a brain drain over the short-term, return migration can contribute to brain gain [54,180]. Moreover, the most recent researches demonstrate that return migrants contribute to the own community's long-term well-being independently by skills they have gained abroad [40].

Regarding the health field, Levitt et al. [111] have investigated dynamics between social practices gained abroad and healthcare. They show that social practices introduced by return migrants positively affect healthcare. These results seem related to the better social conditions of households with links to migrants and return migrants [61]. A different aspect relates to family-related decisions of return migrants. A recent study shows that Egyptian males returning from other Arab countries have more children than average [32], which could be due to the effect of the foreign culture on the decisions of the migrant.

The impact of return migrants on their origin country governance has been examined in [43,28]. Results show that local policies are positively affected by returning migrants since these contribute to increase political participation and enhance political accountability. Political orientation of the home community can also be affected by the migration phenomenon. For instance, for Moldova, a recent study[27] shows how Westbound migration slowly changed the voting behaviour leading to the fall of the communist government in 2009.

Concerning education, research results agree that return migrants can be associated with increases and improvement of educational attainment. Taking the example of Mexico, Montoya et al. [128] have found an increase of 26% in school attendance in households linked to at least a return migrant. This could mean that return migrants give higher priority to education.

Although the study of return migration is a long-standing area, most, if not all, analyses are based on traditional data. There is however great potential in employing novel data types such as mobile data or OSN to study return migration, and it remains an open research area.

5 Discussion and conclusions

We have discussed three lines of research where social big data can complement existing approaches to provide small area and high-time resolution methods for analysis of migration. In terms of estimating flows and stocks, some research already exists trying to use social big data to nowcast immigration. However, models still need to be refined and validated. An important issue here is that a proper gold standard does not exist: exact current immigration rates are unknown, and those in the past can be noisy, so validation of nowcasting models is not straightforward. Finding the relations between policies and immigration could be a step forward in finding means to validate model output. Another big data type that has not been included here and that can help make predictions in terms of migration related to climate is satellite data. To measure migrant integration, we believe that several new data types can be used to introduce novel integration indices, based on retail consumer behaviour, mobile data, OSN language, sentiment and network analysis. Research in this direction is slightly less developed, mostly due to low availability of ready-to-use datasets. Our consortium is making steps in this direction, using existing datasets, participating to data challenges or collecting new data. For the return of migrants, again research is limited, although potential exists in data such as retail, mobile or OSN.

In all three dimensions, research has to carefully consider the issues with the data that is being used. It is important that each study includes a well-planned data

collection phase where available data are analysed to identify gaps, and to devise strategies to fill the gaps by integrating other types of data. This in order to ensure that the problem being studied is thoroughly covered by the data used. In this process, research infrastructures such as SoBigData can be of great help. On one hand they can provide means to catalogue data, so that new datasets are available to the community for integration. On the other hand, they enable the community to share methods and experiences so that gaps identified and the solutions taken to fill these gaps can be reused. This applies not only to traditional data sources, but also to social big data. The complexity of digital demography implies that there is no free lunch with digital traces either [107]. One problem relates to the representativeness of the collected samples. For example, Facebook and Twitter penetration rates are different world-wide and tend to be different depending on the considered age of users [185]. Being unable to track specific categories of users can steer policies on migration in a direction that unwillingly perpetuates discriminations or neglects the needs of the invisible groups. For the above reasons, analytical and technical challenges to extract meaning from this kind of data, in synergy with more traditional data sources, remain an open and very important research area, with some recent efforts made in this direction [94]. Model validation using existing statistics and the relation to migration policies is important. Furthermore, careful data integration could help in overcoming some of the selection bias, resulting in novel, multi-level indices based on big data.

A different issue is that related to the ethics dimension of processing personal data, including sensitive personal data, describing human individuals and activities. As also stated in [188], the first rule that a researcher must follow is to acknowledge that data are people and can do harm. In particular, the context of migration is very sensitive to this problem, since individuals described in the data are often particularly vulnerable: refugees and their families might be persecuted in their home countries, so avoiding their re-identification is a critical matter. Moreover, mass media and social media impact our society and integration itself since a negative tone systematically relates to lower acceptance rates of asylum practices [103], so extreme care has to be taken in publishing results. Nevertheless, migration studies can have a significant impact to improve our society and to help the inclusion process of migrants; thus, encouraging data sharing is one of our main goals for achieving public good.

For all these reasons, it is essential that legal requirements and constraints are complemented by a solid understanding of ethical and legal views and values such as privacy and data protection, composing an actual ethical and legal framework. To this end, a number of infrastructural, organizational and methodological principles have been developed by the SoBigData Project, in order to establish a Responsible Research Infrastructure, allowing users to make full use of the functionalities and capabilities that big data can offer to help us solve our problems, while at the same time allowing them to respect fundamental rights and accommodate shared values, such as privacy, security, safety, fairness, equality, human dignity, and autonomy [67]. In particular, we strongly rely on Value Sensitive Design and Privacy-by-Design methodologies, in order to develop privacy-enhancing technologies, privacy-aware social data mining processes and privacy risk assessment methodologies. These methods are developed mainly in the fields of mobility data (such as GPS trajectories), mobile and retail data, which are some of the (unconventional) Big Data used in our migration studies. Moreover, some other general tools have been implemented to assist researchers in their activities, create a new class of responsible data scientists, and inform the data subjects and the society about our work and our goals, such as an online course, ethics briefs, and public information documents.

Acknowledgements

This work was supported by the European Commission through the Horizon2020 European project "SoBigData Research Infrastructure — Big Data and Social Mining Ecosystem" (grant agreement 654024). The funders had no role in developing the research and writing the manuscript.

Conflict of interest statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Abramitzky, R., Boustan, L.P., Eriksson, K.: Cultural assimilation during the age of mass migration. Tech. rep., National Bureau of Economic Research (2016)
2. ACAPS: Call detail records: The use of mobile phone data to track and predict population displacement in disasters (2013)
3. Adamic, L.A., Glance, N.: The political blogosphere and the 2004 US election: divided they blog. In: Proceedings of the 3rd international workshop on Link discovery, pp. 36–43. ACM (2005)

4. Agliari, E., Barra, A., Contucci, P., Pizzoferrato, A., Vernia, C.: Social interaction effects on immigrant integration. Palgrave Communications **4**(1), 55 (2018)
5. Ahas, R., Silm, S., Tiru, M.: Measuring transnational migration with roaming datasets. In: P. Kiefer, H. Huang, N. Van de Weghe, M. Raubal (eds.) Adjunct Proceedings of the 14th International Conference on Location Based Services, pp. 105 – 108. ETH Zurich (2018-01-15). DOI 10.3929/ethz-b-000225599. 14th International Conference on Location Based Services (LBS 2018); Conference Location: Zurich, Switzerland; Conference Date: January 15-17, 2018
6. Alba, R., Logan, J., Lutz, A., Stults, B.: Only english by the third generation? loss and preservation of the mother tongue among the grandchildren of contemporary immigrants. Demography **39**(3), 467–484 (2002)
7. Alesina, A., Harnoss, J., Rapoport, H.: Birthplace diversity and economic prosperity. Journal of Economic Growth **21**(2), 101–138 (2016)
8. Allport, G.W.: The nature of prejudice. Addison-Wesley (1954)
9. Amir, Y.: Contact hypothesis in ethnic relations. Psychological Bulletin **71**(5), 319–342 (1969)
10. Amuedo-Dorantes, C., Mundra, K.: Social Networks and Their Impact on the Earnings of Mexican Migrants. Demography **44**(4), 849–863 (2007)
11. Andrienko, G., Andrienko, N., Bak, P., Keim, D., Wrobel, S.: Visual Analytics of Movement. Springer (2013). DOI 10.1007/978-3-642-37583-5
12. Andrienko, G., Andrienko, N., Fuchs, G., Wood, J.: Revealing patterns and trends of mass mobility through spatial and temporal abstraction of origin-destination movement data. IEEE Transactions on Visualization and Computer Graphics **23**(9), 2120–2136 (2017). DOI 10.1109/TVCG.2016.2616404
13. Andrienko, N., Andrienko, G., Stange, H., Liebig, T., Hecker, D.: Visual analytics for understanding spatial situations from episodic movement data. KI - Künstliche Intelligenz **26**(3), 241–251 (2012). DOI 10.1007/s13218-012-0177-4. URL `https://doi.org/10.1007/s13218-012-0177-4`
14. Appelt, S., van Beuzekom, B., Galindo-Rueda, F., de Pinho, R.: Which factors influence the international mobility of research scientists? OECD STI Working Papers (2015). DOI http://dx.doi.org/10.1787/5js1tmrr2233-en
15. Aral, S., Alstyne, M.V.: The Diversity-Bandwidth Trade-off. Source: American Journal of Sociology **117**(1), 90–171 (2011)
16. Arnaboldi, V., Conti, M., Passarella, A., Dunbar, R.: Dynamics of personal social relationships in online social networks. In: Proceedings of the first ACM conference on Online social networks - COSN '13, pp. 15–26. ACM Press, New York, New York, USA (2013)
17. Arnaboldi, V., Dunbar, R.I.M., Passarella, A., Conti, M.: Analysis of Co-authorship Ego Networks. In: LNCS - Advances in Network Science, pp. 82–96. Springer, Cham (2016)
18. Arnaboldi, V., Passarella, A., Conti, M., Dunbar, R.: Structure of Ego-Alter Relationships of Politicians in Twitter. Journal of Computer-Mediated Communication **22**(5), 231–247 (2017)
19. Arnaboldi, V., Passarella, A., Conti, M., Dunbar, R.I.M.: Online Social Networks: Human Cognitive Constraints in Facebook and Twitter Personal Graphs. Elsevier (2015)
20. Auriol, L.: Careers of doctorate holders. OECD STI Working Papers **4** (2010). DOI http://dx.doi.org/10.1787/5kmh8phxvvf5-en
21. Avvenuti, M., Bellomo, S., Cresci, S., La Polla, M.N., Tesconi, M.: Hybrid crowdsensing: A novel paradigm to combine the strengths of opportunistic and participatory crowdsensing. In: Proceedings of the 26th international conference on World Wide Web companion, pp. 1413–1421. International World Wide Web Conferences Steering Committee (2017)
22. Azoulay, P., Ganguli, I., Zivin, J.G.: The mobility of elite life scientists: Professional and personal determinants. Research Policy **46**(3), 573 – 590 (2017). DOI https://doi.org/10.1016/j.respol.2017.01.002. URL `http://www.sciencedirect.com/science/article/pii/S0048733317300021`
23. Bachini, V et al.: Demal te niew (go and come back), documentary (2016). URL `http://speciali.espresso.repubblica.it/interattivi-2016/va-e-torna/index.html`
24. Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., Weinstein, J.: Improving refugee integration through data-driven algorithmic assignment. Science **359**(6373), 325–329 (2018). DOI 10.1126/science.aao4408. URL `http://science.sciencemag.org/content/359/6373/325`
25. Barone, G., D'Ignazio, A., de Blasio, G., Naticchioni, P.: Mr. rossi, mr. hu and politics. the role of immigration in shaping natives' voting behavior. Journal of Public Economics **136**, 1–13 (2016)
26. Barra, A., Contucci, P., Sandell, R., Vernia, C.: An analysis of a large dataset on immigrant integration in spain. the statistical mechanics perspective on social action. Scientific reports **4**, 4174 (2014)
27. Barsbai, T., Rapoport, H., Steinmayr, A., Trebesch, C.: The effect of labor migration on the diffusion of democracy: evidence from a former soviet republic. American Economic Journal: Applied Economics **9**(3), 36–69 (2017)
28. Batista, C., Vicente, P.C.: Do migrants improve governance at home? Evidence from a voting experiment. The World Bank Economic Review **25**(1), 77–104 (2011)
29. Bauer, L.: Inferring variation and change from public corpora. The handbook of language variation and change pp. 97–114 (2002)
30. Bengtsson, L., Lu, X., Thorson, A., Garfield, R., Von Schreeb, J.: Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. PLoS medicine **8**(8), e1001083 (2011)
31. Bertoli, S., Cintia, P., Giannotti, F., Madinier, E., Özden, Ç., Packard, M., Pedreschi, D., Rapoport, H., Sîrbu, A., Speciale, B.: Integration of Syrian refugees: insights from D4R, media events and housing market data. In: Guide to Mobile Data Analytics in Refugee Scenarios. Springer (2019)
32. Bertoli, S., Marchetta, F.: Bringing it all back home–return migration and fertility choices. World Development **65**, 27–40 (2015)
33. Blommaert, J., Arnaut, K., Rampton, B., Spotti, M.: Language and superdiversity. Routledge (2016)
34. Blondel, V.D., Decuyper, A., Krings, G.: A survey of results on mobile phone datasets analysis. EPJ Data Science **4**(1), 10 (2015)
35. Blondel, V.D., Esch, M., Chan, C., Clérot, F., Deville, P., Huens, E., Morlot, F., Smoreda, Z., Ziemlicki, C.:

Data for Development: the D4D Challenge on Mobile Phone Data. CoRR **abs/1210.0137** (2012)
36. Böhme, M.H., Gröger, A., Stöhr, T.: Searching for a better life: Predicting international migration with online search keywords. Journal of Development Economics p. 102347 (2019)
37. Boldrini, C., Toprak, M., Conti, M., Passarella, A.: Twitter and the Press. In: Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW'18, pp. 1471–1478. ACM Press, New York, New York, USA (2018)
38. Boyandin, I., Bertini, E., Bak, P., Lalanne, D.: Flowstrates: An approach for visual exploration of temporal origin-destination data. Computer Graphics Forum **30**(3), 971–980 (2011). DOI 10.1111/j.1467-8659.2011.01946.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2011.01946.x`
39. Bubritzki, S., van Tubergen, F., Weesie, J., Smith, S.: Ethnic composition of the school class and interethnic attitudes: a multi-group perspective. Journal of Ethnic and Migration Studies **44**(3), 482–502 (2018)
40. Bucheli, J.R., Fontenla, M., Waddell, B.J.: Return migration and violence. World Development **116**, 113–124 (2019)
41. Bucheli, J.R., Fontenla, M., Waddell, B.J.: Return migration and violence. World Development **116**, 113–124 (2019)
42. Carmon, N.: Immigration and integration in post-industrial societies: Theoretical analysis and policy-related research. Springer (2016)
43. Chauvet, L., Mercier, M.: Do return migrants transfer political norms to their origin country? Evidence from Mali. Journal of Comparative Economics **42**(3), 630–651 (2014)
44. Coletto, M., Esuli, A., Lucchese, C., Muntean, C.I., Nardini, F.M., Perego, R., Renso, C.: Sentiment-enhanced multidimensional analysis of online social networks: Perception of the mediterranean refugees crisis. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, San Francisco, CA, USA, August 18-21, 2016, pp. 1270–1277 (2016)
45. Coletto, M., Esuli, A., Lucchese, C., Muntean, C.I., Nardini, F.M., Perego, R., Renso, C.: Perception of social phenomena through the multidimensional analysis of online social networks. Online Social Networks and Media **1**, 14 – 32 (2017). DOI https://doi.org/10.1016/j.osnem.2017.03.001. URL `http://www.sciencedirect.com/science/article/pii/S246869641630009X`
46. Coletto, M., Lucchese, C., Orlando, S., Perego, R.: Electoral Predictions with Twitter: a Machine-Learning approach. In: IIR 2015, Cagliari, Italy (2015)
47. Coletto, M., Lucchese, C., Orlando, S., Perego, R.: Polarized user and topic tracking in twitter. In: SIGIR 2016, Pisa, Italy (2016)
48. Colizza, V., Flammini, A., Serrano, M.A., Vespignani, A.: Detecting rich-club ordering in complex networks. Nature physics **2**(2), 110–115 (2006)
49. Conover, M., Ratkiewicz, J., Francisco, M.R., Gonçalves, B., Menczer, F., Flammini, A.: Political polarization on Twitter. ICWSM **133**, 89–96 (2011)
50. Contucci, P., Sandell, R., Seyedi, S.: Forecasting the integration of immigrants. The Journal of Mathematical Sociology **41**(2), 127–137 (2017)
51. De Beer, J., Raymer, J., Van der Erf, R., Van Wissen, L.: Overcoming the problems of inconsistent international migration data: A new method applied to flows in europe. European Journal of Population/Revue européenne de Démographie **26**(4), 459–481 (2010)
52. Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D., Tatem, A.J.: Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences **111**(45), 15888–15893 (2014). DOI 10.1073/pnas.1408439111. URL `http://www.pnas.org/content/111/45/15888`
53. Deville, P., Wang, D., Sinatra, R., Song, C., Blondel, V.D., Barabási, A.L.: Career on the move: Geography, stratification, and scientific impact. Scientific Reports **4**, 4770 EP – (2014). URL `http://dx.doi.org/10.1038/srep04770`
54. Docquier, F., Rapoport, H.: Globalization, brain drain, and development. Journal of Economic Literature **50**(3), 681–730 (2012)
55. Doyle, G.: Mapping dialectal variation by querying social media. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 98–106 (2014)
56. Dunbar, R.: Neocortex size as a constraint on group size in primates. Journal of Human Evolution **22**(6), 469–493 (1992)
57. Dunbar, R., Arnaboldi, V., Conti, M., Passarella, A.: The structure of online social networks mirrors those in the offline world. Social Networks **43**, 39–47 (2015)
58. Dunbar, R.I.: The social brain hypothesis. Evolutionary Anthropology **6**(5), 178–190 (1998)
59. Dunbar, R.I.M.: Coevolution of neocortical size, group size and language in humans. Behavioral and Brain Sciences **16**(04), 681 (1993)
60. Dunbar, R.I.M.: Do online social media cut through the constraints that limit the size of offline social networks? Royal Society Open Science **3**(1), 150292 (2016)
61. Duryea, S., López-Córdova, E., Olmedo, A.: Migrant remittances and infant mortality: Evidence from mexico. Washington: Inter-American Development Bank. Mimeo (2005)
62. EU Knowledge Centre on Migration and Demography: KCMD Data Catalogue (Accessed July 2019). URL `https://bluehub.jrc.ec.europa.eu/catalogues/data/`
63. Eurostat: Migration and migrant population statistics (2018). URL `http://ec.europa.eu/eurostat/statistics-explained/index.php/Migration_and_migrant_population_statistics#Migration_flows`
64. EUROSTAT: Asylum and managed migration data (Accessed July 2019). URL `https://ec.europa.eu/eurostat/web/asylum-and-managed-migration/data/database`
65. Everett, M., Borgatti, S.P.: Ego network betweenness. Social Networks **27**(1), 31–38 (2005)
66. Faist, T.: The volume and dynamics of international migration and transnational social spaces. Refugee Survey Quarterly **20**(1) (2001)
67. Forgó, N., Hänold, S., van den Hoven, J., Krügel, T., Lishchuk, I., Mahieu, R., Monreale, A., Pedreschi, D., Pratesi, F., van Putten, D.: SoBigData: a Research Infrastructure for Ethical and Legal Data Science. submitted to JDSA special issue (2019)
68. FRONTEX: Illegal border crossing (Accessed July 2019). URL `https://www.asktheeu.org/en/request/illegal_boarder_crossing#incoming-10314`

69. Gargiulo, F., Carletti, T.: Driving forces of researchers mobility. Scientific Reports **4**(4860) (2014). DOI http://dx.doi.org/10.1038/srep04860
70. Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M.: Quantifying controversy in social media. In: ACM International Conference on Web Search and Data Mining, WSDM '16 (2016)
71. Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., Trasarti, R.: Unveiling the complexity of human mobility by querying and mining massive trajectory data. The VLDB Journal—The International Journal on Very Large Data Bases **20**(5), 695–719 (2011)
72. Glazer, N.: Is Assimilation Dead? The Annals of the American Academy of Political and Social Science **530**(1), 122–136 (1993)
73. Gold, S.J.: Migrant Networks: a Summary and Critique of Relational Approaches to International Migration. In: The Blackwell Companion to Social Inequalities, pp. 257–285. Blackwell Publishing Ltd, Oxford, UK (2007)
74. Gonçalves, B., Perra, N., Vespignani, A.: Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number. PLoS ONE **6**(8) (2011)
75. Gonçalves, B., Sánchez, D.: Crowdsourcing dialect characterization through Twitter. PloS one **9**(11), e112074 (2014)
76. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. Nature **453**(7196), 779–782 (2008)
77. Granovetter, M.: The Strength of Weak Ties: A Network Theory Revisited. Sociological Theory **1**, 201 (1983)
78. Granovetter, M.S.: Getting a Job: A Study of Contacts and Careers, vol. 25. University of Chicago press (2018)
79. Grossi, V., Rapisarda, B., Giannotti, F., Pedreschi, D.: Data science at SoBigData: the European research infrastructure for social mining and big data analytics. International Journal of Data Science and Analytics **6**(3), 205–216 (2018)
80. Guidotti, R., Coscia, M., Pedreschi, D., Pennacchioli, D.: Behavioral entropy and profitability in retail. In: Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on, pp. 1–10. IEEE (2015)
81. Guidotti, R., Coscia, M., Pedreschi, D., Pennacchioli, D.: Going Beyond GDP to Nowcast Well-Being Using Retail Market Data. In: International Conference and School on Network Science, pp. 29–42. Springer (2016)
82. Guidotti, R., Gabrielli, L.: Recognizing Residents and Tourists with Retail Data Using Shopping Profiles. In: International Conference on Smart Objects and Technologies for Social Good, pp. 353–363. Springer (2017)
83. Guidotti, R., Gabrielli, L., Monreale, A., Pedreschi, D., Giannotti, F.: Discovering temporal regularities in retail customers' shopping behavior. EPJ Data Science **7**(1), 6 (2018)
84. Guidotti, R., Monreale, A., Nanni, M., Giannotti, F., Pedreschi, D.: Clustering individual transactional data for masses of users. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 195–204. ACM (2017)
85. Guo, D.: Visual analytics of spatial interaction patterns for pandemic decision support. International Journal of Geographical Information Science **21**(8), 859–877 (2007). DOI 10.1080/13658810701349037
86. Gupta, S., Yan, X., Lerman, K.: Structural properties of ego networks. In: International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, pp. 55–64. Springer (2015)
87. Halla, M., Wagner, A.F., Zweimüller, J.: Immigration and voting for the far right. Journal of the European Economic Association **15**(6), 1341–1385 (2017)
88. Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Ratti, C.: Geo-located Twitter as proxy for global mobility patterns. Cartography and Geographic Information Science **41**(3), 260–271 (2014)
89. Herdagdelen, Amaç, State, B., Adamic, L., Mason, W.: The social ties of immigrant communities in the United States. In: WebSci (2016)
90. Hiir, H., Sharma, R., Aasa, A., Saluveer, E.: Impact of Natural and Social Events on Mobile Call Data Records – An Estonian Case Study. In: International Conference on Complex Networks and their Applications. Springer (2019)
91. Hill, R.A., Dunbar, R.I.M.: Social Network Size In Humans. Human Nature **14**(1), 53–72 (2003)
92. Hofstra, B., Corten, R., van Tubergen, F., Ellison, N.B.: Sources of segregation in social networks: A novel approach using Facebook. American Sociological Review **82**(3), 625–656 (2017)
93. Holten, D., Isenberg, P., van Wijk, J.J., Fekete, J.D.: An extended evaluation of the readability of tapered, animated, and textured directed-edge representations in node-link graphs. In: 2011 IEEE Pacific Visualization Symposium, pp. 195–202 (2011). DOI 10.1109/PACIFICVIS.2011.5742390
94. Iacus, S.M., Porro, G., Salini, S., Siletti, E.: A proposal to deal with sampling bias in social network big data. In: 2nd International Conference on Advanced Reserach Methods and Analytics (CARMA 2018), pp. 29–37. Editorial Universitat Politècnica de València (2018)
95. Ibrahim, H.S., Abdou, S.M., Gheith, M.: Sentiment analysis for modern standard arabic and colloquial. arXiv preprint arXiv:1505.03105 (2015)
96. Instituto Nacional de Estadistica: Ine microdata (Accessed July 2019). URL `https://www.ine.es/en/prodyser/microdatos_en.htm`
97. IPUMS: IPUMS census and survey data (Accessed July 2019). URL `https://ipums.org/`
98. Istituto Nazionale di Statistica: Immigrati.stat: Dati e indicatori su immigranti e nuovi cittadini (Accessed July 2019). URL `http://stra-dati.istat.it/`
99. James, C., Pappalardo, L., Sirbu, A., Simini, F.: Prediction of next career moves from scientific profiles. ArXiv e-prints (2018)
100. Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., Newth, D.: Understanding human mobility from Twitter. PloS one **10**(7), e0131469 (2015)
101. Kadushin, C.: Social density and mental health. In: Social structure and network analysis, pp. 147–158 (1982)
102. Kikas, R., Dumas, M., Saabas, A.: Explaining international migration in the skype network: The role of social network features. In: Proceedings of the 1st ACM Workshop on Social Media World Sensors, pp. 17–22. ACM (2015)
103. Koch, C.M., Moise, I., Donnay, K., Boudemagh, E., Helbing, D.: Dynamics between mass media and asylum acceptance rates. SSRN Electronic Journal (2017). DOI 10.2139/ssrn.2957362
104. Kolchyna, O., Souza, T.T., Treleaven, P., Aste, T.: Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination. Handbook of Sentiment Analysis in Finance (2015)

105. Kulkarni, V., Perozzi, B., Skiena, S.: Freshman or Fresher? Quantifying the Geographic Variation of Language in Online Social Media. In: ICWSM, pp. 615–618 (2016)
106. Labov, W., Ash, S., Boberg, C.: The atlas of North American English: Phonetics, phonology and sound change. Walter de Gruyter (2005)
107. Laczko, F.: Improving Data on International Migration and Development: Towards a Global Action Plan? "Improving Data on International Migration-towards Agenda 2030 and the Global Compact on Migration" (2015)
108. Lamanna, F., Lenormand, M., Salas-Olmedo, M.H., Romanillos, G., Gonçalves, B., Ramasco, J.J.: Immigrant community integration in world cities. PloS one **13**(3), e0191612 (2018)
109. Laumann, E.O., Pappi, F.U.: Networks of collective action : A perspective on community influence systems. Academic Press (1976)
110. Levitt, P., Jaworsky, B.N.: Transnational Migration Studies: Past Developments and Future Trends. Annual Review of Sociology **33**(1), 129–156 (2007)
111. Levitt, P., Lamba-Nieves, D.: Social remittances revisited. Journal of Ethnic and Migration Studies **37**(1), 1–22 (2011)
112. Li, L., Jing, H., Tong, H., Yang, J., He, Q., Chen, B.C.: Nemo: Next career move prediction with contextual embedding. In: Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion, pp. 505–513. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2017). DOI 10.1145/3041021.3054200. URL `https://doi.org/10.1145/3041021.3054200`
113. Lochmann, A., Rapoport, H., Speciale, B.: The Effect of Language Training on Immigrants' Economic Integration-Empirical Evidence from France. European Economic Review **113**, 265–296 (2019)
114. Lu, X., Bengtsson, L., Holme, P.: Predictability of population displacement after the 2010 Haiti earthquake. Proceedings of the National Academy of Sciences **109**(29), 11576–11581 (2012). URL `http://www.pnas.org/content/109/29/11576`
115. Lulli, A., Gabrielli, L., Dazzi, P., Dell'Amico, M., Michiardi, P., Nanni, M., Ricci, L.: Scalable and flexible clustering solutions for mobile phone-based population indicators. International Journal of Data Science and Analytics **4**(4), 285–299 (2017)
116. Mac Carron, P., Kaski, K., Dunbar, R.: Calling Dunbar's numbers. Social Networks **47**, 151–155 (2016)
117. Magdy, A., Ghanem, T.M., Musleh, M., Mokbel, M.F.: Exploiting geo-tagged tweets to understand localized language diversity. In: Proceedings of Workshop on Managing and Mining Enriched Geo-Spatial Data, p. 2. ACM (2014)
118. McCarty, C.: Structure in Personal Networks. Journal of Social Structure **3**, 1–29 (2002)
119. McKenzie, D., Rapoport, H.: Self-selection patterns in Mexico-US migration: the role of migration networks. The Review of Economics and Statistics **92**(4), 811–821 (2010)
120. Mesnard, A.: Temporary migration and capital market imperfections. Oxford economic papers **56**(2), 242–262 (2004)
121. Mesnard, A., et al.: Temporary migration and self-employment: evidence from tunisia. Brussels Economic Review **47**(1), 119–138 (2004)
122. Miller, J.: Pathways in the Workplace: The Effects of Gender and Race on Access to Organizational Resources. Cambridge University Press (1986)
123. Minello, A.: The educational expectations of Italian children: the role of social interactions with the children of immigrants. International studies in sociology of education **24**(2), 127–147 (2014)
124. Minello, A., Barban, N.: The educational expectations of children of immigrants in Italy. The ANNALS of the American Academy of Political and Social Science **643**(1), 78–103 (2012)
125. Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q., Vespignani, A.: The Twitter of Babel: Mapping World Languages through Microblogging Platforms. PloS one **8**(4), e61981 (2013)
126. Moed, H.F., Aisati, M., Plume, A.: Studying scientific migration in scopus. Scientometrics **94**, 929–942 (2013)
127. Moise, I., Gaere, E., Merz, R., Koch, S., Pournaras, E.: Tracking language mobility in the Twitter landscape. In: Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on, pp. 663–670. IEEE (2016)
128. Montoya Arce, J., Salas Alfaro, R., Soberón Mora, J.A.: La migración de retorno desde Estados Unidos hacia el Estado de México: oportunidades y retos. Cuadernos Geográficos (2011)
129. Moore, G.: The structure of a national elite network. American Sociological Review **44**(5), 673 (1979)
130. Newman, M.E.: The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences **98**(2), 404–409 (2001)
131. Nguyen, D., Doğruöz, A.S., Rosé, C.P., de Jong, F.: Computational sociolinguistics: A survey. Computational Linguistics **42**(3), 537–593 (2016)
132. Noorden, R.V.: Global mobility: Science on the move. Nature **490**, 326–329 (2012)
133. Oropesa, R.S., Landale, N.S.: Why do immigrant youths who never enroll in us schools matter? school enrollment among mexicans and non-hispanic whites. Sociology of Education **82**(3), 240–266 (2009)
134. Pak, A., Paroubek, P.: Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: LREc, vol. 10, pp. 1320–1326 (2010)
135. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and trends in information retrieval **2**(1-2), 1–135 (2008)
136. Paparrizos, I., Cambazoglu, B.B., Gionis, A.: Machine learned job recommendation. In: Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11, pp. 325–328. ACM, New York, NY, USA (2011). DOI 10.1145/2043932.2043994. URL `http://doi.acm.org/10.1145/2043932.2043994`
137. Pappalardo, L., Pedreschi, D., Smoreda, Z., Giannotti, F.: Using big data to study the link between human mobility and socio-economic development. In: 2015 IEEE International Conference on Big Data (Big Data), pp. 871–878 (2015). DOI 10.1109/BigData.2015.7363835
138. Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., Barabási, A.L.: Returners and explorers dichotomy in human mobility. Nature Communications **6**, 8166 EP – (2015). URL `http://dx.doi.org/10.1038/ncomms9166`
139. Perra, N., Gonçalves, B., Pastor-Satorras, R., Vespignani, A.: Activity driven modeling of time varying networks. Scientific reports **2** (2012)
140. Pollacci, L., Sîrbu, A., Giannotti, F., Pedreschi, D.: Measuring the Salad Bowl: Superdiversity on Twitter (2019). Submitted

141. Portes, A., Guarnizo, L.E., Landolt, P.: The study of transnationalism: Pitfalls and promise of an emergent research field. Ethnic and Racial Studies **22**(2), 217–237 (1999)
142. Portes, A., Zhou, M.: The New Second Generation: Segmented Assimilation and its Variants. The Annals of the American Academy of Political and Social Science **530**(1), 74–96 (1993)
143. Poulain, M.: Confrontation des statistiques de migrations intra-européennes: Vers plus d'harmonisation? European Journal of Population/Revue européenne de Démographie **9**(4), 353–381 (1993)
144. Poulain, M., Herm, A., Depledge, R.: Central population registers as a source of demographic statistics in europe. Population **68**(2), 183–212 (2013)
145. Prieto Curiel, R., Pappalardo, L., Gabrielli, L., Bishop, S.R.: Gravity and scaling laws of city to city migration. PLOS ONE **13**(7), 1–19 (2018). DOI 10.1371/journal.pone.0199892. URL `https://doi.org/10.1371/journal.pone.0199892`
146. Qian, Z., Glick, J.E., Batson, C.D.: Crossing boundaries: Nativity, ethnicity, and mate selection. Demography **49**(2), 651–675 (2012)
147. Qian, Z., Lichter, D.T.: Social boundaries and marital assimilation: Interpreting trends in racial and ethnic intermarriage. American Sociological Review **72**(1), 68–94 (2007)
148. Quercia, D., Capra, L., Crowcroft, J.: The Social World of Twitter: Topics, Geography, and Emotions. ICWSM **12**, 298–305 (2012)
149. Raymer, J., Wiilekens, F.: Obtaining an overall picture of population movement in the european union. International migration in Europe: Data, models and estimates pp. 209–234 (2008)
150. Ruotsalainen, K.: A census of the world population is taken every ten years (2016). URL `http://www.stat.fi/tup/vl2010/art_2011-05-17_001_en.html`
151. Ryan, L., D'Angelo, A.: Changing times: Migrants' social network analysis and the challenges of longitudinal research. Social Networks **53**, 148–158 (2018)
152. Salah, A.A., Pentland, A., Lepri, B., Letouze, E. (eds.): Guide to Mobile Data Analytics in Refugee Scenarios. Springer International Publishing (2019)
153. Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., de Montjoye, Y.A., Dong, X., Dağdelen, Ö., Vinck, P.: Introduction to the data for refugees challenge on mobility of Syrian refugees in Turkey. In: Guide to Mobile Data Analytics in Refugee Scenarios, pp. 3–27. Springer (2019)
154. Scott, W.R., Laumann, E.O., Knoke, D.: The organizational state: Social choice in national policy domains (1989)
155. Simini, F., Gonzalez, M.C., Maritan, A., Barabasi, A.L.: A universal model for mobility and migration patterns. Nature **484**(7392), 96–100 (2012)
156. Sinatra, R., Wang, D., Deville, P., Song, C., Barabási, A.L.: Quantifying the evolution of individual scientific impact. Science **354**(6312) (2016). DOI 10.1126/science.aaf5239. URL `http://science.sciencemag.org/content/354/6312/aaf5239`
157. Siting, Z., Wenxing, H., Ning, Z., Fan, Y.: Job recommender systems: A survey. In: 2012 7th International Conference on Computer Science Education (ICCSE), pp. 920–924 (2012). DOI 10.1109/ICCSE.2012.6295216
158. Smith, S., Maas, I., van Tubergen, F.: Irreconcilable differences? Ethnic intermarriage and divorce in the Netherlands, 1995–2008. Social Science Research **41**(5), 1126–1137 (2012)
159. Smith, S., Van Tubergen, F., Maas, I., McFarland, D.A.: Ethnic composition and friendship segregation: Differential effects for adolescent natives and immigrants. American Journal of Sociology **121**(4), 1223–1272 (2016)
160. Spörlein, C., van Tubergen, F.: The occupational status of immigrants in western and non-western societies. International Journal of Comparative Sociology **55**(2), 119–143 (2014)
161. State, B., Rodriguez, M., Helbing, D., Zagheni, E.: Migration of Professionals to the U.S. In: SocInfo, pp. 531–543. Springer, Cham (2014)
162. Sugimoto, C.R.: Scientists have most impact when they're free to move. Nature **550**, 29–31 (2017). DOI 10.1038/550029a
163. Sutcliffe, A., Dunbar, R., Binder, J., Arrow, H.: Relationships and the social brain: Integrating psychological and evolutionary perspectives. British Journal of Psychology **103**(2), 149–168 (2012)
164. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational linguistics **37**(2), 267–307 (2011)
165. The OECD: Database on immigrants in oecd and non-oecd countries: Dioc (Accessed July 2019). URL `http://www.oecd.org/els/mig/dioc.htm`
166. The Worldbank: Migration and remittances data (Accessed July 2019). URL `https://www.worldbank.org/en/topic/migrationremittancesdiasporaissues/brief/migration-remittances-data`
167. Thomas-Hope, E.: Return migration to Jamaica and its development potential. International migration **37**(1), 183–207 (1999)
168. Tobler, W.R.: Experiments in migration mapping by computer. The American Cartographer **14**(2), 155–163 (1987). DOI 10.1559/152304087783875273
169. Tosi, D.: Cell phone big data to compute mobility scenarios for future smart cities. International Journal of Data Science and Analytics **4**(4), 265–284 (2017)
170. Turktelekom: Data for refugees turkey (2018). URL `http://d4r.turktelekom.com.tr/`
171. Tversky, B., Morrison, J.B., Betrancourt, M.: Animation: can it facilitate? International Journal of Human-Computer Studies **57**(4), 247 – 262 (2002). DOI https://doi.org/10.1006/ijhc.2002.1017. URL `http://www.sciencedirect.com/science/article/pii/S1071581902910177`
172. United Nations: Recommendations on statistics of international migration. Department of Economic and Social Affairs, Statistics Division, United Nations, New York. (1998)
173. Van Tubergen, F.: Ethnic boundaries in core discussion networks: A multilevel social network study of Turks and Moroccans in the Netherlands. Journal of Ethnic and Migration Studies **41**(1), 101–116 (2015)
174. Van Tubergen, F., Kalmijn, M.: A Dynamic Approach to the Determinants of Immigrants' Language Proficiency: The United States, 1980–2000 1. International Migration Review **43**(3), 519–543 (2009)
175. Van Tubergen, F., Wierenga, M.: The language acquisition of male immigrants in a multilingual destination: Turks and Moroccans in Belgium. Journal of Ethnic and Migration Studies **37**(7), 1039–1057 (2011)
176. Verdery, A.M., Mouw, T., Edelblute, H., Chavez, S.: Communication flows and the durability of a transnational social field. Social Networks **53**, 57–71 (2018)

177. Vertovec, S.: Super-diversity and its implications. Ethnic and racial studies **30**(6), 1024–1054 (2007)
178. Vignoli, D., Pirani, E., Venturini, A.: Female migration and native marital stability: insights from Italy. Journal of family and economic issues **38**(1), 118–128 (2017)
179. Wahba, J.: Selection, selection, selection: the impact of return migration. Journal of Population Economics **28**(3), 535–563 (2015)
180. Wahba, J., Zenou, Y.: Out of sight, out of mind: Migration, entrepreneurship and social capital. Regional Science and Urban Economics **42**(5), 890–903 (2012)
181. Waldinger, R.: 12 networks and niches: the continuing significance of ethnic connections. Ethnicity, social mobility, and public policy: Comparing the USA and UK p. 342 (2005)
182. Wang, D., Pedreschi, D., Song, C., Giannotti, F., Barabasi, A.L.: Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1100–1108. Acm (2011)
183. Wood, J., Dykes, J., Slingsby, A.: Visualisation of origins, destinations and flows with od maps. The Cartographic Journal **47**(2), 117–129 (2010). DOI 10.1179/000870410X12658023467367
184. Zagheni, E., Garimella, V.R.K., Weber, I., et al.: Inferring international and internal migration patterns from Twitter data. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 439–444. ACM (2014)
185. 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)
186. Zhao, Y.: Causes and consequences of return migration: recent evidence from china. Journal of Comparative Economics **30**(2), 376–394 (2002)
187. Zhou, W.X., Sornette, D., Hill, R.a., Dunbar, R.I.M.: Discrete hierarchical organization of social group sizes. Proceedings. Biological sciences / The Royal Society **272**(1561), 439–444 (2005)
188. Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S.P.n., Goodman, A., Hollander, R., Koenig, B.A., Metcalf, J., Narayanan, A., Nelson, A., Pasquale, F.: Ten simple rules for responsible big data research. PLoS Comput Biol **13**(3) (2017). DOI https://doi.org/10.1371/journal.pcbi.1005399