



Essays on International Migration using Big Data Analytics

Doctoral Thesis

by

Jisu Kim

Doctoral Program in Data Science

Supervisor

Giorgio Fagiolo, SSSA

Supervisor

Fosca Giannotti, ISTI-CNR

Supervisor

Hillel Rapoport, PSE-Paris 1

Supervisor

Alina Sîrbu, UniPi

© Jisu Kim, 2021. All rights reserved.

The author hereby grants to Scuola Normale Superiore di Pisa permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

Essays on International Migration using Big Data Analytics

by

Jisu Kim

March 3, 2021

Submitted to the Scuola Normale Superiore di Pisa
in December 2021, in partial fulfillment of the requirements for the
Doctoral Program in Data Science

Abstract

How can social big data help to understand issues related to international migration? Official data such as census, survey and administrative data have been traditionally the main data source to study migration. However, these data have some limitations. They are inconsistent across different nations because countries employ different definitions of international migration and characterisations of migrants. Moreover, collecting traditional data is costly and time consuming, thus tracking instantaneous flows of migrants becomes difficult. This becomes even harder when tracking emigrants because of the lack of motivation from citizens to declare their departure. In recent years, however, we are provided with other alternative data sources for migration. The availability of social big data such as Facebook, and Twitter data allows us to study social behaviours both at large scale and at a granular level, and to peek into real-world phenomena. Although known to suffer from other types of issues, such as selection bias, these data could bring complementary value to standard statistics.

In this work, we employ social big data to study international migration. We try to answer the question through an analysis of various phases of migration, using both traditional data and novel data sources. The first phase includes the journey, and we study migration stocks on Twitter, providing benefits and drawbacks of using such data to study international migration. Here, a generic methodology is developed to identify migrants within the Twitter population. This describes a migrant as a person who has the current residence different from the nationality. The residence is defined as the location where a user spends most of his/her time in a certain year. The nationality is inferred from linguistic and social connections to a migrant's country of origin. This methodology is validated first with an internal gold standard dataset and second with Italian register data and Eurostat, and shows strong performance scores and correlation coefficients.

The second phase concerns the integration of migrants in the destination country and attachments to their home country. We explore Twitter data to build a novel methodology to quantify and understand migrants' different integration types. Here, We describe four different integration types which are *assimilation*, *integration*, *marginalisation* and *separation* using two dimensions: the preservation of links to the home country and culture, i.e. home attachment index, and creation of new

links and adoption of cultural traits from the new residence country, i.e. destination attachment index. The two dimensions are validated by performing a null model analysis. It shows significant differences between the actual indices and the null model indices, confirming that the two indices are not produced at random.

Lastly, We examine the effect of the presence of migrants on political choices of the natives, using a German case study. Specifically, We are interested in understanding whether exposure to reception centres for asylum-seekers in Berlin affected the votes obtained by the radical right AfD in the 2019 European elections, at the electoral district level. We analyse this relationship at a very small scale based on geo-localization techniques and high-resolution spatial data. We study this in a wide range of contextual conditions, including variables such as districts' socio-economic deprivation, the share of established non-European residents, and the geographical location of the districts. Overall the findings show that exposure to reception centres in Berlin is negatively correlated with the AfD vote share. However, the results show remarkable differences between East and West Berlin and between districts characterised by different levels of socio-economic deprivation. Exposure and AfD vote shares are more strongly correlated in Western districts and in better-off districts.

This work is thus aimed at providing a practical contribution to international migration studies by offering novel methods and analyses for identifying, quantifying and understanding dynamics of migration to better shape the policies of international migration.

Keywords: International migration; Big data; Integration; Voting behaviour; Network analysis.

Publications

As main author

1. Jisu Kim, Alina Sîrbu, Fosca Giannotti, and Lorenzo Gabrielli. Digital footprints of international migration on twitter. In *International Symposium on Intelligent Data Analysis*, pages 274–286. Springer, 2020
2. Jisu Kim, Emilio Zagheni, and Ingmar Weber. Improving migration statistics using social media. In *Practitioners' Guide on Harnessing Data Innovation for Migration Policy*. IOM, 2021
3. Jisu Kim, Alina Sîrbu, Giulio Rossetti, Fosca Giannotti, and Hillel Rapoport. Home and destination attachment: study of cultural integration on twitter. *arXiv preprint arXiv:2102.11398*, 2021
4. Jisu Kim, Alina Sîrbu, Fosca Giannotti, and Giulio Rossetti. Characterising different community of twitter users: Migrants and natives. *To be submitted*, 2021

As co-author

1. Alina Sîrbu, Gennady Andrienko, Natalia Andrienko, Chiara Boldrini, Marco Conti, Fosca Giannotti, Riccardo Guidotti, Simone Bertoli, Jisu Kim, Cristina Ioana Muntean, et al. Human migration: the big data perspective. *International Journal of Data Science and Analytics*, pages 1–20, 2020
2. Andrea Pettrachin, Lorenzo Gabrielli, Jisu Kim, Sarah Ludwig-Dehm, Steffen Pötzschke, and Michele Vespe. Did exposure to refugee centres affect the electoral outcome of alternative for germany in berlin? evidence from the 2019 eu elections. *submitted to Journal of Ethnic and Migration Studies: Under review*, 2020
3. Fosca Giannotti, Jisu Kim, Giulio Rossetti, Laura Pollacci, and Alina Sîrbu. Twitter data for migration studies. In *Data Science for Migration and Mobility*.

Oxford: In prepration, 2021

Conferences and talks

1. Conference on Complex Systems (CCS) 2019, Singapore
2. International Forum on Migration Statistics (IFMS) 2020, Cairo, Egypt
3. Symposium on Intelligent Data Analysis (IDA) 2020, Konstanz, Germany
4. International Migration, Integration and Social Cohesion in Europe (IMIS-COE) 2020 in session “Migration citizenship and political participation”, Luxembourg
5. International Migration, Integration and Social Cohesion in Europe (IMIS-COE) 2020 in session: “Studying Migration from a Big Data Perspective”, Luxembourg
6. Conference of the Network Science Society (NetSci) 2020, Rome, Italy
7. Human Migration-Potential areas for combinations of Big Data Workshop at SocInfo2020, Pisa, Italy
8. Invited speaker at BD4M “Harnessing Data Innovations for Migration Policy” webinar series of International Organisation for Migration (IOM)’s Global Migration Data Analysis Centre (GMDAC)
9. Paper at the Association for Digital Humanities and Digital Culture (AIUCD) 2021

Dedicated to my family

Acknowledgments

I cannot express enough how grateful I am to have had Giorgio Fagiolo, Fosca Giannotti, Hillel Rapoport and Alina Sîrbu as my supervisors. I am thankful to have had their continuous supports and invaluable advice. My academic journey throughout the PhD programme would not have been the same without them.

I thank also our programme coordinator of the joint Data Science PhD programme, Dino Pedreschi for making all this possible. My gratitude extends to the members of the PhD board and individuals connected to the programme – Vittorio Romano, Silvia Zappulla, Camilla Sovani, Anna Monreale, Francesca Chiaromonte, Tommaso Cucinotta and Ioanna Miliou – for their kind help and support that they provided throughout the whole PhD programme.

I would also like to thank my panel members Caterina Giusti, and Tiziano Squartini for all the helpful suggestions and comments that they have provided me throughout the PhD programme. I would like to extend my thanks to Giulio Rossetti for his crucial contributions and useful comments for my researches.

I have been lucky to have had also wonderful colleagues who became important friends in my life also – Luca Insolia, Cecilia Panigutti, Giorgio Tripodi, Gevorg Yeghikyan, Tommaso Radicioni, Vasiliki Voukelatou and Elisa Ferrari. They were the cherry on top of my splendid three years in Pisa. Warmest thanks go to Luca for all the kind care and support. Last but not least, my appreciation goes out to my family for their encouragement and unwavering support all through my studies. Thank you for believing in me.

Contents

1	Introduction	15
1.1	International migration studies	15
1.2	Essence of this PhD thesis	16
1.3	Objectives of thesis	18
1.4	Thesis structure	18
2	Background	21
2.1	Key migration terms	21
2.2	Migration statistics-Traditional data sources	23
2.3	Alternative data sources for migration studies	26
2.4	Geo-tagged tweets	29
2.5	Migration network effects	33
2.6	Integration of migrants	35
2.7	Presence of migrants and Shift in voting behaviours of natives	38
3	Digital footprints of international migration on Twitter	40
3.1	Introduction	40
3.2	Related works	43
3.3	Experimental setting for data collection	44
3.4	Identifying migrants	46
3.4.1	Assigning residence	46
3.4.2	Assigning nationality	47
3.5	Evaluation	48
3.5.1	Internal validation: gold standards derived from our data	49
3.5.2	External validations: validation with ground truth data	51
3.6	Case study: topics on Twitter	52
3.7	Conclusion and future work	53
4	Home and destination attachment: study of cultural integration on Twitter	55
4.1	Introduction	55
4.2	Related works	58
4.3	The home and destination attachment indices	59
4.3.1	Data	60
4.3.2	Assigning residence and nationality to users	61
4.3.3	Detecting country-specific topics	62
4.3.4	Computing the home and destination attachment indices	64

4.4	Results	65
4.4.1	Overall distribution of DA and HA values	65
4.4.2	Language as a key factor for integration	67
4.4.3	Country-specific results	68
4.4.4	Hofstede’s cultural dimension scores and other measures	72
4.5	Discussion	75
5	Characterising different communities of Twitter users: Migrants and natives	77
5.1	Introduction	77
5.2	Related works	79
5.3	Data and labelling strategy	80
5.3.1	Data	80
5.3.2	Labelling migrants and natives	80
5.4	Twitter features	83
5.4.1	Profile information	84
5.4.2	Tweets	86
5.5	Network analysis	88
5.5.1	Properties of the network	89
5.5.2	Assortativity analysis	93
5.6	Conclusion	95
6	Presence of migrants and shift in voting behaviours of natives	97
6.1	Introduction	97
6.2	Theory and Hypotheses	102
6.3	Data and Methods	106
6.3.1	Setting	106
6.3.2	Data	108
6.3.3	Variables	111
6.3.4	Methods: Spatial Autoregressive Model or Spatial Lag Model	113
6.4	Findings	114
6.5	Conclusion	120
7	Discussions and conclusion	123
7.1	Summary and conclusion	123
7.2	Ethics and legality issues in using Twitter data	125
7.3	Policy recommendations	126
7.4	Directions for the future research	127
7.5	General conclusion	129
	Bibliography	131
	A Appendix	143

List of Figures

2-1	An example of followerwonk.com search on Twitter	29
2-2	Example of a tweet	31
2-3	Hofstede’s cultural dimensions for Italy and South Korea	37
3-1	Distribution of the number of days (left) and the number of tweets (right) observed in the data per user : on average, our users have tweeted 47 days and 82 tweets in 2018.	45
3-2	Example of calculation of the <i>floc</i> and <i>flang</i> values for a user. The calculation of $floc^{U1}$ and $flang^{U1}$ is based of the <i>floc</i> and <i>flang</i> values for the three friends, showing the distribution of tweets in various countries/languages for each.	47
3-3	Distribution of residences and nationalities of top 30 countries, for all users that possess both residence and nationality labels.	50
3-4	Comparison between the true and predicted data; the first two plots show predicted versus AIRE/EUROSTAT data on European countries. The last plot shows predicted versus AIRE data on non-European countries.	52
3-5	Stream graph: appearance of hashtags related to #Salvini from Italians across 10 selected residence countries in 2018. The discussion continuously appeared in Italy throughout the year and it became more lively employed by Italians overseas as Salvini gained more political attention.	53
4-1	Chord diagram showing migration links between countries. The colour of the chord represents the nationality of the migrants, while the width of the chord represents the number of migrants in our dataset who had the 2018 residence in the corresponding destination country. For visualisation purposes we show only 21 countries: those with at least 10 migrants.	60
4-2	Percentage of Italian emigrants in various destination countries based on AIRE and Eurostat: predicted versus ground truth data.	62
4-3	Entropy distribution in Log scale	64
4-4	Distribution of hashtags’ nationalities	64
4-5	Distribution of HA and DA values, and comparison to null model HA_0 and DA_0 . Means values are: $\bar{HA}_0 = 0.038$ and $\bar{DA}_0 = 0.024$, $\bar{HA} = 0.051$ and $\bar{DA} = 0.034$	65

4-6	Pearson correlation between home and destination attachment indexes for all the migrants in the data: correlation coefficient: -0.13 , p-value: $6.937e^{-14}$	67
4-7	Box plots showing the HA and DA distributions for a group of migrants who speak the language of the host country on the left and a group of migrants who do not speak the language of the host country on the right. Means are $\bar{H}A = 0.034$ and $\bar{D}A = 0.041$ for users who speak the destination language, and $\bar{H}A = 0.072$ and $\bar{D}A = 0.019$ for those who do not speak it.	68
4-8	Left: Box plots for the DA and HA index of immigrants in the United States. Right: Scatter plot of HA vs. DA indicating approximate integration types for immigrants in the US.	69
4-9	Left: Box plots for DA and HA for immigrants residing in the United Kingdom. Right: Scatter plot of HA vs. DA indicating approximate integration types for immigrants in the UK.	71
4-10	Left: Box plots for DA and HA for Italian nationals living abroad. Countries on x-axis are countries of residence of Italians. Bottom: Scatter plot of HA vs. DA indicating approximate integration types for Italian emigrants.	71
5-1	Distribution of top 50 nationalities of natives in log scale	81
5-2	Chord diagram on migration patterns: The number of migrants who have moved from a country to another is represented by the links. The colours represent the nationalities of migrants. We show only countries with at least 10 migrants for the visualisation purpose.	82
5-3	Distribution of DA & HA for migrants and natives	84
5-4	Distributions of profile features: number of days since the account was created until 2018, number of followers, number of friends, and number of tweets published (statuses).	85
5-5	Distribution of tweet locations and languages	87
5-6	Distribution of tweet locations and languages of friends	87
5-7	Top 10 hashtags used by migrants and natives	88
5-8	Degree distribution of the network.	89
5-9	Centrality measures of the network.	90
5-10	Correlation between different centrality measures for network	91
5-11	Summary of labels of users for top 10 central users by different centrality measures.	92
5-12	Stacked histogram of local assortativity: From the top we have local assortativity by nationality, by residence and by migrant/native label. Please note that the histograms are stacked, therefore there is no overlap between the plot bars.	95
6-1	Comparison of covariate means: Panel a. AfD shares of votes in the 2014 EU elections. Panel b. Socio-economic Deprivation (data from 2016). Panel c. Concentration of Established non-European Migrants	109
6-2	Description of Data	110

6-3	Visualisation of data points: Panel a. Visualisation of interaction between EA and SED (model 3), Panel b. Visualisation of interaction between EA and SED (West only). Panel c. Visualisation of interaction between EA and SED (East only). Panel d. Visualisation of interaction between EF and total capacity.	118
A-1	Testing for spatial dependence in the OLS residuals with Exposure to reception facilities: presence of clusters of residuals across different districts in Berlin in the OLS regression model. The colour indicates whether the residual is positive or negative.	144
A-2	Spatial dependence in the SAR model with Exposure to reception facilities: No clusters of residuals in the SAR model. The colour indicates whether the residual is positive or negative.	144
A-3	Testing for spatial dependence in the OLS residuals with Exposure to asylum seekers: presence of clusters of residuals across different districts in Berlin in the OLS regression model. The colour indicates whether the residual is positive or negative.	144
A-4	Spatial dependence in the SAR model with Exposure to asylum seekers: No clusters of residuals in the SAR model. The colour indicates whether the residual is positive or negative.	145

List of Tables

3.1	Average precision, recall and F1 scores, together with scores for the top 7 residences in terms of support size.	49
3.2	Average precision, recall and F1 scores for top 8 nationalities in terms of support numbers	50
4.1	Theories of integration and their relation to HA and DA.	56
4.2	Spearman correlation table for immigrants in the United States: Vigdor’s assimilation scores and DA & HA indices. Significance levels are marked with *** p-value <0.01, ** p-value <0.05, * p-value <0.1.	70
4.3	Correlation table for HA & DA and Hofstede’s cultural dimension scores for migrants at individual level. Significance levels are marked with *** p-value <0.01, ** p-value <0.05, * p-value <0.1.	72
4.4	Correlation table for HA & DA and Hofstede’s cultural dimension scores. Correlation with HA is computed after grouping migrants by nationality, while correlation with DA is computed after grouping by residence. Significance levels are marked with *** p-value <0.01, ** p-value <0.05, * p-value <0.1.	74
6.1	Descriptive Statistics of all Variables	113
6.2	Descriptive Statistics of all Variables by Region	113
6.3	SAR Model 1 (computed with the EF variable). Dependent variable: share of votes for the AfD in EU Elections 2019. N=489. Sources: EU election results for Berlin 2019, reception facility address list, D4I data, socioeconomic data; own calculations. Coefficients, Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1.	115
6.4	SAR Model 2 (computed with the EA variable). Dependent variable: share of votes for the AfD in EU Elections 2019. N=489.	115
6.5	Additional Models (SAR models, the effect reported is the total effect). Dependent variable: share of votes for the AfD in EU Elections 2019. N=489. Sources: EU election results for Berlin 2019, reception facility address list, D4I data, socioeconomic data; own calculations; Coefficients, Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1.	118
6.6	SAR Model 8. Dependent variable: share of votes for the AfD in EU Elections 2019. N=318. (electoral districts with at least one reception centre located within their borders or at max 1,000 meters from their border).	120

A.1 OLS Models (using different exposure measures)	143
--	-----

“How well we understand and respond to the migration context will have ramifications for many years to come”

António Vitorino-IOM director

General

Chapter 1

Introduction

This series of *Essays on International Migration using Big Data Analytics* is a result of my PhD works in the joint PhD program of Data Science with Scuola Normale Superiore of Pisa (SNS), University of Pisa (UniPi), National Research Council (CNR), Sant’Anna School of Advanced Studies (SSSA), and IMT School for Advanced Studies Lucca (IMT).

1.1 International migration studies

International migration studies do not originate from a single field of study but from sociology, economics, history, law, demography, and anthropology. Depending on the field of study, the perspective in which they study migration varies. However one commonality for each field is that migration has important implications.

Regardless of the angle, migrants are human capital who carry knowledge, skills and ideas and are able to produce and contribute to the labour market [64]. Hence, they are considered as disseminators of ideas, diversity and knowledge from one society to another. Not to mention that they are also drivers of globalisation, and development.

Migrants facilitate exchanges between the home and host country through diffusion of preferences and knowledge [77, 137, 138]. Indeed, migration is one of the economic linkage factors that helps to increase connectivity across countries [66]. Other linkage factors include trade and foreign direct investment (FDI) which are

interconnected to each other along with the migration factor. Most importantly, the critical channel for their inter-connectivity is the transmission of knowledge. Migrants possess different types of knowledge than the natives. Among all types, tacit knowledge is the type of knowledge that natives cannot obtain without a direct interaction with immigrants [11]. Hence, acquirement of such knowledge through immigrants is an important vehicle for further economic growth.

As pointed out, study of migration involves consideration of diverse factors. Nowadays, it has also captured the interests of physicists, mathematicians and computer scientists as new theories of complex systems and big data analysis tools have been introduced. As this proves, the study of migration have become a very fertile ground for interdisciplinary studies applying new methods. Considering these crucial aspects, this PhD thesis attempts to explore, compute and understand various migration issues that emerge in different stages of trajectory, by using data science as a new tool for international migration studies.

In what follows in this chapter, we begin by providing a comprehensive summary of this work, followed by objectives and structure of the thesis, including brief abstracts of each chapter.

1.2 Essence of this PhD thesis

What is this thesis about?

Migration is a key driver of demographic and social change, as well as an often hotly debated political issue. Over the past years, an unprecedented displacement of individuals has attracted attention from policymakers, researchers, as well as the general public. At the end of 2019, about 272 million people migrated [2] and about 79.5 million people were forced to dislocate from their homes due to conflicts^a. These numbers have continued to increase till today. In order to cope with the complexity of the matter, the need for up-to-date and rich data to better monitor and manage the situation has become clear. As we live in the digital age, we are provided with complementary and pas-

sively collected data sources from social media platforms, call detailed records and internet activities. Some of them are freely available and enable us to gather real-time data. Furthermore, they provide large scale data at a granular level which enables the study of phenomena hard to study using more coarse-grained data. Together with the complexity of migration phenomenon and availability of big data, analytical tools need to be updated as well.

These are the most important motivations of this PhD thesis. This thesis aims to disentangle the complexity of migration phenomenon by improving existing techniques and proposing different perspectives on various aspects of migration by employing *data analytics as a new tool for migration studies*. Furthermore, it aims to contribute to the empirical analyses of this field of literature, emphasising the implications of complex-network approaches for cross-country comparisons, but also for selected country case-studies.

To be more specific, we propose to guide readers of this thesis through four different phases of migration trajectories that are related to intricate and emerging migration issues;

- *Journey*: tracing digital footprints of migrants
- *Stay*: cultural change through migration
- *Interaction (After arrival)*: digital behaviour of migrants and natives
- *Influence*: causal effect of presence of migrants on voting behaviours of natives

The first three phases of the trajectory are explored using new data-driven models and algorithms by employing social big data to introduce new evidences on migration network effects in a more realistic way. Instead, the last phase proposes to employ a novel data to find causal effects between migration and voting behaviours.

^a<https://www.unhcr.org/globaltrends2019/>

1.3 Objectives of thesis

In the previous section, we introduced the essence of this PhD thesis. In this section, we now present research questions closely related to the four different but relevant phases of migration trajectories.

Journey-International migration statistics

Can social big data be used to measure stocks of international migration?

Stay-Integration of international migrants

Can we quantify how much migrants culturally integrate in the destination country? and Do migrants loose connections to their home country?

Interaction-Migrants and natives on Twitter

What are the distinctive characteristics and behaviours of migrants and natives on Twitter?

Influence-Presence of migrants and their effect on voting behaviours of natives

Did exposure to asylum-seekers and refugees have an impact on electoral support for radical right parties?

This thesis will address these research questions in the following chapters by employing innovative data sources and data-driven models, sometimes by relying on existing works and theories but primarily to overcome limitations of traditional methodologies.

1.4 Thesis structure

In this section, we provide a synopsis of each chapter in this PhD thesis. This is a cumulative thesis consisting of four main contributions produced during the three year program of PhD in Data Science (Chapters 3 4, 5, 6), along with background and conclusions chapters.

Chapter 2 - Background In this chapter, we cover extensive literature review

of both international migration studies and big data science. We begin with key migration terms, followed by migration statistics on both traditional and alternative data. We also go through existing works related to different phases of migration trajectories: migration network effects, integration of migrants and shift in voting behaviours of natives by the presence of migrants. This work was partly published in [145].

Chapter 3 - Digital footprints of international migration on Twitter In

this chapter [96], a generic methodology is developed to identify migrants within the Twitter population. This describes a migrant as a person who has the current residence different from the nationality. The residence is defined as the location where a user spends most of his/her time in a certain year. The nationality is inferred from linguistic and social connections to a migrant's country of origin. The methodology is validated first with an internal gold standard dataset and second with two official statistics.

Chapter 4 - Home and destination attachment: study of cultural integration on Twitter

This chapter [98] is built upon international migration statistics created from the previous chapter. We study cultural integration of migrants described using two dimensions: the preservation of links to the home country and culture, i.e. *home attachment*, and creation of new links and adoption of cultural traits from the new residence country, i.e. *destination attachment*. We introduce a means to quantify these two aspects based on Twitter data. The home and destination attachment indexes are compared with various elements such as language proximity, distance between countries and also with Hofstede's cultural dimension scores.

Chapter 5 - Characterising different communities of Twitter users In this

chapter [97], we study characteristics and behaviours of migrants and natives on Twitter. To do so, we perform general assessment of features including profiles and tweets, and an extensive network analysis including centrality and assortativity measures.

Chapter 6 -Presence of migrants and shift in voting behaviours of natives

This chapter [128] examines whether exposure to reception centres for asylum-seekers in Berlin affected the votes obtained by the radical right AfD in the 2019 European elections, at the electoral district level. We make a significant contribution to the debate about the electoral consequences of the ‘refugee crisis’ and the impact of exposure to migrants on natives’ voting behaviours. We aim to fill two major gaps in the existing literature. First, we analyse this relationship at a very small scale, adopting an innovative methodological approach, based on geo-localization techniques and high-resolution spatial statistics. Second, we analyse how exposure to refugee centres is related to vote shares for the radical right in a wide range of contextual conditions, including variables such as districts’ socio-economic deprivation, the share of established non-European residents, the geographical location of the districts.

Chapter 7 - Conclusion This chapter concludes this thesis with a summary of the main research questions and findings. We also devote a section to mention ethics and legality issues involved in using Twitter data. We present also actions took to address the issues. We then evaluate the values of the works presented in this thesis in policy recommendations. We also present limitations that exist in the studies and future directions of research to overcome the limitations. We then end this thesis with a general conclusion.

“Everyone has the right to a nationality”

Universal Declaration of Human
Rights, Article 15

Chapter 2

Background

In this chapter, an overview of the existing literature on both economics and computer science is introduced. Firstly, we provide definitions of key migration terms. Secondly, we describe traditional data sources used in migration studies. Next, different social big data and methods in the migration studies used in the computer science literature are presented. In the section 2.4, description of geo-tagged tweets and methods to obtain data is provided. In the sections that follows, the effects of migration network in the economics literature, theories of acculturation in both sociology and psychology and studies of political economy on voting behaviours are discussed.

2.1 Key migration terms

Before going into details of related international migration studies, key migration terms need to be clarified. Here, we provide definitions of migration terms in accordance with official definitions provided by the international conventions and recommendations made by the United Nations (UN).

Country of destination (host country) From the perspective of a migrant, the country of destination is the country of arrival that is different from the country of origin.

Country of origin (home country) In the context of migration, the country

of origin is the country of nationality where the usual residence was, before migration took place.

Country of birth The country in which an individual was born in.

Immigrant “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 period of stay determines whether the immigrant is a long-term immigrant or a short-term immigrant. Any period under twelve months is considered a short-term migrant, whereas any period over twelve months is considered a long-term migrant¹.

International migrant “Any person who changes his or her country of usual residence²”. Travels that concern holidays or business purposes do not fit in this definition as usual residence do not change.

Usual residence “The geographical place where the enumerated person usually resides” - the concept used in censuses³.

Emigrant From the perspective of the country of origin, any person who leave to a country other than that of his or her country of nationality is considered as a emigrant.

International migrant stock refers to “the total number of international migrants present in a given country at a particular point in time⁴”.

International migrant flow refers to “the number of migrants entering and leaving (inflow and outflow) a given country over the course of a specific period, usually one calendar year⁵.”

Asylum seeker An individual who seeks international protection for reasons such as wars, conflicts, violence or persecution as one cannot return to home safely⁶.

¹United Nations Department of Economic and Social Affairs (UNDESA), Recommendations on Statistics of International Migration, Revision 1 (1998) para. 36.

²Idem. para. 32.

³Idem. para. 33.

⁴United Nations Department for Economic and Social Affairs Statistical Division (UN SD), Handbook on Measuring International Migration through Population Censuses, p.15, 2017.

⁵Idem, p.10.

⁶United Nations High Commissioner for Refugees (UNHCR), <https://www.unrefugees.org/refugee-facts/what-is-a-refugee/>.

An asylum seeker applies to be recognised as a refugee in the country of arrival. The asylum seeker may or may not be recognised as a refugee.

Refugee A refugee is an individual who fled from his or her country for reasons such as wars, conflicts, violence or persecution, preventing for him or her to return home. A refugee was an asylum seeker before he or she was recognised as a refugee⁷.

2.2 Migration statistics-Traditional data sources

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 organisations to improve quality and harmonise traditional data sources [44, 139, 118]. International organisations such as the United Nations provide also guidelines and suggestions⁸ 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 Nations⁹, countries should collect the data every year that ends with zero in order

⁷<https://www.unrefugees.org/refugee-facts/what-is-a-refugee/>.

⁸Recommendations on Statistics of International Migration, Revision1(p.113). United Nations (UN), 1998.

⁹Idem.

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 billion¹⁰, 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 [141, 54].

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 system¹¹, 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 country¹². 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 [55], the WorldBank [154], Organisation for Economic

¹⁰"The Statistics Portal." Statista. Retrieved from www.statista.com

¹¹"Sources and comparability of migration statistics". Organisation for Economic Co-operation and Development (OECD), Retrieved from <https://www.oecd.org/migration/mig/43180015.pdf>.

¹²Those born outside of Europe

Co-operation and Development (OECD) [153] and other local authorities and research institutions [88, 86, 87, 58, 53]. 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 challenge¹³ 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 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

¹³Data Challenge on Integration of Migrants in Cities (D4I), <https://bluehub.jrc.ec.europa.eu/datachallenge/>

information about asylum seekers¹⁴ and refugees¹⁵. 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 [54]. 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 [133] 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*¹⁶ firstly introduced by UNECE¹⁷ to establish comparability across different nations' data (see for instance [131, 132, 44]). 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.

2.3 Alternative data sources for migration studies

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

¹⁴Asylum seekers are individuals who seek to obtain refugee status

¹⁵Individuals with subsidiary protection are also referred as refugees

¹⁶It 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.

¹⁷United Nations Economic Commission for Europe

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, [161, 76, 116, 96]).

What is Big data?

The definition of Big data is not uniform [60]. Nevertheless, one of the most popular definitions provided by [105] is *high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation*^a To be more specific, the meanings of 3 Vs are:

- Volume: As the name suggests, the size of the data is *Big*, hence the volume of the data.
- Velocity: Big data such as Twitter allow us to stream data at real-time. The rate at which we obtain data is faster than the traditional data sources.
- Variety: Traditional data are mostly structured data. Big data, on the other hand, come in various forms. It can be videos, photos, texts, and audios. It requires a thorough data processing before extracting information/knowledge from it.

^a<https://www.gartner.com/en/information-technology/glossary/big-data>

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 [161], the authors infer migration patterns from Twitter data by looking at where the tweets were posted. Other studies like [116] 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 localise 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 geo-localised Twitter data, with the aim of quantifying diversity in communities, by computing a superdiversity index [129]. 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 [9]. 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.

Skype Ego networks data can also be used to explain international migration patterns [94]. 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.

Big data can also be used to study movements of individuals in the time of crisis. For instance, [15] 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 [27], they study whether the GTI¹⁸ can now-cast the immigrants' journey. However, as authors point out, not every search

¹⁸Google Trend Index, <https://trends.google.com/trends/>



Figure 2-1: An example of followerwonk.com search on Twitter

means that searchers have intention to migrate. To address this issue, they compare Gallup World Poll data¹⁹ with the results obtained with GTI data. The Gallup data is a survey done on more than 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.

2.4 Geo-tagged tweets

Twitter is a freely available data source that can be accessed using an application programming interface (API)²⁰. There are two main methods of accessing data using the Twitter API, which are through the search API or the streaming API. Both APIs return data as JSON objects²¹, which is easy to store and manipulate data. The search API enables us to collect existing tweets and profiles of users. The search can be done on a specific user using either the user ID or the user name. This

¹⁹<http://gallup.com>

²⁰<https://developer.twitter.com/en/docs/twitter-api>

²¹<https://www.json.org/json-en.html>

returns a user object that contains information about individual users' profile such as when the account was created, number of tweets, followers and friends as well as the location that the user declares to be in and more. Otherwise, it can be done on specific keywords or geolocations embedded on tweets. The geolocations can be specified at a country level, place, bounding box or at a point radius from 0.01 km up to about 40 km around specific coordinates²². On the other hand, the streaming API allows us to gather 1% random samples of all new public tweets in near real-time. The streaming API also enables us to specify filter criteria (e.g. keywords, geolocations, and user ID or user names). However, different from the search API, it returns tweets matching the filter criteria as soon as matching tweets are created. The returned content of matching tweet objects includes the tweet text, location information (where present), the language in which the tweet was written in, when the tweet was created, and additional information, such as whether the tweet was part of a thread. It also contains the entity object which lists tweet contents such as hashtags, URL and mentioned IDs. To collect small amounts of data, instead of using the APIs, there are also websites to search for tweets, e.g. on Twitter directly²³, or where detailed searches for particular users can be issued, e.g. on Followerwonk²⁴.

Having obtained the data, there are few notes of caution to consider. First of all, only a small percentage of tweets come with geolocations based on the user opting-in to share their exact position. For instance, [117] showed that only about 3.2% of tweets from the Streaming API are geo-tagged. This means that any given user is unlikely to have geotagged tweets and that, correspondingly, there are challenges related to self-selection bias that need to be addressed. Second, it requires an effort to clean and process the data. Often, the tweets are not directly usable as they are noisy and/or incomplete. For instance, tweets contain repeated characters (e.g. 'wooooooow'), typos, or internet slangs that are not familiar to everyone, and that also pose challenges to standard natural language processing (NLP) tools. Some tweets may also be incomplete in that they require additional context, such

²²<https://developer.twitter.com/en/docs/tutorials/filtering-tweets-by-location>

²³<https://twitter.com/explore>

²⁴<https://followerwonk.com>

as a thread of tweets, to make sense of. Cleaning and removing data may result in considerable loss of data. In addition, there are bots or spam accounts that introduce additional data quality issues. It is also important to make sure that identifying migration events is not interfered with misleading activities. Another limitation is that Twitter does not provide user attributes such as education or income level, which are often helpful for more in-depth migration studies. Nevertheless, certain characteristics, such as age, ethnicity or sex, can often be inferred with reasonable accuracy using the profile image [84, 161]. Lastly, there are privacy issues. It is vital to make sure that no personal information from data is published even if Twitter data is openly available. This requires a proper infrastructure where data can be safely stored in a secured server. In other ways, only the “dehydrated” data can be shared for research or archival purposes. This requires data to be in the form of unique IDs which then can be re-hydrated, in other words, restored to the original data. This gives the user a chance to “opt out” of subsequent studies by deleting their tweet/account. Also note that most of Twitter is public, and accessible by anybody, an individual user might not expect researchers to algorithmically collect and analyse their tweets. How to best address these expectations of data use, which are separate from legal considerations, remains a challenge with answers depending on the specific context.

Tweet object




Figure 2-2: Example of a tweet

When API returns tweet data in a JSON object, it contains a unique ID, an author, a text, a timestamp, and a geo-tag if enabled by the user that are embedded in each tweet. Tweets also produce entity objects that synthesise tweet contents such as hashtags, URLs, and mentioned IDs. Other than the mentioned features so far, tweet object contains over 150 features. For the purpose of the simplicity, here we present only the features that are relevant and that have been used in the literature of migration studies.

lang. This field provides a code-language identifier corresponding to the machine-detected language of the tweet text, if identifiable.

User. The tweet object also produces user object of a user that is related to the tweet. Like the tweet object, the user object also contains many features. Few of the features useful for migration studies are; `user id`, `screen name`, `location`, `description`, `number of friends/tweets/followers`, `verified account` and more^a. `location` represents the user-defined location on the account's profile. `verified account` is an account that is notable and active on either on Twitter or off-Twitter.

Entities. Contains contents of a tweet, such as `#hashtags`, mentions, `$symbols`, URLs, and media.

Extended Entities. Extended Entities contains media contents (photo, video, or GIF). Currently, up to 4 photos can be uploaded on Twitter^b.

Place (Geo). If tweet is geo-tagged, there will be a `Place` object. This includes `coordinates`; geographic location of a tweet as reported by the user application [`longitude`, `latitude`]. Its subgroup object includes also `place_type` for instance, `city`, `country` (`country_code`), and `bounding_box` of coordinates.

Note that there is currently an early access to Twitter API v2^c. The version 2 has some of new features including new and more detailed data objects such as conversation threading, poll results in tweets and pinned tweets on profiles and more which need to be explored further.

^aNote that features like `lang`, `geo_enabled` and `time zone` are set to null since they are deprecated

^b<https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview/extended-entities-object#intro>

^c<https://developer.twitter.com/en/docs/twitter-api/early-access>

2.5 Migration network effects

In the international and development economics' literature, the effects of migration network is well studied. The literature validates that, indeed, the migration network effect exists and it creates positive externality on both the host and home countries in terms of economic growth (see for instance [11, 125, 101, 77, 33, 67, 138]). But as it is also pointed out in the literature, studying the link between the migration and externalities is not an easy task as the causality between migration and externality is difficult to separate. For example, in the studies of migration and trade relationships, one cannot know whether the migrants are attracted to areas with more economic opportunities or whether the existence of migrants in the location stimulated the trade. To deal with this problem, many papers have analysed specific case studies using various methods to show that migrants are partially responsible for growth along with other economic factors of globalisation.

For the migration network to play a role, there are two main channels, which are information and preference effects [67]. First of all, information effect tell us that migrants' previous knowledge about their country of origin can reduce transaction cost associated in economic activities between the host and home country. Migrants do not only speak the language of their country of origin but also have contacts back in their country. They are also familiar with the culture, business conditions as well as the institutional context of the country, which can facilitate processes of economic exchanges between the two countries. [138] studied this effect by analysing ethnic Chinese networks residing in overseas. According to their findings, the network has an important impact on bilateral trade through information channel and better matching and referral services. It was shown that the Chinese network increased bilateral trade by 60%. The amount of information depends on the level of education, professional background as well as the length of stay in their home country. An immigrant who was a businessman back in his home country could better enhance business opportunities between the host and home country. As such, the study of [101] concluded that skilled immigrants, especially, have stronger effects on trade.

On the other hand, preference effect tells us that migrants as well as natives influ-

enced by migrants can generate an additional demand for migrants' home country's goods and services from the host country. From the home country's point of view, the export can be increased thanks to the presence of immigrants in overseas. Such example has been studied by [77] using an augmented gravity equation to the case study of Canada. They found that in Canada the preference effect is somewhat stronger than the information effect. The evidence showed that import creation effect is stronger than the exports; they observed that a 10 percent increase in immigrants is associated with a 3 percent increase in imports and a 1 percent increase in exports. Also [29] found that in Italy, a 10 per cent increase in immigrant stocks leads to a 5.94 per cent increase in import flows. They explained that the effect of immigrants on trade is greater on imports than exports because the export elasticity to immigrants is very low. They stated that in the case of Italy, it is more the transplanted home-bias effect in consumption, i.e., preference effect rather than the information effect that enhanced trade opportunities.

Another interesting paper is written by [125], where the authors employed instrumental variable method using a natural experiment case of Vietnamese refugees in the United States. The distribution of Vietnamese refugees at the time of their arrival showed no correlation with the economic factors of the state. The causality between migration and trade was disentangled. They were able to find a creation of trade between the two countries as the trade barrier was removed in 1994. Equally, in [95], we used the early distribution of Korean immigrants in 1960 (i.e., Korean War brides), as the instrumental variable for the distribution of Korean immigrants in 1995 across states. The Korean War brides were randomly allocated, regardless of the economic conditions of the states. The establishment of a clear causal effect from Korean immigration in 1995 to the U.S states' exports to Korea between 1995 and 2010 showed that Korean immigration has contributed to the increase in the U.S states' exports to Korea. More precisely, the results showed that the elasticity of exports is 0.32; a 10 percent increase in the number of Korean immigrants increases the U.S states' exports to Korea by about 3.2 percent. The result was interpreted by the network effect.

2.6 Integration of migrants

When two different cultures come into contact, adjustment of culture takes place in a society. This often happens to migrants who arrive in a country where their culture is different from the destination country's society. This is called *cultural integration* (or *acculturation*), "the process of group and individual changes in culture and behaviour that result from intercultural contact" [23]. Traditionally, there have been two main streams of integration models; unidimensional and bidimensional. The uni-dimensional model supposes that culture of minority will progressively be weakened, to be absorbed to the majority's culture [65]. The bidimensional model supposes that one's own culture can be maintained but can also adopt new culture. There exist many studies that analyse different forms of integration. For this reason, the definition of culture, or cultural identity also varies slightly from one study to another. For instance, the cultural identity refers to "complex set of beliefs and attitudes that people have about themselves in relation to their culture group membership" [22]. Another way to define culture would be "the beliefs, values, social perspective, traditions, customs, and language shared within a group" [140].

One of the most known study of acculturation using bidimensional model is the "four-fold model of acculturation" [21, 22]. Berry studied acculturation strategies from two different perspectives, one from the perspective of minorities and the other from the perspective of the dominant society. There are four strategies that minority of the society can take; *Assimilation*, *Integration*, *Separation* and *Marginalisation*. These strategies depend on two concepts which are, first "Is it considered to be of value to maintain one's identity and characteristics?" and second, "Is it considered to be of value to maintain relationships with larger society?". If one considers to be of value to maintain one's identity and relationships with larger society, *integration* is the strategy to take. Instead, if one does not consider these values important, *marginalisation* is the suitable strategy. Furthermore, if maintaining one's identity is not considered to be of value but is valuable to maintain relationships with larger society, *assimilation* is the strategy to opt for and by contrast, *separation*. From dominant society's perspective, these strategies become *Melting pot*, *Multicul-*

turalism, Segregation and Exclusion, in the same order. Studies have shown that *Multiculturalism* or *Integration* is the most ideal strategy for both the host country and migrants [24, 56]. The network effects mentioned in the previous section is also proofs of multiculturalism.

However, some studies suggest that the four-fold model of acculturation has limitations. One of which is that it treats acculturation as a static phenomenon. [103] addresses this and includes *Alternation* which supposes that individuals can alter their behaviour adapting to either their original culture or to the one of dominant society, depending on the situation. In the commentary of [82], he also mentions necessity of differentiated approach, for instance, to take into consideration that migrants can also consider sub-group society's culture instead of the majority.

While many studies provide fundamental aspects of integration, studies that quantify these aspects are rather rare due to mainly availability of data. One of the most pro-founding study is done by [80, 81] who accidentally obtained survey of employee values across different world bases by International Business Machines Corporation (IBM) between 1967 and 1973. Using the data, he constructed different cultural dimensions that can be compared from one country to another to understand different cultural values. In his initial study, he designed four cultural dimensions; power, masculinity, individualism, and uncertainty avoidance²⁵. In his later studies, long-term orientation and indulgence²⁶ were added [81]. Using his cultural dimensions, we can observe that, for instance, Italy and South Korea share similarity in power, uncertainty, and indulgence cultural dimensions²⁷ as shown in the figure 2-3. On the other hand, the rest of the dimensions show clear differences. Italians are more individualists, and are driven by competition. Instead, Koreans are more pragmatic than Italians.

Another study that quantifies assimilation is done by [157] on immigrants in the

²⁵Power distance: whether a hierarchical order is accepted among people. Masculinity (vs. Femininity): whether the country is driven by competition, achievement and success. Individualism (vs. Collectivism): how “me-centred” the people are in the country. Uncertainty avoidance: how comfortable people are when faced with uncomfortable and ambiguous situations.

²⁶Long-term (vs. Short term) orientation: whether the importance is given to what has been done already or to the future, Indulgence (vs. restraint): how strict the people are towards their desires.

²⁷<https://www.hofstede-insights.com/country-comparison/italy,south-korea/>

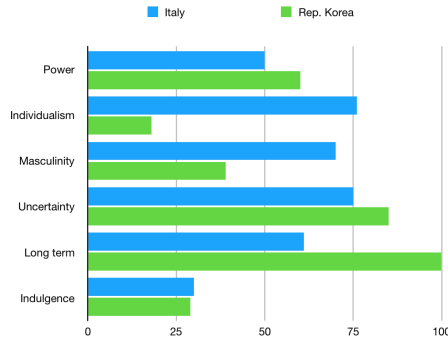


Figure 2-3: Hofstede’s cultural dimensions for Italy and South Korea

United States. Different from [80], he quantified similarities between foreign-borns and natives on three different aspects, i.e., economic, culture, and civic based on Census and American Community Survey data (ACS). To be more specific, economic assimilation index takes employment status, income, education attainment and home ownership into a consideration. The cultural assimilation index looks at intermarriage, the ability to speak English, and the number of children and marital status. The civic index observes military service and citizenship. He also included a composite index which gives an overall score of all three factors. His assimilation index made several interesting observations on immigrants. For instance, he observed that immigrants in the past years have assimilated more rapidly than the immigrants who have arrived a century ago. He also remarked that all of three factors of assimilation do not necessarily happen concurrently.

Thanks to the availability of new innovative data, study of cultural integration has been detailed to more specific areas. [150] is one of the firsts to employ Facebook data to study cultural assimilation. They looked into music preferences of Mexican immigrants to compare it with preferences of the natives in the United States using *likes* on Facebook. They further extended their analysis to understand the differences in assimilation scores between ethnicity and generations across different demographic groups. Using the same data source, [156] have extended the work of [150] to measure cultural distances between countries. In their work, they analysed the diffusion of Brazilian cuisine around the world and estimated cultural distance between countries. They computed a so called interest entropy to measure how the interests are distributed around the world. They showed that the pres-

ence of Brazilian migrants explains, in part, the presence of interests in Brazilian cuisine in the host country. Other related factors were geographical proximity, and linguistic similarity.

2.7 Presence of migrants and Shift in voting behaviours of natives

One of the reasons why migration is hotly debated today is because of rise in anti-immigrant parties. The presence of migrants impacts host country in various fabrics. Natives worry about impact of immigrants on factors like employment, wage, government spending, quality of amenities and more, though some evidences suggest small or no effects (see for instance [39, 34, 147, 57]). For what it is worth, natives' concerns determine the outcome of elections.

Many studies have analysed the relationship between immigration and right-wing parties in different country case scenarios. Nevertheless, no uniform conclusion have been reached yet. Some studies found negative relationship between immigration and votes for right-wing parties and the others found positive impact. The main explanation for negative impact is the “contact theory” which states that inter-group contact can effectively reduce prejudice between majority and minority group members [5]. The opposite is supported by the “group conflict theory” which suggests that conflict between locals and migrants emerges over scarce resources such as access to jobs, housing, public services or education [32].

The studies looked at different countries and cases but also different types of migrants; refugees or immigrants. In studies that looked at share of immigrant, the impact was generally positive [13, 50, 74, 14, 158]. For instance, [158] found that 1 percentage point of Polish immigrants leads up to 3.12 percentage points higher vote for Brexit. Due to the “recent” outbreak of refugee crisis, studies have analysed impact of exposure to refugees on voting behaviours, resulting in rather diverse conclusions. In some of the works studying the exposure to refugees, the impact was mostly negative [49, 149, 155]. For instance, [155] studied whether the resettlement of refugees in Calais Jungle to temporary migrant-centers (CAOs) in

France affected the results of 2017 presidential election. The exposure to refugees was found to reduce vote for Marine Le Pen. However, they also found that this effect can potentially turn positive once a municipality receives a larger number of refugees. Different from other studies, [46] found very strong positive impact of share of refugees on the Golden Dawn's vote share in Aegean islands in Greece.

“Inequalities and discrimination
cause damage to all of society”

Michelle Bachelet, UN High
Commissioner for Human Rights

Chapter 3

Digital footprints of international migration on Twitter

3.1 Introduction

Understanding where migrants are is an important topic because it touches upon multidimensional aspects of the sending and receiving countries' society. It is not only the demographic fabric of countries but also labour market conditions, as well as economic conditions that may alter due to demographic adjustment. Understanding their allocation is essential for both policy makers and researchers to bring the best of its effects.

Official data such as census, survey and administrative data have been traditionally the main data source to study migration. However, these data have some limitations [145]. They are inconsistent across different nations because countries employ different definitions of a migrant. Moreover, collecting traditional data is costly and time consuming, thus tracking instantaneous stocks of migrants becomes difficult. This becomes even harder when tracking emigrants because of the lack of motivation from citizens to declare their departure.

In recent years, however, we are provided with other alternative data sources for migration. The availability of social big data allows us to study social behaviours both at large scale and at a granular level, and to peek into real-world phenomena. Although known to suffer from other types of issues, such as selection bias, these

data could bring complementary value to standard statistics.

Here, we propose a method to identify migrants based on Twitter data, to be used in further analyses. According to the official definition, a migrant¹ is “a person who moves to a country other than that of his or her usual residence for a period of at least a year”. In the context of Twitter, we define a migrant as “*a person who has the current residence different from the nationality.*”

Following this definition, we performed a two step analysis. First, we estimated the current residence for users by examining location information from tweets. The residence is defined as the country where the user spends most of the time in a year. Second, we estimated nationality, by considering the social network of users. In the international literature, nationality is defined as a relationship between a state and an individual, with rights and duties on both sides [72, 1]. Related concepts are ethnicity - in terms of cultural features - and citizenship - in terms of political life. In this paper, we employ the term nationality to define the ensemble of features that make a person feel like they belong to a certain country [47, 8]. This could be the country where a person was born, raised and/or lived most of their lives. By comparing labels of residence and nationality of a user, we were able to understand whether the person has moved from their home country to a host country, and thus if they are a migrant. We validated our estimation internally, from the data itself, and externally, with two official datasets (Italian register and Eurostat data).

One of the advantages of our methodology is that it is generic enough to allow for identification of both immigrants and emigrants. We also overcome one of the limitations of traditional data by setting up a uniform definition of a migrant across different countries. Furthermore, our definition of a migrant is very close to the official definition. We establish the fact that a person has spent a significant period at the current location. Also, we eliminate visitors or short-term stays that do not follow the definition of a migrant. This is also validated by the comparison with official datasets. Another advantage of our method is the fact that it uses only very basic features from the Twitter data: location, language and network information.

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

This is useful since the settings of the freely available Twitter API change constantly. Some of the user attributes that the existing literature use to estimate nationality are no longer available. In addition, we make use of unknown locations of tweets by examining whether they intersect with identified locations. By doing so, we do not neglect any information provided by the tweets from unknown locations which later provide useful information on trending topics of Italian emigrants overseas.

One of the issues with our method is that the migrants that we observed are selected from the Twitter population, and not from the general world population, and it is known that some demographic groups are missing. Nevertheless, we believe that studying the Twitter migrant population can provide important insight into migration phenomena, even if some findings may not apply to the other demographic groups that are not represented in the data.

It is important to note that tracking individual migrants is not the objective of our study, but it is only an intermediate stage to enable further analyses. We simply perform user classification to identify migrants among users in our data, and then aggregate the findings. Further studies we envision are aimed at devising new population-level indices useful to evaluate and improve the quality of life of migrants, through targeted evidence-based policy making. No individual personal information nor migration status is released at any stage during the current analysis, nor in any population-level analysis, which is performed following the highest ethical and privacy standards.

The rest of the paper is organised as follows. In the next section we describe related work that studies migration using big data. In Section 3.3, we provide details of the experimental setting for data collection as well as data pre-processing. We then explain our identification strategy for both residence and nationality in Section 3.4. In Section 3.5, we evaluate our estimation using both internal and external data. Section 3.6 covers a possible application of our method on studying trending topics among Italian emigrants, while Section 3.7 concludes the paper.

3.2 Related works

In the past few years, there have been several works on migration studies using social big data. Most of these employed Twitter data but Facebook, Skype, Email as well as Call Detail Record (CDR) data have also been used to study both international and internal migration [104, 26, 162, 94, 163]. Here, we focus on studies that have employed freely available data. The definition of a migrant varied from one work to another depending on the purpose of the study and the nature of the dataset. Thus, the definitions provided fit under different types of migration such as refugees, internal migrants, seasonal migrants or even visitors.

One example of using Twitter to observe migration flows is [161]. They defined residence as the country where the tweets were most frequently sent out for periods of four months. If one’s residence changed in the following four months period, it was considered that the person has moved. In a more recent work, [111] measure migration flows from Venezuela to neighbouring countries between 2015 and 2019. They look at the bounding boxes and country labels provided by the tweets and identified the most common country of tweets posted monthly. Their definition of a migrant was “any individual leaving Venezuela during the time window of observation” which was observed when an identified Venezuelan resident appeared for the first time in a different country. Our definition of residence is somewhat similar to these works. However, unlike them, we are measuring stocks of migrants, and not flows. Thus, we take into account the aspect of duration of stay. This naturally eliminates short-term trips and visits.

Apart from geo-tagged tweets, there is other information provided by the Twitter API that can help us infer whether a person is a migrant or not. Although [84] did not directly study migrants, but looked at foreigners present in Qatar, it provides important insights to which of the features provided by Twitter is useful in identifying nationality of users. They gathered features from both profile and tweets of users. For features providing information on profile pictures and name, they performed facial recognition and name ethnicity detection. Their final results showed that ethnicity of name, race, language of tweet, language of mention, loca-

tion of followers and friends are the first six features that are useful. In this paper, we purely employ data provided by Twitter for the analysis and therefore, we do not have name, ethnicity and race features. Nevertheless, our work also shows that locations of users and friends are the useful features. The difference here is that we propose to use the social network of users as one of the main features in identifying nationality, which is more flexible than having to perform ethnicity detection on names and profile pictures.

3.3 Experimental setting for data collection

We began with a Twitter dataset collected by the SoBigData.eu Laboratory [41]. We started from a three months period of geo-tagged tweets from August to October 2015. Due to our focus on Italy, we selected from these data the users that tweeted from Italy, obtaining thus 34,160 users. We then crawled the network of geo-enabled friends of these 34,160 users, using the Twitter API. Friends are people that the individual users are following. We focused on friends because we believe that for a user, the information on whom they follow is more informative when it comes to nationality, than who they are followed by.

We concentrated on geo-enabled friends because geo-location is necessary for our analysis. By collecting friends, the list of users crossed our initial geographic boundary, i.e., Italy. At this stage, the number of unique users grew to over 250,000. For all users we also scraped the profile information and the 200 most recent tweets using the Twitter API. During this process, we were able to collect all 200 recent tweets for 97% of users and at least 55 tweets for 99% of users. Our final user network consists of 258,455 nodes and 1,205,133 edges which includes both our initial 34,160 users and their geo-tagged friends.

For the process of identifying migration status, we focus on the core users, i.e., 34,160 users. We assign a residence and a nationality to each user, based on the geo-locations included in the data, the language of tweets and profile information. The final dataset includes 237 unique countries from where individuals have sent out their tweets, including ‘undefined’ location. Even if a user enables geo-tags on their

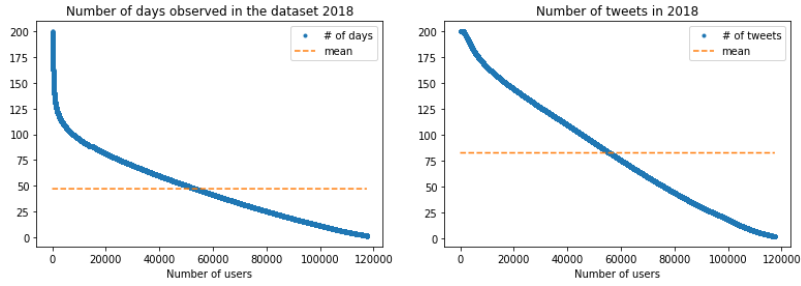


Figure 3-1: Distribution of the number of days (left) and the number of tweets (right) observed in the data per user : on average, our users have tweeted 47 days and 82 tweets in 2018.

tweets, not all tweets are geo-tagged. As a result, 21% of our tweets are ‘undefined’. As for the languages, there are 66 unique languages and 12% of our tweets are in English.

As for the profile features, we observe that 40% of the users have filled out location description. In addition, most of users have set their profile language to English. The number of unique profile languages detected in our data is 58 which is smaller than the languages used, indicating that some users are using languages different from their profile language when tweeting.

In order to assign a place of residence to users, we needed to restrict the observation time period. We have chosen to look at one year length of tweets from 2018, in order to assign the residence label for the 2018 solar year. We selected users that have tweeted in 2018, identifying 128,305 users. To remove bots, we looked at whether a user is tweeting too many times a day. We considered that tweeting more than 50 tweets on average in a single day was excessive and we have eliminated in this way 39 users. In addition, we removed users that were not very active in 2018. If the number of tweets was less than 20, we checked whether the tweeted days were spread out during the year. If the days were not well spread out, we filtered out the user. On the other hand, if it was well spread out, it meant that the user was regularly tweeting, so the user was kept. During this process, we removed 10,764 users. After removing bots and inactive users, we have 117,502 users. For these, we show the distribution of the number of tweets and number of days in which they tweeted in Figure 3-1. On average we see 47 days and 82 tweets.

In addition to the Twitter data, we also collected a list of official and spoken

languages for countries identified in our data².

3.4 Identifying migrants

A migrant is a person that has the residence different from the nationality. We thus consider our core 34,160 Twitter users and assign a residence and nationality based on the information included in our dataset. The difference between the two labels will allow us to detect individuals who have migrated and are currently living in a place different from their home country. The methodology we propose is based on a series of hypotheses: a person that has moved away from their home country stays in contact with their friends back in the home country and may keep using their mother tongue.

3.4.1 Assigning residence

In order for a place to be called residence, a person has to spend a considerable amount of time at the location. Our definition of residence is based on the amount of time in which a Twitter user is observed in a country for a given solar year. More precisely, a residence for each user is the country with the longest length of stay which is calculated by taking into account both the number of days in which a user tweets from a country but also the period between consecutive tweets in the same country. In this work we compute residences based on 2018 data.

To compute the residence, we first compute the number of days in which we see tweets for each country for each user. If the top location is not ‘undefined’, then that is the location chosen as residence. Otherwise, we check whether any tweet sent from ‘undefined’ country was sent on a same day as tweets sent from the second top country. In case at least one date matched between the two locations, we substitute second country as the user’s place of residence. On average, 5 dates matched. This is done under the assumption that a user cannot tweet from two different countries in a day. Although this is not always the case if a user travels, in most of the days of the year this should be true. This approach allowed us to assign a residence in

²Retrieved from <http://www.geonames.org> and <https://www.worlddata.info>

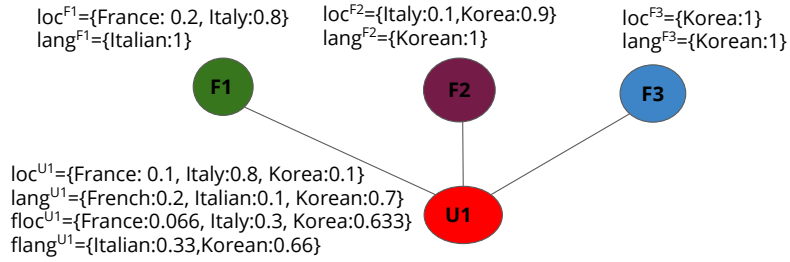


Figure 3-2: Example of calculation of the $floc$ and $flang$ values for a user. The calculation of $floc^{U1}$ and $flang^{U1}$ is based of the $floc$ and $flang$ values for the three friends, showing the distribution of tweets in various countries/languages for each.

2018 to 57,180 users.

For the remaining 60,322 users, a slightly different approach was implemented. We computed the length of stay in days by adding together the duration between consecutive tweets in the same country. We selected the country with the largest length of stay. In case the top country was ‘undefined’, we checked whether ‘undefined’ locations were in between segments of the second top country, in which case the second country was chosen. In this way, an additional 11,046 users were assigned a place of residence. The remaining 49,276 users were neglected because we considered that we did not have enough information to assign a residence.

3.4.2 Assigning nationality

In order to estimate nationalities for Twitter users, we took into account two types of information included in our Twitter data. The first type relates to the users themselves, and includes the countries from which tweets are sent and the languages in which users tweet. For each user u we define two dictionaries loc^u and $lang^u$ where we include, for each country and language the proportion of user tweets in that country/language.

The second type of information used is related to the user’s friends. Again, we look at the languages spoken by friends, and locations from which friends tweet. Specifically, starting from the loc and $lang$ dictionaries of all friends of a user, we define two further dictionaries $floc$ and $flang$. The first stores all countries from where friends tweet, together with the average fraction of tweets in that country,

computed over all friends:

$$floc^u[C] = \frac{1}{|F(u)|} \sum_{f \in F(u)} loc^f[C] \quad (3.1)$$

where $F(u)$ is the set of friends of user u . Similarly, the *flang* dictionary stores all languages spoken by friends, with the average fraction of tweets in each language l :

$$flang^u[l] = \frac{1}{|F(u)|} \sum_{f \in F(u)} lang^f[l] \quad (3.2)$$

Figure 3-2 shows an example of a (fictitious) user with their friends, and the four resulting dictionaries.

The four dictionaries defined above are then used to assign a nationality score to each country C for each user u :

$$N_C^u = w_{loc} loc^u[C] + w_{lang} \sum_{l \in languages(C)} lang^u[l] + \quad (3.3)$$

$$w_{floc} floc^u[C] + w_{flang} \sum_{l \in languages(C)} flang^u[l] \quad (3.4)$$

where $languages(C)$ are the set of languages spoken in country C , while w_{loc} , w_{lang} , w_{floc} and w_{flang} are parameters of our model which need to be estimated from the data (one global value estimated for all users). Each of the w value gives a weight to the corresponding user attribute in the calculation of the nationality. To select the nationality for each user we simply select the country C with maximum N_C : $N^u = \operatorname{argmax}_C N_C^u$.

3.5 Evaluation

To evaluate our strategy for identifying migrants we first propose an internal validation procedure. This defines gold standard datasets for residence and nationality and computes the classification performance of our two strategies to identify the two user attributes. The gold standard datasets are produced using profile information

Table 3.1: Average precision, recall and F1 scores, together with scores for the top 7 residences in terms of support size.

	weighted avg	macro avg	micro avg	IT	KW	US	ID	SG	AU
f1-score	0.858	0.716	0.856	0.928	0.839	0.703	0.945	0.83	0.891
precision	0.879	0.745	0.856	0.935	0.989	0.572	0.949	0.946	0.883
recall	0.856	0.727	0.856	0.921	0.728	0.91	0.941	0.739	0.899
support	3065	3065	3065	343	125	122	119	119	109

as they are provided by the users themselves. We then perform an external validation where we compare the migrant percentages obtained in our data with those from official statistics.

3.5.1 Internal validation: gold standards derived from our data

Residence

To devise a gold standard dataset for residence we consider profile locations set by users. We assume that if users declare a location in their profile, then that is most probably their residence. Very few users actually declare a location, and not all of them provide a valid one, thus we only selected profile locations that were identifiable to country level. Among the user accounts for which we could estimate the residence, 3,065 accounts had a valid country in their profile location. Using these accounts as our validation data, we computed the F1 score to measure the performance of our residence calculation. Table 3.1 shows overall results, and also scores for the most common countries individually. The weighted average of the F1 score is 86%, with individual countries reaching up to 94%, demonstrating the validity of our residence estimation procedure.

Nationality

In order to build a gold standard for nationality, we take into account the profile language declared by the users. The assumption is that profile languages can provide a hint of one’s nationality [151]. However, many users might not set their profile language, but use the default English setting. For this reason, we do not include

Table 3.2: Average precision, recall and F1 scores for top 8 nationalities in terms of support numbers

	weighted avg	macro avg	micro avg	IT	ES	TR	RU	FR	BR	DE	AR
f1-score	0.99	0.98	0.72	0.99	0.96	0.98	0.95	0.94	0.95	0.92	0.97
precision	0.99	0.98	0.73	1	0.94	0.98	0.98	0.9	0.96	0.91	0.98
recall	0.98	0.98	0.75	0.99	0.97	0.99	0.93	0.98	0.94	0.93	0.95
support	12223	12223	12223	10781	302	173	146	118	113	86	59

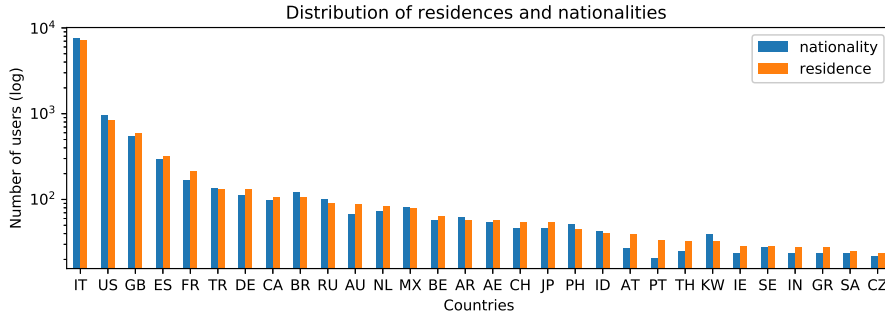


Figure 3-3: Distribution of residences and nationalities of top 30 countries, for all users that possess both residence and nationality labels.

into the gold standard users that have English as their profile language.

The profile language, however, does not immediately translate into nationality. While for some languages the correspondence to a country is immediate, for many others it is not. For instance, Spanish is spoken in Spain and most American countries, so one needs to select the correct one. For this, we look at tweet locations. We consider all countries that match with the profile language and, among these, we select the one with the largest number of tweets, but only if the number of tweets from that country is at least 10% of the total number of tweets of that user. This allows to select the most probable country, also for users who reside outside their native country. If no location satisfies this criterion the user is not included in the gold standard. We were able to identify nationalities of 12,223 users. Due to the fact that during data collection we focused on geo-tags in Italy, the dataset contains a significant number of Italians.

We employed this gold standard dataset in two ways. First, we needed to select suitable values for the w weights from Equation 3.3-3.4. These show the importance of the four components used for nationality computation: own language and location, friends' language and location. We performed a simple grid search and obtained the

best accuracy on the gold standard using values 0 for languages and 2 and 1.5 for own and friends' location, respectively. Thus we can conclude that it is the locations that are most important in defining nationality for twitter users, with a slightly stronger weight on the individual's location rather than the friends. The final F1-score, both overall and for top individual nationalities, are included in table 3.2, showing a very good performance in all cases.

To assign final residences and nationalities to our core users, we combined the predictions with the gold standards (we predicted only if the gold standard was not present). Figure 3-3 shows the final distribution of residences and nationalities of top 30 countries for all users that have both the residence and nationality labels. The difference in the residence and nationality can be interpreted as either immigrants or emigrants.

3.5.2 External validations: validation with ground truth data

In order to validate our results with ground truth data, we study users labelled with Italian nationality and non-Italian residence, i.e. Italian emigrants. We computed the normalised percentage of Italian emigrants resulting from our data for all countries, and compared with two official datasets: AIRE (Anagrafe Italiani residenti all'estero), containing Italian register data, and Eurostat, the European Union statistical office. For comparison we use Spearman correlation coefficients, which allow for quantifying the monotonic relationship between the ground truth data and our estimation by taking ranks of variables into consideration.

Figure 3-4 displays the various values obtained, compared with official data. A first interesting remark is that even between the official datasets themselves, the numbers do not match completely. The correlation between the two datasets is 0.91. Secondly we observed good agreement between our predictions and the official data for European countries. The correlation with AIRE is 0.753, while with Eurostat it is 0.711 when considering Europe. For non-European countries, however the correlation with AIRE data drops to 0.626. We believe the lower performance is due

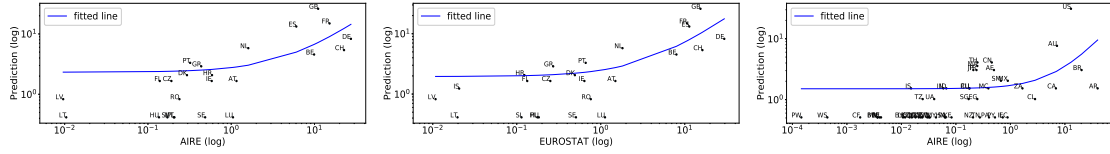


Figure 3-4: Comparison between the true and predicted data; the first two plots show predicted versus AIRE/EUROSTAT data on European countries. The last plot shows predicted versus AIRE data on non-European countries.

to several factors related to sampling bias and data quality in the various datasets. This includes bias on Twitter and in our methods, but also errors in the official data, which could be larger in non-EU countries due to less efficient connections in sharing information.

All in all, we believe our method shows good performance and can be successfully used to build population level indices for studying migration. We do not aim to perform nowcasting of immigrant stocks, but rather to identify a population that can be representative enough for further analyses.

3.6 Case study: topics on Twitter

In this section we show that our methodology can be employed to study how trending topics in Italy are also being discussed among Italian emigrants. As an example, we selected one hashtag that has been very popular in the last years: #Salvini. This refers to the Italian politician Matteo Salvini who served as Deputy Prime Minister and Minister of internal affairs in Italy until recently. To this, we added the top nine hashtags that appear frequently with #Salvini in our data: Berlusconi, Conti, Diciott, DiMaio, Facciamorete, Legga, M5S, Migrant, Ottoemesso. Indeed, they all represent people that are often mentioned together or political parties or other issues that are associated with the hashtag #Salvini.

Figure 3-5 shows an evolution of the usage of the 10 above mentioned hashtags across different Italian communities both within and abroad Italy. The values shown are the number of tweets from Italian nationals residing in each country that include one of the 10 hashtags, divided by the total number of tweets from Italian nationals from that country. Values are computed monthly. Thus, we show

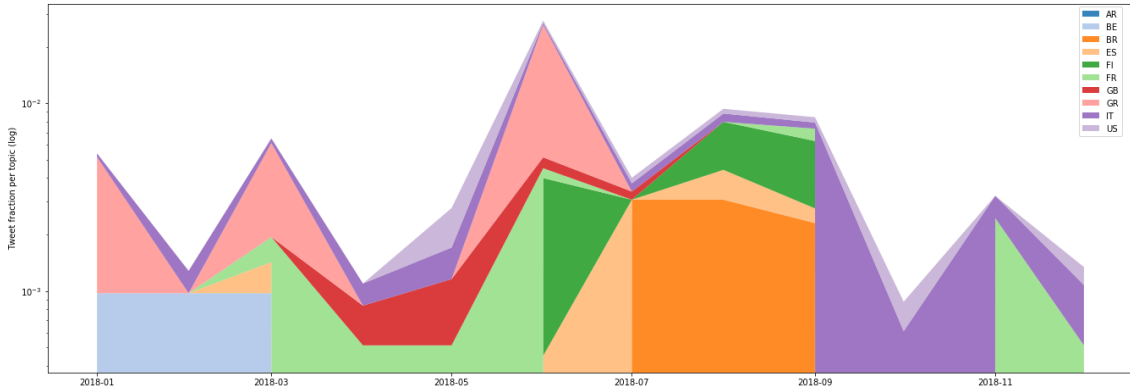


Figure 3-5: Stream graph: appearance of hashtags related to #Salvini from Italians across 10 selected residence countries in 2018. The discussion continuously appeared in Italy throughout the year and it became more lively employed by Italians overseas as Salvini gained more political attention.

the monthly popularity of the topics in each country. In this way, even the tweets from less represented countries are well shown. As the figure shows, the hashtag was continuously used by Italians in Italy. We observed that the hashtag gradually spread over other residence countries as Salvini received more and more attention. We also observe that most of the attention comes from Italians residing in Europe, with non-European countries less represented.

3.7 Conclusion and future work

We have developed a new methodology to provide a snapshot of migrants within the Twitter population. We considered the length of stay in a country as the key factor to define a user's residence. As for the nationality, connections which migrants maintain with their country of origin provided us with a good indication. In particular, the location of friends seemed to be a strong feature in determining nationality, together with the location of the users themselves. Tweet language, on the other hand, was not considered relevant by our model. This is probably due to the fact that English is the dominating language on Twitter, since a language that is widely understood has to be spoken to get more attention from other users. We have validated our results both with internal and external data. The results show good classification performance scores and good correlation coefficients with official datasets.

The constructed dataset can be applied in different scenarios. We have shown how it can be used to study trending topics on Twitter, and how attention is divided between emigrants and non-migrants of a certain nationality. In the future, we plan to analyse social ties, integration and assimilation of migrants [78]. At the same time, one can investigate the strength of the ties with the community of origin.

“Migrants help provide the building blocks for prosperous societies bringing knowledge, support, networks, and skills in countries of origin, transit and destination”

António Guterres-UN Secretary
General

Chapter 4

Home and destination attachment: study of cultural integration on Twitter

4.1 Introduction

The cultural integration of immigrants is a first-order social, political and economic issue. For the individual immigrant, it conditions his or her economic success and overall social integration to the host society. From the viewpoint of the latter, the promotion of the cultural integration of its immigrants has become a political imperative in these times of rising populism and cultural backlash against globalisation in general and immigration in particular (e.g., [119]).¹ However, from both the individual and social perspectives, too much cultural integration (or acculturation) may be detrimental: in terms of immigrants’ subjective wellbeing, and in terms of lost diversity (from the viewpoint of host countries) and of global connections (from the viewpoint of both host and home countries). In other words, it is in the best interests of all stakeholders to find the right balance between acculturation and cultural separatism, between loyalty to the home country and the host country cultures. Successful cultural integration brings new opportunities and, with them, an overall

¹Norris, P. and Inglehart, R.F. (2019): *Cultural backlash: Trump, Brexit, and Authoritarian Populism*, Cambridge University Press).

improvement of living conditions and well-being. Failure to integrate migrants in the host country’s society may result in social conflict and cultural polarisation.

Cultural integration has been long studied by various research communities. These include international economic organisations which have built indicators for integration at different levels, considering socio-economic features such as labour market participation, living conditions, civic engagement and social integration [100, 121, 85]. On the other hand, studies of integration have been mainly done by sociologists, by employing survey data such as World Values Survey, Eurobarometer, and European Social Survey. The main elements used in the studies are often inter-marriage, religion and language [157, 107, 145, 52].

However, studying integration is very complex, as one is “not only attracted to the culture of host society but is also held back from his culture of origin” [124, 142]. The four-fold model reflects this complexity by dividing acculturation into four different classes: *assimilation*, *integration*, *marginalisation* and *separation*. [42, 43, 130, 127, 21]. Integration takes place when a migrant’s and receiving society’s characteristics mutually accommodate. Assimilation on the other hand takes place when a migrant perfectly absorbs the characteristics of the receiving society, losing the connection to the home country. Marginalisation refers to a situation where migrants remain distinguishable from the both of receiving and home society, whereas separation refers to complete rejection of host’s culture. These theories typically consider two dimensions: preservation of links to the home country and cultural traits, which we call here *home attachment*, (*HA*), and formation of new links and adopting cultural traits from the country of migration, that we define as *destination attachment* (*DA*). Based on these two concepts, we can summarise the four integration patterns from the literature, as displayed in Table 4.1.

In this paper we provide a novel method to compute HA and DA from Twitter data, to answer the following questions: *How much do migrants absorb the culture of*

Table 4.1: Theories of integration and their relation to HA and DA.

	Low HA	High HA
Low DA	Marginalisation	Separation
High DA	Assimilation	Integration

their destination society? Do they loose connection with their home country? This is based on the topics that migrants and natives discuss on Twitter, through the analysis of hashtags. The HA index is defined as the fraction of tweets of a migrant that discuss topics related to their home country. Similarly, DA is the fraction of tweets discussing topics related to the destination country. These definitions are based on the idea that the topics discussed provide indications on various aspects of attachment: the amount of information that a person holds about a specific country, the social links to people living in a certain country, the interest in political and public issues of a country, adoption of customs and ideas, all related to integration as a wider concept.

The analytic process that we introduce here includes three stages, and is based on a Twitter dataset containing data on users, their friends and their statuses. The first stage is to identify migrants by assigning a residence and nationality to Twitter users, starting from a previously developed method [96]. The second stage is to determine country-specific topics by assigning nationalities to hashtags. The final stage is to compute the HA and DA indices for each migrant in our data. We examined the two indices in various settings, to demonstrate their validity. First, we analysed the relationship between the two indices and compared them to a null model obtained by shuffling the hashtags in our dataset. Second, we studied different country-specific cases, i.e., immigrants in the United States and the United Kingdom, and emigrants from Italy. The indices were then compared with Hofstede’s cultural dimension scores [80] as well as other related variables such as distance and language proximity measures.

The rest of the paper is organised as follows. In the next section we describe related work that studies integration and acculturation of migrants both in the sociology literature and in recent big data studies. In Section 4.3, we define our methodology to compute the HA and DA indices, including data collection (Section 4.3.1), assigning nationality and residence to users (Section 4.3.2), assigning nationality to hashtags (Section 4.3.3) and calculating the indices (Section 4.3.4). In Section 4.4, we present our results, while Section 4.5 concludes the paper.

4.2 Related works

It has long been in the core interests of sociologists to study cultural identity and integration of migrants. Using survey data, many have studied the complexity of migrants' conversion of cultural identity in the receiving societies. Although a uniform definition of a culture does not exist, one way to define it is the following; "the beliefs, values, social perspective, traditions, customs, and language shared within a group" [140]. Taking the elements stated in the definition into account, studies have looked at language, role of media, inter-marriage and religion² to study whether a migrant is culturally integrated in the society [157, 92, 91]. In particular, language plays an important role in various aspects of integration. It increases labour force participation of migrants and bring positive impacts on practical aspects of life, for example making friends in the class or talking to the teacher [107, 2, 145, 52]. In our work we also underline the relation between language proficiency and our DA index.

In recent years, social big data has been employed to study integration of migrants [70, 48, 150]. Retail data including shopping behaviour in a large supermarket chain was used in [70] to measure the conversion of migrants' consumption behaviour towards that of natives. Through a data-driven approach, they identified 5 groups of migrants that show different trends towards adopting new consumption behaviours. In [48], the authors used data collected from the Facebook Marketing API containing information on the country of origin, age, residence, spoken language and others, including the "likes" of individual users. They quantified assimilation by introducing a score that serves as a proxy for migrants assimilating to local population's interests, using the "likes" used by the Facebook users.

Following the work in [48], [150] studied Mexican immigrants in the U.S and their cultural assimilation in terms of musical taste using Facebook data. They looked at the similarity of immigrants to the host population in terms of musical preferences, also looking at the interests of users. Furthermore, they extended their analysis to understand the differences in assimilation scores between ethnicity and

²<https://migrationdataportal.org/themes/migrant-integration>

generations across different demographic groups. In a more recent work, [156] looked at the diffusion of Brazilian cuisine around the world and estimated cultural distance between countries. They computed a so called interest entropy to measure how the interests are distributed around the world. They showed that the presence² of Brazilian migrants explains, in part, the presence of interests in Brazilian cuisine in the host country. Other related factors were geographical proximity, and linguistic similarity, factors that also appear important in our study.

In this paper, we also employ social big data for the analysis which allows us to overcome some of the limitations of using survey data. For instance, it allows us to cover a wider population throughout broader geographical areas. However, different from Facebook data, Twitter data does not provide interests of individual users in the form of “likes”. We thus build our DA and HA indicators through hashtags as a proxy for their interests. In the process, we also employ the Shannon entropy, but in a different way from [156]: we use it to filter out hashtags that are not country-specific. Learning from the previous studies in Sociology, our analysis also takes into account HA (home attachment), which has not been as widely studied in the literature. In addition, many of the studies have been conducted from the host country’s point of view towards their receiving migrants. Here, we also look at emigrants overseas, allowing the home country to better understand the allocation of their citizens abroad.

4.3 The home and destination attachment indices

We propose to study home and destination attachment through the Twitter lens. We consider the topics discussed by migrants as a proxy to their interests, opinions and also to the amount of information about the context they live in, and define two indices: destination attachment (DA) and home attachment (HA). The methodology includes various stages: data collection, identifying migrant users by automatically assigning a nationality and residence label, identifying country-specific topics by assigning a nationality to Twitter hashtags, and finally the calculation of the indices.

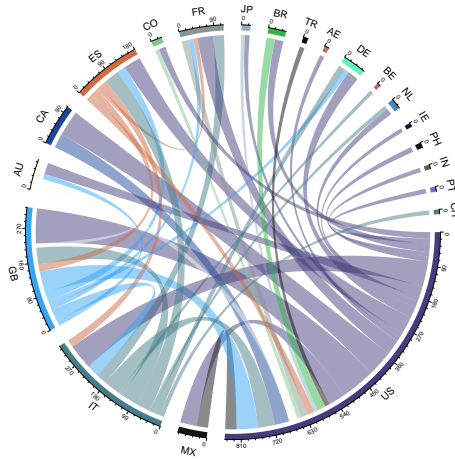


Figure 4-1: Chord diagram showing migration links between countries. The colour of the chord represents the nationality of the migrants, while the width of the chord represents the number of migrants in our dataset who had the 2018 residence in the corresponding destination country. For visualisation purposes we show only 21 countries: those with at least 10 migrants.

4.3.1 Data

Our data collection strategy originated from the methodology developed by [96]. The starting point is a Twitter dataset collected by the SoBigData.eu Laboratory [41]. We extracted from this dataset all the geo-located tweets posted from Italy from August to October 2015. This allowed us to obtain a set of 34,160 individual users that were in Italy in that period, which we call the first layer users. For these users, we downloaded the friends, resulting in 258,455 users that we denominate as second layer users. For all of these users, we have also gathered their 200 most recent tweets. Different from the work of [96], we further extended the dataset to obtain a larger number of migrants by extracting also the friends of the second layer users (i.e. the third layer), and their 200 most recent Tweets. After this process, the total number of users grew to 59,476,205. Our dataset, therefore, consists of three layers: the core first layer users, their friends (second layer users) and the friends of the friends (third layer users). Our analysis concentrates on a subset of these users for which we have information about their friends, resulting in a total of 200,354 users. These are users from the the first and second layers (some overlap was present among the two layers).

4.3.2 Assigning residence and nationality to users

In order to identify migrants in our dataset, we automatically assign to each user u a nationality country $C_n(u)$ and a residence country $C_r(u)$ (for the year 2018) following the methodology in [96]. We define a migrant as “a person who has the residence different from the nationality”, i.e. $C_n(u) \neq C_r(u)$. In order to identify a user’s residence, we look at the number of days spent in each country in 2018 by looking at the time stamps and geo-locations of the tweets. The location where the user spent most of the time in 2018 is considered as the country of residence. On the other hand, the nationality is defined by looking at tweet locations of the user and user’s friends. As shown in the study [96], tweet language was not important in defining the nationality so we set the language weight to 0 here as well. By comparing the country of residence and the nationality labels we were able to determine whether the user was a migrant or not in 2018.

Out of the total 200,354 users, we were able to identify nationalities of 197,464 users. As for the residence, we were able to identify residences of 57,299 users. In total, we have identified both the residences and nationalities for 51,888 users. Among 51,888 users, the total number of individuals users that we have identified as migrants are 4,940 users. We then filtered out users who have used less than 10 hashtags in 2018, leaving us with total of 3,226 migrant users. In Figure 4-1, we display the main migration links in our dataset: the number of migrants for countries that have at least 10 migrants, showing a total of 21 countries. However, overall, we have 128 countries of nationality and 163 countries of residence. From the plot, we see that in terms of nationality, the most present countries are the United States of America, Italy, Great Britain and Spain. This is due to the fact that our first level users were selected among those geo-localised in Italy. In terms of migration patterns, we note that Italy has mostly out-going links whereas countries like the USA and GB has a significant amount of both in and out-going links. France and Germany, on the other hand, have mostly in-coming links.

We chose to employ this methodology because it adopts a definition of a migrant

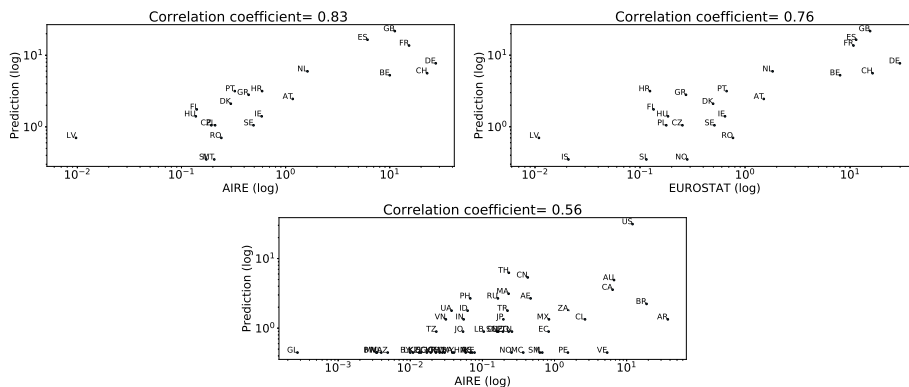


Figure 4-2: Percentage of Italian emigrants in various destination countries based on AIRE and Eurostat: predicted versus ground truth data.

that is close to the official definition³. It also allows us to identify both immigrants and emigrants simply by comparing the nationality and residence labels. It is important to mention that the migration patterns we see here are specific to our dataset, and are not meant to represent a global view of the world’s migration. However we do observe some correlation to official data when looking at individual countries. In figure 4-2, for instance, we show Spearman correlation coefficients between our predicted data and ground truth data for Italian emigrants from AIRE⁴ and Eurostat. For European countries, the correlation with the AIRE data is 0.831 and 0.762 with the Eurostat data. For non-European countries, the correlation stays at 0.56. This gives us reason to believe that this dataset can be used to validate our methodology of studying integration patterns through Twitter.

4.3.3 Detecting country-specific topics

The topics discussed on Twitter can be extracted through the analysis of hashtags. These are phrases that the users add to their tweets to mark the topic. In this analysis phase we detect country-specific topics by assigning nationalities to all the hashtags in our data. To do this, for each hashtag we extract the list of users who use it, and we study the distribution of the nationality of all the users that are not labelled as migrants in the first stage (i.e. users who have the residence equal

³Recommendations on Statistics of International Migration, Revision 1(p.113). United Nations, 1998, defines a 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”.

⁴Anagrafe degli italiani residenti all’estero (AIRE) is the Italian register data.

to the nationality). For those hashtags that appear mostly in one country (small entropy of the country distribution), we assign the nationality to the most frequent country. The hashtags that display a heterogeneous distribution across countries are not considered, since they are deemed international.

We begin by performing simple word processing for all the hashtags we have in the dataset. We selected all the hashtags used by non migrant users in 2018. We converted all the hashtags to lower case and removed signs such as comma, quotes, semicolons, and slashes. We removed also single characters. After the data cleaning process, we obtained a total of 639,494 hashtags that were used by non-migrants in 2018. For each hashtag h , we define a dictionary where we store P_h , the distribution of the nationalities of the users using hashtag h . Hence P_h is a vector where for each country c we have $P_h(c)$, the fraction, among all non-migrant users that use hashtag h , of users with nationality c . Provided with this probability distribution, we compute the normalised entropy for each hashtag following Equation 4.1, where $|P_h(c)|$ is the cardinality of the dictionary $P_h(c)$, i.e. the number of countries where the hashtag is used.

$$H(h) = \frac{-\sum_c P_h(c) \log P_h(c)}{\log(|P_h(c)|)} \quad (4.1)$$

Figure 4-3 displays the distribution of normalised entropy values across all hashtags in our dataset. We note that a majority of hashtags have zero entropy, hence they are mentioned in one country only, while a few show very high entropy levels, indicating they are international topics.

To filter out international topics we select a threshold for the normalised entropy, that we here fix at the value 0.5. After applying the threshold, 81,941 hashtags were categorised as international topics and 557,552 were given nationality labels. In other words, about 13% of the total hashtags are considered international. The entropy threshold chosen is rather strict, and it eliminates a large number of hashtags, maintaining mostly those for which we are sure they are specific to a nation. Figure 4-4 displays the distribution of the hashtag nationalities obtained. We note that the American specific hashtags are in lead, followed by Italian and Great Britain,

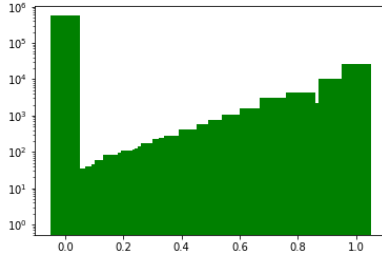


Figure 4-3: Entropy distribution in Log scale

following the distribution of the number of users from figure 4-1. Examples of Italian specific topics that we have identified are the following: *Salvini, Lavoro, Immigrazione, Caffè, Renzi, Trenitalia, Epifania*. Moreover, examples of some of the international topics we have identified are *Trump, EU, Immigration, Refugee, Coffee, and Fiat*.

4.3.4 Computing the home and destination attachment indices

Provided with the nationality of hashtags, we can define for each 3,226 migrant user the home and destination attachment, HA and DA. Consider user u with the country of nationality denoted as $C_n(u)$ and country of residence denoted as $C_r(u)$. To define the home attachment of user u , $HA(u)$, we consider $HT(u, C_n(u))$ the number of hashtags used by user u specific to their country of origin, divided by $HT(u)$ the total number of hashtags of user u . For example, for an Italian national living in Korea, what fraction of their hashtags is Italian?

$$HA(u) = \frac{\# C_n(u) \text{ hashtags}}{\# \text{ total hashtags}} = \frac{HT(u, C_n(u))}{HT(u)} \quad (4.2)$$

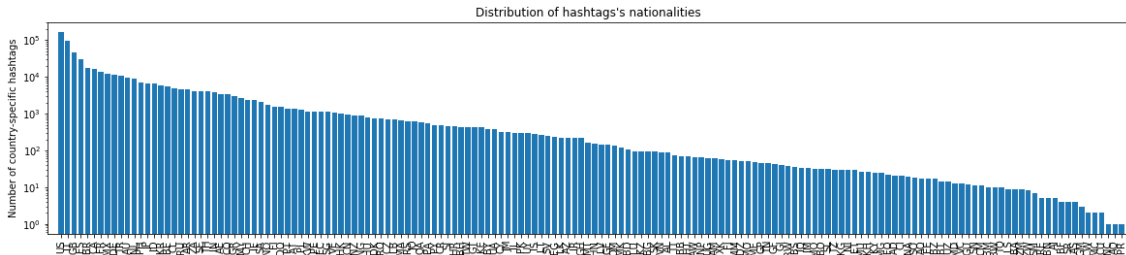


Figure 4-4: Distribution of hashtags' nationalities

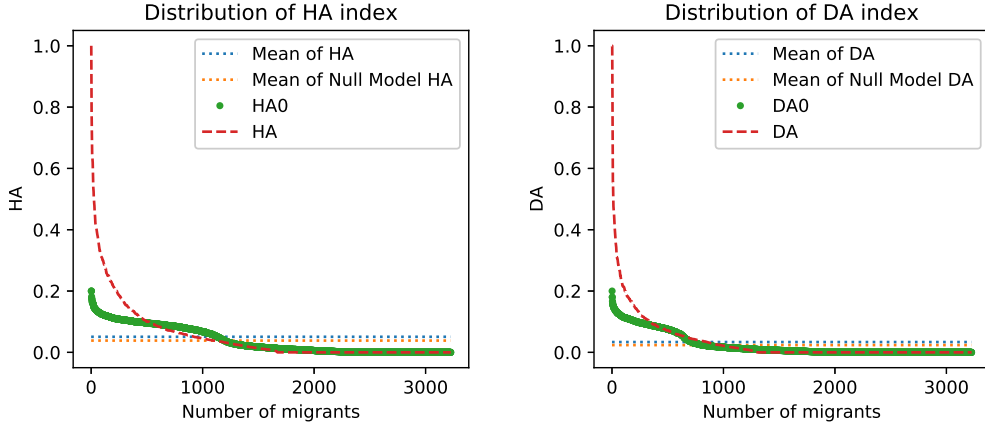


Figure 4-5: Distribution of HA and DA values, and comparison to null model HA_0 and DA_0 . Means values are: $\bar{HA}_0 = 0.038$ and $\bar{DA}_0 = 0.024$, $\bar{HA} = 0.051$ and $\bar{DA} = 0.034$.

Similarly, the destination attachment index DA is the fraction of hashtags they use that are labelled with their country of residence:

$$DA(u) = \frac{\# C_r(u) \text{ hashtags}}{\# \text{ total hashtags}} = \frac{HT(u, C_r(u))}{HT(u)} \quad (4.3)$$

Following the previous example, what is the fraction of Korean specific hashtags that the Italian emigrant is using?

Both indices vary from 0 to 1. If they are equal to 1, it means that a migrant is either fully attached to the destination country or fully attached to home country. In contrast, indexes equal to 0 means that a migrant is either not attached to the destination country or not attached to the home country. The sum of the two indices is always ≤ 1 : a user cannot be fully attached to both home and destination, but has to ‘divide’ their attention among the various countries they are interested in.

4.4 Results

4.4.1 Overall distribution of DA and HA values

The distributions of the home and destination attachment indices are shown in Figure 4-5. The HA index is 0.051 on average and the DA index is 0.034 on average for all the migrants we have in our dataset regardless of the nationality or the place

of residence. We observe that some users have relatively high values for the two indices, however the majority are under 0.2 in both cases. In the same figure, we compare these values with a null model analysis where the hashtags of individual users were randomly re-distributed five times. The null model tells us what the DA and HA values would be if users chose their topics of discussion randomly, i.e. there was no influence from the country of residence or nationality. We observe that in general the null model DA_0 and HA_0 are smaller than the actual index values, with lower means for the null model distributions.

To statistically validate the difference between the null model, and DA and HA, we also computed two non-parametric tests: Wilcoxon and Kolmogorov–Smirnov (KS) tests. The results for the Wilcoxon test show that for both the DA and HA, their distributions are significantly different from the distribution of the DA_0 and HA_0 with p-values of $5.16e^{-07}$ and 0.014, respectively. We obtained similar results from the KS tests, with p-values of $1.18e^{-51}$ for DA and $2.98e^{-56}$ for HA. Although not reported here, the results for KS-tests for sub-populations split by country of residence and country of origin equally show that the null model and the actual index values have different distributions.

To understand the relationship between the DA and HA, we computed the Pearson correlation among them. Figure 4-6 displays the HA versus DA values for all users. A weak negative relation is found with $r = -0.13$, and p-value = $6.937e^{-14}$, indicating that in general the more a migrant is attached to his country of origin, the less the migrant is attached to the host country and vice versa. However, we can observe various different patterns for individual users, leading to different acculturation types as mentioned in Table 4.1. In the same figure, the red curve provides an approximate indication of users' acculturation type. We underline the fact that we do not aim to provide a specific categorisation of acculturation types in this paper. Instead, we aim to provide a broad picture where the angle of each individual from the x/y-axis gives us an indication of the acculturation type. Thus, a migrant close to the x-axis is most probably going through an assimilation process, a migrant close to the y-axis is undergoing separation, while those in between are undergoing integration or marginalisation. The distinction between integration and marginali-

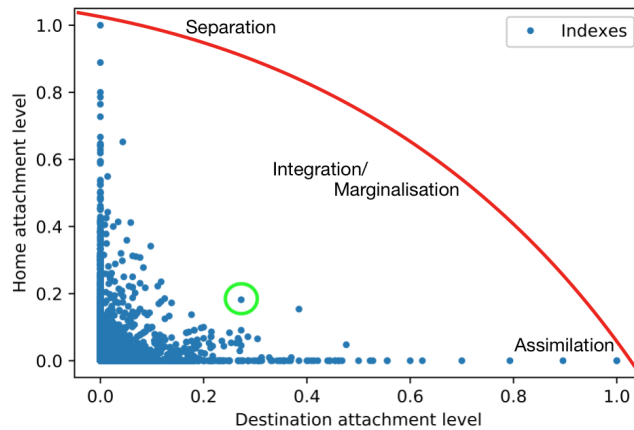


Figure 4-6: Pearson correlation between home and destination attachment indexes for all the migrants in the data: correlation coefficient: -0.13 , p -value: $6.937e^{-14}$.

sation depends on the length of the distance of data point from the origin. In other words, marginalisation is when the data point is close to 0 and integration is when the data point is point further away from 0. The data point circled in green would be a good example of an integrated migrant, who keeps good links with both home and destination country.

4.4.2 Language as a key factor for integration

One possible candidate factor to explain the DA and HA values observed is language. As previously studied, language is considered to be a key factor in integration and our indexes reflect this importance as well. In Figure 4-7 we display the distribution of the DA and HA for two user groups: a group that speaks the language of the host country (i.e. over 90% of their tweets are in that language) and a group that very rarely speaks the language of the host country (under 10% of their tweets are in that language). Here, we are looking at all the migrants we have in the dataset regardless of the country of origin or the country of residence. We observe that the group that speaks the language of the destination country shows in general higher DA compared to the non-speaking group, confirming the significance of the language for integration in the host country. In addition, we observe that users who do not speak the language of the destination country tend to be more attached to their home country compared to those speaking the destination language. Hence,

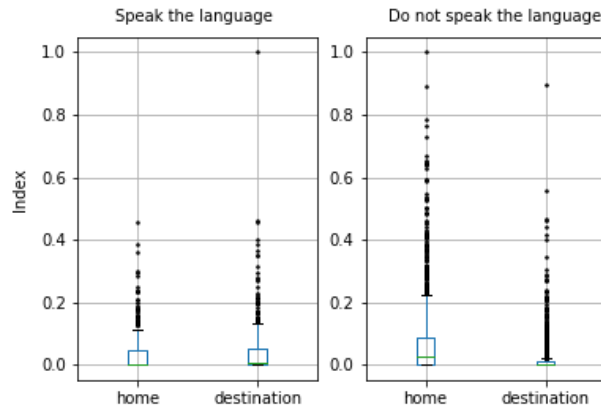


Figure 4-7: Box plots showing the HA and DA distributions for a group of migrants who speak the language of the host country on the left and a group of migrants who do not speak the language of the host country on the right. Means are $\bar{H}A = 0.034$ and $\bar{D}A = 0.041$ for users who speak the destination language, and $\bar{H}A = 0.072$ and $\bar{D}A = 0.019$ for those who do not speak it.

interestingly, destination language proficiency seems to affect both destination and home attachment levels. When comparing DA and HA within groups, the groups that speak the destination language have the two indices comparable, while for those who do not speak it, HA is much larger than DA, indicating a pattern of separation. However, we do not mean to generalise, what we observe are population level patterns. When looking at individual level, we do observe all four acculturation types discussed in Table 4.1.

4.4.3 Country-specific results

In this section, we provide country-specific results. One of the advantage of using our methodology is that we can look at different countries simply by changing the labels. Hence, here we look at different country cases to understand how immigrants in a specific country behave and to know how emigrants from a certain country of origin behave in different countries. We selected three study cases which had the largest number of users in our data: immigrants in the US and UK, and emigrants from Italy. Here we consider only the migrant groups with at least 10 users. The square brackets in the figures below show the number of users we have for each country of origin.

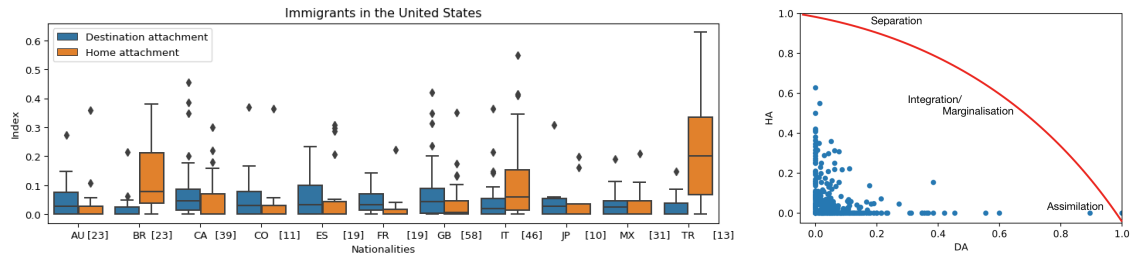


Figure 4-8: Left: Box plots for the DA and HA index of immigrants in the United States. Right: Scatter plot of HA vs. DA indicating approximate integration types for immigrants in the US.

Immigrants in the United States

In Figure 4-8 on the left, we observe different destination and home attachment indices of 17 groups of immigrants from different countries of origin. Overall, we observe that for many groups of immigrants in the United States DA is larger than HA. Immigrants from Canada have the highest DA followed by Colombian and English immigrants. On the other hand, immigrants from Turkey have the highest HA followed by Brazilian and Italian immigrants. In the right figure, we observe data points individually on a scatter plot of HA vs. DA. It tells us that immigrants in the US are integrated and assimilated in general.

We also compared our indexes to the work of Vigdor [157] that measures the degree of similarity between foreign-borns from different countries and natives in the United States. They measure three factors of assimilation: economic, cultural, civic, and their combination. The economic factor looks at employment status, income, education attainment and home ownership. The cultural factor looks at intermarriage, the ability to speak English, number of children and marital status. The civic factor looks at military service and citizenship. The composite factor is the overall score of the all three factors. Table 4.2 shows the Spearman correlation between our indices and the four factors of assimilation, trying to understand whether the attachment levels we see for each individual are similar to the assimilation levels Vigdor [157] found for nationals from the same countries. The table shows that our DA and HA are most correlated with the cultural factor, followed by the economic factor. It is interesting to remark that DA is positively correlated whereas HA is negatively correlated with the cultural factor of assimilation. This tells us that for

	DA	HA	Composite	Economic	Cultural	Civic
DA	1.0***	-0.231***	0.087	0.185***	0.198***	0.045
HA	-0.231***	1.0***	0.129**	-0.145**	-0.2***	0.159***
Composite	0.087	0.129**	1.0***	0.628***	0.406***	0.916***
Economic	0.185***	-0.145**	0.628***	1.0***	0.766***	0.551***
Cultural	0.198***	-0.2***	0.406***	0.766***	1.0***	0.218***
Civic	0.045	0.159***	0.916***	0.551***	0.218***	1.0***

Table 4.2: Spearman correlation table for immigrants in the United States: Vigdor’s assimilation scores and DA & HA indices. Significance levels are marked with *** p-value <0.01, ** p-value <0.05, * p-value <0.1.

those nationalities for which Vigdor observed high cultural assimilation, we observe high DA and low HA, which is exactly how we propose to use our indices to describe assimilation (see Table 4.1 above). A similar relation can be seen with the economic factor: nationalities with high economic assimilation levels also show high DA and low HA. Interestingly, the civic factor does not show the same relation: foreign-borns of nationalities that appear to be well assimilated from the civic point of view in Vigdor’s work tend to show a high HA in our work, and no relation with DA. It appears thus that civic assimilation in the destination country corresponds also to a tighter relation with the home country of a migrant.

A caveat in looking at this table is that here we are looking at identified migrants and hashtags in 2018 and comparing them to the assimilation scores of 2006. There could be possible changes in immigrants’ behaviours between 2006 and 2018. A second caveat is that we are computing correlations at individual level, while Vigdor’s scores are based on groups of migrants. Since there is variability among individuals, it is likely the case that two US immigrants with the same nationality will have different DA and HA scores in our data, while the Vigdor data will contain an unique score for them. This inevitably decreases correlations.

Immigrants in the United Kingdom

Figure 4-9 shows the indices for the immigrants residing in the United Kingdom. Only four groups are shown, corresponding to those that have at least 10 migrants. Overall, UK immigrants in our data are more attached to home than to the destination country. On average, the DA is 0.04 and the HA is 0.063. From the figure on

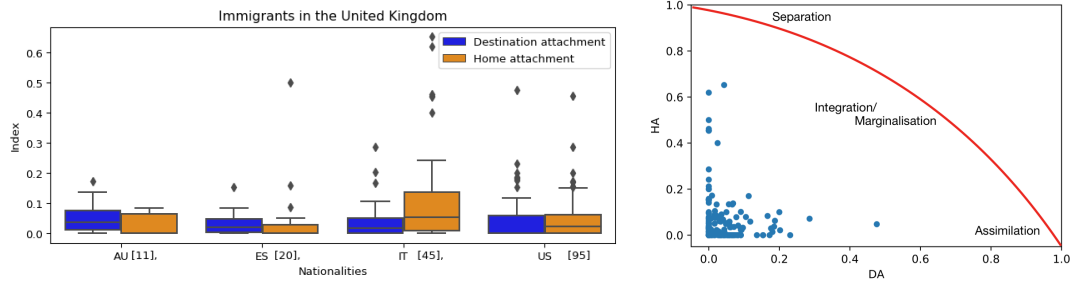


Figure 4-9: Left: Box plots for DA and HA for immigrants residing in the United Kingdom. Right: Scatter plot of HA vs. DA indicating approximate integration types for immigrants in the UK.

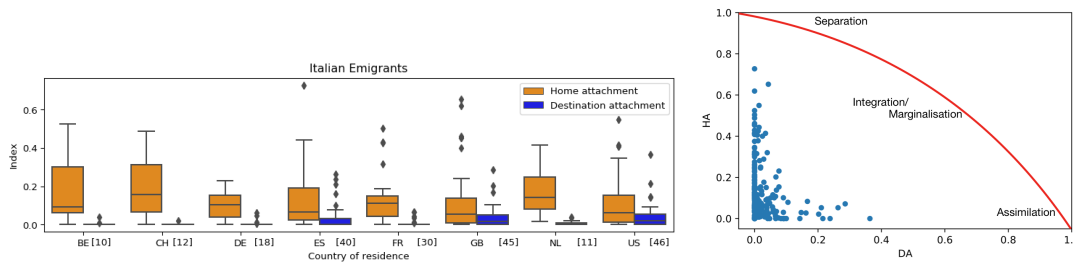


Figure 4-10: Left: Box plots for DA and HA for Italian nationals living abroad. Countries on x-axis are countries of residence of Italians. Bottom: Scatter plot of HA vs. DA indicating approximate integration types for Italian emigrants.

the left, it is clear that immigrants from Italy have the highest HA index. On the other hand, we observe that immigrants from Australia that share long historical ties with the UK have the highest DA index. Looking at the figure on the right, we can observe that immigrants are mostly in the area of marginalisation/integration.

Italian emigrants

Figure 4-10 displays the DA and HA indices for Italian emigrants across different countries of residence. In general, we observe that Italians are more attached to their home country than to their destination country. Switzerland, Belgium and Netherlands are the three countries where Italian emigrants are most attached to home. On the other hand, Italians tend to show higher DA levels in English speaking countries: the US and in the UK. Among the higher DA levels we also observe Spain, probably due to the language similarity. In the figure on the right, we also observe that Italian emigrants have higher HA level compared to DA level. This data points indicate that they are in general close to the *separation* type of acculturation.

	DA	HA	Power	Individualism	Masculinity	Uncertainty	Orientation	Indulgence	contig	comlang_off	distcap	csl	cnl
DA	1.0***	-0.153***	-0.054***	0.155***	0.133***	-0.046***	-0.041**	0.016	0.003	0.069***	0.034*	0.083***	0.099***
HA	-0.153***	1.0***	0.029	-0.092***	-0.113***	-0.014	0.026	0.03*	0.063***	-0.012	-0.074***	0.023	0.021

Table 4.3: Correlation table for HA & DA and Hofstede’s cultural dimension scores for migrants at individual level. Significance levels are marked with *** p-value <0.01, ** p-value <0.05, * p-value <0.1.

4.4.4 Hofstede’s cultural dimension scores and other measures

To further validate our indices, we have also compared our results with Hofstede’s six cultural dimensions, plus various other language proximity measures and geographical distances [80, 81, 110, 113]. Hofstede’s cultural dimensions are well known measures of culture, initially studied to better design the organisational context of business [80]. According to his initial studies, cultures can be studied along four dimensions: power, masculinity, individualism, and uncertainty avoidance⁵. In his later studies, long-term orientation and indulgence⁶ were added to the cultural dimensions [81]. To compare our indices with Hofstede’s cultural dimensions, we computed the differences of scores between the home and the destination countries of migrants, as a measure of the cultural distance among countries. We then computed the correlation between our HA and DA indices and the cultural distances obtained. Hofstede’s data contain a total of 114 countries, while our nationalities and residences cover 128 and 163 countries, respectively. Therefore we considered only users for which both nationality and residence were among the 114 countries, resulting in 3,082 users. In addition to Hofstede’s scores, we also added the following variables: distance between the capitals of the countries (*distcap*), common native language (*cnl*), common spoken language (*csl*), and two dummy variables on whether the countries are sharing borders (*contig*) and common official language (*comlang_off*). The *cnl* and *csl* variables vary at a scale between 0 to 1, indicating 0 if there are no commonality and 1 if they share full commonality.

⁵Power distance: whether a hierarchical order is accepted among people. Masculinity (vs. Femininity): whether the country is driven by competition, achievement and success. Individualism (vs. Collectivism): how “me-centred” the people are in the country. Uncertainty avoidance: how comfortable people are when faced with uncomfortable and ambiguous situations.

⁶Long-term (vs. Short term) orientation: whether the importance is given to what has been done already or to the future, Indulgence (vs. restraint): how strict the people are towards their desires.

Table 4.3 shows the Pearson correlations computed at individual level. The first interesting remark is that in general our DA and HA indices behave differently across the six cultural dimensions, language and distance variables. This means that, when correlations are significant, when HA shows a positive relation, DA shows a negative one and vice-versa. This is compatible with the fact that HA and DA are negatively correlated among themselves, meaning that, in general, as migrants becomes more attached to the destination they lose links to the home country. Among the cultural dimensions, Individualism correlates the most with the DA index, with the correlation coefficient of 0.155. This means that higher the difference between the home and the destination country in terms of individualism, the higher a migrant's DA level. The same can be observed for masculinity: higher cultural differences result in higher DA. A contrasting picture is provided for the HA index: we see that it is significantly negatively correlated with individualism and masculinity. This means that the higher the difference between the home and the destination country in terms of individualism and masculinity, the less a migrant remains attached to their home country.

Among the other variables, in general absolute correlations are rather low. The distance appears to be significantly related to both of our DA and HA indices: the further the destination country is to the country of origin, the higher the DA level and the lower the HA level. Also, the correlation between *contig* and HA indicates that immigrants in destination countries where they share the border with their country of origin have higher HA levels. This makes sense since having the home country close means more possibilities to go back home frequently resulting in higher HA levels. For the variables concerning language, the DA index is significantly positively correlated with all of them. The positive relationship between the DA index and the *csl* highlights that the ease of communication is as important as having common native language or common official language for higher DA.

As already noted, absolute correlation values above are quite low, albeit significant. This is most probably due to individual differences within groups of migrants with the same nationality and residence, which decrease the correlations. To account for this, we repeat the correlation analysis, after grouping the migrants. Specifically,

	Power	Individualism	Masculinity	Uncertainty	Orientation	Indulgence	contig	comlang_off	distcap	csl	cnl
DA	-0.032	0.215**	0.281***	0.164*	-0.09	0.028	0.427***	0.121	0.194*	0.138	0.257**
HA	0.126	-0.301***	-0.164*	-0.094	-0.03	-0.159	0.343***	0.215*	0.061	0.257**	0.385***

Table 4.4: Correlation table for HA & DA and Hofstede’s cultural dimension scores. Correlation with HA is computed after grouping migrants by nationality, while correlation with DA is computed after grouping by residence. Significance levels are marked with *** p-value <0.01, ** p-value <0.05, * p-value <0.1.

we group the migrants by nationality to compute correlations with HA levels, and by residence to compute correlations to DA levels. This allows us to have, for each home and destination country, an average HA and DA level, computed over a group of migrants.

The correlations obtained are shown in table 4.4. We note that grouping increased the correlations observed, confirming that the previous low correlations were due to individual variability, which averages out when grouping. Among the cultural dimensions, Individualism and Masculinity remain the most correlated, with the sign of the relation from the individual analysis confirmed. We observe an additional positive relation between Uncertainty and DA: the higher the difference in uncertainty the more the migrants are attached to the destination country. Regarding the other variables, grouping the migrants also increased the correlations significantly, and now the picture is clearer. It appears that the closer the home and destination countries are in terms of language, the higher the DA and HA levels. This confirms what we saw earlier, language is not important only for DA, but also for HA. In this case, having a common spoken/national/official language with the destination country allows migrants to maintain stronger links also with their home country. The same applies when home and destination countries share borders: both HA and DA are higher. In terms of geographical distance between capitals, we observe a weaker positive correlation with DA significant at 5% level. This would indicate that the larger the distance among capitals, the more migrants become attached to the destination. While this could appear to contradict the results obtained with the contig variable, this is not necessarily the case: it may be very well possible that neighbouring countries have large distances among capitals (especially non European countries) and vice-versa non neighbouring countries have small distances between capitals.

4.5 Discussion

In this work, we have developed a novel method to study cultural integration patterns of migrants through Twitter. Different from the existing literature, here we introduced hashtags from Twitter as a proxy for links to cultural traits of the country of origin or of the country of destination, which we call *home attachment (HA)* and *destination attachment (DA)*, respectively. The HA and DA were defined by taking the proportions of country-specific hashtags that either belongs to the country of residence (DA) or the country of nationality (HA). The null model analysis performed to validate the indices showed a significant difference between the actual indices and the null model indices, confirming the validity of our approach. The comparison between the indices and other related variables allowed us to discover interesting relations. First, the proficiency of the language of the host country facilitates higher DA level. Having a common native language with the destination country also contributes to higher DA levels. Interestingly, common languages also increase HA levels, which is a less explored result. Second, we saw that in general, sharing borders also increases both the DA and HA level. At the same time, the further the destination country, the higher the DA level. Through the comparison with Hofstede’s cultural dimensions, we found that the higher the differences between the origin and destination countries in terms of individualism, masculinity and uncertainty, the higher the DA level is. These relationships are found to be the opposite with the HA index.

Having employed social big data for our analysis came with several advantages. We were able to observe real-world social behaviour in an uncontrolled environment, avoiding the risk of having evasive answers, or/and misinterpretation of questions when completing a survey. In addition, unlike surveys which often are incomparable across countries, we were able to conduct a cross-country study of cultural integration of international migrants. It is important to note, however, that employing big data also has its drawbacks. Although we began with a total of about 60 million users, we ended up working with only 3,226 identified international migrants mainly due to the lack of geo-tagged tweets. This shows that such a study

requires very extensive resources to be completed. This analysis also suffers from sampling bias. The Twitter population is different from the real one, hence not all the demographic groups are covered in the analysis. Importantly, privacy and ethical aspects are often raised when using big data that contain personal information, even if the information was made public by the individuals themselves. This becomes particularly important when dealing with specific populations of minorities such as migrants. In this work, neither personal information nor migration status of individuals has been released at any stage of the analysis. The data was securely stored and accessed. All results are aggregated at national level and presented in such a way that re-identification is not possible. In addition, we need to underline the fact that the findings of this paper cannot be generalised. They apply solely to a small sample of the population, and not to larger groups. The study has passed ethics approval before publication.

Chapter 5

Characterising different communities of Twitter users: Migrants and natives

5.1 Introduction

Twitter is one of the microblogging platforms that attracted many users. Unlike some of the other platforms, Twitter is widely used to communicate in real-time and share news among different users [102]. On Twitter, users follow other accounts that interest them to receive updates on their messages, called “tweets”. Tweets can include photos, GIFS, videos, hashtags and polls. Amongst them, hashtags are widely used to facilitate cross-referencing contents. The tweets can also be retweeted by other users who wish to spread the information among their networks. This involves sometimes adding new information or expressing opinion on the information stated. Despite the limit on maximum 280 characters of tweets¹, users are able to effectively communicate with others.

But above all, Twitter has become a useful resource for research. Twitter data can be accessed freely through an application programming interface (API)². On top of this, the geo-tagged tweets are widely used to analyse real-world behaviours. One

¹<https://developer.twitter.com/en/docs/counting-characters>

²<https://developer.twitter.com/en/docs/twitter-api>

of fields of research that makes use of geo-tagged tweets is migration studies. Typically, migration studies have relied on traditional data such as census, survey and register data. However, provided with alternative data sources to study migration statistics in the recent period, many studies have developed new methodologies to complement traditional data sources (See for instance, [96, 75, 76, 161, 112, 145]). While these studies have successfully shown advantages of alternative data sources, distinguishing characteristics and behaviours of migrants and natives on Twitter have not been fully understood.

Here, we aim to study the characteristics and behaviours of two different communities on Twitter: migrants and natives. We plan to do so through a general assessment of features of individual users from profiles and tweets and an extensive network analysis to understand the structure of the different communities. For this, we identified 4,940 migrant users and 46,948 native users across 174 countries of origin and 186 countries of residence using the methodology developed by [96]. For each user, we have their profile information which includes account age, whether the account is a verified account, number of friends, followers and tweets. We also have information extracted from the public tweets which includes language, location (at country level) and hashtags. With these collected data, we explore how each of the communities utilises Twitter and their interests in both the world- and local-level news using the method developed by [98]. Furthermore, we also explore their social links by studying the properties of the mixed network between migrants and natives. We study centrality and assortativity of the nodes in the network.

We discovered that migrants tend to have more followers than friends. They also tweet more and from various locations and languages. The assortativity scores show that users tend to connect based on nationality more than country of residence, and this is true more for migrants than natives. Furthermore, both natives and migrants tend to connect mostly with natives.

The rest of the article is organised as follows: we begin with related works, followed by Section 5.3 on data and the identification strategy for labelling migrants and natives on Twitter. Section 5.4 focuses on statistics on different features of Twitter and Section 5.5 deals with analysis of the different networks. We then

conclude the paper in Section 5.6.

5.2 Related works

Many studies exist that analyse different networks on microblogging platforms. Twitter is one of the platforms that has been studied extensively as it enables us to collect directed graphs unlike Facebook for instance. We can study various types of relationships defined by either a friendship (followers or friends³), conversation threads (tweets and retweets) or semantics (tweets and hashtags). Performing network analysis on these allows us to study properties, structures and dynamics of various types of social relationships.

One of the the first quantitative studies on topological characteristics of Twitter and its role in information sharing is [102]. From this study onward, many have found distinguished characteristics of Twitter’s social networks. According to the study, Twitter has a “non-power-law follower distribution, a short effective diameter, and low reciprocity”. The study showed that unlike other microblogging platforms that serve as mainly social networking platforms, Twitter acts as a news media platform where users follow others to receive updates on others’ tweets. A further study of the power of Twitter in information sharing and role of influencers is [37]. The authors focused on three different types of influence: indegree, retweets and mentions of tweets. They found that receiving many in-links does not produce enough evidence for influence of a user but the content of tweets created, including the retweets, mentions and topics matter equally. The same authors extended the work to observe information spreaders on Twitter based on certain properties of the users which led to a natural division into three groups: mass media, grassroots (ordinary users) and evangelists (opinion leaders) [36]. Furthermore, by looking at the six major topics in 2009 and how these topics circulated, they found different roles played by each group. For example, mass media and evangelists play a major role in spreading new events despite of their small presence. On the other hand,

³Followers are users that follow a specific user and friends are users that a specific user follows. <https://developer.twitter.com/en/docs/twitter-api/v1/accounts-and-users/follow-search-get-users/overview>

grassroots users act as gossip-like spreaders. The grassroots and evangelists are more involved to form social relationships.

Studies that appear in the latter years focused on characteristics on Twitter networks and properties in various scenarios, e.g. political context, social movements, urban mobility and more (See for instance [160, 136, 115]). For instance, [68] studied the network of followers on Twitter in the digital humanities community and showed that linguistic groups are the main drivers to formation of diverse communities. Our work contributes to the same line of these works. But unlike any precedent works, here we explore new types of communities that, to the best of our knowledge, have not yet been explored, i.e., migrants and natives.

5.3 Data and labelling strategy

5.3.1 Data

The dataset used in this work is similar to the one used in [98]. We begin with Twitter data collected by [41], from which we extract all geo-tagged tweets from August 2015 to October 2015 published from Italy, resulting in a total of 34,160 individual users (that we call first layer users). We then searched for their friends, i.e. other accounts that first layer users are following which added 258,455 users to the dataset (called second layer users). We further augmented our data by scraping also the friends of the 258,455 users. The size of the data grew extensively up to about 60 million users. To ensure sufficient number of geo-tagged tweets, all of these users' 200 most recent tweets were also collected. To synthesise the dataset, we focus on a subset of these users for whom we have their social network, and which have published geo-located tweets. This results in total of 200,354 users from the first and second layers with some overlaps present among the two layers.

5.3.2 Labelling migrants and natives

The strategy for labelling migrants and natives originates from the work of [96]. It involves assigning a country of nationality $C_n(u)$ and a country of residence $C_r(u)$

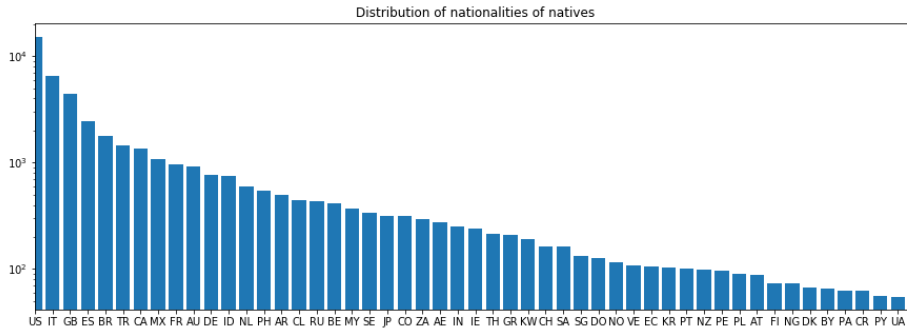


Figure 5-1: Distribution of top 50 nationalities of natives in log scale

to each user u , for the year 2018. The definition of a migrant is “a person who has the residence different from the nationality”, i.e. $C_n(u) \neq C_r(u)$. The strategy to assign a user’s residence requires observing the number of days spent in different countries in 2018 through the time stamps of the tweets. In other words, the country of residence is the location where the user remains most of the time in 2018. To assign nationality, we analyse the tweet locations of the user and user’s friends. In this work, we took into account the fact that tweet language was not considered important in defining the nationality as found in the study of [96]. Thus, the language was not considered here as well. By comparing the labels of country of residence and the nationality, we determined whether the user was a migrant or a native in 2018.

Some users could not be labelled since the procedure outlined in [96] only assigns labels when enough data is available. As a result, we identified nationalities of 197,464 users and the residence 57,299 users. Amongst them, the total number of users that have both the nationality and residence labels are 51,888. Most importantly, we were able to identify 4,940 migrant users and 46,948 natives from our Twitter dataset. In total, we have identified 163 countries of nationalities for natives. From the figure 5-1, we see that the most present countries are the United States of America, Italy, Great Britain and Spain in terms of nationality. This is due to several factors. First because Twitter’s main users are from the United States. Second, we have large number of Italian nationalities present due to the fact that we initially selected the users whose geo-tags were from Italy. Figure 5-2 displays the main migration links in our dataset for the countries that have at least 10 mi-

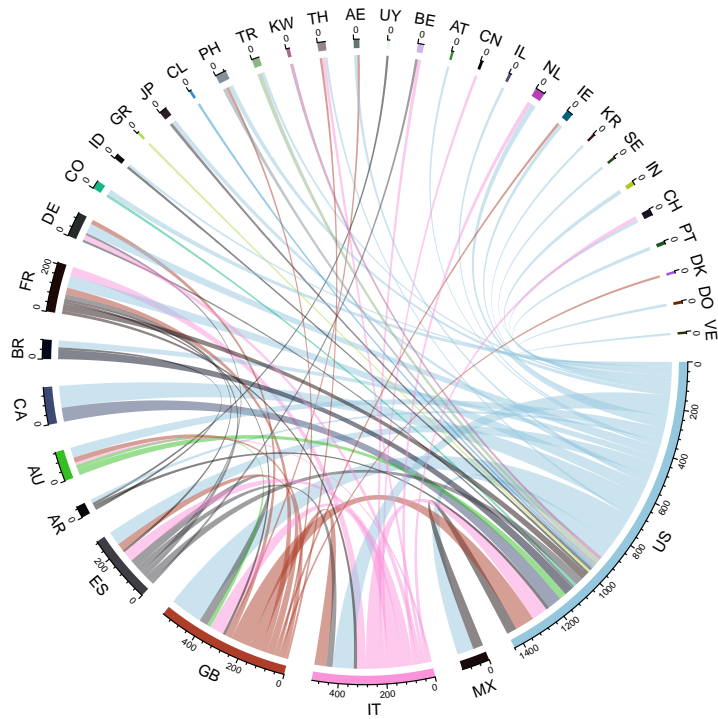


Figure 5-2: Chord diagram on migration patterns: The number of migrants who have moved from a country to another is represented by the links. The colours represent the nationalities of migrants. We show only countries with at least 10 migrants for the visualisation purpose.

grants. Overall, we have identified 144 countries of nationalities and 169 countries of residences for the migrants. In terms of migration patterns, it is interesting to also remark from our data that the U.S. and U.K have significant number of in and out-going links. In addition, France and Germany have mainly in-coming links.

Here, we emphasise that through our labelling process we do not intend to reflect a global view of the world’s migration patterns but simply what is demonstrated through our dataset. However as it is also shown in the work of [98], the predicted data correlate fairly with official data when looking at countries separately. For instance, when comparing predicted data with Italian emigration data of AIRE⁴, we observed a correlation coefficient of 0.831 for European countries and 0.56 for non-European countries. When compared with Eurostat data on European countries, the correlation coefficient was 0.762. This provides us the confidence to employ this dataset to analyse characteristics of different communities through Twitter.

⁴Anagrafe degli italiani residenti all’estero (AIRE) is the Italian register data.

5.4 Twitter features

In this section we look at the way migrants and natives employ Twitter to connect with friends and produce and consume information.

Home and Destination Attachment index

A first analysis concentrates on the types of information that users share, from the point of view of the country where the topics are discussed. In particular, we compute two indices developed by [98] : Home Attachment (HA) and Destination Attachment (DA), which describe how much users concentrate on topics from their nationality and residence country, respectively. We compute the two indices for both migrants and natives; obviously, for natives the residence and nationality are equal and thus the two indices coincide.

To compute HA and DA , we first assign nationalities to hashtags by considering the most frequent country of residence of natives using the hashtags. A few hashtags are not labelled, if their distribution across countries is heterogeneous (as measured by the entropy of the distribution). The HA is then computed for each user as the proportion of hashtags specific to the country of nationality. Similarly, the DA is the proportion of hashtags specific to the country of residence. Thus, the HA index measures how much a user is interested in what is happening in his/her country of nationality and the DA index reflects how much a user is interested in what is happening in his/her country of residence.

As shown in the figure 5-3, the indices clearly behave differently for the two groups: migrants and natives. Similar to [98], we observe that migrants have, on average, very low level of DA and HA . When looking at natives, this index distribution is wider and has an average of 0.447 which is surely higher than the average of migrants. Without a doubt, this shows that natives are more attached to topics of their countries, while migrants are generally less involved in discussing the topics, both for the home and destination country. However, we observe that a few migrant users do have large HA and DA showing different cultural integration patterns, as detailed in [98]. At the same time, some natives show low interest in the country's

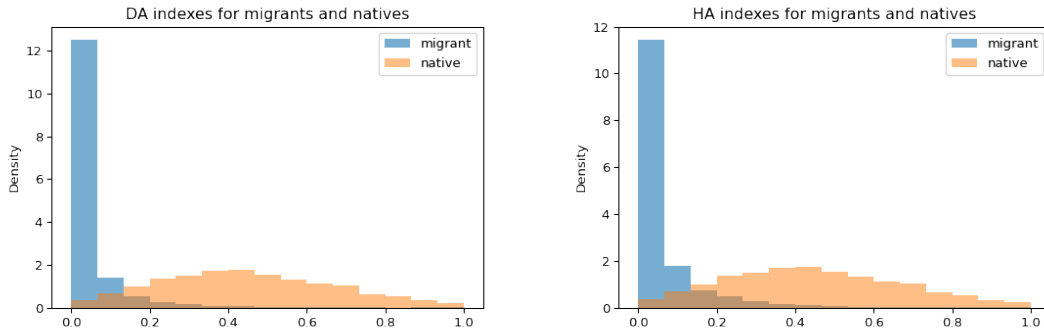


Figure 5-3: Distribution of DA & HA for migrants and natives

topics, which could be due to interest in world-level topics rather local-level topics.

5.4.1 Profile information

Can we find any distinctive characteristics of migrants and natives from the profiles of users? Here, we look at public information provided by the users themselves on their profiles. We examine the distribution of profile information and perform Kolmogorov–Smirnov (KS) test to compare the distributions for migrants and natives.

On the profile, various information are declared by the users themselves such as the joined date, location, bio, birthday and more. We begin by looking at the age of the Twitter accounts from the moment they created their accounts till 2018, as shown in the figure 5-4. We observe that migrants and natives have similar shape of distributions, providing information that there is no earlier or later arrival of one group or another on Twitter. The KS test with high p-value of 0.404 also confirms that the two distributions are indeed very similar.

The other criteria we study show some differences. First, we generally observe that natives have slightly more friends than migrants. On average, migrants follow about 1,160 friends and 1,291 friends for the natives. We can also see from the figure 5-4 that the range of this number is much wider for the natives, ranging from 0 to maximum of 436,299 whereas for the migrants, this range ends at 125,315. The KS test yields a p-value of $1.713e^{-23}$, confirming that the two distributions are different.

Secondly, we observe that the migrants have a larger number of followers. On average, migrants have 10,972 followers versus 7,022 followers for natives (KS p-

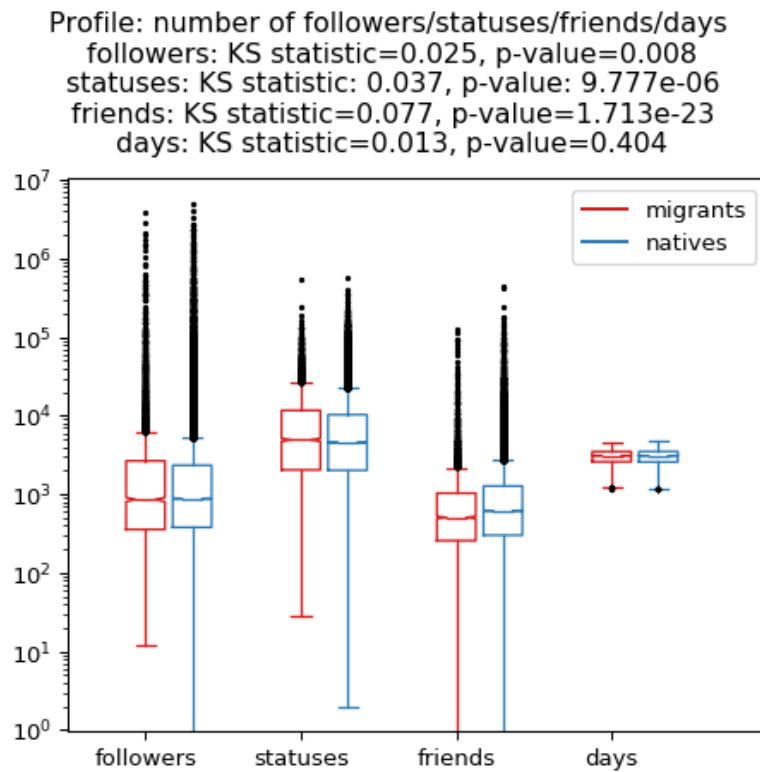


Figure 5-4: Distributions of profile features: number of days since the account was created until 2018, number of followers, number of friends, and number of tweets published (statuses).

value of 0.008). This tells us that there are more users on average that are waiting to get updates on migrant users' tweets. Interestingly, when it comes to the number of tweets (statuses) that users have ever tweeted since the account was created, the number is about 9% higher for the migrants than the natives: average values of 9,836 for migrants and 9,016 for natives, p-value of $9.777e^{-06}$.

We also look at the number of accounts that are classified as verified accounts. The verified accounts are usually well-known people such as celebrities, politicians, writers, or directors and so on. Indeed when looking at the proportion of verified accounts, we observe that this proportion is higher among migrants than natives which partly explains also the higher number of followers and tweets for this group. To be more specific, 5% of the users' accounts are verified accounts among migrants and 3.7% of the accounts are verified accounts among natives.

5.4.2 Tweets

Tweets also provide useful information about user behaviour. We are interested in the locations (country level) and languages a user employs on Twitter. Hence, we look at the number of languages and locations that appear in the users' 200 most recent tweets and computed also the KS statistics to compare the differences between the distributions of migrants and natives. As shown in Figure 5-5, we note that migrants tweet in a wider variety of languages and locations. The two distributions for migrants and natives are different from each other as the KS tests show low p-values; $2.36e^{-194}$ for location and $1.412e^{-38}$ for language.

Since we possess network information, we also studied the tweet language and location information for a user's friends. In Figure 5-6, the two distributions show smaller differences among natives and migrants, compared to Figure 5-5. However, the p-value of the KS test tells us that the distributions are indeed different from one another, where the p-value for location and language distribution for migrants and natives are $3.246e^{-05}$ and 0.005 respectively. Although the differences are small, we observe that the friends of migrants tweet in more numerous locations than those of natives, with average of 29.6 for migrants and 27.4 for natives. However, although the two distributions are different from each other from the KS p-value, the actual difference between average values is very small in the case of the number of languages of friends. In fact, the average for migrants is 30.22 and 30.43 for natives.

These numbers indicate that the migrants have travelled in more various places and hence write in diverse languages than the natives. The friends of migrants tend to have travelled more also. However, no large differences were observed for the number of languages that friends can write in for both migrants and natives.

Popular hashtags

What were the most popular hashtags used by natives and migrants in 2018? In Figure 5-7 we display the top 10 hashtags used by the two communities, together with the number of tweets using those hashtags, scaled to $[0, 1]$. We observe that natives and migrants share some common interests but they also have differences.

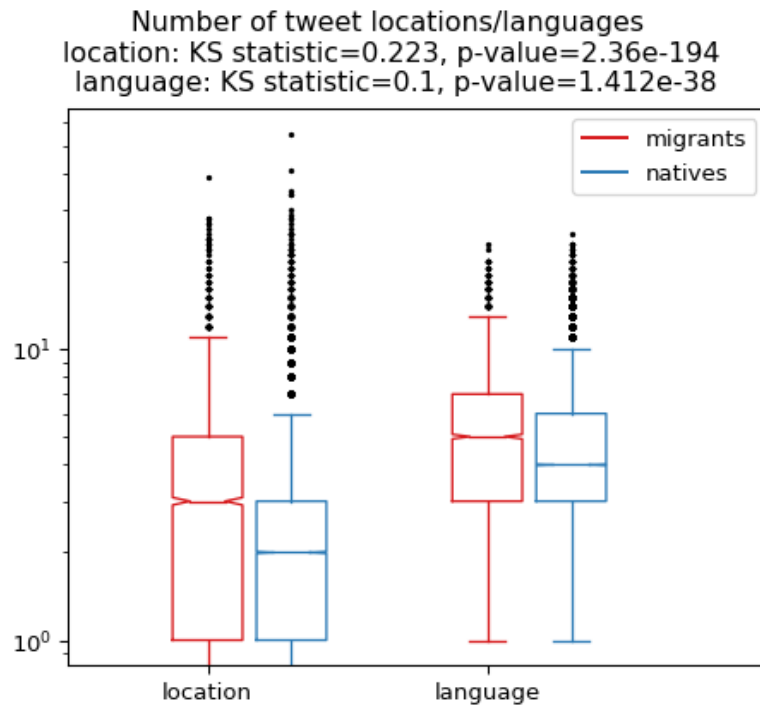


Figure 5-5: Distribution of tweet locations and languages

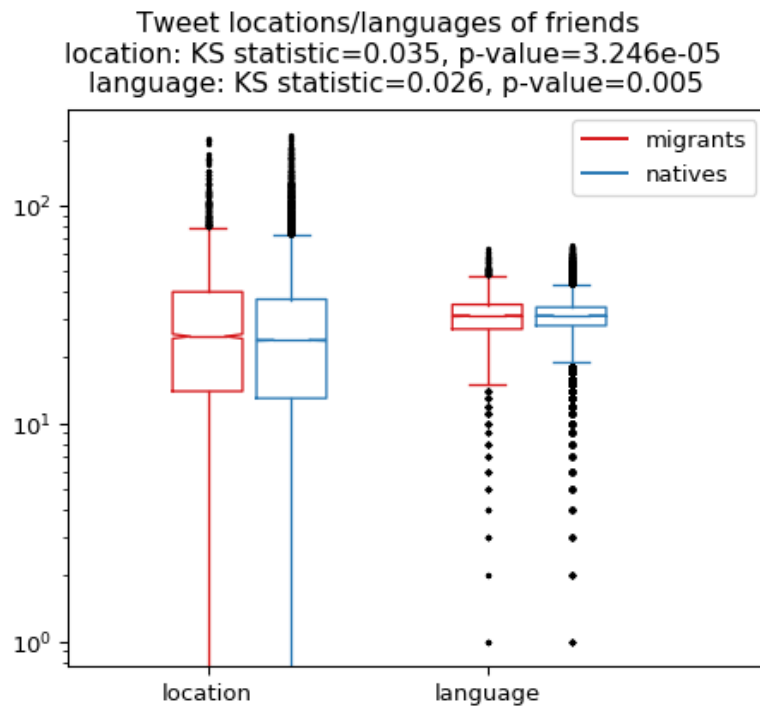


Figure 5-6: Distribution of tweet locations and languages of friends

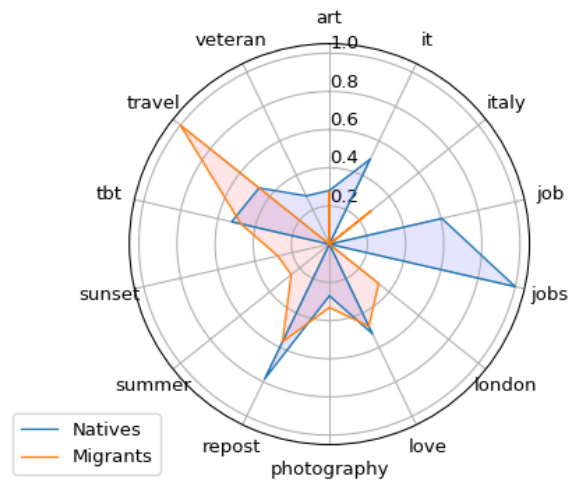


Figure 5-7: Top 10 hashtags used by migrants and natives

For instance, some of the common hashtags between natives and migrants are #tbt, #love and #art. Other hashtags such as #travel, and #repost are in the top list but the usage of these hashtags is much higher in one of the groups than the other. For instance, the hashtag #travel is much more used by migrants than the natives. This is interesting because the number of tweet locations of migrants also reflect their tendency to travel, more than natives. Followed by the hashtag #travel, migrants also used other hashtags such as #sunset, #photography, #summer, and hashtags for countries which show their interests in travelling. On the other hand, natives are more focused on hashtags such as #job, #jobs, and #veteran.

5.5 Network analysis

In this section we perform social network analysis on the social graph of our users to examine the relationships between and within the different communities, i.e., migrants, and natives. Initially, our network consisted of 45,348 nodes and 232,000 edges. We however focus on the giant component of the network which consists of 44,582 nodes and 231,372 edges. Each node represents either a migrant or a native and the edges are directed and represent friendship on Twitter (in other words, our source nodes are following the target nodes). Since we have migrants and natives labels, our network allows us to study the relationship between migrants and natives.

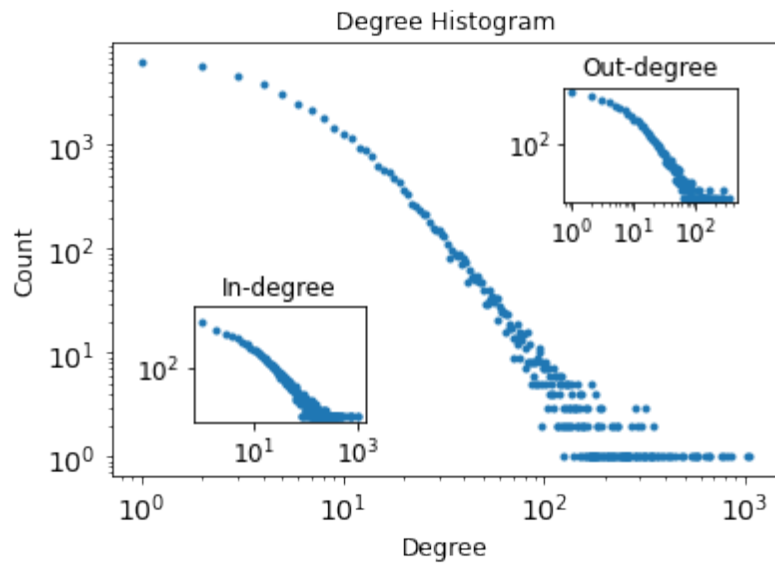


Figure 5-8: Degree distribution of the network.

5.5.1 Properties of the network

In this section we start by looking at density, reciprocity, and shortest path length for the network, and then study node centrality including degree distribution. The average density score of our network tells us that on average each node is connected to other 5.2 nodes. The reciprocity coefficient is low and indicates that only 23.8% of our nodes are mutually linked. This is normal on Twitter as most of the users follow celebrities but the other way around does not happen in many cases. Within the network, the average shortest path length is 2.42, which means we need on average almost 3 hops to receive information from one node to another.

We also compute 7 measures of centrality to study. The measures include all-, in- and out-Degree (Figure 5-8) plus Closeness, Betweenness, Pagerank and Eigenvector centrality measures (shown in Figure 5-9). The Degree centrality measures the number of connections that a particular node has, which can either be an in- or out-going connection. The Pagerank measure considers that nodes with low out-degree are more important. The Betweenness centrality looks at nodes that serve as a bridge from one part of a graph to another. On the other hand, the Closeness centrality looks at how the node is in a most favourable position to control and acquire vital information within the network. Lastly, the Eigenvector measure considers that a

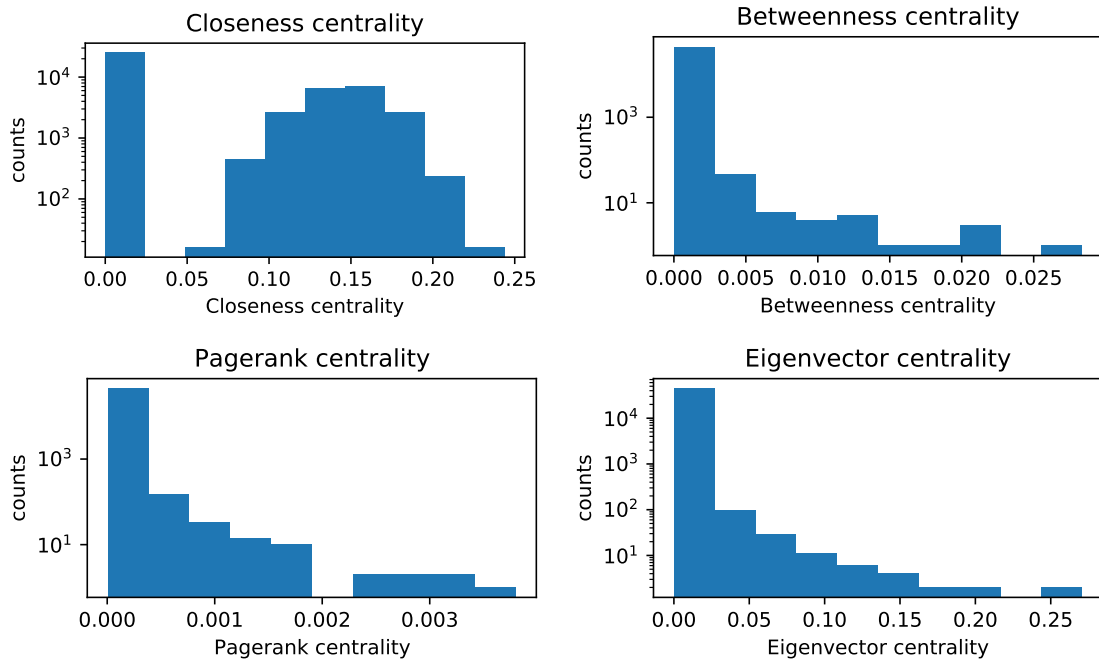


Figure 5-9: Centrality measures of the network.

node is important if the node is connected to other highly connected nodes.

As we can observe from Figure 5-8, the degree distribution follows a power-law distribution with alpha equal to 2.9. This means that a minority of the nodes are highly connected to the rest of the nodes. From Figure 5-9, we observe that most of the users have low centrality while a small number of users show higher centrality values. This is true for all measures, however for closeness the number of users who show higher centrality is larger than for the other measures. This means that many users are well-embedded in the core of the network, and are in a good position to receive information. The distribution of Betweenness, on the other hand, tells us that small part of the users are situated in the most crucial points in case of a diffusion process. We also note that that range of betweenness values is rather narrow, even users with the largest betweenness show a small value. This indicates that information is flowing rather uniformly through the network, and no nodes are particularly important in the process. This was also shown previously by the low average shortest path length. A similar situation arises with the Pagerank measure of centrality: a small minority of users show higher values, however they are all rather low which indicates that generally the network is quite uniform. Lastly, the

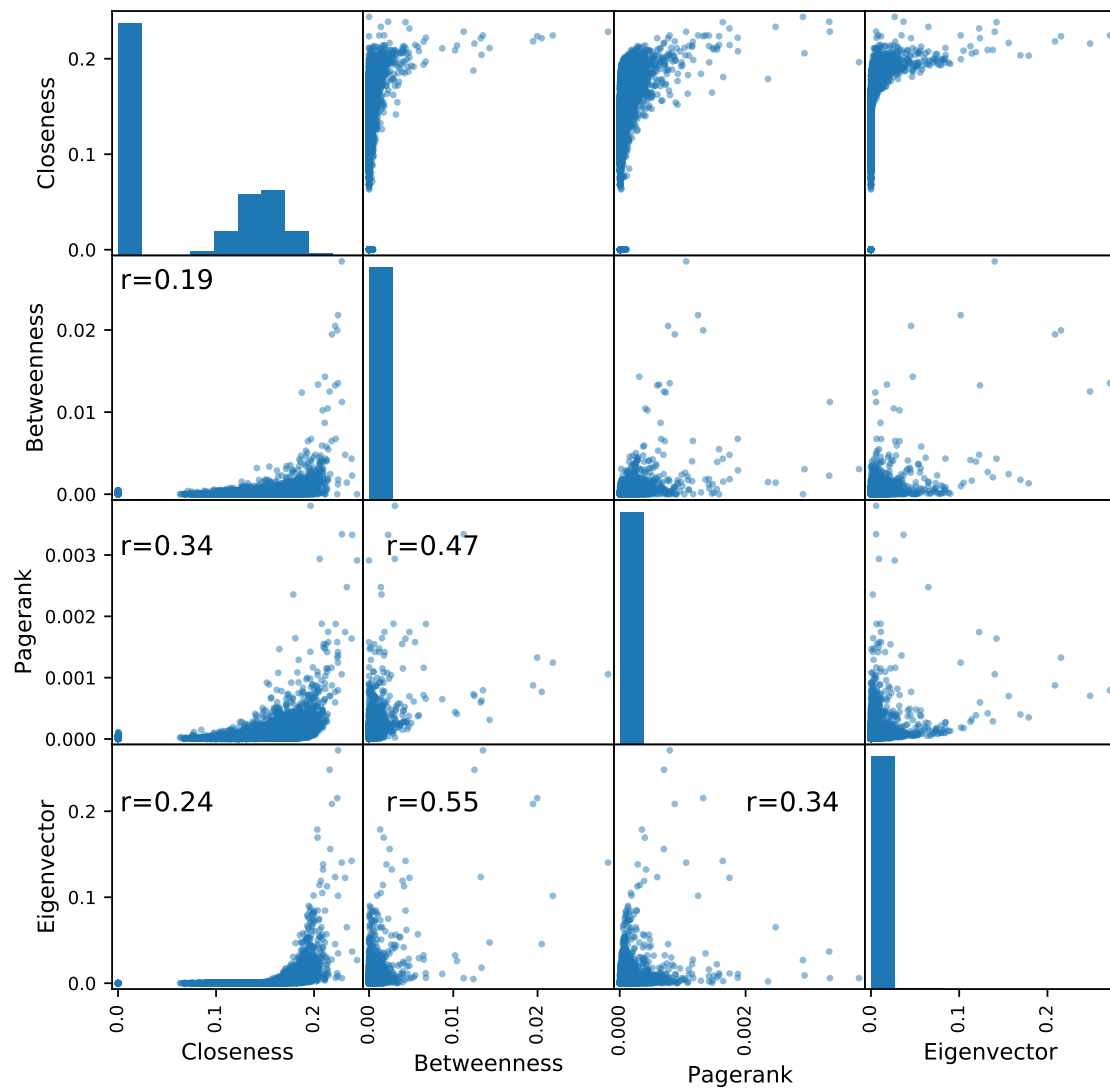


Figure 5-10: Correlation between different centrality measures for network

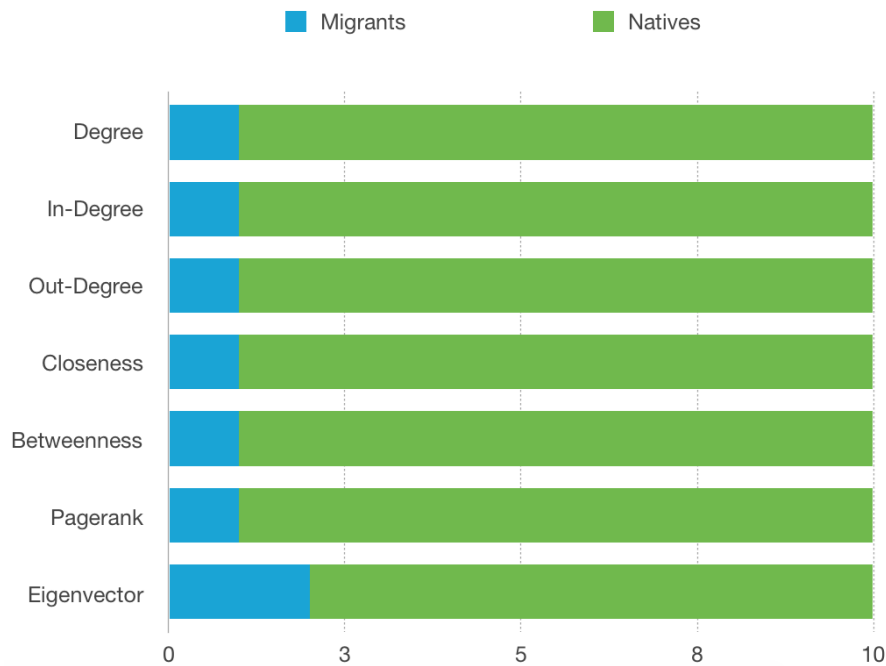


Figure 5-11: Summary of labels of users for top 10 central users by different centrality measures.

Eigenvector centrality reveals that a small part of users has influence even beyond the nodes that are directly connected to. Overall, the centrality measures seem to indicate that while the topological structure of the network is heterogeneous with some nodes showing higher connection and centrality, from the point of view of the flow of information the user in our networks have similar roles.

We continue to examine the centrality of users by computing the correlation between the different measures as shown in Figure 5-10. First of all, we observe a positive relationship among all measures, which is expected, as it means that users who are central from one point of view are also central from another. The Betweenness and Eigenvector centrality measures correlate the most ($r=0.55$). This tells us that users that serve as a bridge between two parts of graphs are also likely to be the most influential user in the network. On the other hand, Betweenness and Closeness centrality measures have the lowest correlation with $r=0.19$. However, the scatterplot shows that those few users who have larger Betweenness also have a large Closeness. The low correlation is determined by the fact that a large majority of users show almost null Betweenness, however Closeness is heterogeneous among

this group. A similar observation can be made for the relation between Closeness on one side and Pagerank and Eigenvector centrality on the other: high Pagerank and Eigenvector centralities always correspond to high Closeness, however for users with low Pagerank and Eigenvector centrality the Closeness values vary.

When checking the labels, in terms of migrant or native, of the most central users, we see that in general these are mostly natives. To be more specific, in Figure 5-11 we show the labels of the top 10 users for each centrality measure. We observe that among the top 10, 8 or 9 users are natives. In other words, most of the nodes have majority of in- and out-going links directed to natives' accounts. This is somewhat expected since in our network only 10% of users are migrants. However, we note that a migrant user is always in the top 3 in Closeness, Pagerank and Eigenvector centrality measures. This tells us that this migrant user has a crucial influence over the network around itself but also beyond its connections.

5.5.2 Assortativity analysis

We now focus on measuring assortativity of nodes by different attributes of individuals, i.e., migrants or natives, country of residence and country of nationality. Assortativity tells us whether the network connections correlate in any way with the given node attributes. In our case this analysis allows us to infer whether and in what measure the network topology follows the nationality or residence of the users, or whether the migrant/native status is relevant when building online social links.

We begin with global assortativity measures. First, the degree assortativity coefficient of -0.046 shows no particular homophily behaviour from the point of view of the node degree. That means high degree nodes do not link with other high degree nodes. However, when we measure the assortativity by different attributes, we observe different results. When looking at the coefficient by the country of residence, the score of 0.55 shows a very good homophily level. The score improves slightly when we examine the behaviour through the attributes of country of nationality (0.6). These values tell us that nodes tend to follow other nodes that share same country of residence and country of nationality, with a stronger effect for the latter. However, when looking at the coefficient by the migrant/native label, we observe no

particular correlation (0.037).

The global assortativity scores are susceptible to be influenced by the size of the data and the imbalance in labels, which is our case especially for the migrant/native labels. Therefore we continue to examine the assortativity at local level, allowing us to overcome the possible issues at global level. We thus compute the scores based on an extension of Newman's assortativity introduced by [126]. In Figure 5-12 we show the distribution of node-level assortativity of migrants and natives, for the three attributes (nationality, residence and migrant/native label). We observe again good homophily for all attributes at local level. However, we remark different behaviour patterns for migrants and natives. Specifically, we see that migrants tend to display lower homophily compared to natives, when looking at the assortativity of nodes by country of residence and migrant/native labels. This tells us that migrant users tend to consider less the country of residence when following other users. Instead, most natives tend to connect with users residing in the same country. When looking at nationality, this effect is less pronounced. While natives continue to display generally high homophily, with a small proportion of users with low values, migrants show a flatter distribution compared to the nationality. Again, a large part of migrants show low homophily, however a consistent fraction of migrant users show higher nationality homophily, as opposed to what we saw for the residence. This confirms what we observed at global level: there is a stronger tendency to follow nationality labels when creating social links. As for the assortativity of nodes by migrant/native labels, we observe that migrants and natives clearly have distinctive behaviours. While natives tend to form connections with other natives, migrants tend to connect with natives as well. This could also be due to the fact that migrants are only about 10% of our users so naturally many friends will be natives (from either residence, nationality or other country). This result is different from what we observed at global level and confirms that the global assortativity score was influenced by the size of the data and the imbalance in labels.

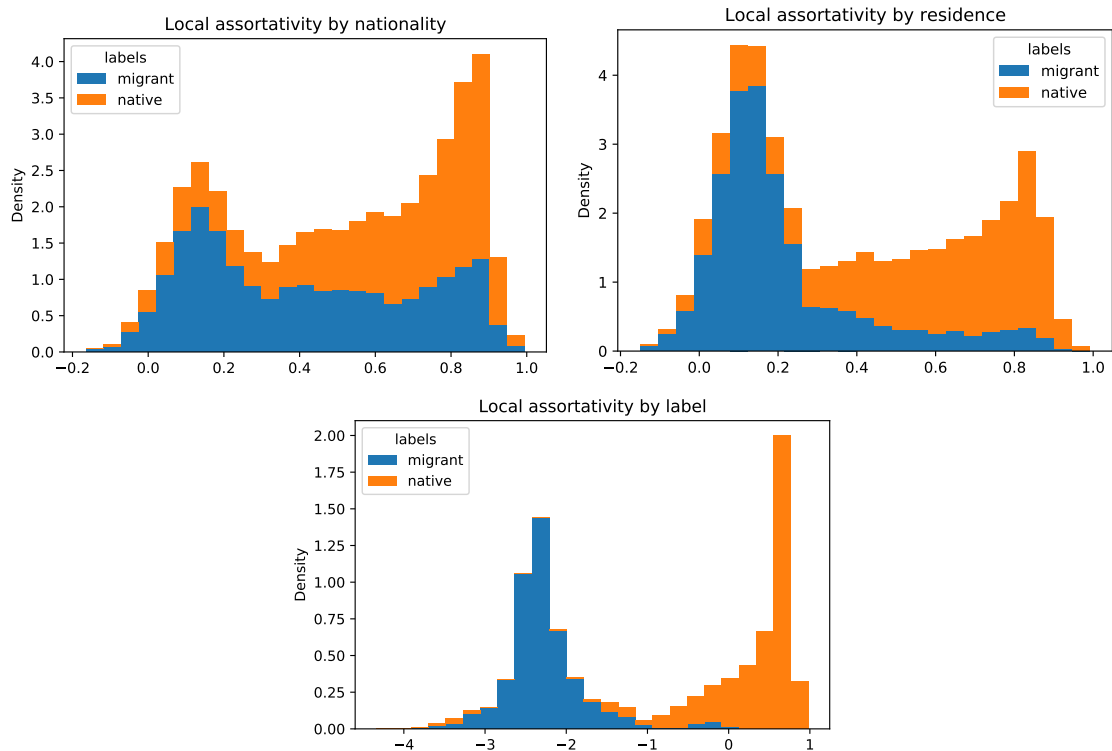


Figure 5-12: Stacked histogram of local assortativity: From the top we have local assortativity by nationality, by residence and by migrant/native label. Please note that the histograms are stacked, therefore there is no overlap between the plot bars.

5.6 Conclusion

We studied the characteristics of two different communities; migrants and natives observed on Twitter. Analysing profiles, tweets and network structure of these communities allowed us to discover interesting differences. We observed that migrants have more followers than friends. They also tweet more often and in more various locations and languages. This is also shown through the hashtags where the most popular hashtags used among migrants reflect their interests in travels. Furthermore, we detected that Twitter users tend to be connected to other users that share the same nationality more than the country of residence. This tendency was relatively stronger for migrants than for natives. Furthermore, both natives and migrants tend to connect mostly with natives.

As mentioned previously, we do not intend to generalise the findings of this work as only a small sample of individual Twitter data was used. However, we believe that by aggregating the individual level data, we were able to extract information

that is worthwhile to be investigated further. To this extent, we simply intend to present what is demonstrated through out dataset. In spite of this drawback, we were able to observe social interactions between migrants and natives thanks to the availability of the Twitter data. In the future, we plan to analyse semantic networks of these users' tweets and hashtags to understand core interests of their discussions and how each community gets involved in different discussions.

“Remember, remember always, that all of us, and you and I especially, are descended from immigrants and revolutionists.”

Franklin D. Roosevelt

Chapter 6

Presence of migrants and shift in voting behaviours of natives

6.1 Introduction

While 540,000 asylum-seekers were living in Germany at the end of 2013 this number rose to nearly 1.6 million five years later [30]. The sudden increase in the number of asylum-seekers in the EU and the difficulties experienced by European governments in coping with their reception resulted in important political repercussions. In Germany, the Alternative für Deutschland (AfD), previously an ‘outfit party’ that ‘combined soft euroscepticism with economic liberalism and socially conservative policies’, transformed into a radical right party [7], and it increasingly focused its discourse and electoral campaigns on immigration, framed as a threat to German’s security and identity [109]. These transformations resulted in the AfD, as many other radical right anti-immigration parties in Europe, increasing its electoral support after 2015 [7]. The EU elections of 2019, which took place at the end of the ‘refugee crisis’, were seen by many experts as the ultimate opportunity for the radical right to take over Europe by exploiting the salience of the immigration issue and Europeans’ negative attitudes towards immigration [10], which had become the main motivation for radical right voting in the previous years [7]. While the predicted far-right ‘surge’ ended in a ‘ripple’ [159], radical right parties increased their consensus in most European countries. In Germany, the AfD obtained 11 per cent

of the votes, resulting in the fourth most voted party, and it became the first party in many of the Eastern regions, and in many electoral districts in the Eastern part of the city of Berlin [31], which revealed a deep, new, East-West divide in German politics [16].

Did exposure to asylum-seekers and refugees have an impact on electoral support for radical right parties such as the AfD? An increasing number of scholars has explored this research question in the last five years [6, 46, 49, 62, 62, 122, 148], and this literature so far has produced contradictory conclusions. Some of these recent works find evidence that exposure to asylum-seekers' reception centres during the 'refugee crisis' increased support for radical right parties [93, 46, 62, 49]. These scholars argue that this happened because these parties did successfully convert prevalent negative attitudes to immigration [12] into vote share during the 'crisis'. Other scholars, conversely, found negative effects of exposure to asylum-seekers' facilities on support for radical right parties in Italy [59], Austria [148], Finland [108] and France [155]. Their findings lend support to the so-called 'contact hypothesis', according to which inter-group contact can effectively reduce prejudice between majority and minority group members (in this case: locals and asylum-seekers), thus decreasing the votes for radical right parties [5].

In addition to assessing the overall impact of exposure to reception facilities on vote shares for the radical right, most of the existing studies have identified a number of contextual variables that contribute to influence this main relationship. Many scholars, for instance, have focused on how the economic context influences the effect of asylum-seeking immigration on votes for the radical right, identifying two opposite mechanisms. [28] for instance, shows that 'under local conditions of material deprivation, measured by the local unemployment rate', the effect of immigration inflows on municipalities' radical right vote share is 'amplified'. [74] and [13] reach similar findings in Austria and Italy. Locals' increased hostility towards newcomers in worse-off areas is explained by these authors as an effect of an increased labour market competition. Conversely, [108] provides evidence for an opposite effect, arguably due to 'the distributional effect of immigration through transfers and taxes', meaning that in areas characterized by high levels of unskilled

immigration and a redistributive tax system, ‘political support for immigration tends to decrease with individual income’ [122].

A few studies have explored the influence of pre-existing immigration levels (stocks) on the relationship between recent immigration flows and electoral outcomes. [93, 146, 13, 38] find that (pre-existing) geographic proximity to immigrants tends to dilute negative reactions to more recent migrant flows and radical-right voting (a finding in line with the contact hypothesis). Conversely, [49] find that larger shares of established migrants within local communities increased the threat natives perceive from recent asylum-seeking flows leading to more opposition to refugees (this finding is in line with group conflict theories, see also: Quillian, 1995; Lahov, 2004). Few studies have also included non economic determinants of individual attitudes on immigration into the analysis. Some scholars, for instance, include in their analysis data about xenophobic feelings [122], religious diversity [13] and local increases of foreign children [122, 74], expected to foster radical-right votes due to locals’ concerns about compositional changes in kindergartens and schools.

By specifically looking at the case of the AfD and radical-right voting in the city of Berlin in the 2019 EU elections, we contribute to this literature in four main respects. First, to the best of our knowledge, all the existing studies analyse the relationship between refugee flows and voting outcomes at the level of municipalities. Unlike them¹, we focus on a lower level of analysis, that of the electoral city district (which, in the case analysed, covers an average area of 1.6 square kilometres). While shifting to this lower level of analysis might not be essential for analyses of small rural municipalities, we do claim that it is of importance when testing the contact hypothesis in bigger cities. It has been shown that the relationship between contact with refugees and votes for the radical right depends on the intensity of contact [144]. Additionally, in bigger cities contact with migrants can be very uneven across neighbourhoods due to the location and size of reception centres. Therefore, we argue that the more aggregated approach adopted by existing studies, i.e. using larger

¹To our best knowledge, the only work on the electoral effects of immigration that focused on the city district level is the one published by Otto and Steinhardt (2014). This article, however, does not focus on refugee migration (but rather on changing concentrations of migrants regardless of the type of migration) and focuses on a very different time period (elections between 1987 and 1998).

districts or municipalities to determine the relationship between the two variables, is not useful in the case of a big city, like Berlin, since this implies assuming that the population in different neighbourhoods experiences the same exposure and contact to refugees, although in reality it might be highly different.

Second, in addition to this focus on the electoral district level, we test the ‘contact hypothesis’ adopting an innovative approach, based on geo-localization techniques and high-resolution spatial statistics, providing a methodological contribution to the existing literature. To the best of our knowledge, none of the existing contributions on the electoral effects of asylum-seeking migration – which mostly develop regression analyses on data aggregated at the municipal level – takes the spatial structure of the data into account. Unlike these existing works, we do not define our independent variable as the mere share of asylum-seekers in the population [144, 155] or the mere presence of refugees in the municipality or any other unit of analysis [148], but we rather construct it as a spatial exposure variable, which depends on the distance of all reception facilities to the centroid of voting districts and the capacity of reception facilities. To test the relationship between these exposure variables and voting outcomes we use spatial regression models to take the spatial structure of our data into account, generating findings that provide support for the ‘contact hypothesis’.

Third, to the best of our knowledge this is the first study that provides comprehensive evidence about the influence of the size or capacity of reception centres within the same city on the relationship between the exposure to refugee migration and votes for the radical-right. To do so, we rely on data about the capacity of our geolocalised reception facilities, showing that bigger reception centres are correlated with less negative effects of exposure on vote shares for the AfD compared to small reception centres. In providing such evidence, we complement findings produced by Schneider-Strawczynski on the influence of the intensity of contact between asylum-seekers and locals – which this author however measures by ‘the number of places available in refugee centres relative to the municipality’s population’ (p.25) – on the relationship between radical right voting and exposure to asylum-seekers. Our contribution also connects with the findings of [61], showing that the size of reception centres (in Southern Italy) influenced locals’ attitudes to immigration.

Fourth, this paper provides evidence for a possible additional mechanism to explain variations in the relationship between immigration and radical-right voting. Our research shows that the negative effect of exposure on radical-right voting is lower in worse-off districts than in better-off districts, a finding in line with the so-called ‘labour market channel’. However, we also identify an additional variable that significantly affects the relationship between migration flows and radical-right voting, and that is linked to the socio-cultural history of the context analysed: the former East-West divide. The socio-cultural history of countries is recognised as a key determinant of public attitudes to immigration [89, 69], and of the electoral support for the radical right [120, 73], but has so far been neglected in the literature on the electoral effects of immigration. [89], for instance, have shown that the diffuse political context or climate in which a cohort of individuals came of age² – and more specifically the ‘contextual exposure to principles of equality and tradition – is central to the formulation of a person’s attitudes towards immigration later in life’ (p.1), with the prevalence of the principle of equality typical of liberal democracies affecting immigration attitudes in adulthood positively, and the principle of tradition typical of non-democratic countries doing so negatively. This finding, these authors point out, might explain the increasing gap in public attitudes to immigration in areas, such as Eastern and Western Europe, which are geographically close but have a different socio-cultural history [114]. As all the analysed studies on the electoral effects of immigration reviewed in this paper focus on single countries, typically in Western Europe, or even on single regions within these countries, they could not investigate this additional mechanism. The unique case of the city of Berlin, due to its peculiar history, represents an ideal setting to assess this under-explored dimension and explore whether, *ceteris paribus*, the relationship between exposure to refugee centres and votes for the radical right varies in contexts with a different socio-cultural history. Our findings indeed point to remarkable differences between East and West Berlin (exposure and AfD vote shares are more strongly correlated in Western districts) which provides evidence for this additional mechanism.

²The scholars define political climate as ‘an ensemble of normative principles, beliefs, ideals, and values that prevail in the political zeitgeist and which are reflected in the views of the ruling political elites’ [89].

The paper is organized as follows. Section 2 reviews the existing literature on the electoral effects of immigration, with a specific focus on works produced during the so-called ‘refugee crisis’, and derives a number of hypotheses, grounded on the existing scholarship and the main theories developed over time by scholars. Section 3 describes our setting, our data, and the variables we constructed to analyse our dataset. Section 4 elaborates our findings and Section 5 discusses the theoretical implications of our results.

6.2 Theory and Hypotheses

Several theories have been proposed in the literature to explain the relationship between exposure to refugees and voting behaviour of the majority population. One strand of theory, referred to as ‘contact theory’ [5], builds upon the argument that interaction and contact between different groups (in this case: locals and asylum-seekers) lead to more tolerance between and positive perceptions of the groups [71, 45, 51], and so decreases the votes for radical right parties [146]. In contrast, ‘group conflict theory’ argues that majority group members will feel threatened by the presence of another racial or ethnic group [106], leading to negative attitudes towards refugees and immigration in general and, therefore, to an increase in votes for the far-right [14].

In recent years, a growing number of researchers has investigated whether and how refugee or migrant inflows affect political behaviour and more specifically, voting behaviour, of the majority population [6, 46, 62, 122, 148]. There is, however, no scholarly agreement in the research literature on the political effects of immigration: some studies find that exposure to asylum-seekers’ reception centres during the ‘refugee crisis’ increased support for radical right parties [93, 46, 62, 49]. Other scholars, conversely, found negative effects of exposure to asylum-seekers’ facilities on support for radical right parties [59, 148, 108, 155]. Interestingly, [144] observes that evidence for group conflict theory is mostly provided by studies that focus on large increases in refugee inflows while evidence for contact theory is mostly provided by scholars who focus on exposure to small-scale refugee inflows, suggesting that

the intensity of contact might influence the relationship between locals' exposure to refugees and electoral outcomes. Several factors have an impact on the intensity of contact between refugees and the rest of the population, like the number of refugees present in an area, the cultural distance between the two groups, or the media coverage of this topic [144]. The cultural distance between the two groups as well as the media coverage of the refugee topic, are two factors situated on a macro level, meaning that the population of Berlin was exposed in the same degree regarding these two aspects. In contrast, the number of refugees also affects people at the micro level [148]: individuals living in neighbourhoods of Berlin in which more refugees were hosted are more exposed to refugees and are more likely to have more intense contact with them than individuals living in neighbourhoods with no reception centres. In our study, we rely on data on the electoral district level for the electoral results as well as on point data for the refugee reception centres. We, therefore, assume that, in our study, exposure to refugees involves a more intense degree of contact, and hence, hypothesize that:

H1: Exposure to reception centres is expected to negatively affect the vote shares for the AfD in Berlin.

How exactly exposure to refugees affects electoral outcomes is also determined by several other contextual factors. Many studies have identified variables related to the neighbourhood and the larger context that contribute to influence the main relationship. One example is the socio-economic status of the neighbourhood. The conditions under which individuals live might affect how they react to the exposure to refugee reception centres and asylum seekers. Although, as hypothesized, exposure to refugees might have a negative effect on vote shares for the AfD, individuals might still feel threatened under some circumstances [25, 135]. For example, the perception of threat might emerge because of the view that refugees are competitors on the job market or in a fight over other limited resources, like financial support [106, 134]. Especially in areas or neighbourhoods that are in a poor socio-economic situation, marked by unemployment and poverty, natives might feel more threatened by refugees [40] and, hence, be more likely to vote for radical right parties. For instance, [28] shows that 'under local conditions of material deprivation, measured

by the local unemployment rate’, the effect of immigration inflows on municipalities’ radical right vote share is ‘amplified’. [74] and [13] reach similar findings in Austria and Italy. Therefore, our second hypothesis states:

H2: The negative foreseen impact of exposure to reception centres on votes for the AfD is expected to be bigger in rich or better-off districts compared to poor or worse-off districts.

In addition to the neighbourhood’s socio-economic status, the share of immigrants who are already residing within a neighbourhood is an important aspect to take into account [146, 13]. If an immigrant group is already present in a context, any positive effects of the arrival of new refugee groups on far-right voting behaviour might be reduced. The presence of many different groups lessens the salience of any single group [106]. Therefore, geographic proximity to already settled immigrants tends to dilute negative reactions to more recent migrant flows and radical-right voting. This hypothesis is supported by findings provided by [146, 13, 38]. [146] finds, for example, that effects to immigration, like increased support for the extreme right or cultural anxieties, are only present during the initial phase of migration. Once the migrant share reaches a certain threshold, additional immigration does not further influence the support for the extreme right, because people are directly exposed to immigrants on a daily basis [146]. [13] reach similar conclusions in the case of big Italian cities. Therefore, our third hypothesis states:

H3: The negative foreseen impact of exposure to reception centres on votes for the AfD is expected to be bigger in districts with a high share of established non-European residents compared to districts with a low share of established non-European residents.

However, following [83] – who have studied effects on attitudes to immigration of the interaction between the concentration of immigrants and neighbourhoods’ socioeconomic conditions – we do expect this effect hypothesised in H3 to vary depending on the socio-economic status of our districts and to disappear in the most deprived areas.

The specific setting of our study, Berlin, also allows us to introduce another context factor into our analysis: the East-West difference. Research shows that

the diffuse political context or climate, in which a cohort grows up, influences their attitudes towards immigration later in life [89]. Historically, Eastern Germany and Western Germany have experienced different political systems and contexts, just like East Berlin and West Berlin. This suggests that also the attitudes towards immigration of residents in East and West Berlin differ. Studies in fact show an East-West divide in Germany in immigration opinions that has been stable over time [152] and that immigration is a much more important determinant of electoral support for the AfD in East Germany than in West Germany [120]. Based on these arguments, our fourth hypothesis states:

H4: The negative foreseen impact of exposure to reception centres on votes for the AfD is expected to be bigger in West Berlin compared to East Berlin.

Given the different salience of the migration issue in contexts with different socio-economic characteristics [28], we also hypothesise that the effect predicted in H4 might vary depending on districts' socio-economic status.

Apart from contextual variables that can influence the relationship between exposure to reception centres and vote shares for the AfD, aspects concerning the amount of exposure are also important to consider. In other words, the link between contact or exposure to refugees and voting for the far right does not have to be linear [25]. It is possible, for example, that a threshold effect or a tipping point exists, where the negative effect of exposure to refugee centres on right-wing voting, lessens or becomes positive. Scholars have argued that the increase of out-group members (in this case: refugees) can have very different effects depending on the total out-group size [135, 45]. These thresholds or tipping points were, for example, found in studies analysing residential mobility [143, 35, 3] showing that the relationship between exposure to refugees and radical right voting depends on the perceived contact intensity, i.e. on how much contact is perceived as potentially disruptive. A recent study on attitudes towards immigration in Southern Italy [61] suggests that the capacity of reception centres might characterise this disruptive contact, demonstrating that the bigger the size of asylum-seekers' reception centres, the stronger their effect on locals' negative attitudes to immigration. We therefore include in our analysis another important variable, which is the size of reception centres.

H5: The negative foreseen impact of exposure to reception centres on votes for the AfD is expected to be smaller in districts which contain big reception centres within them, compared to districts that contain small reception centres.

6.3 Data and Methods

6.3.1 Setting

Our study focuses on the city of Berlin, the capital of Germany and the German city that received the highest number of asylum-seekers during the ‘refugee crisis’ [90]. In 2014 Berlin had around 3.5 million inhabitants [18] and between 2014 and 2018 the number of asylum-seekers hosted in the city rose from 33,000 to 98,270. Once sent to the city by the national government, asylum-seekers were dispersed by the authorities in reception centres, providing group accommodation for several hundred individuals³. Decisions about the location of reception centres were taken by a dedicated administrative unit of the Senate of Berlin the Landesamt für Flüchtlingsangelegenheiten (State Office for Refugee Affairs)⁴. Official sources suggest that, besides considerations regarding the quality of proposed housing units and infrastructural aspects, decisions about the location of refugee facilities were to a large part driven by the immediate availability of suitable buildings [17]. As a result, some districts of the city hosted several reception centres, while other districts did not host any reception facility, as shown in Figure 6-2a and 6-2b. Approximately the same number of reception facilities was located in East and West Berlin (45 in East Berlin and 40 in West Berlin) and facilities in the East and West have a similar

³Reception centres for asylum-seekers are of two types. The so-called *Erstaufnahmeeinrichtungen* provide accommodation to newly arrived asylum-seekers for a period of 6 months, while the so-called *Gemeinschaftsunterkünfte* are group housing facilities meant to offer accommodation to asylum-seekers that do not find any private accommodation after the end of the six months (SPI 2017). While these accommodations have to fulfill state-specified quality criteria and need to be accredited by the State Office for Refugee Affairs (Landesamt für Flüchtlingsangelegenheiten, LAF) they are usually run by private enterprises or non-governmental organizations.

⁴In Germany, after asylum-seekers were registered and had officially engaged in the process of claiming asylum, they were dispersed across the sixteen German Federal States in accordance with a fixed proportional system, so-called ‘Königstein key’ (Juran and Broer 2017). Federal States are, in a second step, responsible for housing these asylum claimants within their jurisdiction and, consequently, for their distribution across individual facilities.

average capacity (average capacity of 332 in whole Berlin, 335 in East Berlin, and 329 in West Berlin). However, the distribution of reception facilities across electoral districts within East and West Berlin was not homogeneous.

Importantly for our analysis, in order to exclude endogeneity and self-selection issues, we need to check whether policymakers, when deciding where to locate reception facilities, took into consideration the local support for the radical right, for instance deciding to disperse asylum-seekers in areas where AfD support was lower, deliberately avoiding putting refugees where there was pre-existing anti-refugee sentiment. As graphically shown in panel 6-2a – where we have mapped the location of reception centres and the share of votes obtained by the AfD in the 2014 EU election⁵ (which took place before the so-called ‘refugee crisis’) – reception facilities seem indeed not to have been located in districts where support for AfD was higher in 2014. To further corroborate these insights, we have compared covariate means between treated and untreated units, i.e. between electoral districts with and without reception facilities. Findings from this analysis are illustrated in Figure 6-1. The first panel of the figure shows that on average, support for the AfD in the 2014 elections was slightly higher in treated units compared to untreated ones, which suggests that when selecting the districts where the refugee facilities are located, policymakers did not take the pre-existing support for the AfD into account. The second and third panels of Figure 6-1 show that on average, the treated and untreated units, did not differ in terms of the levels of socio-economic deprivation or the number of established non-European residents (see the following section for the definition of these variables). T-tests confirm that the variables do not significantly differ between treated and untreated units. Based on these considerations, we therefore assume that reception centres were distributed across the city on a quasi-random basis, depending on the availability of buildings across the city (a similar

⁵Unfortunately, some significant changes in the geographies of electoral districts for the EU elections between 2014 and 2019 made it impossible for us to conduct our analysis with a focus on the shift in votes for the AfD between the two elections as our main independent variable. In any case, this might have been problematic for other reasons, considering the radical changes in the electoral programme of the party that took place between the two elections and the very different composition of the party’s electorate: as already mentioned, the AfD, previously an ‘outfit party’ that ‘combined soft euroscepticism with economic liberalism and socially conservative policies’, transformed into a radical right anti-immigration party only after 2014 [7].

assumption is made by other studies e.g. [49]).

6.3.2 Data

We use data from various sources. Our four main data sources are: the results of the Elections to the European parliament (in the following: EU elections) for Berlin of the year 2019, a list with addresses and information of housing facilities in Berlin [79], the Data for Integration (D4I) dataset [4] and a number of socio-economic data [20]. Our unit of analysis is the voting district for the EU elections or, more precisely, the absentee voting districts (henceforth: voting districts), based on the availability of the EU election results [19]. Of the 489 absentee voting districts available, 317 were located in West Berlin and 172 were located in East Berlin. Figure 6-2b, illustrates the 489 voting districts, showing that the AfD obtained more votes in Eastern districts (on average: 12.1%) compared to Western districts (on average: 8.3%). The list of reception centres (derived from: [20]) contains information on 85 facilities, including their specific locations as well as information on their capacity (the smallest centre hosted 90, the biggest hosted 1,024 asylum-seekers). To test the accuracy of the data, these have been double checked with official data for Pankow [123], one of the districts of Berlin.

The D4I dataset is based on the statistics of the 2011 Census and provides data about the share of non-European residents in 100 meters by 100 meters cells [4]. These cells have been aggregated in order to derive data about the share of migrant residents in each electoral district. Non-European residents represented 7.9 percent of Berlin's population in 2011. Looking separately at East and West Berlin significant differences emerge. While non-European residents made up only about 4.2 percent of the population in East Berlin, they comprised on average 9.8 percent of the population in West Berlin. Figure 6-2c shows the percentage of non-European residents in 2011 (henceforth: established non-European residents) across all districts. Districts with high shares of non-European residents are mostly clustered in areas in central Berlin, with the large majority of them being located in West Berlin. Finally, we use a number of socio-economic data available at the district level [20], to construct a variable which describes the socio-economic deprivation of

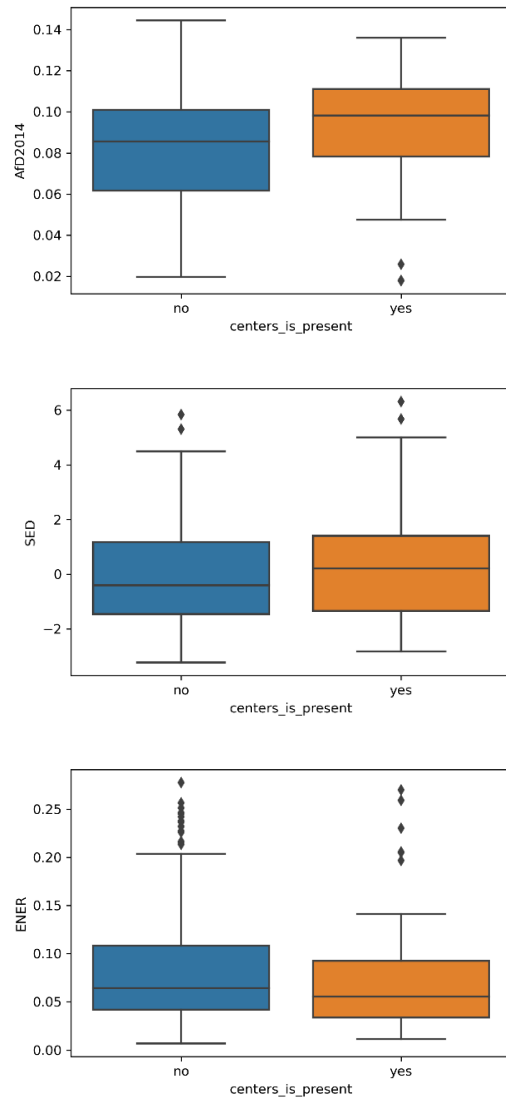
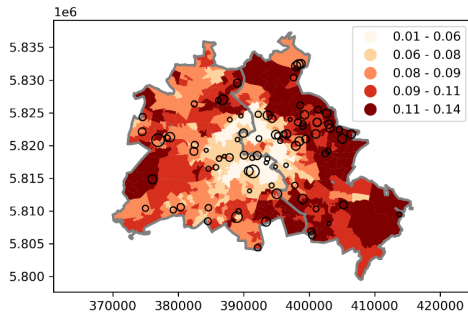
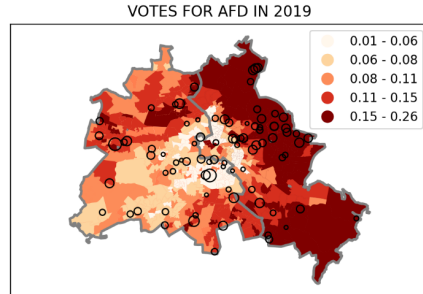


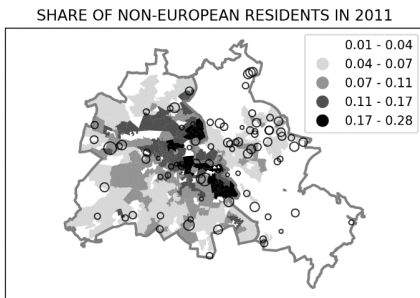
Figure 6-1: Comparison of covariate means: Panel a. AfD shares of votes in the 2014 EU elections. Panel b. Socio-economic Deprivation (data from 2016). Panel c. Concentration of Established non-European Migrants



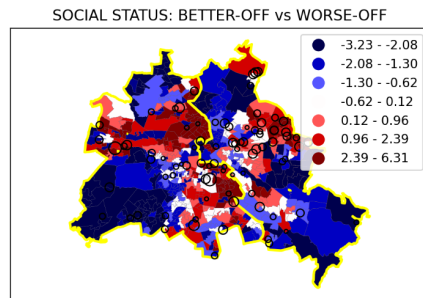
(a) Location of Reception Facilities.



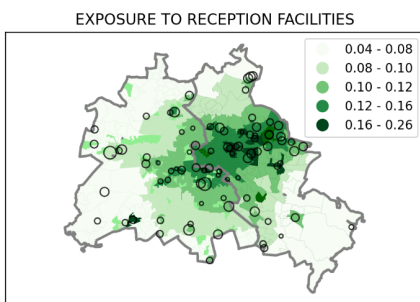
(b) AfD share in 2019 EU elections.



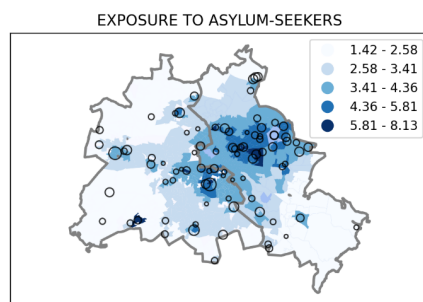
(c) Share of established non-European residents in 2011.



(d) Districts' socio-economic deprivation (blue=better-off; red= worse-off).



(e) Exposure to Reception Facilities.



(f) Exposure to Asylum-seekers.

Figure 6-2: Description of Data

each district (see next section). As illustrated in figure 6-2d, worse-off districts are concentrated in the central areas of the city in both East and West.

6.3.3 Variables

Main Variables

Our dependent variable is the share of votes for the AfD in the EU election in 2019, at the electoral district level. We are interested in investigating the correlation between these vote shares and the exposure of electoral districts to reception centres for asylum-seekers. Based on the information available, we created two main independent variables to measure this exposure. With the first variable, we measure the exposure to reception facilities (EF). Our exposure to reception facilities measure is calculated as follows:

$$EF_i = \sum_{f=1}^n \frac{1}{D_f} \quad (6.1)$$

where i represents the voting district, f the reception facility, and D_f the distance of the reception facility to the centroid of the voting district. EF_i can be interpreted as the mere exposure within a voting district to all reception facilities in Berlin, irrespective of their size. The higher the value of EF , the more exposure a voting district has to reception facilities in the city. Our second exposure variable has been constructed to account for the different capacity or size of reception centres. We call this second variable exposure to asylum-seekers. This variable is calculated as follows:

$$EA_i = \sum_{f=1}^n \frac{N_f}{D_f} \quad (6.2)$$

where i represents the voting district, f the reception facility, N_f the capacity of the reception facility, and D_f the distance of the reception facility to the centroid of the voting district. EA_i can be interpreted as the exposure within a voting district to asylum-seekers hosted in reception centres in Berlin. The value of EA for a specific voting district increases if that voting district is close to reception facilities

hosting, in total, a high number of asylum-seekers. Therefore, a higher value of *EA* means that a voting district is exposed to a highest number of asylum-seekers hosted in reception centres, while a smaller value means that a voting district has less exposure to asylum-seekers hosted in reception centres in Berlin. Figures 6-2e and 6-2f show the two exposure variables across all districts in Berlin. Looking at the exposure to reception facilities (Panel 6-2e), we see that higher values on this variable are clustered towards the central and north-eastern parts of city. Exposure to asylum-seekers (Panel 6-2f) is higher in the centre of Berlin and in a number of districts in East Berlin.

Other Variables

We include additional variables in our models to test certain mechanisms that might be of importance in explaining the association between exposure to reception facilities and AfD vote shares. The first of these variables is the share of established non-European residents in each district in 2011 ('ENER'), measured as their percentage of the total population within each voting district in 2011. The second variable has been constructed to provide information about the socio-economic deprivation of each district. Using principal component analysis, the socio-economic deprivation variable ('SED') integrates four different socio-economic aspects for each voting district: the unemployment rate, the rate of long-term unemployment, the rate of welfare recipients, and the child poverty rate [20]⁶. Third, we include the region of each voting district, distinguishing between East and West Berlin. The table 6.1 shows the descriptive statistics of all variables and Table 6.2 shows the descriptive statistics by region (East and West Berlin). Fourth, in order to test our fifth hypothesis, we have introduced a new variable which we call "total capacity" and which measures the number of asylum-seekers hosted within reception centres located within the district or at a maximal distance of 1,000m from the borders of the district.

⁶These data are available only until 2016.

	Total (N=489)					
	Value	%	Mean	SD	Min	Max
Share of votes for AfD in 2019			0.1	0.051	-0.014	0.257
Exposure to asylum-seekers			3.36	1.24	1.419	20.38
Exposure to reception facilities			0.00	1	-1.682	11.112
Share of established non-European residents			0.079	0.054	0.007	0.277
Socio-economic Deprivation			0	1.954	-3.229	6.313
Total capacity			593	448.137	89	2086
Region	1–West Berlin	64.83				
	2–East Berlin	35.17				

Source: EU election results for Berlin 2019, reception facility address list, D4I data, socioeconomic data; own calculations.

Table 6.1: Descriptive Statistics of all Variables

	East Berlin (N=172)				West Berlin (N=317)			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Share of votes for AfD in 2019	0.126	0.061	0.025	0.257	0.086	0.038	0.014	0.186
Exposure to asylum-seekers	3.96	1.613	1.452	20.38	3.035	0.814	1.419	7.299
Exposure to reception facilities	0.5	1.22	-1.63	11.11	-0.26	0.742	-1.68	3.36
Share of established non-European residents	0.042	0.026	0.007	0.133	0.098	0.056	0.020	0.277
Socio-economic Deprivation	-0.463	1.49	-3.23	3.87	0.251	2.122	-3.12	6.313

Table 6.2: Descriptive Statistics of all Variables by Region

6.3.4 Methods: Spatial Autoregressive Model or Spatial Lag Model

We investigate the effect of our independent variables exposure to reception facilities EF_i and exposure to asylum-seekers EA_i to vote shares for the AfD V_i using spatial autoregressive models which take the following form:

$$V_i = pWS_i + \beta_1 * EF_i + X'\beta_3 + \epsilon_i \quad (6.3)$$

$$V_i = pWS_i + \beta_1 * EA_i + X'\beta_3 + \epsilon_i \quad (6.4)$$

where V_i is the spatially lagged dependent variable and p is the spatial auto-regressive parameter representing the effect of neighbourhoods' share of votes on the district's own share of votes. X is our set of control variables and β_1, β_3 are the parameters to be estimated. The RMSE⁷ is 0.0863 for the model including the EF variable, and 0.0862 for the model computed with the EA variable. We choose to use the spatial

⁷The Root Mean Squared Error (RMSE) is the standard deviation of the residuals (or the differences between predicted values and observed values). It indicates how close the data is to the line of best fit.

auto-regressive model after careful examination of other spatial models. From Figures A-1, A-2, A-3, A-4 in Appendix, which plot model residuals, it is clear that we have clusters of districts showing evidence of spatial auto-correlation. The results of the Moran's I test also reject the null hypothesis that the value is independently normally distributed. Moreover, the result of the Lagrange multiplier diagnostics for spatial dependence shows that the spatial lag model is the suitable spatial regression model to be applied. Additionally, to the spatial lag model, we also computed OLS models with robust standard errors. The results obtained with the OLS models are similar to the ones obtained using the spatial lag models.

6.4 Findings

Tables 6.3 and 6.4 show our spatial regression models (the corresponding OLS models are included in Table A.1 in the Appendix). Model 1 shows the results of our independent variable 'exposure to reception facilities' (EF), while controlling for several possible confounders. Three different types of effects are shown in the table. Three different types of coefficients are shown in the table. The direct coefficient reports the effect of each independent variable on the share of votes for the AfD within each district, whereas the indirect coefficient reports the effect of each independent variable on the share of votes for the AfD of the neighbouring districts. Finally, the total coefficient presents the combined impact of each independent variable from within the district and from the neighbouring districts. The model shows that EF is significantly and negatively related to the vote share of AfD in 2019. This means that in districts that were more exposed to reception facilities during the 'refugee crisis' (independently of their size), AfD on average obtained less votes compared to districts that were less exposed to reception facilities. This finding provides some evidence in support of our hypothesis H1a, suggesting that exposure to reception facilities is negatively correlated with the share of votes for the AfD. Most of the control variables show expected effects in this first model: the AfD has obtained less votes in districts located in West Berlin compared to districts located in East Berlin. The districts' socioeconomic deprivation also plays an important role. In

	Direct	Indirect	Total
Exposure to reception facilities (EF)	-0.011**	-0.048**	-0.059**
Share of established non-European residents (ENER)	-0.157***	-0.707***	-0.864***
West (East=0)	-0.007**	-0.034**	-0.041**
Socio-economic deprivation (SED)	0.007***	0.03***	0.037***
AIC			-2483.8
AIC for lm			-1876.3

Table 6.3: SAR Model 1 (computed with the EF variable). Dependent variable: share of votes for the AfD in EU Elections 2019. N=489. Sources: EU election results for Berlin 2019, reception facility address list, D4I data, socioeconomic data; own calculations. Coefficients, Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

	Direct	Indirect	Total
Exposure to asylum-seekers (EA)	-0.008*	-0.038*	-0.046*
Share of established non-European residents (ENER)	-0.172***	-0.801***	-0.973***
West (East=0)	-0.006*	-0.028*	-0.034*
Socio-economic deprivation (SED)	0.007***	0.032***	0.039***
AIC			-2481.5
AIC for lm			-1856.5

Table 6.4: SAR Model 2 (computed with the EA variable). Dependent variable: share of votes for the AfD in EU Elections 2019. N=489.

districts with a higher socioeconomic deprivation (or ‘worse-off’ districts), the AfD gained more votes than in ‘better-off’ districts. Remarkably, the share of established non-European residents in each district also contributes to explain the share of votes obtained by the AfD. The negative coefficient for this variable means that in districts with a higher percentage of established non-European residents, the AfD has obtained less votes compared to districts with a lower percentage of established non-European residents in 2011.

Model 2 in Table 6.4 shows the results of an alternative model to the one just presented, in which we replaced the exposure to reception facilities (*EF*) with our second exposure measure, the exposure to asylum-seekers (*EA*), while controlling for the same possible confounding variables. The total coefficient of the *EA* (-0.046) is similar but, in its absolute value, slightly smaller than the coefficient of the *EF* variable in Model 1 (-0.059). The control variables show very similar effects and significance as in the first model. Therefore, even when using the *EA* instead of the *EF* variable, we find some support for our hypothesis H1.

In order to test our second, third and fourth hypotheses, we have created five

additional models, whose findings are illustrated in Table 6.5. These new models introduce a number of interactions between our independent variables, in order to provide insights on how the relationship between *EA* and votes obtained by the AfD in 2019 varies in different contextual conditions (all models have been also computed replacing the variable *EA* with *EF*, and they provided very similar findings). These new models provide several interesting insights. To make sense of the interaction effects of the multivariate models, in Figure 6-3, we have elaborated some simple visualizations of the data points regarding the relationship between our main variables.

First, to test our second hypothesis, in Model 3, we introduce an interaction between our main independent variable – the exposure to asylum-seekers – and the socio-economic deprivation variable. The interaction is statistically significant, meaning that overall, in Berlin, the effect of exposure to asylum-seekers on the votes for the AfD is different depending on the socio-economic status of the districts. Panel a of Figure 6-3, in which districts with a high or low SED values have been marked with different colours, helps us making sense of the coefficient of the interaction term. It suggests that in better-off districts higher values of exposure are associated with lower shares of votes for the AfD. The trend is much less noticeable or essentially non-existent in the case of worse-off districts. Our findings, therefore, do provide support for our second hypothesis, according to which exposure to reception centres is correlated with votes for the AfD more positively or less negatively in worse-off districts compared to better-off districts. In model 4, we have included instead an interaction between the exposure to asylum-seekers (*EA*) and the share of established non-European residents (*ENER*), which allows us to explore our third hypothesis. The interaction term is not significant, meaning that overall, in Berlin, the effect of the *EA* on the vote share for the AfD does not vary depending on the share of established non-European residents. To further corroborate this finding, we have also included a fifth model in which we introduced a triple interaction between *EA*, *ENER* and *SED*, to test whether the *ENER* variable interacted with the *EA* and *SED* variables in a way which influenced the vote shares for the AfD. This interaction term also proved not to be significant. Expectations based on our third

hypothesis, therefore, are not supported by our findings.

Models 6 and 7, have been computed to investigate our fourth hypothesis, regarding the effect of the geographical location of districts across the West/East divide on the main relationship between exposure to asylum-seekers and vote shares for the AfD. In model 6 we simply introduced an interaction term between exposure to asylum-seekers and the West/East variable. This interaction term is not statistically significant, meaning that, overall, the correlation between exposure and votes for the AfD does not depend on the geographical location of the districts across East and West Berlin. In Model 7, instead, we have introduced a triple interaction between *EA*, *SED* and the West/East variable, to test whether the West/East variable interacted with the *SED* and *EA* variable in a way that influenced our dependent variable (the vote shares for the AfD). This interaction term proves to be statistically significant. Model 7, therefore, suggests that the West/East variable interacts with the *SED* variable in influencing the effect of exposure to asylum-seekers on vote share for the AfD. In this model To make sense of the insights produced by Model 7, we illustrate the relationship between our main variables in Panels 6-2b and 6-2c, which illustrate trends in West and East Berlin, respectively. These graphs allow us to better understand the correlation between exposure to asylum-seekers and vote shares for the AfD in better-off and worse-off districts in both East and West Berlin. As can be seen from the figure, the exposure to asylum-seekers shows a negative relationship with the AfD vote shares in West Berlin, independently on the socio-economic status of the district. Conversely, in East Berlin, a negative relationship is evident only in better-off districts. In Eastern worse-off districts, higher values of the exposure to asylum-seekers seem to be correlated with only slightly higher votes for the AfD (the line is almost flat in the graph).

Finally, we focus on our fifth and final hypothesis, according to which exposure to bigger reception centres is expected to affect votes for the AfD more positively or less negatively compared to exposure to smaller reception centres. To test this hypothesis, we have followed a different strategy, compared to the one adopted so far. First, we selected those electoral districts that had at least one reception centre situated within their borders or within a distance of 1,000 meters from their borders.

	Model 3	Model 4	Model 5	Model 6	Model 7
Exposure to asylum-seekers (EA)	-0.032	-0.057	-0.044	-0.034	-0.033
Share of established non-European residents (ENER)	-1.024***	-1.253*	-0.022	-0.937***	-0.655***
West (East=0)	-0.031*	-0.033*	-0.060***	-0.004	-0.006
Socio-Economic Deprivation (SED)	0.007	0.039***	-0.011	0.039***	-0.056**
EA:SED	0.028***		0.061***		-0.093***
EA:ENER		0.207	-0.345		
EA:SED:ENER			-0.093		
EA:West(East=0)				-0.027	-0.024
SED:West(East=0)					0.093***
EA:SED:West(East=0)					-0.093***
AIC	-2486.8	-2479.7	-2505.5	-2479.9	-2508.8

Table 6.5: Additional Models (SAR models, the effect reported is the total effect). Dependent variable: share of votes for the AfD in EU Elections 2019. N=489. Sources: EU election results for Berlin 2019, reception facility address list, D4I data, socioeconomic data; own calculations; Coefficients, Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1.

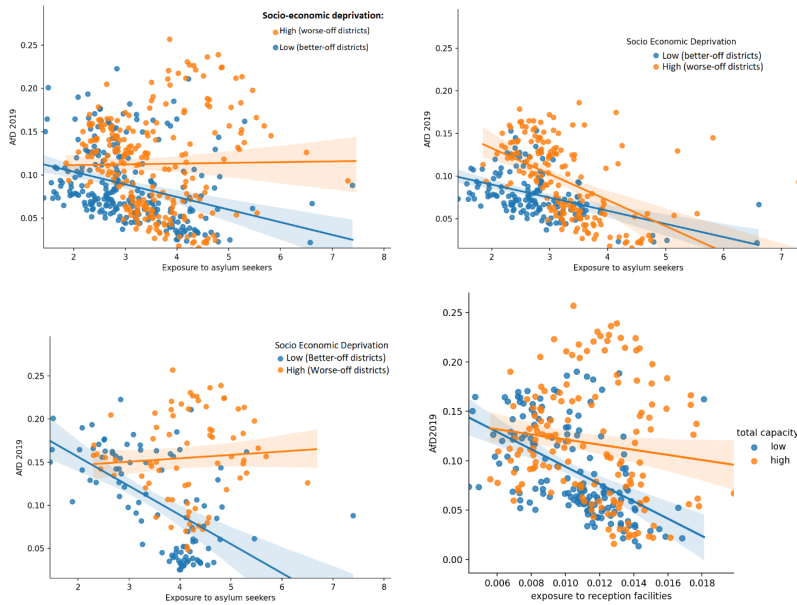


Figure 6-3: Visualisation of data points: Panel a. Visualisation of interaction between EA and SED (model 3), Panel b. Visualisation of interaction between EA and SED (West only). Panel c. Visualisation of interaction between EA and SED (East only). Panel d. Visualisation of interaction between EF and total capacity.

318 of our 489 electoral districts fulfilled this criterion. Second, we have computed our main model (with share of votes for the AfD as main dependent variable and EF as main independent variable⁸) in this subset of districts adding an additional variable, which we call total capacity, which is the number of asylum-seekers hosted within the above mentioned reception centre(s), located within or very close to the borders of the district. If more than one reception centre was linked to one district, the sum of their capacities has been selected. We assume here that two reception centres that are located very close to each other tend to be perceived by locals as two branches of the same reception centre. The model is presented in Table 6.6 and provides some interesting findings. On the one hand, in this model, as in all our previous models, exposure to reception facilities is negatively correlated with vote shares for the AfD, meaning that also in this subset of districts, the higher the exposure to reception facilities (or the exposure to asylum-seekers, as our additional model computed with the EA variable shows), the lower the votes for the AfD. On the other hand, the same model shows that the new variable we introduced (total capacity) is positively correlated with the share of votes for the AfD, meaning that the bigger the reception centre located within the district, the higher the number of votes for the AfD. Crucially, the model also shows that the interaction between the Exposure to reception facilities (EF) and total capacity has a positive coefficient and is statistically significant. This means that the effect of EF on vote shares for the AfD depends on the size of reception centres. As shown in panel d of Figure 3, which illustrates the relationship between our variables, our findings provide support for our fifth hypothesis, suggesting that the bigger the reception centres located within a district, the lower the negative effect of exposure on votes shares for the AfD.

On the one hand, the model suggests that the new variable we introduced (total capacity) is positively correlated with the share of votes for the AfD, meaning that the higher the number of asylum-seekers hosted within or nearby a district (mostly: within one single centre), the higher the number of votes for the AfD.

⁸We chose to run this model with the EF variable rather than the EA variable, since the EA variable contained itself some information on the number of asylum-seekers hosted within reception centres which risked distorting our findings due to problems of multicollinearity. OLS models have been also computed and they provide very similar findings.

	Direct	Indirect	Total
Exposure to reception facilities (EF)	-0.03862***	-0.0841***	-0.12267***
Share of established non-European residents (ENER)	-0.19909***	-0.43325***	-0.63234***
West (East=0)	-0.01368***	-0.02977***	-0.04344***
Socio-economic Deprivation (SED)	0.00798***	0.0174***	0.02540***
Total capacity	0.00013***	0.00029***	0.00043***
EF:Total capacity	0.00003***	0.00006***	0.00009***
AIC			-1532.5
AIC for lm			-1228.8

Table 6.6: SAR Model 8. Dependent variable: share of votes for the AfD in EU Elections 2019. N=318. (electoral districts with at least one reception centre located within their borders or at max 1,000 meters from their border).

More importantly, the model also shows that the interaction between the Exposure to reception facilities (EF) and total capacity has a positive coefficient and is statistically significant. These findings provide support for our fifth hypothesis, suggesting that exposure to bigger reception centres affects votes for the AfD slightly more positively compared to exposure to smaller reception centres.

This model has also been computed with an alternative variable – the number of asylum-seekers hosted in the biggest of the reception centres located within or very close to the district – providing very similar findings. Spatial models with the EA variable and OLS models have been also computed and, again, they provide very similar findings.

6.5 Conclusion

Unlike previous studies of this type, we examine the relationship between our main variables at the level of electoral districts rather than at the municipality level. Using high-resolution spatial statistics and geolocalization techniques, we create measures of the exposure to asylums-seekers and reception facilities, rather than using the mere share of asylum-seekers as done by most of the existing studies. Applying spatial regression models, we find that in electoral districts which were more exposed to centres for asylum-seekers during the ‘refugee crisis’, electoral support for the radical right AfD was lower than in electoral districts which were less exposed to reception facilities. With this finding we contribute to an ongoing debate on the electoral effects of asylum-seeking immigration to Europe, challenging the conclu-

sions of scholars who found that, especially during the so-called ‘refugee crisis’, these flows triggered votes for the radical right [46]. Rather, we provide support for those works that rather lend support to the ‘contact theory’, according to which the contact between migrants and the local population alleviates locals’ negative attitudes towards migrants themselves and their inclination to vote for anti-immigration parties [148]. Remarkably, however, we also find that the presence of bigger rather than smaller reception centres within a district lowers the negative effect of exposure on radical-right voting. This finding is consistent with scholarly works suggesting that the relationship between radical-right voting and immigration inflows depends on the type of contact between natives and migrants and, specifically, by how much contact is perceived as potentially disruptive by locals [135, 49, 45].

Furthermore, our additional regression models suggest that a number of specific contextual variables influence the relationship between our two main variables: in particular, the negative correlation mentioned above is distinctive of Western districts and of better-off Eastern districts, while in Eastern worse-off districts exposure to reception facilities seems to be correlated with slightly higher votes for the AfD. Conversely, different to our expectations, no difference is found between districts with a higher or lower share of established non-European residents, meaning that the relationship between exposure to reception facilities and vote shares for the AfD seems not to be influenced by the share of established European residents of electoral districts. These findings provide support for the strand of the literature that argues that the relationship between radical-right voting and asylum-seeking flows in Berlin is influenced by districts’ socio-economic status. In addition to that, however, our study identifies a so-far underexplored variable which *ceteris paribus* does influence the relationship between radical right voting and asylum-seeking flows: the East-West divide, which is a proxy for different socio-cultural histories. Most of the existing studies in this field conclude by acknowledging the lack of external validity. [148], as many others, states that his findings ‘do not necessarily contradict previous findings since the differences might be driven by the specific context under study’. While our study does not solve the issue of external validity, our findings point to the importance of including into the analysis specific contextual historical and cultural

characteristics, like the democratic tradition of the context analysed, when examining the relationship between immigration and voting patterns. This finding is in line with the findings of research on the drivers of individual attitudes to immigration, according to which the political context in which individuals are socialised is a key determinant of their attitudes to immigration later in life [89, 69]. Finally, the article makes a methodological contribution to the existing literature on the electoral consequences of immigration, revealing the potential of innovative methodological approaches based on high-resolution spatial statistics and geo-localization techniques on the study of these phenomena. We invite scholars in this field to further explore our research questions relying on these methods, in other geographical settings and socio-cultural contexts. Future research might also use experimental approaches to draw more robust conclusions on issues of causality and reverse causality adopting a similar research design. Our cross-sectional design is based on the assumption that reception centres were distributed across the city on a quasi-random basis, depending on the availability of buildings across the city. For this reason, we cannot claim to have provided definite evidence about the causal direction between our main variables, although the comparison of covariate means between treated and untreated units that we conducted, however, led us to exclude that refugee centres were de facto located predominantly where policymakers expected less resistance from the local population (as also in districts characterised by higher numbers of established migrants or significantly higher levels of socio-economic deprivation), which might have been another alternative explanation for the negative relation between AfD vote share and exposure to asylum-seekers that we show. Future research might further explore this issue.

“Public policy on migration needs to come to terms with this complexity.”

Paul Collier

Chapter 7

Discussions and conclusion

7.1 Summary and conclusion

In this chapter, we summarise and conclude findings of this thesis. In chapter 1, and 2, we provided motivations and related existing researches necessary to justify objectives of this thesis. Breaking down migration phenomenon into different trajectories, we then covered four related research questions.

In chapter 3, we built migration statistics using Twitter data to provide more up-to-date and rich data to better monitor migration stocks. To be more specific, we identified migrants on Twitter using both the linguistics and social networks that link migrants back to their home country from host country. We have validated our results both with internal and external data. The results show good classification performance scores and good correlation coefficients with official statistics. Different from other works, we set a definition of a migrant that is closed to the official definition. Our methodology also enables us to identify both immigrants and emigrants which allows us to further explore data in other related issues in migration.

In chapter 4, we analysed topics of interest for migrants on Twitter to provide a novel method to measure cultural integration of migrants. The cultural integration of migrants is computed using two dimensions: preservation of links to the home country and cultural traits, which we call home attachment (HA), and formation of new links and adopting cultural traits from the country of migration, that we define as destination attachment (DA). This work was able to compute these indices

for individuals and compare adjustability of migrants in the host country but also connections back to migrants' home country. Different from other works, a cross-country study of integration was possible thanks to the employment of social big data, not to mention that it overcame limitations of survey data.

Chapter 5 examined social networks of migrant and native communities which enabled us to discover distinctive characteristics of migrants and natives on Twitter. We observed that migrants have more followers than friends. They also tweet more often and in more various locations and languages. This is also shown through the hashtags where the most popular hashtags used among migrants reflect their interests in travels. Most interestingly, we detected that Twitter users tend to be connected to other users that share the same nationality more than the country of residence. This tendency was relatively stronger for migrants than for natives. Furthermore, both natives and migrants tend to connect mostly with natives.

Chapter 6 studied the relation between locals' exposure to reception facilities and asylum-seekers during the so-called 'refugee crisis' and vote shares obtained at the district level by the radical right AfD in the 2019 EU elections. Unlike previous studies of this type, we examined this relationship within an urban context, and at the level of electoral districts, relying on high-resolution spatial statistics and geo-localisation techniques. Crucially, we exploited the unique history of the city of Berlin to analyse the impact of different socio-political and historical contexts. As results, we found negative relationship between the exposure variable and electoral support for the radical party. Our additional regression models suggested, however, that a number of specific contextual variables influence the relationship between our two main variables: in particular, the negative correlation mentioned above is distinctive of Western districts and of better-off Eastern districts, while in Eastern worse-off districts exposure to reception facilities seems to be correlated with slightly higher votes for the AfD. Conversely, the relationship between exposure to reception facilities and vote shares for the AfD seems not to be influenced by the share of established non-European residents of electoral districts. Remarkably, our findings suggested that the size of the reception centres matters also.

7.2 Ethics and legality issues in using Twitter data

Although the privacy issues were not treated in depth in the chapters, it is one of the crucial limitations of using social media data. Twitter data is public, but whether users are aware of researchers listening to their conversations and activities is unknown. It would be similar to having conversations on a bus where no one expects neighbours to eavesdrop on their conversations. To manage privacy issues, it is necessary to take security measures such as pseudonymization, or anonymization to prevent re-identification of individuals and to protect personal data. Furthermore it is also essential for the researchers to secure the collected data to ensure that both the raw and processed data are not in the hands of wrong people. This would involve storing data in a secured server or limiting shared access to the data.

For this thesis, to comply with ethics and privacy regulations, the various analysis stages follow strict procedures:

- Data collection: The data is collected in accordance with the terms and conditions of the Twitter API¹.
- Data storage: The data is stored and processed for the sole purpose of the research. The data is retained on servers of CNR-ISTI. The Data has not been transferred or lost in the process.
- Data processing: The data processing is carried out by electronic, automated and manual instruments, with methods and tools to ensure maximum security and confidentiality, by authorised personnel in compliance with the regulations in force and following the operating instructions provided for by the regulations of the structure.
- Data protection:
 - The data collected from Twitter is not published or reported to public or available to any third party.

¹<https://twitter.com/en/tos>

-
- No personal information is published in any of the research outputs.
 - Mock examples are provided to explain the methodology instead of using real examples from the data.
 - All the results reported in the research outputs are aggregated to country or world level.
 - All the researchers involved in the produced works have signed a non-disclosure agreement.

The process of identifying migrants requires processing of information published by Twitter users, in particular their location and profile information. In compliance of Twitter’s privacy policy, solely the public information on Twitter is collected and stored. The European General Data Protection Regulation includes very strict rules for user profiling used to make automated decisions that target individuals and may have legal or other types of personal consequences. However, our analysis does not have any component that involves decision making at the individual level. It is only aimed at studying the process at the population level, just as much and in the same way integration indices are extracted from census data. Thus, researchers involved in our analysis are not permitted to disclose any personal information, and all results are presented in a fully aggregated manner, in both space and time, in a manner similar to publication of census data. The figures shown throughout the thesis are examples of results that can be presented in research outputs. As shown, no individual personal information is disclosed, and anonymity is preserved throughout. We also respected the Twitter’s privacy policy according to information provided on <https://developer.twitter.com/en/developer-terms/policy> and <https://help.twitter.com/en/rules-and-policies#twitter-rules>.

These strict procedures were approved by the SoBigData Board for Operational Ethics and Legality.

7.3 Policy recommendations

The findings of this PhD thesis provide important policy recommendations:

-
- Improvement in migration statistics to be able to provide timely and reliable data.
 - It is recommended for countries to ameliorate migration-relevant questions in their surveys and census to gain more insights into motivations and links to arrival country. Improvement of migration-relevant questions can allow countries to uniformalise definitions of key migration terms in line with international recommendations.
 - Further exploration of big data should be encouraged to complement information provided by traditional data, and not to replace them fully.
 - Countries should develop measures related to progress of sustainable development as migration contributes many of its factors such as poverty, well-being, economic growth, sustainable communities, peace, justice and strong institutions. This can be done with a help of big data which can explore areas not monitored by traditional data.
 - Promotion of data-driven approaches to support evidence-based policymaking.

7.4 Directions for the future research

Limitations of each work in this thesis have been addressed in each chapter. Here, we try to summarise limitations presented and provide directions for the future research. As the findings suggest, they open us to further interesting research questions, which we consider important to be discussed here.

In this thesis, we have studied how social media data can be used in migration studies. Yet, little attention was given to how biased our data is. We validated our measures with official statistics and golden standard datasets that we built from information provided by the users themselves on Twitter. While the validations show good correlation levels, the representativity of the population remains unclear, as we lack information on socio-economic demographics to say more about the identified users. Furthermore, in this thesis, we were able to identify migrants from many

different countries. However the usage of Twitter varies greatly from one country to another. In 2019, Twitter had 152 million users² worldwide which is less than 2% of the world population³. Not to mention that the internet penetration rate also differ significantly. This being the case, it is only natural to further investigate in this line to address the selection bias.

Another interesting research question that I intend investigate in the future is knowledge diffusion channels through conversation threads of users on Twitter. In particular, conversation networks, in other words, tweets and retweets on a specific topic would be interesting to follow up. Different from the friendship network we observed in this present work, the conversation network would capture how a topic flows from one to another. Having observed how topics are associated with some country more than the others, it would be natural to follow up how these topics move from one community to another and roles played by migrants and natives in diffusing a specific topic. This would require new collection of data and a different strategy. Furthermore, if the right data can be found, this could potentially lead to another interesting question on the relationship between knowledge diffusion and economic development such as trade, and foreign direct investments (FDI). This would allow us to concretely observe the migration network effect that existing works have observed only through recorded data.

Last but not least, I intend to carry out an online survey on Twitter which would also be a great asset as a ground truth data of Twitter on nationality of users. The online survey on Twitter was attempted before. Nevertheless, the results were not reported in the work as not enough responses were collected. Obtaining a ground truth data would enhance greatly the quality of researches produced using big data, enabling us to generalise some of the findings. One of the promising method would be to employ Amazon Mechanical Turk where crowdsourcing is possible.

²www.statista.com

³United Nations Department of Economic and Social Affairs (UNDESA), *World Population Prospects 2019*

7.5 General conclusion

The need for better data on migration has been pointed out. A way to do so is to design migration statistics from innovative data. As this thesis has shown, however, innovative data alone cannot completely replace traditional data but complement them.

Twitter data are freely available through public APIs but Twitter requires considerably many API calls as individual, not aggregate, data needs to be collected. Collecting large amounts of Twitter data is also becoming more and more challenging due to increased scrutiny, resulting in regular changes in rate limits. However, once obtained, Twitter data is relatively straightforward to interpret than for instance, Facebook which is similar to a “black box”. With Twitter data longitudinal studies are also possible as it supports the collection of historical data. Despite its drawbacks, there are unexplored potentials in employing big data. One of the idealist directions of this area of research would be to combine the two data sources as some of the researchers have already begun to investigate. For instance, the dataset built by the Joint Research Centre (JRC) of the European Commission, which was employed in chapter 6, combined the census data and satellite data to map stocks of migrants, providing us with high-resolution spatial data (100m by 100m). Notwithstanding the drawbacks, this thesis along with existing studies, shows that big data in migration studies is a promising area of research which is worth the attention that is receiving.

Having initialised my first steps of research in migration studies as mentioned, I look forward to continue to pursue my research interests and to contribute to this line of research.

Bibliography

- [1] Castillo petruzzi case, 1999.
- [2] Richard Alba, John Logan, Amy Lutz, and Brian Stults. 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.
- [3] Lina Aldén, Mats Hammarstedt, and Emma Neuman. Ethnic segregation, tipping behavior, and native residential mobility. *International Migration Review*, 49(1):36–69, 2015.
- [4] A Alessandrini, F Natale, F Sermi, and M Vespe. High resolution map of migrants in the eu. *JRC Technical Report EUR*, 28770, 2017.
- [5] GW Allport. *The Nature of Prejudice*. Beacon Press, 1954.
- [6] Onur Altındağ and Neeraj Kaushal. Do refugees impact voting behavior in the host country? evidence from syrian refugee inflows to turkey. *Public Choice*, pages 1–30, 2020.
- [7] Kai Arzheimer and Carl C Berning. How the alternative for germany (afd) and their voters veered to the radical right, 2013–2017. *Electoral Studies*, 60:102040, 2019.
- [8] Munzoul AM Assal. Nationality and citizenship questions in sudan after the southern sudan referendum vote. *Sudan Report*, 2011.
- [9] Marco Avvenuti, Salvatore Bellomo, Stefano Cresci, Mariantonietta Noemi La Polla, and Maurizio Tesconi. 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*, pages 1413–1421. International World Wide Web Conferences Steering Committee, 2017.
- [10] Nahlah Ayed. For far-right populists, eu election a battle to 'save europe', May 2019.
- [11] Dany Bahar and Hillel Rapoport. Migration, knowledge diffusion and the comparative advantage of nations. *The Economic Journal*, 128(612):F273–F305, 2018.

-
- [12] Kirk Bansak, Jens Hainmueller, and Dominik Hangartner. How economic, humanitarian, and religious concerns shape european attitudes toward asylum seekers. *Science*, 354(6309):217–222, 2016.
- [13] Guglielmo Barone, Alessio D’Ignazio, Guido de Blasio, and Paolo Naticchioni. Mr. rossi, mr. hu and politics. the role of immigration in shaping natives’ voting behavior. *Journal of Public Economics*, 136:1–13, 2016.
- [14] Sascha O Becker, Thiemo Fetzer, et al. Does migration cause extreme voting? *Center for Competitive Advantage in the Global Economy and The Economic & Social Research Council*, pages 1–54, 2016.
- [15] Linus Bengtsson, Xin Lu, Anna Thorson, Richard Garfield, and Johan Von Schreeb. 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.
- [16] Katrin Bennhold. German elections reveal, and deepen, a new east-west divide, Aug 2019.
- [17] Sozialpädagogisches Institut Berlin, 2017.
- [18] Statistics Berlin-Brandenburg. Statistics for berlin and brandenburg, 2015.
- [19] Statistics Berlin-Brandenburg. Statistics for berlin and brandenburg, 2019.
- [20] Berlin.de. Senatsverwaltung für stadtentwicklung und wohnen / land berlin.
- [21] John W Berry. Immigration, acculturation, and adaptation. *Applied psychology*, 46(1):5–34, 1997.
- [22] John W Berry. A psychology of immigration. *Journal of social issues*, 57(3):615–631, 2001.
- [23] John W Berry. *Acculturation: a personal journey across cultures*. Cambridge University Press, 2019.
- [24] John W Berry, Jean S Phinney, David L Sam, and Paul Vedder. Immigrant youth: Acculturation, identity, and adaptation. *Applied psychology*, 55(3):303–332, 2006.
- [25] Hubert M Blalock. Per cent non-white and discrimination in the south. *American Sociological Review*, 22(6):677–682, 1957.
- [26] Joshua E Blumenstock. Inferring patterns of internal migration from mobile phone call records: evidence from rwanda. *Information Technology for Development*, 18(2):107–125, 2012.
- [27] Marcus H. Böhme, André Gröger, and Tobias Stöhr. Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, page 102347, 2019.

-
- [28] Diane Bolet. Local labour market competition and radical right voting: Evidence from france. *European Journal of Political Research*, 2020.
- [29] Massimiliano Bratti, Luca De Benedictis, and Gianluca Santoni. On the pro-trade effects of immigrants. *Review of World Economics*, 150(3):557–594, 2014.
- [30] Statistisches (Destatis) Bundesamt. Persons seeking protection (länder, reference date, sex/age years/marital status). table 12521-0110. database of the federal statistical office of germany.
- [31] Bundeswahlleiter, 2019.
- [32] Donald T Campbell. Ethnocentric and other altruistic motives. In *Nebraska symposium on motivation*, volume 13, pages 283–311, 1965.
- [33] David Card. Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. *Journal of Labor Economics*, 19(1):22–64, 2001.
- [34] David Card. Is the new immigration really so bad? *The economic journal*, 115(507):F300–F323, 2005.
- [35] David Card, Alexandre Mas, and Jesse Rothstein. Tipping and the dynamics of segregation. *The Quarterly Journal of Economics*, 123(1):177–218, 2008.
- [36] Meeyoung Cha, Fabrício Benevenuto, Hamed Haddadi, and Krishna Gummadi. The world of connections and information flow in twitter. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 42(4):991–998, 2012.
- [37] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, P Krishna Gummadi, et al. Measuring user influence in twitter: The million follower fallacy. *Icwsm*, 10(10-17):30, 2010.
- [38] Effrosyni Charitopoulou and Javier García-Manglano. Fear of small numbers? immigrant population size and electoral support for the populist radical right in switzerland. *Journal of Ethnic and Migration Studies*, 44(5):849–869, 2018.
- [39] Xavier Chojnicki. The fiscal impact of immigration in f rance: A generational accounting approach. *The World Economy*, 36(8):1065–1090, 2013.
- [40] Marcel Coenders, Marcel Lubbers, Peer Scheepers, and Maykel Verkuyten. More than two decades of changing ethnic attitudes in the netherlands. *Journal of Social Issues*, 64(2):269–285, 2008.
- [41] Mauro Coletto, Andrea Esuli, Claudio Lucchese, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, and Chiara Renso. Perception of social phenomena through the multidimensional analysis of online social networks. *Online Social Networks and Media*, 1:14–32, 2017.

-
- [42] Amelie F Constant and Klaus F Zimmermann. Measuring ethnic identity and its impact on economic behavior. *Journal of the European Economic Association*, 6(2-3):424–433, 2008.
- [43] Council of Europe Directorate of Social and Economic Affairs. *Measurement and indicators of integration*. Council of Europe, 1997.
- [44] Joop De Beer, James Raymer, Rob Van der Erf, and Leo Van Wissen. 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.
- [45] Robert DeFina and Lance Hannon. Diversity, racial threat and metropolitan housing segregation. *Social forces*, 88(1):373–394, 2009.
- [46] Elias Dinas, Konstantinos Matakos, Dimitrios Xeferis, and Dominik Hangartner. Waking up the golden dawn: does exposure to the refugee crisis increase support for extreme-right parties? *Political analysis*, 27(2):244–254, 2019.
- [47] Ruth Donner. *The regulation of nationality in international law*. Brill Nijhoff, 1994.
- [48] Antoine Dubois, Emilio Zagheni, Kiran Garimella, and Ingmar Weber. Studying migrant assimilation through facebook interests. In *International Conference on Social Informatics*, pages 51–60. Springer, 2018.
- [49] Christian Dustmann, Kristine Vasiljeva, and Anna Piil Damm. Refugee migration and electoral outcomes. *The Review of Economic Studies*, 86(5):2035–2091, 2019.
- [50] Anthony Edo, Yvonne Giesing, Jonathan Öztunc, and Panu Poutvaara. Immigration and electoral support for the far-left and the far-right. *European Economic Review*, 115:99–143, 2019.
- [51] Christopher G Ellison and Daniel A Powers. The contact hypothesis and racial attitudes among black americans. *Social Science Quarterly*, 1994.
- [52] Hartmut Esser. *Migration, language and integration*. Citeseer, 2006.
- [53] EU Knowledge Centre on Migration and Demography. KCMD Data Catalogue, Accessed July 2019.
- [54] EUROSTAT. Migration and migrant population statistics, March 2018.
- [55] EUROSTAT. Asylum and managed migration data, Accessed July 2019.
- [56] Jo Ann M Farver, Bakhtawar R Bhadha, and Sonia K Narang. Acculturation and psychological functioning in asian indian adolescents. *Social Development*, 11(1):11–29, 2002.

-
- [57] Rachel M Friedberg and Jennifer Hunt. The impact of immigrants on host country wages, employment and growth. *Journal of Economic perspectives*, 9(2):23–44, 1995.
- [58] FRONTEX. Illegal border crossing, Accessed July 2019.
- [59] Matteo Gamalerio, Mario Luca, Alessio Romarri, and Max Viskanic. Is this the real life or just fantasy? refugee reception, extreme-right voting, and broadband internet. *Refugee reception, extreme-right voting, and broadband internet (August 14, 2020)*, 2020.
- [60] Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, 35(2):137–144, 2015.
- [61] Federica Genovese, Margherita Belgioioso, and F Kern. The political geography of migrant reception and public opinion on immigration: Evidence from italy. *Unpublished manuscript*, 2016.
- [62] Theresa Gessler, Gergő Tóth, and Johannes Wachs. No country for asylum seekers? how short-term exposure to refugees influences attitudes and voting behavior in hungary. 2019.
- [63] Fosca Giannotti, Jisu Kim, Giulio Rossetti, Laura Pollacci, and Alina Sirbu. Twitter data for migration studies. In *Data Science for Migration and Mobility*. Oxford: In prepration, 2021.
- [64] Claudia Goldin. The human capital century and american leadership: Virtues of the past. Technical report, National Bureau of Economic Research, 2001.
- [65] Milton Myron Gordon. *Assimilation in American life: The role of race, religion, and national origins*. Oxford University Press on Demand, 1964.
- [66] David Gould, Dror Y Kenett, and Georgi Panterov. *Multidimensional connectivity: Benefits, risks, and policy implications for Europe and Central Asia*. The World Bank, 2018.
- [67] David M Gould. Immigrant links to the home country: empirical implications for us bilateral trade flows. *The Review of Economics and Statistics*, pages 302–316, 1994.
- [68] Martin Grandjean. A social network analysis of twitter: Mapping the digital humanities community. *Cogent Arts & Humanities*, 3(1):1171458, 2016.
- [69] Maria Teresa Grasso, Stephen Farrall, Emily Gray, Colin Hay, and Will Jennings. Thatcher’s children, blair’s babies, political socialization and trickle-down value change: An age, period and cohort analysis. *British Journal of Political Science*, 49(1):17–36, 2019.
- [70] R Guidotti et al. Measuring immigrants adoption of natives shopping consumption with machine learning. *ECML PKDD: In press*, 2020.

-
- [71] Birte Gundelach and Markus Freitag. Neighbourhood diversity and social trust: An empirical analysis of interethnic contact and group-specific effects. *Urban Studies*, 51(6):1236–1256, 2014.
- [72] Kay Hailbronner. *Nationality in public international law and European law*. JSTOR, 2006.
- [73] Jens Hainmueller and Daniel J Hopkins. Public attitudes toward immigration. *Annual Review of Political Science*, 17, 2014.
- [74] Martin Halla, Alexander F Wagner, and Josef Zweimüller. Immigration and voting for the far right. *Journal of the European Economic Association*, 15(6):1341–1385, 2017.
- [75] Ricardo Hausmann, Julian Hinz, and Muhammed A Yildirim. Measuring venezuelan emigration with twitter. Technical report, Kiel Working Paper, 2018.
- [76] Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271, 2014.
- [77] Keith Head and John Ries. Immigration and trade creation: econometric evidence from canada. *Canadian journal of economics*, pages 47–62, 1998.
- [78] Amaç Herdağdelen, Bogdan State, Lada Adamic, and Winter Mason. The social ties of immigrant communities in the united states. In *Proceedings of the 8th ACM Conference on Web Science*, pages 78–84, 2016.
- [79] Berlin Hilft. Unterkünfte berlin. netzwerk berlin hilft. informieren - vernetzen - helfen.
- [80] Geert Hofstede. *Culture’s consequences: International differences in work-related values*, volume 5. sage, 1984.
- [81] Geert Hofstede. Dimensionalizing cultures: The hofstede model in context. *Online readings in psychology and culture*, 2(1):8, 2011.
- [82] Gabriel Horenczyk. Immigrants’ perceptions of host attitudes and their reconstruction of cultural groups. *Applied Psychology*, 46(1):34–38, 1997.
- [83] Rezart Hoxhaj and Carolina V Zuccotti. The complex relationship between immigrants’ concentration, socioeconomic environment and attitudes towards immigrants in europe. *Ethnic and Racial Studies*, 44(2):272–292, 2021.
- [84] Wenyi Huang, Ingmar Weber, and Sarah Vieweg. Inferring nationalities of twitter users and studying inter-national linking. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 237–242. ACM, 2014.

-
- [85] Thomas Huddleston, Jan Niessen, and Jasper Dag Tjaden. Using eu indicators of immigrant integration. *Final Report for Directorate-General for Home Affairs. Brussels: European Commission*, 2013.
- [86] Instituto Nacional de Estadística. Ine microdata, Accessed July 2019.
- [87] IPUMS. IPUMS census and survey data, Accessed July 2019.
- [88] Istituto Nazionale di Statistica. Immigrati.stat: Dati e indicatori su immigranti e nuovi cittadini, Accessed July 2019.
- [89] Anne-Marie Jeannet and Lenka Dražanová. Cast in the same mould: How politics during the impressionable years shapes attitudes towards immigration in later life. *Robert Schuman Centre for Advanced Studies Research Paper No. RSCAS*, 79, 2019.
- [90] Sabrina Juran and P Niclas Broer. A profile of germany’s refugee populations. *Population and Development Review*, pages 149–157, 2017.
- [91] Anneli Kaasa, Maaja Vadi, and Urmas Varblane. Regional cultural differences within european countries: evidence from multi-country surveys. *Management International Review*, 54(6):825–852, 2014.
- [92] Anneli Kaasa, Maaja Vadi, and Urmas Varblane. A new dataset of cultural distances for european countries and regions. *Research in International Business and Finance*, 37:231–241, 2016.
- [93] Paul D Kenny and Charles Miller. Does asylum seeker immigration increase support for the far right? evidence from the united kingdom, 2000–2015. *Journal of Ethnic and Migration Studies*, pages 1–18, 2020.
- [94] Riivo Kikas, Marlon Dumas, and Ando Saabas. 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*, pages 17–22. ACM, 2015.
- [95] Jisu Kim. Korean immigrants and their network effect on trade-korean war brides. Master’s thesis, Université Paris 1 Panthéon-Sorbonne, June 2017.
- [96] Jisu Kim, Alina Sirbu, Fosca Giannotti, and Lorenzo Gabrielli. Digital footprints of international migration on twitter. In *International Symposium on Intelligent Data Analysis*, pages 274–286. Springer, 2020.
- [97] Jisu Kim, Alina Sirbu, Fosca Giannotti, and Giulio Rossetti. Characterising different community of twitter users: Migrants and natives. *To be submitted*, 2021.
- [98] Jisu Kim, Alina Sirbu, Giulio Rossetti, Fosca Giannotti, and Hillel Rapoport. Home and destination attachment: study of cultural integration on twitter. *arXiv preprint arXiv:2102.11398*, 2021.

-
- [99] Jisu Kim, Emilio Zagheni, and Ingmar Weber. Improving migration statistics using social media. In *Practitioners' Guide on Harnessing Data Innovation for Migration Policy*. IOM, 2021.
- [100] Katarzyna Kraszewska, Bettina Knauth, and D Thorogood. Indicators of immigrant integration—a pilot study. *Luxembourg, Luxembourg: Eurostat, European Commission, Publications Office of the European Union*, 2011.
- [101] Maurice Kugler, Hillel Rapoport, et al. Migration, fdi and the margins of trade. Technical report, Center for International Development at Harvard University, 2011.
- [102] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. AcM, 2010.
- [103] Teresa LaFromboise, Hardin LK Coleman, and Jennifer Gerton. Psychological impact of biculturalism: Evidence and theory. *Psychological bulletin*, 114(3):395, 1993.
- [104] Fabio Lamanna, Maxime Lenormand, María Henar Salas-Olmedo, Gustavo Romanillos, Bruno Gonçalves, and José J Ramasco. Immigrant community integration in world cities. *PloS one*, 13(3):e0191612, 2018.
- [105] Doug Laney. 3d data management: Controlling data volume, velocity and variety, 2001.
- [106] Jennifer Lee and Frank D Bean. *The diversity paradox: Immigration and the color line in twenty-first century America*. Russell Sage Foundation, 2010.
- [107] Alexia Lochmann, Hillel Rapoport, and Biagio Speciale. The effect of language training on immigrants' economic integration: Empirical evidence from france. *European Economic Review*, 113:265–296, 2019.
- [108] Jakub Lonsky. Does immigration decrease far-right popularity? evidence from finnish municipalities. *Journal of Population Economics*, 34(1):97–139, 2021.
- [109] Matthias Mader and Harald Schoen. The european refugee crisis, party competition, and voters' responses in germany. *West European Politics*, 42(1):67–90, 2019.
- [110] Thierry Mayer and Soledad Zignago. Notes on cepii's distances measures: The geodist database. *CEPII*, 2011.
- [111] Mattia Mazzoli, Boris Diechtiareff, Antònia Tugores, Willian Wives, Natalia Adler, Pere Colet, and José J Ramasco. Migrant mobility flows characterized with digital data. *arXiv preprint arXiv:1908.02540*, 2019.
- [112] Mattia Mazzoli, Boris Diechtiareff, Antònia Tugores, Willian Wives, Natalia Adler, Pere Colet, and José J Ramasco. Migrant mobility flows characterized with digital data. *Plos one*, 15(3):e0230264, 2020.

-
- [113] Jacques Melitz and Farid Toubal. Native language, spoken language, translation and trade. *Journal of International Economics*, 93(2):351–363, 2014.
- [114] Observatory Migration. Opam - observatory of public attitudes to migration.
- [115] Sahar Mirzaee and Qi Wang. Urban mobility and resilience: exploring boston’s urban mobility network through twitter data. *Applied Network Science*, 5(1):1–20, 2020.
- [116] Izabela Moise, Edward Gaere, Ruben Merz, Stefan Koch, and Evangelos Pournaras. Tracking language mobility in the twitter landscape. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 663–670. IEEE, 2016.
- [117] Fred Morstatter and Huan Liu. Discovering, assessing, and mitigating data bias in social media. *Online Social Networks and Media*, 1:1–13, 2017.
- [118] United Nations. Recommendations on statistics of international migration. volume Statistical Papers Series M, No. 58, Rev.1. Department of Economic and Social Affairs, Statistics Division, United Nations, New York., 1998.
- [119] Pippa Norris and Ronald Inglehart. *Cultural backlash: Trump, Brexit, and authoritarian populism*. Cambridge University Press, 2019.
- [120] Lukáš Novotný and Pavel Maškarinec. More to the people, less to brussels! alternative for germany and the 2014 european parliament elections. *Politologická revue (Czech Political Science Review)*, (2):5–23, 2018.
- [121] OECD, and European Commission. Settling in 2018. main indicators of immigrant integration, 2018.
- [122] Alkis Henri Otto and Max Friedrich Steinhardt. Immigration and election outcomes—evidence from city districts in hamburg. *Regional Science and Urban Economics*, 45:67–79, 2014.
- [123] Bezirksamt Pankow. Flüchtlingsunterkünfte in pankow (stand november 2018). *Politik und Verwaltung. Integration*.
- [124] Robert E Park. Human migration and the marginal man. *American journal of sociology*, 33(6):881–893, 1928.
- [125] Pierre-Louis Parsons, Christopher and Vézina. Migrant networks and trade: The vietnamese boat people as a natural experiment. *The Economic Journal*, 128(612):F210–F234, 2018.
- [126] Leto Peel, Jean-Charles Delvenne, and Renaud Lambiotte. Multiscale mixing patterns in networks. *Proceedings of the National Academy of Sciences*, 115(16):4057–4062, 2018.

-
- [127] MJA Penninx et al. Integration. the role of communities, institutions, and the state. *The Migration Information Source (on-line)*, 2003.
- [128] Andrea Pettrachin, Lorenzo Gabrielli, Jisu Kim, Sarah Ludwig-Dehm, Steffen Pötzschke, and Michele Vespe. Did exposure to refugee centres affect the electoral outcome of alternative for germany in berlin? evidence from the 2019 eu elections. *submitted to Journal of Ethnic and Migration Studies: Under review*, 2020.
- [129] Laura Pollacci, Alina Sîrbu, Fosca Giannotti, Dino Pedreschi, Claudio Lucchese, and Cristina Ioana Muntean. Sentiment spreading: an epidemic model for lexicon-based sentiment analysis on twitter. In *Conference of the Italian Association for Artificial Intelligence*, pages 114–127. Springer, 2017.
- [130] Alejandro Portes and Robert L Bach. *Latin journey: Cuban and Mexican immigrants in the United States*. Univ of California Press, 1985.
- [131] Michel Poulain. Confrontation des statistiques de migrations intra-européennes: Vers plus d’harmonisation? *European Journal of Population/Revue européenne de Démographie*, 9(4):353–381, 1993.
- [132] Michel Poulain, Anne Herm, and Roger Depledge. Central population registers as a source of demographic statistics in europe. *Population*, 68(2):183–212, 2013.
- [133] Rafael Prieto Curiel, Luca Pappalardo, Lorenzo Gabrielli, and Steven Richard Bishop. Gravity and scaling laws of city to city migration. *PLOS ONE*, 13(7):1–19, 07 2018.
- [134] Robert D Putnam. E pluribus unum: Diversity and community in the twenty-first century. *Scandinavian Political Studies*, 30(2):137–174, 2007.
- [135] Lincoln Quillian. Prejudice as a response to perceived group threat: Population composition and anti-immigrant and racial prejudice in europe. *American sociological review*, pages 586–611, 1995.
- [136] Tommaso Radicioni, Elena Pavan, Tiziano Squartini, and Fabio Saracco. Analysing twitter semantic networks: the case of 2018 italian elections. *arXiv preprint arXiv:2009.02960*, 2020.
- [137] James E Rauch. Business and social networks in international trade. *Journal of economic literature*, 39(4):1177–1203, 2001.
- [138] James E Rauch and Vitor Trindade. Ethnic chinese networks in international trade. *Review of Economics and Statistics*, 84(1):116–130, 2002.
- [139] James Raymer and Frans Wiilekens. Obtaining an overall picture of population movement in the european union. *International migration in Europe: Data, models and estimates*, pages 209–234, 2008.

-
- [140] Helen Mele Robinson. Dynamics of culture and curriculum design: Preparing culturally responsive teacher candidates. In *Early Childhood Development: Concepts, Methodologies, Tools, and Applications*, pages 343–365. IGI Global, 2019.
- [141] Kaija Ruotsalainen.
- [142] Mirna Safi. The immigrant integration process in france: Inequalities and segmentation. *Revue française de sociologie*, 49(5):3–44, 2008.
- [143] Thomas C Schelling. Dynamic models of segregation. *Journal of mathematical sociology*, 1(2):143–186, 1971.
- [144] Sarah Schneider-Strawczynski. Hosting refugees and voting for the far-right: Evidence from france. 2020.
- [145] Alina Sîrbu, Gennady Andrienko, Natalia Andrienko, Chiara Boldrini, Marco Conti, Fosca Giannotti, Riccardo Guidotti, Simone Bertoli, Jisu Kim, Cristina Ioana Muntean, et al. Human migration: the big data perspective. *International Journal of Data Science and Analytics*, pages 1–20, 2020.
- [146] Rune Jørgen Sørensen. After the immigration shock: The causal effect of immigration on electoral preferences. *Electoral Studies*, 44:1–14, 2016.
- [147] Biagio Speciale. Does immigration affect public education expenditures? quasi-experimental evidence. *Journal of public economics*, 96(9-10):773–783, 2012.
- [148] Andreas Steinmayr. Exposure to refugees and voting for the far-right:(unexpected) results from austria. 2016.
- [149] Andreas Steinmayr. Contact versus exposure: Refugee presence and voting for the far-right. *Review of Economics and Statistics*, pages 1–47, 2020.
- [150] Ian Stewart, René D Flores, Timothy Riffe, Ingmar Weber, and Emilio Zagheni. Rock, rap, or reggaeton?: Assessing mexican immigrants’ cultural assimilation using facebook data. In *The World Wide Web Conference*, pages 3258–3264. ACM, 2019.
- [151] Bruce Stokes. Language: The cornerstone of national identity. *Pew Research Center’s Global Attitudes Project*, 2017.
- [152] Teresa Talò. Public attitudes to immigration in germany in the aftermath of the migration crisis. 2017.
- [153] The OECD. Database on immigrants in oecd and non-oecd countries: Dioc, Accessed July 2019.
- [154] The Worldbank. Migration and remittances data, Accessed July 2019.
- [155] Paul Vertier, Max Viskanic, Matteo Gamalerio, et al. Dismantling the’jungle’: Relocation and extreme voting in france. Technical report, Sciences Po, 2020.

-
- [156] Carolina Vieira, Filipe Ribeiro, Pedro Olmo Vaz de Melo, Fabricio Benevenuto, and Emilio Zagheni. Using facebook data to measure cultural distance between countries: The case of brazilian cuisine. In *Proceedings of The Web Conference 2020*, pages 3091–3097, 2020.
- [157] Jacob L Vigdor. Measuring immigrant assimilation in the united states. civic report no. 53. *Manhattan Institute for Policy Research*, 2008.
- [158] Max Viskanic. Fear and loathing on the campaign trail: did immigration cause brexit? *Available at SSRN 2941611*, 2017.
- [159] Shaun Walker. European elections: far-right “surge” ends in a ripple. *The Guardian*, 2019.
- [160] Ying Xiong, Moonhee Cho, and Brandon Boatwright. Hashtag activism and message frames among social movement organizations: Semantic network analysis and thematic analysis of twitter during the# metoo movement. *Public relations review*, 45(1):10–23, 2019.
- [161] Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, and Bogdan State. Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 439–444. ACM, 2014.
- [162] Emilio Zagheni, Kivan Polimis, Monica Alexander, Ingmar Weber, and Francesco C Billari. Combining social media data and traditional surveys to nowcast migration stocks. In *Annual Meeting of the Population Association of America*, 2018.
- [163] Emilio Zagheni and Ingmar Weber. You are where you e-mail: using e-mail data to estimate international migration rates. In *Proceedings of the 4th annual ACM web science conference*, pages 348–351. ACM, 2012.

Appendix A

Appendix

Variable	Model 1 (Exposure to reception facilities)	Model 2 (Exposure to asylum-seekers)
Exposure to reception facilities (EF)	-0.1235*** (-0.0165)	
Exposure to asylum-seekers (EA)		-0.0980*** (-0.0174)
West(East=0)	-0.0340*** (-0.005)	-0.0278*** (-0.005)
Share of established non-European residents	-0.5370*** (-0.0512)	-0.6208*** (-0.0492)
Socio-economic Deprivation	0.0180*** (-0.0011)	0.0189*** (-0.0011)
Intercept	-0.0831** (-0.0335)	0.2166*** (-0.0097)
R-squared	0.5361 0.5399	0.5144 0.5184

Table A.1: OLS Models (using different exposure measures)

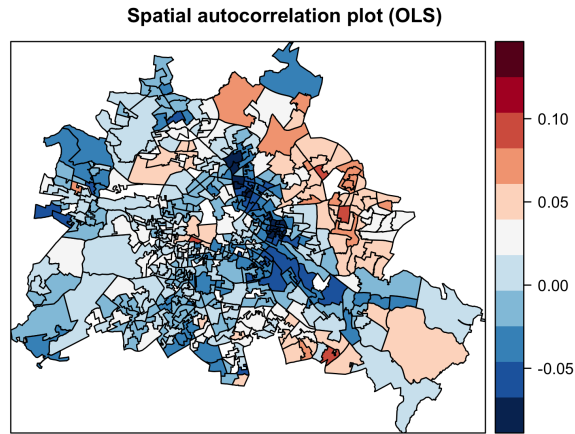


Figure A-1: Testing for spatial dependence in the OLS residuals with Exposure to reception facilities: presence of clusters of residuals across different districts in Berlin in the OLS regression model. The colour indicates whether the residual is positive or negative.

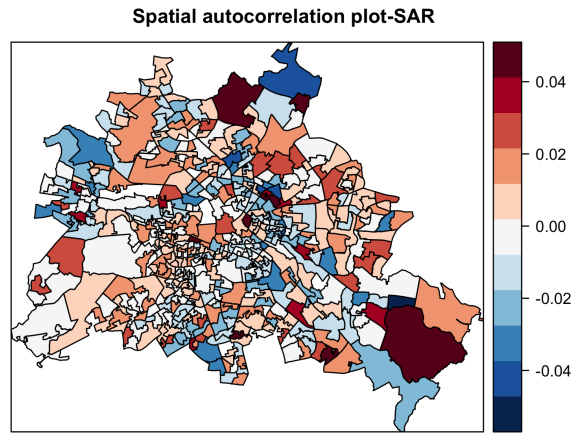


Figure A-2: Spatial dependence in the SAR model with Exposure to reception facilities: No clusters of residuals in the SAR model. The colour indicates whether the residual is positive or negative.

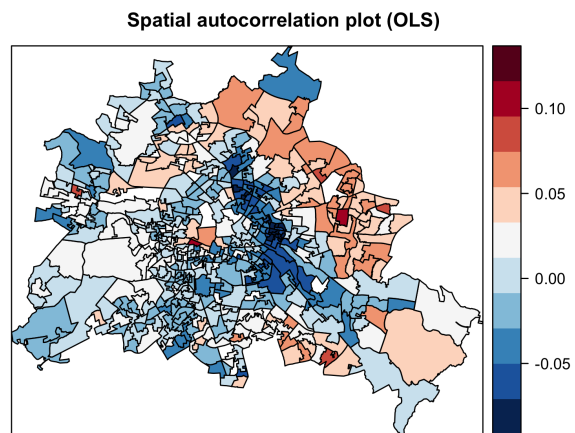


Figure A-3: Testing for spatial dependence in the OLS residuals with Exposure to asylum seekers: presence of clusters of residuals across different districts in Berlin in the OLS regression model. The colour indicates whether the residual is positive or negative.

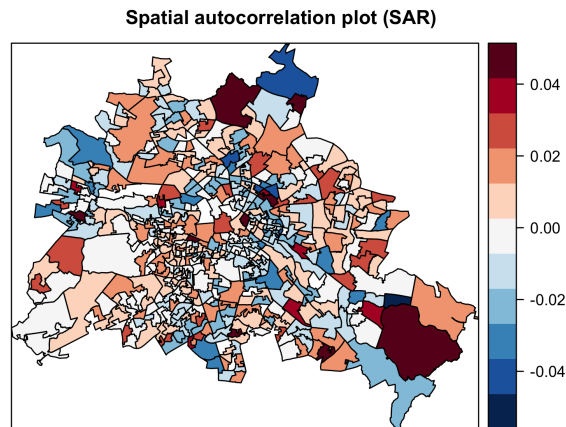


Figure A-4: Spatial dependence in the SAR model with Exposure to asylum seekers: No clusters of residuals in the SAR model. The colour indicates whether the residual is positive or negative.