Kim et al. *EPJ Data Science* (2022) 11:55
https://doi.org/10.1140/epjds/s13688-022-00363-5

EPJ Data Science
a SpringerOpen Journal

REVIEW Open Access

Origin and destination attachment: study of cultural integration on Twitter

Jisu Kim[1,2*], Alina Sîrbu[3], Fosca Giannotti[4], Giulio Rossetti[4] and Hillel Rapoport[5]

*Correspondence:
kim@demogr.mpg.de
[1]Scuola Normale Superiore, Pisa, Italy
[2]Max Planck Institute for Demographic Research, Rostock, Germany
Full list of author information is available at the end of the article

Abstract

The cultural integration of immigrants conditions their overall socio-economic integration as well as natives' attitudes towards globalisation in general and immigration in particular. At the same time, excessive integration—or assimilation—can be detrimental in that it implies forfeiting one's ties to the origin country and eventually translates into a loss of diversity (from the viewpoint of host countries) and of global connections (from the viewpoint of both host and home countries). Cultural integration can be described using two dimensions: the preservation of links to the origin country and culture, which we call *origin attachment*, and the creation of new links together with the adoption of cultural traits from the new residence country, which we call *destination attachment*. In this paper we introduce a means to quantify these two aspects based on Twitter data. We build origin and destination attachment indices and analyse their possible determinants (e.g., language proximity, distance between countries), also in relation to Hofstede's cultural dimension scores. The results stress the importance of language: a common language between origin and destination countries favours origin attachment, as does low proficiency in the host language. Common geographical borders seem to favour both origin and destination attachment. Regarding cultural dimensions, larger differences among origin and destination countries in terms of Individualism, Masculinity and Uncertainty appear to favour destination attachment and lower origin attachment.

Keywords: International migration; Cultural integration; Big data; Twitter

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 immigrants' cultural integration has become a political imperative in times of rising populism and cultural backlash against globalisation in general and immigration in particular (e.g., [1]).[1] However, too much cultural integration (or assimilation) may be detrimental in terms of immigrants' subjective well-being as well as in terms of lost diversity (from the

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

viewpoint of host countries) and of global connections (from the viewpoint of both host and home countries). Successful cultural integration brings new opportunities and, with them, an overall 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 long been 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 [2–4]. On the other hand, studies of integration have been mainly performed by sociologists, relying largely on survey data such as the World Values Survey, the Eurobarometer, or the European Social Survey. The main dimensions used in these studies are inter-marriage, religion and language [5–8]. 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" [9] (in reference to [10]). The four-fold model below reflects this complexity by dividing assimilation into four different classes: *assimilation*, *integration*, *marginalisation* and *separation*. [11–15]. 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 origin country. Marginalisation refers to a situation where migrants remain distinguishable from the both of receiving and sending society, whereas separation refers to complete rejection of host's culture. These theories typically consider two dimensions: preservation of links to the origin country and cultural traits, which we call here *origin attachment (OA)*, 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 1. In our study, however, we change the term *marginalisation* to *globalisation* to reflect that some migrants focus their attention towards international topics rather than any of their origin or destination countries' topics as it can be the case for many Twitter users discussing mainly about global issues [16]. Importantly, in the four-fold model, the two dimensions are independent so migrants can be both highly *Origin*-attached and *Destination*-attached and provide full attention to both. In our work, however, these two dimensions are in fact competing. In other words, migrants have to 'divide' their attention between the origin and destination countries, which we believe is a more realistic assumption.

In this paper we provide a novel method to compute OA and DA from Twitter data, to answer the following questions: *How much do migrants absorb the culture of their destination society? Do they lose connection with their origin country?* This is based on the topics that migrants and natives discuss on Twitter, through the analysis of hashtags. The OA index is defined as the fraction of tweets of a migrant that discuss topics related to their origin 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

Table 1 Theories of integration and their relation to OA and DA

	Low OA	High OA
Low DA	Globalisation (Marginalisation in the four-fold model)	Separation
High DA	Assimilation	Integration

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 [17]. The second stage is to determine country-specific topics by assigning nationalities to hashtags. The final stage is to compute the OA and DA indices for each migrant in our data. We examined the two indices in various settings, to demonstrate their validity. First, we analyse the relationship between the two indices and compare them to a null model obtained by shuffling the hashtags in our dataset. Second, we study 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 [18] as well as other related variables such as geographic distance and language proximity measures.

The rest of the paper is organised as follows. In the next section we describe related work on integration and assimilation of migrants both in the sociology literature and in recent big data studies. In Sect. 3, we present our methodology to compute the OA and DA indices, including data collection (Sect. 3.1), assigning nationality and residence to users (Sect. 3.2), assigning nationality to hashtags (Sect. 3.3) and calculating the indices (Sect. 3.4). In Sect. 4, we present our results, while Sect. 5 concludes the paper.

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 "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" [19]. Taking the elements stated in the definition into account, studies have looked at language, role of media, inter-marriage and religion[2] to study whether a migrant is culturally integrated in the society [5, 20, 21]. 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 [6–8, 22]. 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 [23–25]. Retail data including shopping behaviour in a large supermarket chain was used in [23] 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 [24], 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

[2] https://migrationdataportal.org/themes/migrant-integration

users. Following the work in [24, 25] 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 generations across different demographic groups. In a more recent work, [26] 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 presence 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 OA indicators through hashtags as a proxy for their interests. In the process, we also employ the Shannon entropy, but in a different way from [26]: 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 OA (origin 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 origin country to better understand the allocation of their citizens abroad.

3 The origin and destination attachment indices

We propose to study origin 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 origin attachment (OA). 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.

3.1 Data

Our data collection strategy originated from the methodology developed by [17]. The starting point is a Twitter dataset collected by the SoBigData.eu Laboratory [27]. 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 [17], 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 first and second layers (some overlap was present among the two layers).

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 [17]. 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 both tweet locations and languages of a user and the user's friends. Here, we compute the fraction of tweet locations and languages for the user of interest and the average of the fractions of tweet locations and languages for the user's friends. By summing up these fractions for each country identified, the country with the largest value would be the nationality of that user. However, as shown in the study [17], 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 4940 users. We then filtered out users who have used less than 10 hashtags in 2018, leaving us with total

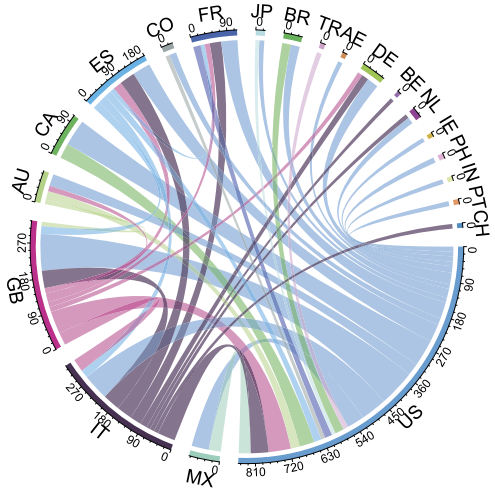

Figure 1 Chord diagram showing migration links between countries from our data. Each country is represented by an arc in the circle, with a country-specific colour. For instance, the largest arc corresponds to the US. Each chord has a starting point in the nationality country, with the same colour of the chord, and the end point in the destination country, with a different colour. The width of the chord represents the relative number of the 2018 migrants with a specific nationality and residence. For example, the widest light blue chord leaving the US reaches Great Britain, marked by the arc coloured in pink. For visualisation purposes, we show only 21 countries: those with at least 10 migrants. Note that this diagram is not representative of global patterns but only shows the structure of our data

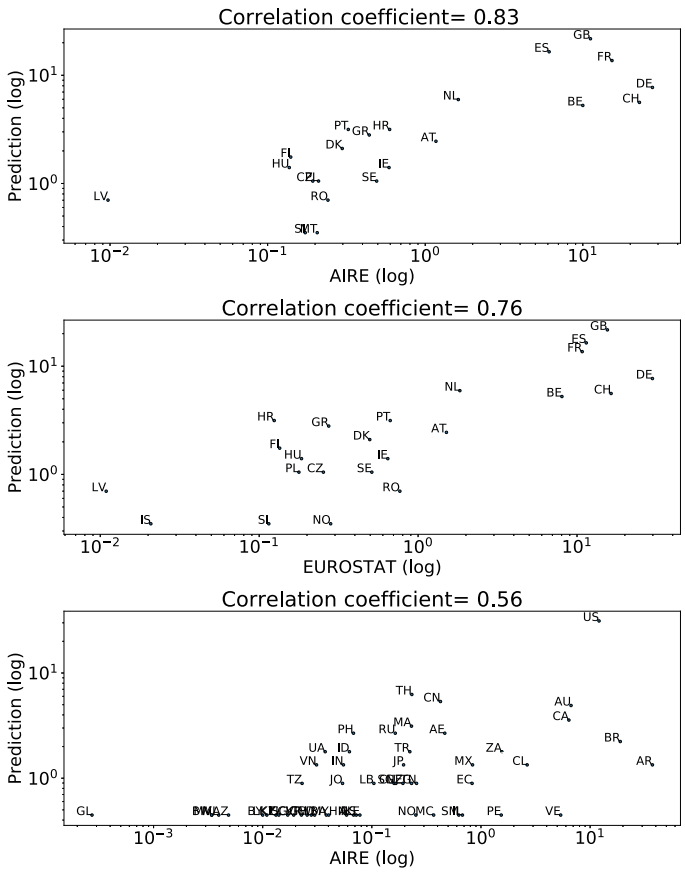

Figure 2 Percentage of Italian emigrants in various destination countries based on AIRE and Eurostat: predicted versus ground truth data. The first two figures show the results for EU countries and the last figure shows the result for non-EU countries

of 3226 migrant users. In Fig. 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 that is close to the official definition.[3] 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 Fig. 2, for instance, we show

[3]Recommendations 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".

Spearman correlation coefficients between our predicted data and ground truth data for Italian emigrants from AIRE[4] 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. Here, we computed the correlation coefficients using the ground truth data for Italian emigrants as it is the largest user population in this dataset. However, we believe that this method can also be adopted to different geographic levels where countries share similar characteristics to our dataset in terms of Twitter population size and penetration rate. These results allow us to believe that this dataset can be used to validate our methodology of studying integration patterns through Twitter.

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 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 (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.

$$H(h) = \frac{-\sum_c P_h(c) \log P_h(c)}{\log(|P_h(c)|)}. \tag{1}$$

Figure 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

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

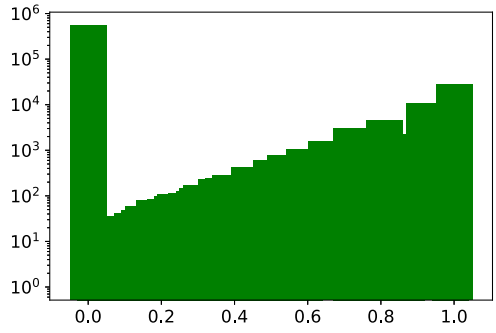

Figure 3 Entropy distribution in Log scale

we are sure they are specific to a nation. We note that the American specific hashtags are in lead, followed by Italian and Great Britain, following the distribution of the number of users from Fig. 1. Examples of Italian specific topics that we have identified are the following: *Salvini, Lavoro, Immigrazione, Caffe, Renzi, Trenitalia, Epifania*. Moreover, examples of some of the international topics we have identified are *Trump, EU, Immigration, Refugee, Coffee, and Fiat.*

3.4 Computing the origin and destination attachment indices

Provided with the nationality of hashtags, we can define for each 3226 migrant user the origin and destination attachment, OA 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 origin attachment of user u, $OA(u)$, 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?

$$OA(u) = \frac{\#\ C_n(u)\ hashtags}{\#\ total\ hashtags} = \frac{HT(u, C_n(u))}{HT(u)}. \tag{2}$$

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

$$DA(u) = \frac{\#\ C_r(u)\ hashtags}{\#\ total\ hashtags} = \frac{HT(u, C_r(u))}{HT(u)}. \tag{3}$$

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 fully attached to the destination/origin country. In contrast, indices equal to 0 means that a migrant is not attached at all to the destination/origin country. The sum of the two indices is always $\leq$1: a user cannot be fully attached to both origin and destination, but has to 'divide' their attention among the various countries they are interested in.

4 Results

4.1 Overall distribution of DA and OA values

The distributions of the origin and destination attachment indices are shown in Fig. 4. We observe that for both OA and DA, the distribution is power-law-like where 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 OA values would be if users chose their topics of discussion randomly, i.e. there was no influence from the country of residence or nationality. Here, we observe that the null model DA_0 and OA_0 also follow a power-law-like pattern. However, we note that the null model DA_0 and OA_0 values are smaller than the actual index values.

To statistically validate the difference between the null model, and DA and OA, 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 OA, their distributions are significantly different from the distribution of the DA_0 and OA_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 OA. 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 OA, we computed the Pearson correlation among them. Figure 5 displays the OA 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/her 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 assimilation types as mentioned in Table 1. In the same figure, the red lines provide an approximate indication of users' assimilation type. We underline the fact that we do not aim to provide a specific categorisation of assimilation types in this paper. Instead, we aim to provide a broad picture where the angle of each individual from the x/y-axis, together with the distance from the origin, gives us an indication of the assimilation 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 globalisation. The distinction between integra-

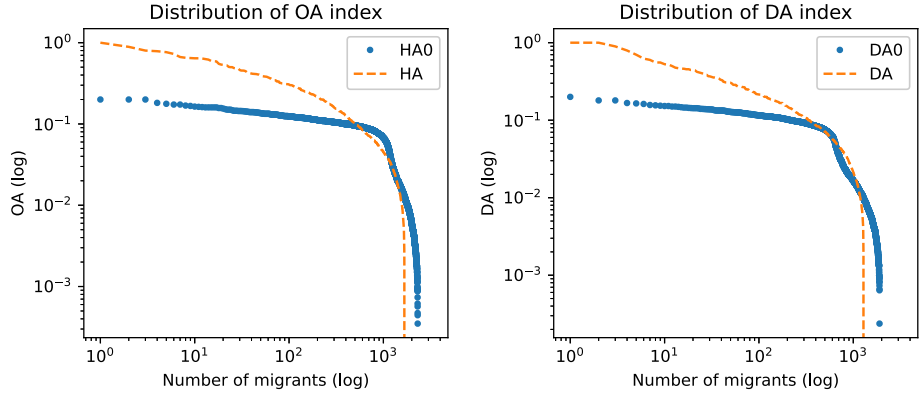

Figure 4 Distribution of OA and DA values, and comparison to null model OA_0 and DA_0 (in log-log scale)

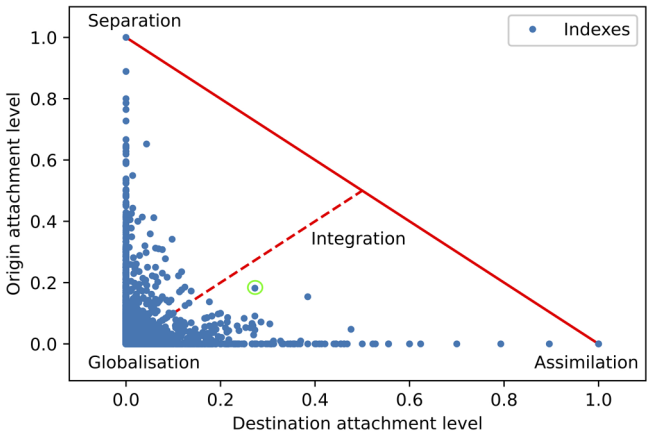

Figure 5 Pearson correlation between origin and destination attachment indices for all the migrants in the data: correlation coefficient: –0.13, p-value: $6.937e^{-14}$

tion and globalisation depends on the length of the distance of data point from the origin. In other words, globalisation is when the data point is close to 0 and integration is when the data point is point further away from 0 along the red dashed line shown in the Fig. 5. The data point circled in green would be a good example of an integrated migrant, who keeps good links with both origin and destination country. We note that a large number of data points are categorised as *globalised* in our data. This tells us that many migrants have little interest in both preserving their origin culture and absorbing the new culture. It could also be that these immigrants choose to talk about international topics and do not choose to display a strong identification with either culture. This supports our decision to introduce a new category of assimilation, globalisation, where individuals become so-called "citizens of the world". In a related work we also found that migrant users tend to speak more languages and travel to more destination, supporting further the globalisation theory [28].

4.2 Language as a key factor for integration

One possible candidate factor to explain the DA and OA values observed is language. As previously studied, language is considered to be a key factor in integration and our indices reflect this importance as well. In Fig. 6 we display the distribution of the DA and OA 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 origin country compared to those speaking the destination language. Hence, interestingly, destination language proficiency seems to affect both destination and origin attachment levels. When comparing DA and OA within groups, the groups that speak the destination language have the two indices comparable, while for those who do not speak it,

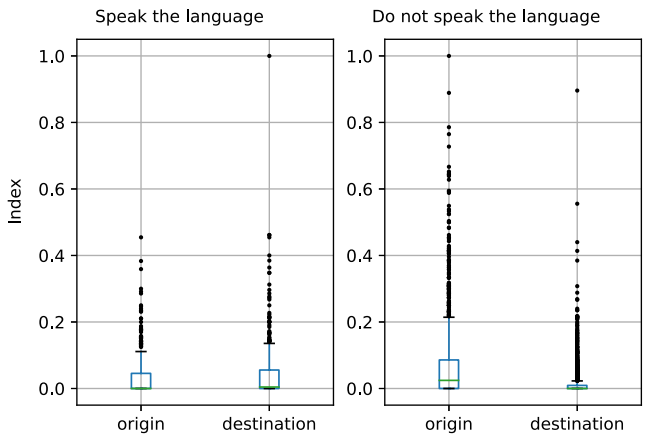

Figure 6 Box plots showing the OA and DA distributions for a group of migrants who speak the language of the host country on the left and a group of migrants who do not speak the language of the host country on the right. Means are $\bar{OA} = 0.034$ and $\bar{DA} = 0.041$ for users who speak the destination language, and $\bar{OA} = 0.072$ and $\bar{DA} = 0.019$ for those who do not speak it

OA 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 assimilation types discussed in Table 1.

4.3 Country-specific results

In this section, we provide country-specific results as an illustration of the methodology. 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. Note that for some countries the sample size is relative small. For these countries, caution is required when interpreting the results.

4.3.1 Immigrants in the United States

In the top panel of the Fig. 7, we observe different destination and origin 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 OA. Immigrants from Canada have the highest DA followed by Colombian and English immigrants. On the other hand, immigrants from Turkey have the highest OA followed by Brazilian and Italian immigrants. In the bottom panel of the same figure, we observe data points individually on a scatter plot of OA vs. DA. It tells us that immigrants in the US are integrated and assimilated in general.

In the case of the United States, we additionally compared our indices to the work of Vigdor [5], which provides three different perspectives of assimilation. In his work, he 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,

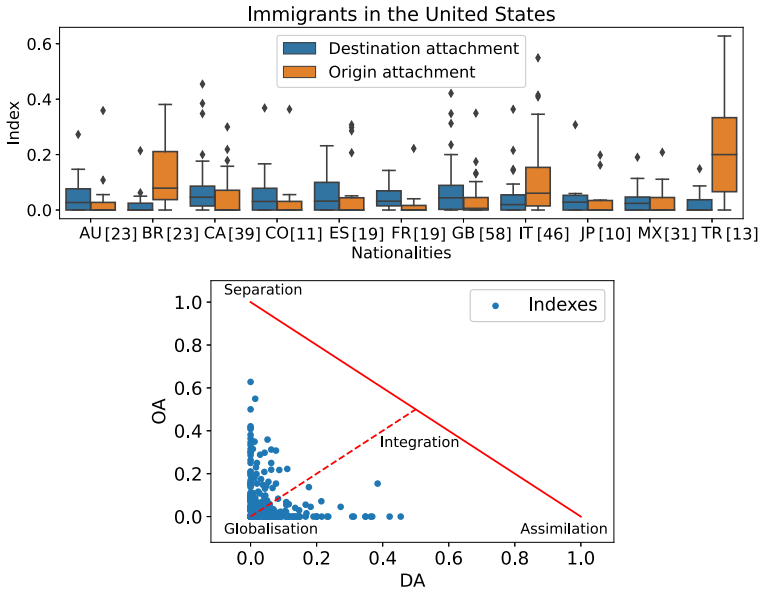

Figure 7 Top: Box plots for the DA and OA index of immigrants in the United States. Bottom: Scatter plot of OA vs. DA indicating approximate integration types for immigrants in the US

Table 2 Spearman correlation table for immigrants in the United States: Vigdor's assimilation scores and DA & OA indices ($n = 292$)

	DA	OA	Composite	Economic	Cultural	Civic
DA	1.0***	−0.231***	0.087	0.185***	0.198***	0.045
OA	−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***

Significance levels are marked with ***p-value < 0.01, **p-value < 0.05, *p-value < 0.1.

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 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 [5] found for nationals from the same countries. The table shows that our DA and OA are most correlated with the cultural factor, followed by the economic factor. It is interesting to remark that DA is positively correlated whereas OA is negatively correlated with the cultural factor of assimilation. This tells us that for those nationalities for which Vigdor observed high cultural assimilation, we observe high DA and low OA, which is exactly how we propose to use our indices to describe assimilation (see Table 1 above). A similar relation can be seen with the economic factor: nationalities with high economic assimilation levels also show high DA and low OA. 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 OA 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 origin 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 OA scores in our data, while the Vigdor data will contain an unique score for them. This inevitably decreases correlations. On the same note, here we report only the correlations at individual level as the correlations at aggregated level yields insignificant p-values except for DA index and Economic assimilation score. This is due to variance that we observe within the same nationality groups, together with the reduced amount of country pairs observed.

4.3.2 Immigrants in the United Kingdom

Figure 8 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 origin than to the destination country. On average, the DA is 0.04 and the OA is 0.063. From the top panel of the Fig. 8, it is clear that

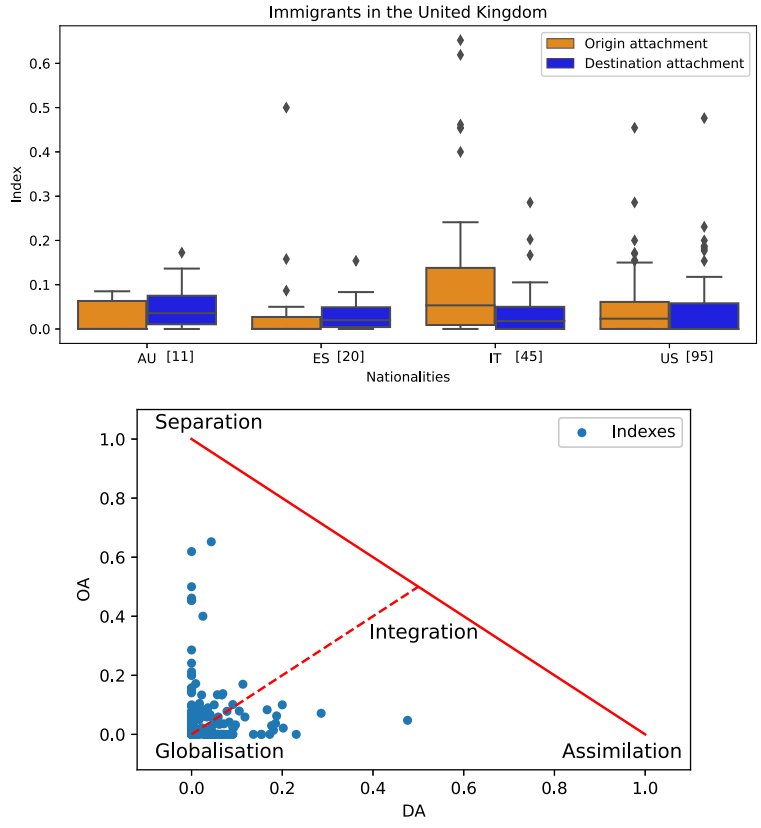

Figure 8 Top: Box plots for DA and OA for immigrants residing in the United Kingdom. Bottom: Scatter plot of OA vs. DA indicating approximate integration types for immigrants in the UK

immigrants from Italy have the highest OA 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 bottom, we can observe that immigrants are mostly in the area of globalisation/integration.

4.3.3 Italian emigrants

Figure 9 displays the DA and OA indices for Italian emigrants present in our dataset across different countries of residence. In general, we observe that our Italian emigrants are more attached to their origin country than to their destination country. Switzerland, Belgium and Netherlands are the three countries where Italian emigrants are most attached to the country of origin. 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 bottom, we also observe that Italian emigrants have higher OA level compared to DA level. This data points indicate that they are in general close to the *separation* type of assimilation.

The high OA levels that we observe, compared to DA levels, could be in part due to the data collection procedure we employed. By using an initial seed location in Italy, in 2015, we probably included in our data some Italian migrants who happened to be in Italy (so manifesting their origin attachment) during the seed period. Our analysis concentrates on 2018, so the migrants we observe could have left Italy either before 2015 or between 2015 and 2018. Those who left before 2015 are more likely to be origin attached. Therefore, again, the results we observe are relevant to the Twitter population in our data and are

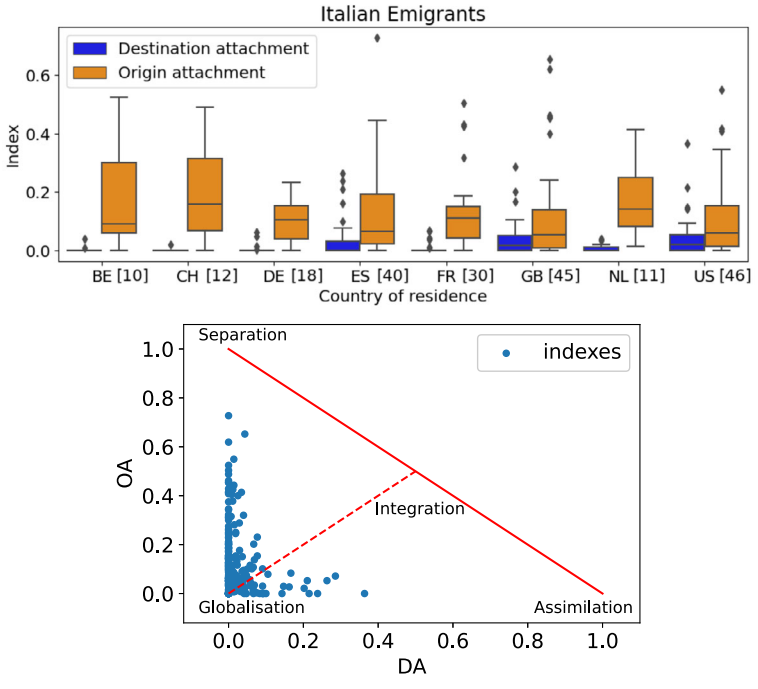

Figure 9 Top: Box plots for DA and OA for Italian nationals living abroad. Countries on *x*-axis are countries of residence of Italians. Bottom: Scatter plot of OA vs. DA indicating approximate integration types for Italian emigrants

not necessarily generalisable to all Italian emigrants. However, we expect that differences among countries of residence are more generalisable.

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 [18, 29–31]. Hofstede's cultural dimensions are well known measures of culture, initially studied to better design the organisational context of business [18]. According to his initial studies, cultures can be studied along four dimensions: power, masculinity, individualism, and uncertainty avoidance.[5] In his later studies, long-term orientation and indulgence[6] were added to the cultural dimensions [29]. To compare our indices with Hofstede's cultural dimensions, we computed the differences of scores between the origin and the destination countries of migrants, as a measure of the cultural distance among countries. We then computed the correlation between our OA 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 3082 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.

Table 3 in Additional file 1 shows the Pearson correlations computed at individual level. The first interesting remark is that in general our DA and OA indices behave differently across the six cultural dimensions, language and distance variables. This means that, when correlations are significant, when OA shows a positive relation, DA shows a negative one and vice-versa. This is compatible with the fact that OA and DA are negatively correlated among themselves, meaning that, in general, as migrants becomes more attached to the destination they lose links to the origin 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 origin 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 OA index: we see that it is significantly negatively correlated with individualism and masculinity. This means that the higher the difference between the origin and the destination country in terms of individualism and masculinity, the less a migrant remains attached to their origin 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 OA indices: the further the destination country is to the country of origin, the higher the DA level and the lower the OA

[5] Power 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.

[6] Long-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.

level. Also, the correlation between *contig* and OA indicates that immigrants in destination countries where they share the border with their country of origin have higher OA levels. This makes sense since having the origin country close means more possibilities to go back home frequently resulting in higher OA 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, we group the migrants by nationality to compute correlations with OA levels, and by residence to compute correlations to DA levels. This allows us to have, for each origin and destination country, an average OA and DA level, computed over a group of migrants.

The correlations obtained are shown in Table 4 in Additional file 2. 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 origin and destination countries are in terms of language, the higher the DA and OA levels. This confirms what we saw earlier, language is not important only for DA, but also for OA. In this case, having a common spoken/national/official language with the destination country allows migrants to maintain stronger links also with their origin country. The same applies when origin and destination countries share borders: both OA 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 obtain with the contig variable, this is not necessary 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.

5 Discussion and conclusions

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 *origin attachment (OA)* and *destination attachment (DA)*, respectively. The OA 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 (OA). 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. By definition of our indicators, high OA and DA cannot be

achieved independently, as opposed to the theory, but rather the attention between the origin and destination countries has to be 'divided', which is possibly a more realistic setting. 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 corresponds to higher DA level, as does having a common native language with the destination country. Interestingly, common languages also correlate with large OA levels, which is a less explored result. Second, we saw that in general, sharing borders increases both the DA and OA 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 OA index.

It is important to mention that detecting causality in the results above may be difficult. For example, the proficiency of the language of the host country could facilitate higher DA levels. But this relationship could also be true the other way around. Although the causality issue cannot be disentangled in our analysis, we believe that through this work, we were able to shed light to important relationships between several important elements of culture and migrants' attachments to the host and home countries, thus highlighting different cultural integration processes through data.

An important aspect is that assimilation is a dynamic process, while here we only look at a snapshot, without taking into account the length of the residence period in a certain country, which is surely important in determining DA and OA levels. This could be in principle calculated from Twitter data, however a larger data sample would be needed. This would also help understand biases in data collection, for instance if OA levels observed in Italian emigrants in Sect. 4.3 are increased by our methodology. We will investigate this aspect in future work.

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 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 3226 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. As a related issue, we have seen in the country-specific results that the sample sizes for some countries were small despite the initial size of the data. Regarding this issue, although we eliminated countries with less than 10 users, interpreting these results is difficult as they cannot represent the population of the country in interest. However, we can still observe some trends that can then be confirmed in future work.

Another drawback of using big data is that this analysis may suffer from sampling bias. It is known that Twitter is positively selective on highly educated individual users [32] which is most likely reflected in our data also. The Twitter population is different from the real one, hence not all the demographic groups are covered in the analysis. Our findings apply mostly to the upper-tail of immigration. While integration of low-skill immigrants is much discussed as their economic and cultural integration is an important political issue,

being able to integrate diverse high-skill immigrants is also critical in terms of competitiveness, nation-building, political inclusiveness, and long-run economic growth (e.g., [33]). Our work, therefore, can be implemented to monitor integration processes of high skilled immigrants in real-time allowing researchers and policy makers to study how these immigrants relate back to the culture, language and politics of their origin country. On the same note, it is also possible that we are neglecting well-integrated migrants due to the way we assign nationality in this work. In particular, a well-integrated migrant whose friends are mainly located at the destination country will not be assigned the nationality of their origin, but would appear as a native at their destination. Our methodology only sees migrants who maintain some links, even if not active, with people at home. However, we believe that in such case, these neglected migrant users would be permanent migrants who have been in the country for a long period of time, or even have been naturalised at the point of analysis.

In this project, we relied on information from Twitter such as language of tweets and hashtags to determine the levels of OA and DA. However, since the exact means through which Twitter identifies the language of tweets are unknown, in the future, one could also detect languages using existing algorithms to improve language identification. Moreover, topic identification can also be improved in the future by using existing topic modelling methods.

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 we do not intend to generalise the findings of this paper. They apply solely to a small sample of the population, and not to larger groups. The study has undergone ethics screening before publication.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-022-00363-5.

Additional file 1. Table 3. Correlation table for OA & DA and Hofstede's cultural dimension scores for migrants at individual level. (CSV 1 kB)
Additional file 2. Table 4. Correlation table for OA & DA and Hofstede's cultural dimension scores. (CSV 1 kB)

Acknowledgements
This work received a first prize award for the flash talk at the conference "New Data for the new challenges of population and society", organised by the Department of Statistical Sciences of the University of Padova, within the Project of Excellence "Statistical methods and models for complex data".

Funding
This work was supported by the European Commission through the Horizon2020 European projects "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (grant agreement no 871042) and "HumMingBird—Enhanced migration measures from a multidimensional perspective" (Research and Innovation Action, grant agreement no 870661). Open Access funding enabled and organized by Projekt DEAL.

Abbreviations
OA, Origin Attachment index; DA, Destination Attachment index.

Availability of data and materials
Anonymised data is openly available at https://doi.org/10.6084/m9.figshare.19348058.v3 which includes OA and DA indexes and number of followers, friends and tweets of individual users.

Declarations

Ethics approval and consent to participate
All the procedures on use, process and storage of data were reviewed by the SoBigData Board for Operational Ethics and Legality.

Competing interests
The authors declare that they have no competing interests.

Author contribution
Conceptualization: JK, AS, FG and HR; Data collection: JK and GR; Methodology: JK, AS, FG, GR and HR; Formal analysis and investigation: JK, and AS; Writing: JK, AS, FG, GR and HR; Funding acquisition: FG, HR and AS. All authors read and approved the final manuscript.

Author details
[1]Scuola Normale Superiore, Pisa, Italy. [2]Max Planck Institute for Demographic Research, Rostock, Germany. [3]University of Pisa, Pisa, Italy. [4]Istituto di Scienza e Tecnologie dell'Informazione, National Research Council of Italy, Pisa, Italy. [5]Paris School of Economics, Université Paris 1 Panthéon-Sorbonne, & CEPII, Paris, France.

Publisher's Note

Received: 13 October 2021 Accepted: 31 August 2022 Published online: 18 November 2022

References

1. Norris P, Inglehart R (2019) Cultural backlash Trump, Brexit, and authoritarian populism. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781108595841
2. Eurostat, European Commission (2011) Indicators of immigrant integration. A pilot study. Eurostat/European Commission, Brussels
3. OECD and European Commission (2018) Settling in 2018. Main Indicators of Immigrant Integration. OECD Publishing, Paris
4. Huddleston T, Niessen J, Tjaden JD (2013) Using EU indicators of immigrant integration. Final report for directorate-general for home affairs. European Commission, Brussels
5. Vigdor JL (2008) Measuring immigrant assimilation in the united states. Civic report no. 53, Manhattan Institute for Policy Research
6. Lochmann A, Rapoport H, Speciale B (2019) The effect of language training on immigrants' economic integration: empirical evidence from France. Eur Econ Rev 113:265–296
7. Sîrbu A, Andrienko G, Andrienko N, Boldrini C, Conti M, Giannotti F, Guidotti R, Bertoli S, Kim J, Muntean CI et al (2021) Human migration: the big data perspective. Int J Data Sci Anal 11:341–360
8. Esser H (2006) Migration, language and integration. Citeseer
9. Safi M (2008) The immigrant integration process in France: inequalities and segmentation. Rev Fr Sociol 49(5):3–44
10. Park RE (1928) Human migration and the marginal man. Am J Sociol 33(6):881–893
11. Constant AF, Zimmermann KF (2008) Measuring ethnic identity and its impact on economic behavior. J Eur Econ Assoc 6(2–3):424–433
12. Werth M, Delfs S, Stevens W (1997) Measurement and indicators of integration, Council of Europe. Directorate of Social Economic Affairs
13. Portes A, Bach RL (1985) Latin journey: Cuban and Mexican immigrants in the United States. University of California Press, Berkeley
14. Penninx M et al (2003) Integration. The role of communities, institutions, and the state. The Migration Information Source (on-line)
15. Berry JW (1997) Immigration, acculturation, and adaptation. Appl Psychol 46(1):5–34
16. Rapoport H (2016) Migration and globalization: what's in it for developing countries? Int J Manpow 37(7):1209–1226
17. Kim J, Sîrbu A, Giannotti F, Gabrielli L (2020) Digital footprints of international migration on Twitter. In: International symposium on intelligent data analysis. Springer, Berlin, pp 274–286
18. Hofstede G (1984) Culture's consequences: international differences in work-related values. SAGE, Beverly Hills
19. Robinson HM (2019) Dynamics of culture and curriculum design: preparing culturally responsive teacher candidates. IGI Global, Hershey
20. Kaasa A, Vadi M, Varblane U (2016) A new dataset of cultural distances for European countries and regions. Res Int Bus Finance 37:231–241
21. Kaasa A, Vadi M, Varblane U (2014) Regional cultural differences within European countries: evidence from multi-country surveys. Manag Int Rev 54(6):825–852
22. Alba R, Logan J, Lutz A, Stults B (2002) Only English by the third generation? Loss and preservation of the mother tongue among the grandchildren of contemporary immigrants. Demography 39(3):467–484
23. Guidotti R et al (2021) Measuring immigrants adoption of natives shopping consumption with machine learning. In: Machine learning and knowledge discovery in databases. Applied data science and demo track. ECML PKDD 2020. Lecture notes in computer science, vol 12461. Springer, Cham

24. Dubois A, Zagheni E, Garimella K, Weber I (2018) Studying migrant assimilation through Facebook interests. In: International conference on social informatics. Springer, Berlin, pp 51–60
25. Stewart I, Flores RD, Riffe T, Weber I, Zagheni E (2019) Rock, rap, or reggaeton? Assessing Mexican immigrants' cultural assimilation using Facebook data. In: The world wide web conference. ACM, New York, pp 3258–3264
26. Vieira C, Ribeiro F, Vaz de Melo PO, Benevenuto F, Zagheni E (2020) Using Facebook data to measure cultural distance between countries: the case of Brazilian cuisine. In: Proceedings of the web conference 2020, pp 3091–3097
27. Coletto M, Esuli A, Lucchese C, Muntean CI, Nardini FM, Perego R, Renso C (2017) Perception of social phenomena through the multidimensional analysis of online social networks. Online Soc Netw Media 1:14–32
28. Kim J, Sîrbu A, Rossetti G, Giannotti F (2021) Characterising different communities of Twitter users: migrants and natives. In: International conference on complex networks and their applications. Springer, Berlin, pp 130–141
29. Hofstede G (2011) Dimensionalizing cultures: the Hofstede model in context. Online Read Psychol Cult 2(1):8
30. Mayer T, Zignago S (2011) Notes on cepii's distances measures: the geodist database. CEPII
31. Melitz J, Toubal F (2014) Native language, spoken language, translation and trade. J Int Econ 93(2):351–363
32. Wojcik S, Hughes A (2019) Sizing up Twitter users. PEW research center 24
33. Alesina A, Harnoss J, Rapoport H (2016) Birthplace diversity and economic prosperity. J Econ Growth 21(2):101–138