Immigrant community integration in world cities

Fabio Lamanna,1,* Maxime Lenormand,2 María Henar Salas-Olmedo,3 Gustavo Romanillos,3 Bruno Gonçalves,4 and José J. Ramasco1
1Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), Campus UIB, ES-07122 Palma de Mallorca, Spain
2Irstea, UMR TETIS, 500 rue JF Breton, FR-34093 Montpellier, France
3Departamento de Geografía Humana, Facultad de Geografía e Historia, Universidad Complutense de Madrid, 28040, Madrid, Spain
4Center for Data Science, New York University, New York, 10011 NY, USA

Migrant and hosting communities face long-term challenges in the integration process. Immigrants must adapt to new laws and ways of life, while hosts need to adjust to multicultural societies. Integration impacts many facets of life such as access to jobs, real state and public services and can be well approximated by the extent of spatial segregation of minority group residence. Here we conduct an extensive study of immigrant integration in 53 world cities by using Twitter language detection and by introducing metrics of spatial segregation. In this way, we quantify the "Power of Integration of cities (their capacity to integrate diverse cultures), and characterize the relations between cultures when they act in the role of hosts and immigrants.

INTRODUCTION

Immigrant integration is a complex process comprehend many different factors such as employment, housing, education, health, language, legal recognition as well as the built of a new social fabric. In the last years, there have been advances in the definition of a common framework concerning immigration studies and policies [1], although the approach to this issue remains strongly country-based [2]. The outcome of the process actually depends on the culture of origin, the one of integration and the policies of the hosting country government [3]. Traditionally, spatial segregation in the residential patterns of a certain community has been taken as an indication of ghettoization or lack of integration [4]. While this applies to immigrant communities, it can also affect to minorities within a single country [5]. The spatial isolation reflects in the economic status of the segregated community and in social relationships of its members [6].

In global terms, while international migration flows have remained almost stable over the last 20 years [7] political and economic developments such as the Arab Spring and the Syrian civil war have brought the problem of migrants and their integration to the forefront of world news. A good part of newcomers concentrates in cities, and particularly in the large metropolises known as Global Cities; these are centers that attract specialized immigration, driving important social and cultural transformations in cities worldwide [8]. The concept of Global or World Cities emerged in the 80s [9, 10] as strategic territories that articulate the international economic structure. According to Sassen [9], Global Cities are not only characterized by growing multiculturalism but also by a rising social polarization, which was finally materialized into an increasing social spatial segregation and gentrification processes.

[This assertion is still under debate requiring further empirical evidence [11, 12].

Immigrant integration has been the focus of many research studies, but most of them have been conducted from national perspectives, especially in European countries and the USA [2, 3, 5, 13, 14], and still lack of information sources beyond national census [15]. In parallel, in the last few years we have witnessed a paradigm shift in the context of socio-technical data. Human interactions are being digitally traced, recorded and analysed in large scale. Sources are as varied and different as mobile phone data [16], credit card transactions [17], or Twitter data [18]. Going beyond the urban scale, Twitter data have been used to detect the diffusion of human mobility [19, 20] and on the languages spoken. Language identification related to the spatial location of Twitter users has been investigating [21, 22], towards a more complete characterization of spatial local dialects [23, 24]. Finally, Twitter has been used as a statistical database for representations of demographical characteristics of users [25] and language identification patterns [26]. Several attempts have been made in order to identify, characterize and group international communities in cities based on Information and Communication Technologies (ICT) data [27, 28] and to perform social segregation analyses [29, 30].

Here we present a novelty approach to quantify the spatial integration of immigrant communities in urban areas worldwide by using social media information collected through the Twitter microblogging platform; first we characterize immigrants through their digital spatio-temporal communication patterns, defining their residence place and most probable native language; then we perform a spatial distribution analysis through a modified entropy metric, as a quantitative measure of the spatial integration of each community in cities and the corresponding relevance within countries.
Figure 1. Dataset and framework description. The cities passed through the lens (a) of our framework are mostly distributed over four continents (Africa has been not considered due to the lack of data available). We cover each city with a square grid in order to keep the distribution of the users over the whole urban area (b), selecting resident users and their most frequent location thanks to their activity over space and time (c). We assign users most probable native language (d) and perform a spatial analysis over the cities (e) in order to get information about the population distribution in function of the spoken language of the users.

MATERIALS AND METHODS

Dataset description

We select 53 of the most populated cities in the world (Fig. 1a) and analyzed all the geo-localized tweets originating in each city between October 2010 and December 2015 through the Twitter API. Several items were extracted from each tweet: user ID, geographical coordinates (Latitude and Longitude), date and time information and the text of the tweet. In order to get the correct time frames of data among all locations, we converted the Twitter UTC time into the local timezone for each city. Working with such huge databases implies the need of filter out non-human users from the Twitter dataset; this is fundamental in this kind of studies in order to prevent results given by bot or cyborg automatic tweets generators, which are indeed frequent in those kind of social networking communications [31]. We found in the database several tweets generated at the same time (with the precision of the second) by the same account. Moreover, we discard users who tweet more than three times per minute. Finally, we detect the speed of users moving through consecutive locations in order to filter out those who travel faster than the normal velocity admitted in an urban area.

City Definition

In order to fairly compare the different cities, we must agree on a city definition that can be applied around the world and is large enough to include the whole metropolitan area. Unfortunately, generic definitions such as the Larger Urban Zone (LUZ) definition in Europe (introduced in 2004 by Eurostat) does not exist at the world scale. There are plenty of different ways of defining cities, with, for example, methods based on urban growth, percolation, attraction or fractal theory. All these methods require third party data such as population, built-up area or flows of commuters that is not easily available in a consistent form everywhere. To side step this difficulty, we use a very pragmatic definition based only on the Euclidean distance and consider all activity within a frame of $60 \times 60$ km centered on the barycenters listed in Table S1 to belong to the city itself, dividing each city area using an equally spaced grid of 500 x 500 meters (Fig. 1b). In order to check the stability of the
results in function of the size of the city boundaries chosen, we evaluate the relative error among the edge weights of the Bipartite Spatial Integration Network for each city for different sizes of the frame (20, 40, 60, 80 and 100 km), taking as reference the original 60×60 km frame grid. The results are stable for frame sizes ranging from 20 to 80 (Fig. S7 in the appendix section); above these values, we face a different behavior in terms of integration process since we extend the influence area to cover a higher number of users.

Users’ residency and language characterization

For each city, users’ place of residence was identified as the most frequented cell during night time (Fig. 1c). The source code used to extract most visited locations from individual spatio-temporal trajectories is available online [32]. It is important to note at this stage that all cities have at least 1000 users with identifiable home location (Table S2). Finally, a language has been assigned to each user (Fig. 1d), based on those detected in their communications (Table S3). A user tweeting in any language other than the local one (Table S4) or in English is assigned to the corresponding minority community. The presence of these communities as immigrants in the city is validated against external data sources such as the census information (Fig. S2 and Fig. S3). Further information on data preparation and validation can be found in the appendix section.

RESULTS

Bipartite Spatial Integration Network

In order to quantify the spatial segregation of each immigrant community in each city, we build a Bipartite Spatial Integration Network, $H$ (Fig. 2), where every language, $l$, is connected to each city, $c$, where the corresponding immigrant communities is detected. The edge weight $h_{l,c}$ represents the degree of integration based on the Shannon entropy-like descriptors [33] but modifying the metric to cope with finite sampling. Since the number of detected users in each city (Table S2) and per language (Table S5) is very different, we had to consider a normalized metric in order to get real information about the distribution of the languages spoken in the cities, no matter the number of users involved. To do this, we take as null model a situation where the $l$ users detected in a language are drawn at random $R$ times over the city cells according to the total distribution of users (population) in each cell, and evaluate the corresponding average entropy as:

$$h^\text{ran}_{l,c} = \frac{1}{R} \sum_{r=1}^{R} \sum_{i=1}^{N} p_{l,i}^r \log(p_{l,i}^r),$$  \hspace{1cm} (1)
Figure 3. Clusters of cities and Power of Integration. Three groups of cities show similar behaviour in the number of communities detected and in their levels of integration. The length of the vectors represent the number of languages (communities) detected in each city; the color scale is representative of the decay of the entropy metric; the Power of Integration metric leads us to evaluate the potential of each city in uniformly integrating immigrant communities within its own urban area according to entropy values.

In order to highlight similarities in the way the different cities integrate foreign cultures, we perform a clustering analysis based on the distribution of edge weight $h_{l,c}$ for each city $c$. For each city, the distribution of edge weights are sorted in descending order and stored into a vector $E_c$. The size of $E_c$ is equal to the total number of languages detected in the dataset $L_{max}$. If the number of languages $L_c$ detected in city $c$ is lower than $L_{max}$, then the last $L_{max} - L_c$ elements of vector $E_c$ are set to 0. We then perform a clustering analysis to cluster together cities exhibiting similar distribution of edge weights. The vectors have been clustered together using the k-means algorithm based on Euclidean distances. We also performed the clustering analysis using a Hierarchical Clustering Algorithm but we obtained the same results (Fig. S10).

Power of Integration

In order to highlight similarities in the way the different cities integrate foreign cultures, we perform a clustering analysis based on the distribution of edge weight $h_{l,c}$ for each city $c$. For each city, the distribution of edge weights are sorted in descending order and stored into a vector $E_c$. The size of $E_c$ is equal to the total number of languages detected in the dataset $L_{max}$. If the number of languages $L_c$ detected in city $c$ is lower than $L_{max}$, then the last $L_{max} - L_c$ elements of vector $E_c$ are set to 0. We then perform a clustering analysis to cluster together cities exhibiting similar distribution of edge weights. The vectors have been clustered together using the k-means algorithm based on Euclidean distances. We also performed the clustering analysis using a Hierarchical Clustering Algorithm but we obtained the same results (Fig. S10).

Figure 4. Decays of entropy for clusters. Each line corresponds to the normalized decay of the entropy metric for each cluster. To compare the level of entropy of cities integrating different numbers of communities, we normalized the vector values by the total number of languages/city and plot the corresponding values of $h'$. Figure 3 shows the three clusters (C1, C2, C3) ob-
tained after applying the clustering algorithm. According to the decay of entropy values (Fig. 4), three types of clusters are identified: C1 is the lower limit, where many communities are generally spatially distributed closely to the local population; C3 is the upper limit, comprising cities that both (spatially) well-integrate/high-segregate immigrants. Cities where a smaller amount of groups of population are uniformly distributed, with skews of segregation phenomena (C2) stays in the middle. We also introduce a new metric in order to summarize the distribution of entropy and to assess the city’s Power of Integration (Table S8). This metric is defined, for each city, as:

\[ P_c = \frac{L_c}{L_{\text{max}}} Q_2(1 - Q_2), \]  

where \( L_c \) is the number of languages spoken in city \( c \) and \( L_{\text{max}} \) is the maximum number of languages across the whole set of cities, \( Q_2 \) is the median value of entropy and IQR the Interquartile range used as a measure of dispersion.

Language Integration Network

We project the Bipartite Spatial Integration Network into the language side, in order to obtain insights about the level of integration of languages into countries (Table S9). We link the languages, along with their proper level of integration, into the country where they are integrated in. We take the mean value of Entropy when the same language is linked with multiple cities in the same countries, obtaining the Language Integration Networks shown in Fig. 5 and 6. We do not consider English language in the network, because of its role as lingua franca [34] in social networking, and the distribution of their levels of integration in the network (Fig. S11). We select two thresholds of levels of integration in countries: in the top set (Fig. 5) the strong Power of Integration of UK cities (London and Manchester) sets its dominant role in uniformly spatial integrating several communities. Several patterns of uniform spatial integration appear, such as the Italian community influence in Venezuela, and the Spanish-speaking in Germany, Singapore and Turkey; the latter country shows uniform distributed communities of Spanish people (due to historical migrations of Spanish Jews dating as far back as the 15th century), and Kurdish (largest ethnic minority in Istanbul). South-Slavic and East-Slavic communities keep their traditional presence in Russia and Germany. Increasing the threshold of links, UK leads its role in hosting several communities within its local population and some other patterns emerge, such as the German presence in Japan and UK. By contrast (Fig. 6), Arabic arises as the most common spatially segregated community within countries followed by French communities that appear to be spatially concentrated in other European countries such as Germany and Turkey. Increasing the thresholds, results in more forms of segregation appearing in Canada (East-Slavic, French and Tagalog), Australia (Malay and Japanese), Brazil (French) and Philippines (Italian and Spanish).

DISCUSSION

People are constantly moving within cities and countries, looking for jobs, experiences or just for better life conditions, facing the fact of the integration in habits and laws of new local cultures. Migration flows have been extensively studied so far, by means
of widely available data that range from the number of people living outside their country of birth to place of residency or labour market data. Rather than using the latter sources, in this study we proposed new findings on how Twitter users language might be a direct connection to their hometown and/or their nationality, how spatial and linguistic characteristics of people vary within dimensions and constraints of the cities they are living in and how far cities are able to handle diversity of languages and cultures. Despite some natural biases (due to the social network penetration rate through social hierarchies, age and countries), Twitter is able to reproduce spatial patterns and mobility profiles like other data sources (like census) do [18]. This is a general method that helps measure how well different communities are spatially integrated/segregated within urban areas. Our findings provide a new way to observe the patterns of historically permanent migration of people in urban areas, and any potential changes that might arise. We are able to move beyond the estimation of past, current and foreshadowed global flows, and towards a better comprehension of the integration phenomena on a city scale. Residents online communications let us know that the hometown roots have been permanently kept inside communities, although impacted on different levels by local welcoming and hosting policies.

**ACKNOWLEDGMENTS**

Partial financial support has been received from the Spanish Ministry of Economy (MINECO) and FEDER (EU) under the project ESOTECOS (FIS2015-63628-C2-2-R), and from the EU Commission through project INSIGHT. The work of M-HS-O was supported in part by a post-doctoral fellowship of MINECO at Universidad Complutense de Madrid (FPDI 2013/17001). JJR acknowledges funding from the Ramn y Cajal program of MINECO. BG thanks the Moore and Sloan Foundations for support as part of the Moore-Sloan Data Science Environment at New York University.

---

[1] Alastair Ager and Alison Strang. Understanding integration: A conceptual framework. *Journal of Refugee Studies*, 21:166–191, 2008.
[2] Han Entzinger and Renske Biezeveld. Benchmarking in Immigrant Integration. *Managing Integration. The European Union's Responsibilities Towards Immigrants*, (August):123–136, 2005.
[3] Tol Gonul. A Comparative Study of the Integration of the Turks in Germany and the Netherlands. *Migration Letters*, 9:25–32, 2012.
[4] Douglas S. Massey and Nancy A Denton. *American apartheid: segregation and the making of the underclass*. Cambridge, MA, USA, harvard un edition, 1993.
[5] Douglas S. Massey and Nancy A Denton. Trends in the Residential Segregation of Blacks, Hispanics, and Asians: 1970-1980. *American Sociological Review*, 52(6):802–825, 1987.
[6] Masayoshi Oka and David W.S. Wong. Spatializing segregation measures: an approach to better depict social relationships. *Citiescape*, (17):97–113, 2015.
[7] Guy J. Abel and Nikola Sander. Quantifying Global International Migration Flows. *Science*, 343(1520), 2014.
[8] Jonathan Beaverstock. Lending Jobs to Global Cities: Skilled International Labour Migration, Investment Banking and the City of London. Urban Studies, 33(8):1377–1394, 1996.

[9] Saskia Sassen. The global city: introducing a concept. The Brown Journal of World Affairs, XI(2):27–40, 2005.

[10] John Friedmann. The World City Hypothesis. Development and change, 17(1):69–83, 1986.

[11] Michael Samers. Immigration and the Global City Hypothesis: Towards an Alternative Research Agenda. International Journal of Urban and Regional Research, 26(2):389–3402, 2002.

[12] Chris Hamnett. Social Polarisation in Global Cities: Theory and Evidence. Urban Studies, 31(3):401–424, 1994.

[13] Sako Musterd. Social and Ethnic Segregation in Europe: Levels, Causes, and Effects. Journal of Urban Affairs, 27(3):331–348, 2005.

[14] Gillian Stevens Frank D. Bean. America’s Newcomers. Immigration and the Dynamics of Diversity. Russell Sage Foundation, 2003.

[15] Karen Phalet and Marc Swyngedouw. Measuring migrant integration: The case of Belgium. Studi Emigrazione, (152):773–804, 2003.

[16] Thomas Louail, Maxime Lenormand, Oliva G. Cantú-Ros, Miguel Picornell, Ricardo Herranz, Enrique Frais-Martinez, José Javier Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. Scientific reports, 4:5276, jun 2014.

[17] Maxime Lenormand, Thomas Louail, Oliva G. Cantú-Ros, Miguel Picornell, Juan Murillo Arias, Marc Barthelemy, and Maxi San Miguel. Influence of sociodemographic characteristics on human mobility. Scientific Reports, 5:1–23, 2015.

[18] Maxime Lenormand, Miguel Picornell, Oliva G. Cantú-Ros, Antonia Tugores, Thomas Louail, Ricardo Herranz, Marc Barthelemy, Enrique Frais-Martínz, and José Javier Ramasco. Cross-checking different sources of mobility information. PloS one, 9(8), 2014.

[19] Maxime Lenormand, Bruno Gonçalves, Antónia Tugores, and José Javier Ramasco. Human diffusion and city influence. Journal of The Royal Society Interface, 12(109), 2015.

[20] 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, 00(Gesenhues):1–12, 2014.

[21] Delia Mocanu, Andrea Barouchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. The Twitter of Babel: Mapping World Languages through Microblogging Platforms. PLoS ONE, 8(4):e61981, jan 2013.

[22] Raja Jurda, Kun Zhao, Jiajun Liu, Maurice Abou-Jaoude, Mark Cameron, and David Newth. Understanding human mobility from Twitter. PLoS ONE, 10(7):35, dec 2015.

[23] Bruno Gonçalves and David Sanchez. Crowdsourcing dialect characterization through twitter. PLoS ONE, 9(11):1–10, 2014.

[24] Gabriel Doyle. Mapping dialectal variation by querying social media. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014.

[25] Alan Mislove, Sune Lehmann, Yong-yeol Ahn, Jukka-Pekka Omela, and J Niels Rosenquist. Understanding the Demographics of Twitter Users. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, pages 554–557, 2011.

[26] Daniel Arribas-Bel. The spoken postcodes. Regional Studies, Regional Science, 2(1):458–461, jan 2015.

[27] Paolo Bajardi, Matteo Delfino, André Panisson, Giovanni Petri, and Michele Tizzoni. Unveiling patterns of international communities in a global city using mobile phone data. EPJ Data Science, 4(1):3, 2015.

[28] Amir Magdy, Thanas M. Ghanem, Mashaal Musleh, and Mohamed F. Mokbel. Exploiting Geo-tagged Tweets to Understand Localized Language Diversity. In Proceedings of Workshop on Managing and Mining Enriched Geo-Spatial Data - GeoRich’14, pages 1–6, 2014.

[29] Alexander Amini, Kevin Kung, Chaogui Kang, Stanislav Sobolevsky, and Carlo Ratti. The impact of social segregation on human mobility in developing and industrialized regions. EPJ Data Science, 3(1):6, 2014.

[30] Eszter Bokáinyi, Dániel Kondor, László Dobos, Tamás Sebk, József Stéger, István Csabai, and Gábor Vattay. Race, religion and the city: twitter word frequency patterns reveal dominant demographic dimensions in the United States. Palgrave Communications, 2:16010, apr 2016.

[31] Zi Chu, Steven Gionvecchio, Haining Wang, and Sushil Jajodia. Who is Tweeting on Twitter: Human, Bot, or Cyborg? In Proceedings of the 26th Annual Computer Security Applications Conference on - ACSAC ’10, page 21, 2010.

[32] https://github.com/maximelenormand/Most-frequented-locations.

[33] Michael J White. Segregation and Diversity Measures in Population Distribution. Population Index, 2014.

[34] Shahar Ronen, Bruno Gonçalves, Kevin Z. Hu, Alessandro Vespignani, Steven Pinker, and César a. Hidalgo. Links that speak: The global language network and its association with global fame. Proceedings of the National Academy of Sciences, (111):E5616, dec 2014.

[35] These spatial operations were made using the commercial software ESRI ArcGIS 10.3.

[36] Luc Anselin. Local indicators of spatial association-LISA. Geographical analysis, 27(2):93–115, 1995.

[37] Luc Anselin, Ihnui Syabri, and Younglun Kho. GeoDa: An introduction to spatial data analysis. Geographical Analysis, 38(1):5–22, 2006.

[38] Andrea Lancichinetti, Filippo Radicchi, José Javier Ramasco, and Santo Fortunato. Finding statistically significant communities in networks. PLoS ONE, 6, 2011.

[39] Martin Rosvall, Daniel Axelsson, and Carl T. Bergstrom. The map equation. The European Physical Journal Special Topics, 178(1):13–23, 2009.
APPENDIX

DATA PREPROCESSING

The methodology for estimating the integration of immigrants communities in cities through Twitter data is described in this section. Firstly we describe the dataset, with a special section dedicated to the language identification framework and the living patterns of users; then we review the spatial analysis of the data over the cities and the validation of the data; finally we present the statistical analyses that lead us to introduce the network approach to describe the relationships among languages and cultures.

Definition of the users place of residence

We must identify the likely place of residence of each user. For each city, we define a spatial grid composed of square cells (500 × 500 m$^2$). Equidistant Cylindrical Projection) and define the home location of a user to be the grid cell where most of his or hers activity occurs between 8pm and 8 am. Only users who posted more than one tweet from within the city boundaries are considered. To ensure that a user shows enough regularity and that he/she is really living in the city, and not just a visitor for a small period of time, we applied three filters: $C$ the minimum number of consecutive months of activity, $N$ the minimum number of hours spent by the user in their Most Frequent Location, MFL, and $\Delta$ as the ratio between $N$ and the total number of hours of activity for each user.

Users who are active within a given city for at least three consecutive months are considered to reside within the city. The values of the other two parameters, were determined empirically. In Fig. S1 we plot the evolution of the number of reliable users in the dataset as a function of $\Delta$ according to $N=\{5,10,15,20\}$ for $C=3$ for each of the 53 cities. As the shape of the curves is similar for different values of $N$ and does not display any natural features that would allow us to define a clear cutoff, we fix $\Delta$ to 0.2 and $N=5$, as a trade-off that both allows us to keep as many users as possible while assuring us that users exhibiting irregular mobility patterns during the time period considered is minimized. In Table S2 we list the final number of residents for each city after this data cleaning procedure. Note that there are at least 1,000 reliable users in each city.

| City       | Latitude   | Longitude   |
|------------|------------|-------------|
| Amsterdam  | 52.370216  | 4.895168    |
| Atlanta    | 33.748995  | -84.387982  |
| Bandung    | -6.914744  | 107.609811  |
| Bangkok    | 13.727980  | 100.524123  |
| Barcelona  | 41.385064  | 2.173403    |
| Berlin     | 52.519171  | 13.406091   |
| Bogota     | 4.598056   | -74.075833  |
| Boston     | 42.358431  | -71.059773  |
| Brussels   | 50.85034   | 4.35171     |
| Buenos Aires | -34.603723 | -58.381593  |
| Caracas    | 10.491016  | -66.902061  |
| Chicago    | 41.878114  | -87.629798  |
| Dallas     | 32.78014   | -96.800451  |
| Detroit    | 42.331427  | -83.045754  |
| Jakarta    | -6.211544  | 106.845172  |
| Dublin     | 53.349805  | -6.26035    |
| Guadalajara| 20.67359   | -103.343803 |
| Houston    | 29.760193  | -95.36939   |
| Istanbul   | 41.05027   | 28.97696    |
| Kuala Lumpur| 3.139003  | 101.68655   |
| Lima       | -12.047816 | -77.062203  |
| Lisbon     | 38.725299  | -9.150036   |
| London     | 51.511214  | -0.119824   |
| Los Angeles| 33.95      | -118.14     |
| Madrid     | 40.416775  | -3.70379    |
| Manchester | 53.479324  | -2.248485   |
| Manila     | 14.599512  | 120.984219  |

Table S1. Coordinates of the centers of the frame for each city.
Figure S1. Number of reliable users as a function of $N$ and $\Delta$. Each line represents the trend of each city in the number of users according to the ratio between $N$ and the total number of hours of activity for each user ($\Delta$). Set as $C=3$ the number of months for consecutive activities, (a) refers to $N=5$, (b) to $N=10$, (c) to $N=15$ and (d) to $N=20$.

Language detection framework

We detect the languages used in each tweet using version 2.0 of the "Chromium Compact Language Detector" (CLD2) which returns the languages detected along with a confidence assessment. CLD2 is a C++ library with Python wrappers that implements a Bayesian classifier for detecting language from UTF-8 text. The code was originally written by Google employees and later extracted from the open source Chromium browser source code. Thanks to Google’s efforts, it has a very fast detection and a large library of known languages (75+).

Twitter entities (urls, mentions, hashtags) that might impair our language detection efforts were removed and only the remaining text was given as input to CLD2. In order to obtain the more reliable results as possible, we kept only tweets for which the detector returned a confidence level of at least 90%. To take into account “mutually intelligible” languages and phenomena like “dialectal continuum” among dialects, we aggregated languages that show the latter phenomena. In particular, we group Norwegian, Swedish, Danish, Icelandic and Faroese language into one main group called "North Germanic" ($np$); then we build three main groups related to the three branches of the Slavic language in Europe: the "East Slavic Language" ($esl$) group, which includes Belarusian, Russian and Ukrainian; the West Slavic Language ($wsl$) which includes Czech, Slovak and Polish and, finally, the "South Slavic Language" ($ssl$) branch formed by Macedonian, Croatian, Serbian, Serbo-Croatian, Bosnian, Slovenian and Bulgarian. Table S3 shows the complete aggregation process of languages according to the rules presented above.

Assigning a language to every user

We built a framework able to assign the most probable language to each user in each city. To do this, we created a dictionary of the occurrences of each language in each users’ tweets pattern. Some users in the cities tweeted in more than one language during the time periods. English is, most of the times, one of the most frequent language per user, because of its diffusion as "lingua franca" for spreading information to a higher number of Twitter followers. Since we are interested in finding the language representative of users’ hometown origin, we propose a framework in order to identify the user language from his/her dictionary. Let us define as Local the official language of each city. There are cases where the unique local language is not easily identifiable
| City          | Resident Users | City          | Resident Users |
|--------------|----------------|--------------|----------------|
| Amsterdam    | 4986           | Mexico City  | 18079          |
| Atlanta      | 8474           | Miami        | 7754           |
| Bandung      | 27818          | Milan        | 4243           |
| Bangkok      | 25659          | Montreal     | 2613           |
| Barcelona    | 9957           | Moscow       | 9673           |
| Berlin       | 1301           | Nagoya       | 7589           |
| Bogota       | 19353          | NewYork      | 34325          |
| Boston       | 8989           | Osaka        | 16348          |
| Brussels     | 2325           | Paris        | 19757          |
| Buenos Aires | 48934          | Philadelphia | 13679          |
| Caracas      | 4613           | Phoenix      | 9259           |
| Chicago      | 15397          | Rio de Janeiro | 37177        |
| Dallas       | 15549          | Rome         | 1994           |
| Detroit      | 10652          | Saint Petersburg | 3819       |
| Jakarta      | 98997          | San Diego    | 5014           |
| Dublin       | 5480           | San Francisco | 25504         |
| Guadalajara  | 3459           | Santiago     | 10066          |
| Houston      | 11413          | Sao Paulo    | 21862          |
| Istanbul     | 101556         | Seoul        | 3099           |
| Kuala Lumpur | 41084          | Singapore    | 20997          |
| Lima         | 2003           | Stockholm    | 2668           |
| Lisbon       | 4321           | Sydney       | 4751           |
| London       | 37402          | Tokyo        | 75929          |
| Los Angeles  | 70592          | Toronto      | 8737           |
| Madrid       | 15447          | Vancouver    | 2298           |
| Manchester   | 23836          | Washington   | 10147          |
| Manila       | 20093          |              |                |

**Table S2. Total number of residents users detected in the cities.**

due to more than one language that coexists in the same city. This is the case for cities like Barcelona and Brussels, where Spanish and Catalan on one side, and French and Flemish on the other, live, respectively, together in each city; the same occur for Dublin and Singapore; see Table S4 for more details. We assigned to each user its most frequent language. In case of bilingual users, we set as user’s language the one which differs from English and Local except when both of them are in the same dictionary: in the latter case, we defined that user as speaking the Local language. In case of three languages spoken by the same user, we adopted the same hypothesis, assigning to the user the “third” language spoken apart when only one or both between English and Local are in the dictionary. The final number of reliable users per language and per city is displayed Table S5.
| Detected language | Aggregated group | Detected Language | Aggregated Group |
|-------------------|-----------------|-------------------|-----------------|
| Albanian          | Albanian        | Kurdish           | Kurdish         |
| Arabic            | Arabic          | Lettonian         | Baltic          |
| Belarusian        | East Slavic     | Lithuanian        | Baltic          |
| Bosnian           | South Slavic    | Macedonian        | South Slavic    |
| Bulgarian         | South Slavic    | Malay             | Malayo          |
| Catalan           | Catalan         | Norwegian         | Northern European |
| Chinese           | Chinese         | Polish            | West Slavic     |
| Croatian          | South Slavic    | Portuguese        | Portuguese      |
| Czech             | West Slavic     | Romanian          | Roman          |
| Danish            | Northern European | Russian         | East Slavic     |
| Dutch             | Dutch (including Flemish) | Serbian       | South Slavic  |
| English           | English         | Serbo-Croatian    | South Slavic    |
| Faroese           | Northern European | Slovak           | West Slavic     |
| Finnish           | Finnish         | Slovenian         | South Slavic    |
| French            | French          | Southern Sotho    | Southern Sotho  |
| German            | German          | Spanish           | Spanish         |
| Greek             | Greek           | Swahili           | Swahili         |
| Haitian           | Haitian         | Swedish           | Northern European |
| Hungarian         | Hungarian       | Sundanese         | Sundanese       |
| Icelandic         | Northern European | Tagalog         | Tagalog         |
| Indonesian        | Indonesian      | Thai              | Thai            |
| Irish             | Irish           | Turkish           | Turkish         |
| Italian           | Italian         | Ukrainian         | East Slavic     |
| Japanese          | Japanese        | Vietnamese        | Vietnamese      |
| Javanese          | Javanese        |                   |                 |
| Korean            | Korean          |                   |                 |

Table S3. Language aggregation process. A main Aggregated group has been associated to each language detected in the framework, to overlap “mutually intelligible” issues in the detection.

| City          | Local Culture |
|---------------|---------------|
| Amsterdam     | Dutch         |
| Atlanta       | English       |
| Bandung       | Indonesian    |
| Bangkok       | Thai          |
| Barcelona     | Spanish/Catalan |
| Berlin        | German        |
| Bogota        | Spanish       |
| Boston        | English       |
| Brussels      | French/Flemish |
| Buenos Aires  | Spanish       |
| Caracas       | Spanish       |
| Chicago       | English       |
| Dallas        | English       |
| Detroit       | English       |
| Jakarta       | Indonesian    |
| Dublin        | English/Irish |
| Guadalajara   | Spanish       |
| Houston       | English       |
| Istanbul      | Turkish       |
| Kuala Lumpur  | Malay         |
| Lima          | Spanish       |
| Lisbon        | Portuguese    |
| London        | English       |
| Los Angeles   | English       |
| Madrid        | Spanish       |
| Manchester    | English       |
| Manila        | Tagalog       |
| Mexico City   | Spanish       |
| Miami         | English       |
| Milan         | Italian       |
| Montreal      | French/English|
| Moscow        | East-Slavic   |
| Nagoya        | Japanese      |
| New York      | English       |
| Osaka         | Japanese      |
| Paris         | French        |
| Philadelphia  | English       |
| Phoenix       | English       |
| Rio de Janeiro| Portuguese    |
| Rome          | Italian       |
| Saint Petersburg| East-Slavic  |
| San Diego     | English       |
| San Francisco | English       |
| Santiago      | Spanish       |
| Sao Paulo     | Portuguese    |
| Seoul         | Korean        |
| Singapore     | Malay/Chinese/English/Tamil |
| Stockholm     | Northern-European |
| Sydney        | English       |
| Tokyo         | Japanese      |
| Toronto       | English       |
| Vancouver     | English       |
| Washington    | English       |

Table S4. Cities and local languages. Each city has been associated to its main local language; Barcelona, Brussels, Dublin, Montreal and Singapore have been related to more than one language due to the coexistence of multiple languages in the same urban area.
| City                  | Number of resident users per language and per city. |
|----------------------|-----------------------------------------------------|
| Amsterdam            | 4382 217 34 51 15 31                             |
| Atlanta              | 38 714 146 144 144 305                             |
| Bandung              | 152 2184 116 312 2464 2661                      |
| Bangkok              | 55 490 40 47 554 398                             |
| Barcelona            | 1516 153 72 57 7955                               |
| Berlin               | 49 119 98 179 89 37                               |
| Bogota               | 45 1906 271 31 1024                               |
| Brussels             | 1039 30 7572 156 112 173 91 218 459 39              |
| Brussel              | 1445 81 632 3021 3027                             |
| Buenos Aires         | 45 301 357 226 3267                               |
| Caracas              | 95 178 4331                                        |
| Chicago              | 55 12877 255 197 206 182 1539 84 36 31            |
| Dallas               | 32 19994 222 218 1188 74 80                       |
| Davao                | 35 8664 73 107 367 47                             |
| Dublin               | 17 4232 142 58 52 57 44 180 760 39 31             |
| Dusseldorf           | 3427                                              |
| Espoo                | 44 8158 183 1200 57                               |
| Istanbul             | 207 33 90 152 452 110 99462                       |
| Jakarta              | 1932 6973 99 238 52 752 113                        |
| Kaohsiung            | 55 1208 1134 12751 48 27970 62 71                  |
| Lima                 | 1935                                              |
| London               | 83 596 138 99 29072 123 618 357 46 62 814 532 04   |
| Los Angeles          | 955 44 102 90537 1194 269 98 323 885 6451 650 39    |
| Madrid               | 152 146 146 1065 55 23079 36                      |
| Manchester           | 108 66 21709 522 239 243 287 94 4176 284 46 202 74 |
| Manila               | 1044 53 37 52 1620                               |
| Mexico City          | 90 93 146 1764                                    |
| Milan                | 4588 110 81 57 74 249 155 1911 37                 |
| Montreal             | 92 96 5786                                        |
| Moscow               | 596 139 200 3511 39 93                            |
| Nagoya               | 48 44 7203 37 75 41                               |
| New York             | 101 119 62 20702 723 400 624 305 79 146 180 3491 220 |
| Nice                 | 112 95 1882 138 44                                 |
| Paris                | 149 37 290 8255 39 127 36 37 213 41 363 87        |
| Philadelphia         | 74 11910 217 272 72 207 112 217 4179 73 33        |
| Phoenix              | 8264 92 1177                                       |
| Rio de Janeiro       | 73 276 10950 902                                 |
| Rome                 | 124 32 1088 105 54                                |
| Saint Petersburg     | 2180 46 1145                                       |
| San Diego            | 142 5933 59 46 46 67 600 44                       |
| San Francisco        | 72 49 94 45 20961 614 598 297 149 50 89 151 367 1811 |
| Santiago             | 41 35 368 3571                                    |
| Sao Paulo            | 94 91 20970 606                                   |
| Seoul                | 100 47 145 2428 123                               |
| Singapore            | 1043 3014 70 82 5932 149 100 68 424 104 1956 2336 |
| Stockholm            | 157 3820 102 46 61 87 69 33 90 103 99 39          |
| Sydney               | 52 580 140 147 48 157 50 2895 429 59 90 120 196 193 |
| Tokyo                | 82 35 93 661 89 149 94 178 54 363 238 86 51       |
| Vancouver            | 1958 166 31 57 78                                |
| Washington           | 129 6320 200 154 676 88                           |

Table S5. Number of resident users per language and per city.
DATA VALIDATION

Population data by country of origin was extracted from the Continuous Register Statistics of the Municipal Register, which is published by the National Institute of Statistics of Spain on an annual basis, regarding the cities of Madrid and Barcelona. The smallest spatial units for this dataset are census tracks, of which the latest available geometrical boundaries for both study areas are the corresponding to 2013. It is well known that census tracks cover all the territory (not only populated areas) and that their size depends on the population density of an area, i.e. the more population density, the smallest the size and vice versa, in order to ensure that all census tracks have a similar number of inhabitants. This means that low density census tracks are larger than those corresponding to the city center, thus integrating non populated territory. For this reason, complementary data about the exact location of the residential areas is needed in order to properly geo-reference population data from census track statistics. In this research, information was extracted from the "Downloads of data and cartography by town” service of the SEC, the point of access to electronic services provided by the Directorate General of Land Registry of Spain. This data was transformed in order to obtain the surface devoted to each land use in each urban parcel. Some data treatment was required in order to obtain the number of people residing in each 500 x 500 m\(^2\) grid cell according to the main language spoken in the country of origin. Table S6 shows the correspondence between country of origin and the languages detected in the main part of this paper. It is important to notice that not all the countries in the world are present in the original table.

| Language   | Country of Origin |
|------------|-------------------|
| German     | Germany           |
| South Slavic | Bulgaria         |
| French     | France            |
| Italian    | Italy             |
| West Slavic | Poland           |
| Portuguese | Portugal, Brazil |
| English    | United Kingdom    |
| Romanian   | Romania           |
| East Slavic | Russia, Ukraine  |
| Arabic     | Morocco, Algeria  |
| Spanish    | Spain, Argentina, Bolivia |
|            | Colombia, Cuba, Chile, Ecuador |
|            | Paraguay, Peru, Dominican Republic |
|            | Uruguay, Venezuela |
| Chinese    | China             |
| Urdu       | Pakistan          |

Table S6. Correspondence between languages detected in Twitter users and country of origin.

The second step in data treatment was to locate where people in each census track actually live according with the location of residential land. We selected the blocks containing some surface devoted to residential use from the cadastral dataset. With the use of a Geographic Information System (GIS)[35] we were able to intersect these polygons with the census track boundaries, and to assign the population of each census track to its residential land, proportional to the size of each residential polygon within each census track. Finally, the resulting dataset was intersected with the grid used in the previous parts of this research in order to obtain the estimated number of residents of each language in each grid cell.

Anselin Local Morans $I$ [36] is a well-known statistic that provides information on the location and size of four types of clusters: a) high-high clusters of significant high values of a variable that are surrounded by high variables of the same variable; b) high-low clusters of significant high values of a variable surrounded by low values of the same variable; c) low-high clusters of significant low values of a variable surrounded by high values of the same variable; and d) low-low clusters of significant low values of a variable surrounded by low values of the same variable. While the typical tools available in most GIS software solutions allow for univariate analysis, GeoDa is an open source product that also allows the computation of bivariate analysis [37], thus enabling the identification of spatial clusters in which high values of one variable are surrounded by high values of the second (i.e. lagged) variable (high-high clusters) and so on.

Bivariate global Morans I (Table S7) indicates the existence of positive spatial autocorrelation between the location of tweets and residential areas. In general terms, there is a high positive spatial correlation in both study areas (Morans $I = 0.6$). The z and p values have been evaluated through 99 permutations. This value remains high for local language (Spanish in Madrid and Spanish and Catalan in Barcelona). The spatial autocorrelation of foreign languages is a bit lower, which might be in part due to the inconsistencies between Twitter language and available countries of origin in the official statistics (i.e. United Kingdom is the only country of origin for English speakers and so are Morocco and Algeria for Arabic languages). Anyway, Arabic is
| Language | City      | $I$     | Z-value | p-value | Spatial Autocorrelation |
|---------|-----------|---------|---------|---------|-------------------------|
| Total   | Barcelona | 0,63    | 236,51  | 0,01    | Positive               |
|         | Madrid    | 0,62    | 268,62  | 0,01    | Positive               |
| Spanish | Barcelona | 0,62    | 216,99  | 0,01    | Positive               |
|         | Madrid    | 0,62    | 267,29  | 0,01    | Positive               |
| English | Barcelona | 0,50    | 230,53  | 0,01    | Positive               |
|         | Madrid    | 0,38    | 190,62  | 0,01    | Positive               |
| French  | Barcelona | 0,37    | 151,51  | 0,01    | Positive               |
|         | Madrid    | 0,32    | 159,25  | 0,01    | Positive               |
| Italian | Barcelona | 0,28    | 125,84  | 0,01    | Positive               |
|         | Madrid    | 0,26    | 146,32  | 0,01    | Positive               |
| Portuguese | Barcelona | 0,32 | 151,20  | 0,01    | Positive               |
|          | Madrid    | 0,44    | 204,95  | 0,01    | Positive               |
| Arabic  | Barcelona | 0,08    | 89,88   | 0,01    | Random                 |
|         | Madrid    | 0,07    | 41,50   | 0,01    | Random                 |
| East-Slavic | Barcelona | 0,21 | 112,83  | 0,01    | Positive               |
|          | Madrid    | 0,06    | 37,66   | 0,01    | Random                 |

**Table S7. Data Validation.** Global Moran’s $I$

The only language whose tweets show a random spatial pattern in relation with the location of resident population from Morocco or Algeria in both cities, whereas tweets in English in Barcelona and tweets in Portuguese in Madrid are highly positively spatially correlated with resident population from the UK and Portugal or Brazil, respectively. Figures S2 and S3 show the visual comparison of the spatial distributions of the Spanish and Portuguese speaking communities in Madrid, through statistical data and through our framework, respectively.

**Figure S2. Data Validation (1/2).** Distribution of Spanish native users in Madrid, according to official statistics and to our framework of language detection process.
Figure S3. Data Validation (2/2). Distribution of Portuguese native users in Madrid, according to official statistics and to our framework of language detection process.
BIPARTITE SPATIAL INTEGRATION NETWORK

In order to quantify the spatial segregation of each immigrant community in each city, we built a Bipartite Spatial Integration Network, $H$, where every language $l$, is connected to each city $c$, where the corresponding immigrant communities have been detected.

Weight of the links

We introduce a user-independent metric based on the concept of Shannon entropy, later extended to spatial analysis. The original Shannon entropy was proposed in information theory, related to the average expected value of the information gained over a series of $n$ events; the well-known formula can be written as:

$$H = -\sum_i p_i \log(p_i)$$  \hspace{1cm} (1)

where $p_i$ is the probability of the occurrence of a single event over a series of $n$. In our case, we consider a spatial regular grid of $N$ cells; Let $n_l^i$ be the number of users speaking the language $l$ in the cell $i$ of the grid. The probability to observe a user speaking language $l$ living in cell $i$ can be expressed as:

$$p_l^i = \frac{n_l^i}{\sum_{k=1}^N n_k^i},$$  \hspace{1cm} (2)

for each city. We take as null model a situation where the $l$ users detected in a language are drawn at random $R$ times over the city cells according to the total distribution of users (population) in each cell, and evaluate the corresponding average entropy as:

$$\bar{h}_{l,c}^{ran} = -\frac{1}{R} \sum_{r=1}^R \sum_{i=1}^N p_{l,i}^r \log(p_{l,i}^r)$$  \hspace{1cm} (3)

where $p_{l,i}^r$ corresponds to the fraction of users associated to language $l$ found in cell $i$ at realization $r$. Our final metric is $h_{l,c}$ evaluated for each language detected in each city as:

$$h_{l,c} = -\frac{\sum_{i=1}^N p_{l,i} \log(p_{l,i})}{\bar{h}_{l,c}^{ran}}$$  \hspace{1cm} (4)

we evaluate also the entropy normalized by the entropy of the total number of users, defined as:

$$h_{tot}^l = \frac{\sum_{i=1}^N p_{l,i} \log(p_{l,i})}{\sum_i N p_i \log(p_i)}$$  \hspace{1cm} (5)

Figure S4. Entropy metrics as a function of the number of users. Each point of the scatterplot in (a) and (b) is related to each single community in all cities, together with the corresponding value of $h^l$ and $h_{tot}^l$, respectively.

in this case, $p_i$ is the probability of occurrence of each single user in a cell over the total number of users in the city. Since we were looking for a user-independent metric, for verification purposes we plot the trend of
both metrics, for each language and for each city in function of the number of users associated to each language. Figure S4a shows that after 30 users, there is no more clear dependency between \( h_l \) and the number of users. On the other hand, as it can be observed in Fig. S4b, the \( h_{tot} \) metric might lead to a wrong estimation of the distribution of users in the cities since it strongly depends on the number of users.

| Cluster | City         | Q1       | Q2       | Q3       | IQR        | \( P_c \) |
|---------|--------------|----------|----------|----------|------------|----------|
| C1      | London       | 0.81     | 0.91     | 0.95     | 0.13       | 0.789    |
| C1      | Manchester   | 0.91     | 0.95     | 0.96     | 0.06       | 0.543    |
| C1      | Los Angeles  | 0.87     | 0.93     | 0.96     | 0.09       | 0.518    |
| C1      | San Francisco| 0.77     | 0.83     | 0.92     | 0.15       | 0.522    |
| C1      | Tokyo        | 0.71     | 0.80     | 0.87     | 0.16       | 0.413    |
| C2      | Philadelphia | 0.88     | 0.90     | 0.92     | 0.04       | 0.375    |
| C2      | Paris        | 0.76     | 0.81     | 0.90     | 0.14       | 0.336    |
| C2      | Singapore    | 0.81     | 0.86     | 0.95     | 0.15       | 0.319    |
| C2      | New York     | 0.31     | 0.64     | 0.85     | 0.54       | 0.180    |
| C2      | Kuala Lumpur | 0.83     | 0.87     | 0.90     | 0.07       | 0.246    |
| C2      | San Diego    | 0.82     | 0.88     | 0.93     | 0.12       | 0.236    |
| C2      | Boston       | 0.65     | 0.80     | 0.88     | 0.23       | 0.241    |
| C2      | Chicago      | 0.53     | 0.82     | 0.84     | 0.31       | 0.247    |
| C2      | Dublin       | 0.57     | 0.79     | 0.87     | 0.29       | 0.220    |
| C2      | Sydney       | 0.27     | 0.65     | 0.75     | 0.48       | 0.161    |
| C2      | Washington   | 0.72     | 0.82     | 0.84     | 0.13       | 0.217    |
| C2      | Madrid       | 0.61     | 0.91     | 0.94     | 0.33       | 0.159    |
| C2      | Nagoya       | 0.76     | 0.86     | 0.97     | 0.22       | 0.146    |
| C2      | Bangkok      | 0.40     | 0.77     | 0.84     | 0.43       | 0.133    |
| C2      | Berlin       | 0.44     | 0.77     | 0.90     | 0.46       | 0.108    |
| C2      | Jakarta      | 0.42     | 0.63     | 0.82     | 0.40       | 0.099    |
| C3      | Amsterdam    | 0.30     | 0.52     | 0.74     | 0.44       | 0.063    |
| C3      | Bandung      | 0.36     | 0.66     | 0.77     | 0.41       | 0.068    |
| C3      | Barcelona    | 0.36     | 0.60     | 0.80     | 0.44       | 0.044    |
| C3      | Bogota       | 0.25     | 0.50     | 0.75     | 0.50       | 0.011    |
| C3      | Brussels     | 0.12     | 0.24     | 0.62     | 0.50       | 0.011    |

Table S8. Power of Integration of Cities.

Normalizing the weights

In order to make the metric comparable across cities, we decided to normalize each edge weight \( h_{l,c} \) of the bipartite network by the value obtained for the local languages spoken in city \( c \) (Table S4). The value of the weight \( h_{l,c} \) for every couple (city, language) is available in the matrix shown in Fig. S5. Degree and strength distributions of \( H \) are plotted in Fig. S6.

Influence of the city definition on the Bipartite Spatial Integration Network

Figures S8 and S9 show the Bipartite Spatial Integration matrices for each value of the frame side. Since we set a threshold of a minimum of 30 users to compare the spatial distribution of communities within the urban area, matrices column numbers may vary since the larger the city definition, the higher number of users we detect. We select the 60×60 km frame grid in order to minimize the error on each city, knowing that considering larger grid sizes may lead to detect a different number of communities; however, since our metric is user-independent, the variation in the number of users on communities detected on different grids do not influence the results in terms of spatial integration.
Figure S5. Bipartite Spatial Integration Network’s edge weight. The color of the cells is a measure of the distribution of the values of $h_{i,e}$. 
Figure S6. Distribution of degree and weights in H. Languages are connected to the cities where the corresponding community of immigrants has been found. The distribution of the weights (a) is heterogeneous, but the level of Integration of the languages in cities is long-tailed for cities and for languages (b). In (c) and (d), strength distributions of the nodes in each projection is a qualitative visual measure of how well and how many languages each city is able to handle within its boundaries in function of their presence and integration in the cities.

Figure S7. Boxplots of the relative error for different side sizes of the city definition. We evaluate the norm for each row (i.e. city) of the matrix (||v_c||), and then plot the value of the (||v_c||−||v_c^ref||)/||v_c^ref|| ratio for each city over each frame size, where ||v_c^ref|| is the norm of each vector for the reference frame size. We compare the norm of each row (i.e. vector of the city) of the Bipartite Spatial Integration Network for each side value, and plot the corresponding relative error, taking as reference value the frame of 60×60 km.
Figure S8. Bipartite Spatial Integration Network (1/2). Matrices in function of different grid frame size. Here we compare the 20 x 20 km (a) and the 40 x 40 km (b). Matrices have the same size (number of columns) that corresponds to the same amount of communities detected.
Figure S9. Bipartite Spatial Integration Network (2/2). Matrices in function of different grid frame size. Here we compare the 80 × 80 km (a) and the 100 × 100 km (b). The Matrix related to the 100 × 100 km size comprehends a different number of communities (languages) detected because of the larger area that allows to detect more users which overrun our original threshold set for the 60 × 60 km size.
CLUSTERING ANALYSIS BASED ON THE DISTRIBUTION OF EDGE WEIGHTS

We chose the number of clusters with the Silhouette index and the evolution of the percentage of variance explained. We obtained three clusters of cities (C1, C2 and C3) exhibiting common characteristics. Figure S10 shows a comparison of the cities belonging to each cluster detected for each method, along with the results of the Elbow and Silhouette methods.

![Figure S10. Comparison of a k-means and hierarchical clustering algorithms over the vectors of the Bipartite Spatial Integration Network. C1, C2 and C3 are the clusters obtained through both algorithms, choosing as 3 the initial number of clusters to assign to the k-means analysis.](image)

LANGUAGE INTEGRATION NETWORK

In order to discover more properties of the bipartite network $G = (L, C)$ of Languages ($L$) and cities ($C$) over the language side, we projected the network over the set of the languages alone assigning, to each city, its own main Local language as a measure of the Country in which other languages had to integrate with (Table S9). This means that languages in the same city are connected to the own Country, with they respective levels of integration (entropy). We remove from the network minority languages (e.g. Irish and Catalan), in order to get an insight of the integration of main communities only. We analyze the network with and without English (related to the corresponding community of immigrants). Looking at the distributions of values (Fig. S11), the role of the latter language is dominant only in the worst links in terms of integration. Since we normalized the values in the range [0,1], English points, most of the times, to countries with levels of entropy equal to 0 (minimum value). For the above reasons, and since it has the role of "lingua franca" in Twitter, we do not consider it as a trace of English-speaking immigrant communities in the integration network.

Since the same language might be connected with more than one city in the the same country, the resulting network $M$ of Integration of Cultures has been defined as a Multi Directed Graph among languages. We then reduced it into a single Directed Graph $N$ having as new weights the mean value of the common entropies among languages. Self-loops have been removed from the graph. Similarly to what already defined for the cities level, we have defined the Language Integration Network through its weighted Language Integration Matrix.
Figure S11. Distribution of degree and weights in H with and without English. Distribution of weights for the full network (a) and in the network obtained removing English from the nodes of the Bipartite Spatial Integration Network (b). As shown, English is dominant in the worst links in terms of spatial integration.

\[ L = (h_{i,j})_{i,j=1,...,L} \] where \( h_{i,j} \) is the measure of integration (entropy) of the \( i \) language into the \( j \) country and \( L \) is the total number of languages living in the cities. We perform several clustering analysis over the networks, running both OSLOM [38] and Infomap [39] algorithms, which are able to identify overlapping communities of nodes taking into account the direction of the edge and the weight. Due to the relative small size of the network, we were not able to discover any particular structure in them.

| City          | Local Culture | City          | Local Culture |
|---------------|---------------|---------------|---------------|
| Amsterdam     | Netherlands   | Mexico City   | Mexico        |
| Atlanta       | USA           | Miami         | USA           |
| Bandung       | Indonesia     | Milan         | Italy         |
| Bangkok       | Thailand      | Montreal      | Canada        |
| Barcelona     | Spain         | Moscow        | Russia        |
| Berlin        | Germany       | Nagoya        | Japan         |
| Bogota        | Colombia      | New York      | USA           |
| Boston        | USA           | Osaka         | Japan         |
| Brussels      | Belgium       | Paris         | France        |
| Buenos Aires  | Argentina     | Philadelphia  | USA           |
| Caracas       | Venezuela     | Phoenix       | USA           |
| Chicago       | USA           | Rio de Janeiro| Brazil        |
| Dallas        | USA           | Rome          | Italy         |
| Detroit       | USA           | Saint Petersburg| Russia   |
| Jakarta       | Indonesia     | San Diego     | USA           |
| Dublin        | Ireland       | San Francisco | USA           |
| Guadalajara   | Mexico        | Santiago      | Chile         |
| Houston       | USA           | Sao Paulo     | Brazil        |
| Istanbul      | Turkey        | Seoul         | Korea         |
| Kuala Lumpur  | Malaysia      | Singapore     | Singapore     |
| Lima          | Peru          | Stockholm     | Sweden        |
| Lisbon        | Portugal      | Sydney        | Australia     |
| London        | UK            | Tokyo         | Japan         |
| Los Angeles   | USA           | Toronto       | Canada        |
| Madrid        | Spain         | Vancouver     | Canada        |
| Manchester    | UK            | Washington    | USA           |
| Manila        | Philippines   |               |               |

Table S9. City/Country Correspondence.