PLOS ONE

RESEARCH ARTICLE

Immigrant community integration in world cities

Fabio Lamanna[1], Maxime Lenormand[2], María Henar Salas-Olmedo[3], Gustavo Romanillos[3], Bruno Gonçalves[4], José J. Ramasco[1]*

1 Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), Campus UIB, 07122 Palma de Mallorca, Spain, **2** Irstea, UMR TETIS, 500 rue JF Breton, 34093 Montpellier, France, **3** Departamento de Geografía Humana, Facultad de Geografía e Historia, Universidad Complutense de Madrid, 28040, Madrid, Spain, **4** Center for Data Science, New York University, New York, 10011 NY, United States of America

* jramasco@ifisc.uib-csic.es

Citation: Lamanna F, Lenormand M, Salas-Olmedo MH, Romanillos G, Gonçalves B, Ramasco JJ (2018) Immigrant community integration in world cities. PLoS ONE 13(3): e0191612. https://doi.org/10.1371/journal.pone.0191612

Editor: Renaud Lambiotte, University of Oxford, UNITED KINGDOM

Received: September 20, 2017

Accepted: January 8, 2018

Published: March 14, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. Interested researchers can reproduce the data underlying this study by consulting the Supporting Information file that contains a sample of the python code used to query Twitter's API.

Funding: 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 (611307). The work of M-HS-O was supported in

Abstract

As a consequence of the accelerated globalization process, today major cities all over the world are characterized by an increasing multiculturalism. The integration of immigrant communities may be affected by social polarization and spatial segregation. How are these dynamics evolving over time? To what extent the different policies launched to tackle these problems are working? These are critical questions traditionally addressed by studies based on surveys and census data. Such sources are safe to avoid spurious biases, but the data collection becomes an intensive and rather expensive work. Here, we conduct a comprehensive study on immigrant integration in 53 world cities by introducing an innovative approach: an analysis of the spatio-temporal communication patterns of immigrant and local communities based on language detection in Twitter and on novel metrics of spatial integration. We quantify the *Power of Integration* of cities –their capacity to spatially integrate diverse cultures– and characterize the relations between different cultures when acting as hosts or immigrants.

Introduction

Immigrant integration is a complex process involving a multitude of aspects such as religion, language, education, employment, accommodation, legal recognition and many others. Its study counts with a long tradition in sociology through concepts such as immigrant assimilation [1], structural assimilation [2] or immigrant acculturation and adaptation [3]. Over the last years, there have been advances in the definition of a common framework concerning immigration studies and policies [4], although the approach to this issue remains strongly country-based [5]. The outcome of the process actually depends on the culture of origin, the one of integration and the policies of the hosting country government [6]. Traditionally, spatial segregation in the residential patterns of a certain community has been taken as an indication of ghettoization or lack of integration [7]. While this applies to immigrant communities, it can also affect to minorities within a single country [8]. The spatial isolation reflects in the economic status of the segregated community and in social relationships of its members [9].

part by a post-doctoral fellowship of MINECO at Universidad Complutense de Madrid (FPDI 2013/17001). BG thanks the Moore and Sloan Foundations for support as part of the Moore-Sloan Data Science Environment at New York University.

Competing interests: The authors have declared that no competing interests exist.

In global terms while international migration flows have remained almost stable over the last 20 years [10, 11], political and economic upheavals 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 and even the academic press [12, 13]. A good part of newcomers concentrates in cities, and particularly in the large metropolises known as World Cities. These are centers that attract specialized immigration, driving important social and cultural transformations in cities worldwide [14]. The concept of Global or World Cities emerged in the 80s [15, 16] as strategic territories that articulate the international economic structure. According to Sassen [15], 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 in the area of social sciences, requiring its settlement further empirical evidence [17, 18]. Furthermore, immigrant integration has been the focus of many research studies, most of which conducted from national perspectives especially in European countries and the USA [5, 6, 8, 19, 20], and it is still in dare need of information sources beyond national census [12, 13, 21].

In parallel, the last few years have brought a paradigm shift in the context of socio-technical data. Human interactions are being digitally traced, recorded and analyzed in large scale. Sources as varied as mobile phone records [22–34], credit card transactions [35], or Twitter data [36–38] have been used to study mobility and land use in urban areas. Most of these works have been carried out in the zones where data was available, mostly inside cities or single countries. Twitter data has, however, the particularity of extending beyond national borders and, therefore, it allows researchers to analyze mobility and city hierarchies at an international level [36, 39]. Besides activity and mobility, the content of the tweets bears also a wealth of information starting by the language in which the text is written. The spatial distribution of languages has been investigated in Refs. [40–42], exploring as well the relations between languages trough multilingual individuals, and in Refs. [43, 44], where the spatial extension of Spanish and English dialects was examined. Of course, one of the weak points of Twitter as data source is its representativeness. This question has been boarded in Refs. [37, 45–47], finding acceptable coverage for the American, British and Spanish populations in terms of geographic allocation, race, religion and mobility, although the data shows a bias towards younger individuals. In this context, it is of special interest the mix of location and language detection. This issue opens the door to characterize foreign users in short visits, temporal or permanent stays. Arribas-Bel [48] published a first exploratory work on this direction using Twitter and census data in Amsterdam. Contemporarily, the use of phone call records to foreign countries has provided a picture of communities with external connections in the area of Milan [49]. When it comes to immigrant integration, there are less works but one that deserves mention is a study recently published by [50] who looked at the social ties (friendships and affinities) between immigrant communities by using data from Facebook. There have been diverse attempts to measure the degree of immigrant integration over the last years [51] by introducing a quantitative index, the Composite Assimilation Index (CAI), that quantifies the degree of similarity between native- and foreign-born adults in the United States, based on US census data. In [50], a similar measure of integration is considered based on the relative proportion of ties between immigrant people born in the US, compatriots living in the US, and inter-group friendships with immigrants from other countries.

In this work, we introduce a novel approach to quantify the spatial integration of immigrant communities in urban areas worldwide. By analyzing language in Twitter data, we are able to assign languages to each user paying special attention to those corresponding to migrant communities in the city considered. The individuals' digital spatio-temporal communication patterns allow us to define as well areas of residence. With this information, we perform a spatial

distribution analysis through a modified entropy metric, as a quantitative way to measure the spatial integration of each community. The metric can be expressed in a bipartite network with the culture of origin in one side and the hosting cities, countries and languages in the other. These results lead us to categorize the cities according to how well they integrate immigrant communities and also to quantify how well hosting countries integrate people from other cultures.

Materials and methods

We selected 53 of the most populated cities in the world (see Fig 1a) and analyzed the geolocalized tweets originating in each city between October 2010 and December 2015 as captured from the Twitter API (see S2 File of the Supporting Information for an example of a query). The data was collected respecting Twitter's terms of service and privacy conditions. Several items are extracted from each tweet: user ID, geographical coordinates (latitude and longitude), date and time and the text of the tweet. In order to get a coherent picture in the different

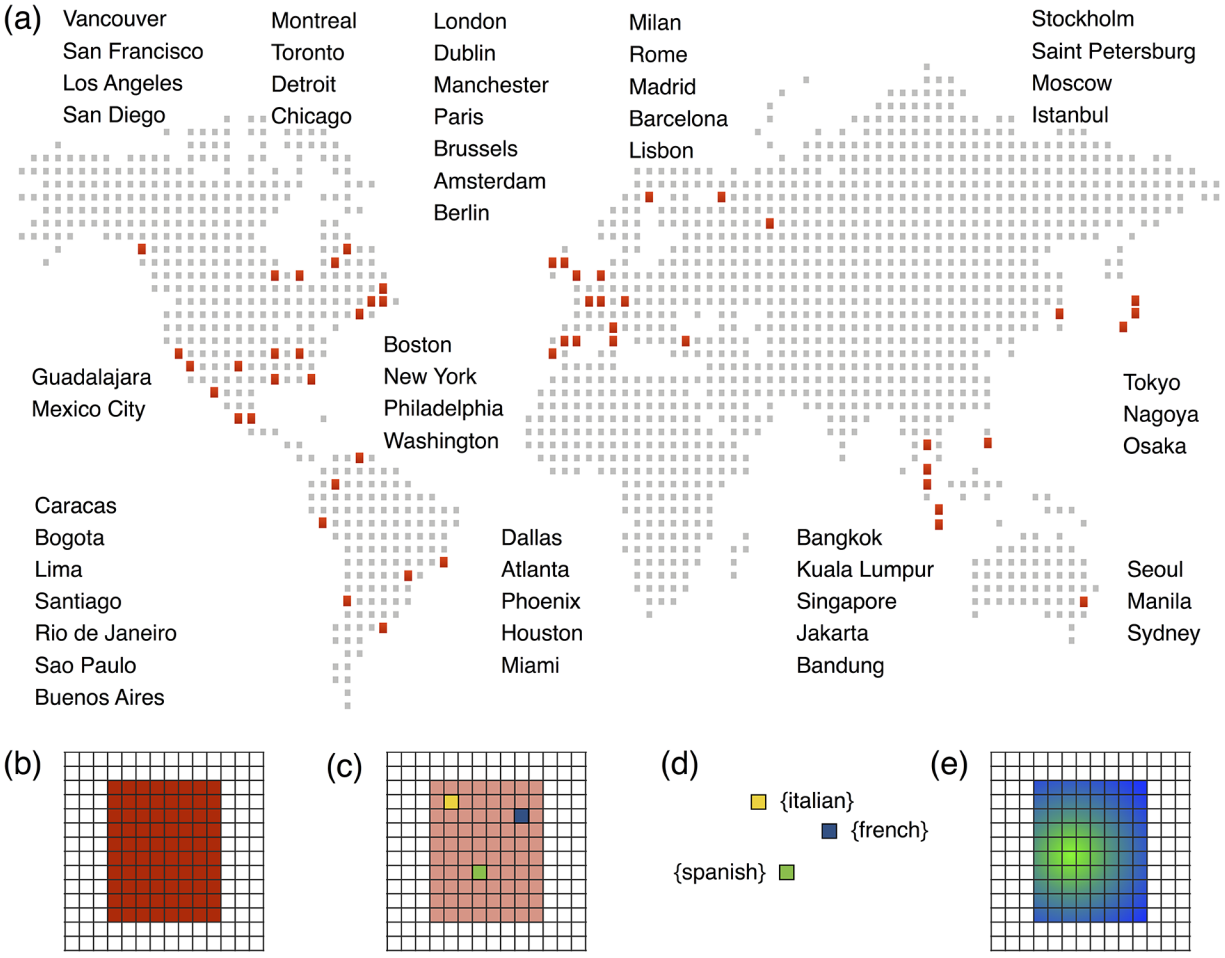

Fig 1. Dataset and framework description. The cities passed through the lens of our analysis are mostly distributed over four continents (a). Africa has been not considered due to the lack of data. We cover each city with a square grid in order to keep a homogeneous spatial division over the whole urban area where the users are going to be distributed (b), selecting resident users and their most frequent location thanks to their activity over space and time (c). In addition, we assign the users' most probable native language (d) and perform a spatial analysis over the cities (e) to get information about the population distribution in function of the language spoken by the users.

https://doi.org/10.1371/journal.pone.0191612.g001

time zones, we convert the Twitter UTC time into the local timezone for each city. Before starting with the analysis, it is necessary to filter out non-human users from the dataset. This is fundamental in order to prevent result pollution by signals coming from automatic tweet generators (bots), which are not rare in social networks [52]. We found and disregarded 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 traveling faster than a reasonable speed in urban areas ($100km/h$ or $62mph$). This procedure leaves us with a total of 350.9 millions of tweets posted by 14.5 millions of users in the 53 cities (see Table A in the S1 File of the Supporting Information (SI) for detailed numbers per city).

We will propose below a metric to assess spatial segregation of immigrant communities that is not highly sensitive to the specific borders of the area studied. However, everything has its limits. The mix of local and immigrant population is different in urban and rural areas. It is important thus to attain a balance and ensure that the region considered contains the city, where the signal on immigrants is stronger, but it does not extend unnecessarily far from it. This means that we should agree on a city definition that can be applied around the world and it is large enough to include the whole metropolitan area. Unfortunately, generic definitions such as the Larger Urban Zone (LUZ) definition of Eurostat for Europe does not exist at the global 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 60×60 km^2 centered on the barycenters listed in Table B of the S1 File of the SI to belong to the city itself, dividing each city area using an equally spaced grid of 500×500 meters (Fig 1b).

Definition of the user's place of residence

As represented in Fig 1c, the place of residence of every user is defined as the most frequented grid cell between 8pm and 8am local time. 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: a minimum number of consecutive months of activity C, a minimum number of hours spent by the user in the most frequented cell N measured out of his/her consecutive tweets, and Δ as the ratio between N and the total number of hours of activity for each user (number of hours during which he/she has posted at least one tweet). The source code used to extract most visited locations from individual spatio-temporal trajectories is available online (https://github.com/maximelenormand/Most-frequented-locations).

Users who are active within a given city for at least three consecutive months are considered to be residents, so this establishes the first condition $C \geq 3$ months. The values of the other two parameters were determined empirically. In Fig A of S1 File(Supporting Information), we plot the evolution of the number of users left in the dataset as a function of Δ for different values of $N = [5, 10, 15, 20]$ in each of the 53 cities. As the shape of the curves is similar for different values of N, it does not seem to be a natural features that would allow us to define a clear cutoff. We fix $\Delta \geq 0.2$ and $N \geq 5$, as a trade-off between being relatively sure about the users' residence area and keeping enough number of users to have proper statistics. Table C in the S1 File of the Supporting Information lists the final number of residents per city after this data cleaning procedure. Note that there are at least 1000 reliable users per city.

Language assignment

At this point, we are interested in introducing a method to determine which languages each user speaks, or at least in which languages he/she tweets. If any of these languages is proper of an immigrant community, this most likely will identify the user as a member of that community. To do this, the language in each tweet is detected using the version 2.0 of the *Chromium Compact Language Detector* (CLD2), which returns the languages detected along with a confidence assessment. CLD2 implements a Bayesian classifier for detecting language from UTF-8 text. Twitter entities (urls, mentions, hashtags) that may difficult our language detection efforts are removed, and only the remaining text was given as input to CLD2. To obtain reliable results, we keep only tweets for which the detector returned a language with confidence level of at least 90%. Also, we aggregate close languages to take into account the uncertainty in the identification of *mutually intelligible* languages and dialectal varieties (see Table D in S1 File of the Supporting Information for more details).

As can be expected, there are users tweeting in more than one language. We create a dictionary of the occurrences of each language in each users' tweets pattern. English is one of the most frequent language per user, because of its diffusion as *lingua franca* for spreading information to the highest number of Twitter followers. Still since we are interested in finding the language representative of users' community of origin, we propose a language algebra in order to extract this information from the user's dictionary. Let us define as *Local* the official language of each city. There are cases where there can be more than one Local language coexisting in the same city, like Catalan and Spanish in Barcelona, French and Flemish in Brussels or French and English in Montreal. The same occurs for Dublin and Singapore (see Table E in the S1 File of the SI for a complete list of cities and languages). After defining the Local languages in each city, we assign to each user its most frequent language. In case of bilingual/multilingual users, we set as user's language the one which differs from English or the Local unless these are the only two languages in the dictionary. In this latter case, we define the user as speaker of 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. In general, we take the most popular language in the dictionary other than English and the Local ones. If there are only Local languages and English, we keep the Local. English can be only assigned if it is the only one in the dictionary. The final number of users left for the analysis with a reliable residence cell, per language identified and per city are displayed in Table F of the S1 File of the Supporting Information. We consider languages in each city with **30** users or more.

Results

Bipartite spatial integration network

To quantify the spatial segregation of each immigrant community in every city, we build a bipartite spatial integration network H (see Fig 2). Every language is connected to the cities where the corresponding immigrant communities has been detected. The weight of an edge between language l and city c, $h_{l,c}$, corresponds to the level of spatial integration measured with a new metric inspired by the Shannon entropy, but modified to take into account the finite character of the sampling of communities in our Twitter database. Shannon entropy-like descriptors have been used before in this context especially when considering the spatial segregation of ethnic minorities in the US cities [53]. Recalling that the cities have been divided in equal area grid cells and focusing first only on one generic city c, we can directly calculate from the data the fraction of users of a certain community l having their residence at cell i, $p_{l,i}$. This

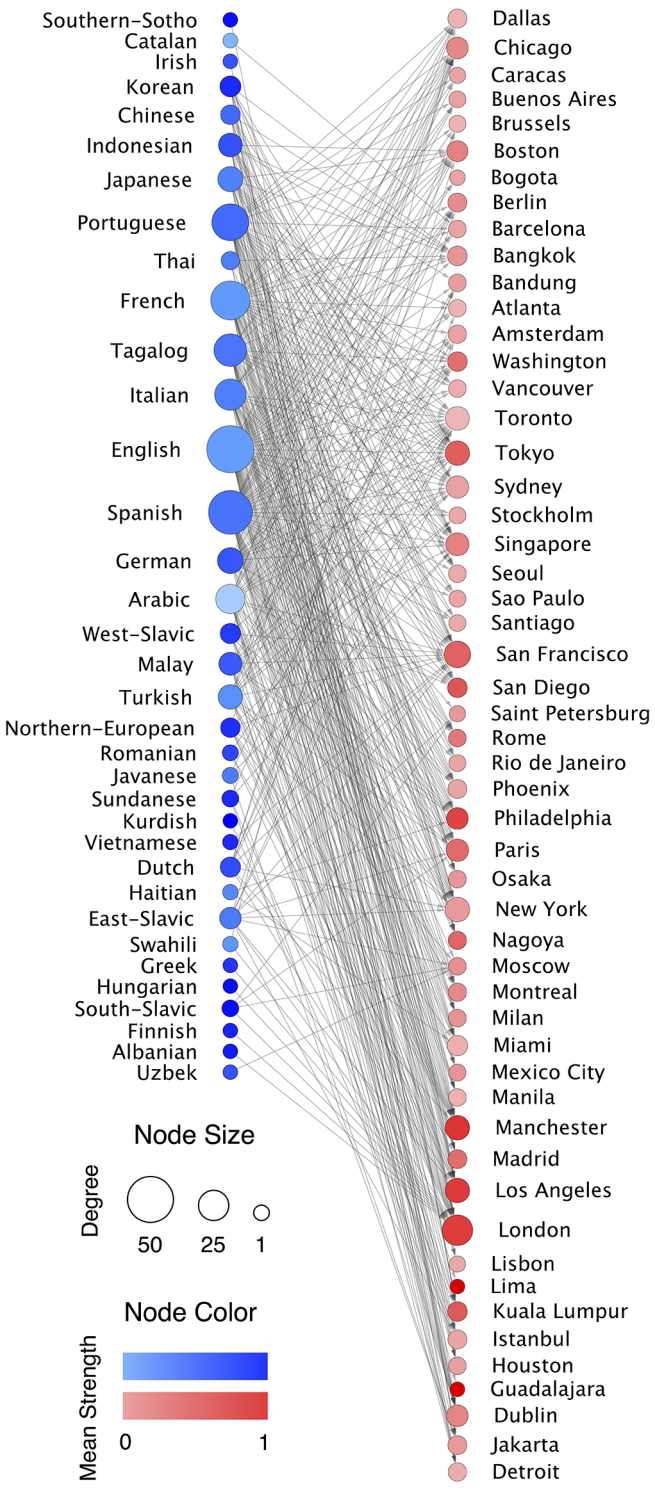

Fig 2. Bipartite spatial integration network. The network comprises of two sets: **L** of Languages and **C** of cities; the languages detected are connected to the cities set where the corresponding community of immigrants has been found. The weight of the edge corresponds to the values of $h_{l,c}$. The size of the nodes is proportional to its degree and the color to its mean strength.

https://doi.org/10.1371/journal.pone.0191612.g002

allows us to define an entropy per language community l:

$$s_{l,c} = -\sum_{i=1}^{N} p_{l,i} \log(p_{l,i}/\Delta x^2), \quad (1)$$

where N is the total number of cells and the index i runs over all the cells. Δx^2 is the area of the cells, it is added to make the entropy stable against changes of spatial scale as proposed in Ref. [54]. We take as unit the area our 500×500 m^2 cells and, thus, a change in cell size as those shown in the Supporting Information for 1×1 and 2×2 square kilometers requires a correction factor 4 and 16, respectively, as expressed in Eq (1). The distribution of the population is generally heterogeneous, so $s_{l,c}$ by itself is not telling us anything about characteristic features of the community l. To overcome this and also to take into account the finite sampling size, we introduce next a random null model. The $n_{l,c}$ users associated to language l in city c are drawn at random over the city cells according to the total distribution of users to obtain new fractions $p^r_{l,i}$ for language l in each cell i, and then we evaluate the following entropy:

$$s^{rand}_{l,c} = -\sum_{i=1}^{N} p^r_{l,i} \log(p^r_{l,i}/\Delta x^2). \quad (2)$$

This process is repeated R times to smooth out fluctuations and in this way we obtain an average $\langle s^{rand}_{l,c} \rangle$. Here, we are interested in the limit of large number of realizations, R, in which the users speaking language l would be distributed at random within the local population (fully integrated). The reason to repeat the procedure instead of using in a single run the distribution of the full population is to maintain the effect of the finite number of users speaking l. The speakers of this community l can be more or less concentrated in certain areas than the general population. To assess this effect, we define for each city c and detected language l the ratio:

$$\hat{h}_{l,c} = \frac{s_{l,c}}{\langle s^{rand}_{l,c} \rangle}. \quad (3)$$

To make the metric further comparable across cities, we further normalized $\hat{h}_{l,c}$ by the value obtained for the local language(s) spoken in city c, $\hat{h}_{loc,c}$ (Table E in the S1 File of the Supporting Information). If more than one local language is present in the city, the data for all these languages is aggregated to obtain a joint value of $\hat{h}_{loc,c}$. The final definition of the ratio of entropies is thus:

$$h_{l,c} = \hat{h}_{l,c}/\hat{h}_{loc,c}. \quad (4)$$

In this way, the information provided takes as baseline the local population and will inform us whether a specific group is spatially segregated or not. According to this definition, low values of $h_{l,c}$ are symptoms of segregation, whereas local languages and those distributed spatially in a similar manner are characterized by $h_{l,c}$ values close to unit. The values of this normalized ratio $h_{l,c}$ constitute the weights of the links in the bipartite network displayed in Fig 2.

The stability of the spatial entropy in function of different cells sizes (different scales Δx^2) is studied in the Supporting Information. We evaluate the relative error among the links of the bipartite network in function of Δx taking as reference the unit-like cell with 500 m side frame. Results are quite stable taking into account the spatial component of entropy related to the side size of the cells of 1000 and 2000 meters, respectively, as shown in the Fig F of the S1 File of the SI.

Evaluation of the migrant communities spatial distribution's accuracy

Twitter has the advantage of being a global source of data, but also the disadvantage of having several uncontrollable biases. Young people are usually over-represented [45, 47], and most likely the people belonging to the diverse communities are adopting the technology in different ways. If the use of geolocated Twitter is widespread in the host country, this will depend on the maturity of the migrant community: second and third generations are more likely to behave as locals and to adopt generalized technologies in the host population than first generations. On the other hand, things may vary if the technology is already commonly accepted in the country of origin of the community. Certainly, there are communities that are not detected. According to the National Institute for Statistics of Spain (INE, http://www.ine.es), a total of 45,728 and 54,599 Chinese citizens are residing in Barcelona and Madrid provinces, respectively, in 2016. Provinces are territorial divisions that enclose the urban areas and that loosely correspond to the area of analysis taken for our Twitter data. However, the number of users detected tweeting in Chinese is below the threshold of 30 with a valid residence cell and, therefore, this community does not appear in either of these cities. In the case of this particular group, there may be various reasons for this situation including the relative novelty of Chinese migration to Spain with most of this people belonging to the first generation, as well as the existence of alternatives to Twitter in China such as Sina Weibo. The important question here is thus not whether we find all the communities, but whether we are able to say something meaningful about those detected.

Going step by step, let us consider first the influence of the geographical area chosen on the structure of the bipartite network between language communities and cities paying special attention to the weights of its links. For this, recall that we have selected areas of 60×60 km^2 around the barycenter of the 53 cities considered. These areas have been further divided in cells of 500×500 m^2, which are the basic units of the analysis. The 53 cities are large megalopolis, still one can wonder if a square frame of 60 km side is enough to cover all of them, or whether we are including rural areas that could pollute the results. To check the stability of the network in function of the size of the city boundaries, we evaluate the relative error among the edge weights for different side sizes (20, 40, 80 and 100 km) using as reference the original 60×60 km^2 frame. In particular, the relative change $\epsilon_{l,c}$ of the link weights in the bipartite spatial integration network taking as reference the 60 km side frame is computed as follows,

$$\epsilon_{l,c} = \frac{h_{l,c} - h_{l,c}^{ref}}{h_{l,c}^{ref}} \tag{5}$$

where $h_{l,c}^{ref}$ represents the edge weight for $60 \times 60\ km^2$ frame. Box plots displaying the distribution $\epsilon_{l,c}$ values for different frame side sizes can be found in Fig 3a. The network weights are stable for frame side sizes ranging from 40 to 80 km. Beyond these values, the differences are increasing, the influence zone is too limited or extended far away from the center into rural areas or other neighboring cities. The value of 60 km for the side size is thus a safe choice. It is also worth nothing than the number of detected languages increases with the size of the frame. This number is however quite stable for box sizes ranging from 40 to 80 km (± 6% of the reference value). We perform the same analysis over the cell side size, taking as reference the 500 m side frame. Results are still quite stable increasing the size to 1000 and 2000 meters, respectively, as shown in the Fig G in the S1 File of the SI.

A next question to consider concerns the minimum number of users needed to obtain a stable measure of $h_{l,c}$. The number of users for whom we can detect a residence area per community are not very high (Table F in the S1 File of the SI), and in addition we have set a threshold

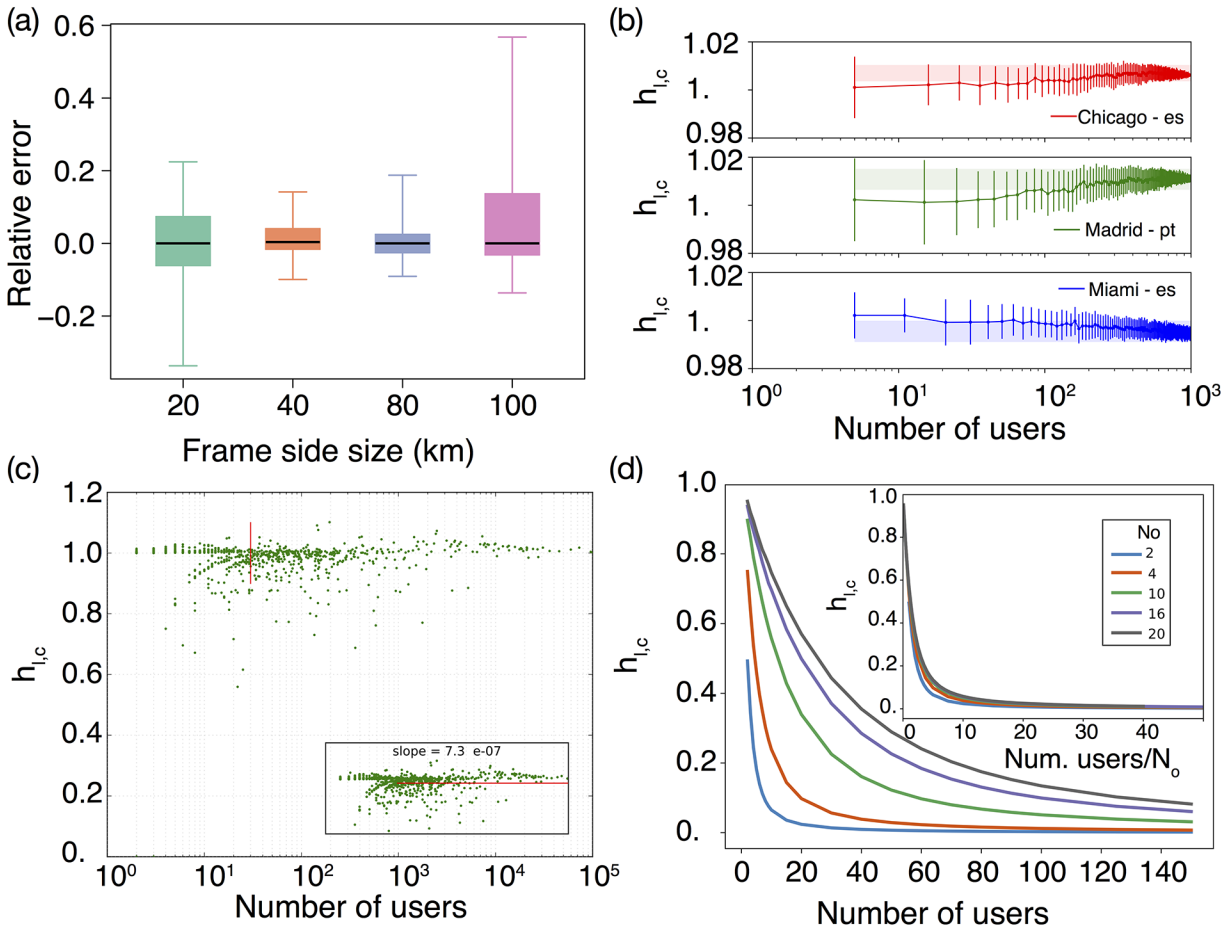

Fig 3. Evaluation of the migrant communities spatial distribution. (a) Box plots of the relative change $\epsilon_{l,c}$ of the link weights in the bipartite spatial integration network taking as reference the 60 km side frame. (b) The entropy ratio $h_{l,c}$ for three examples of communities with more than 1000 detected users (Spanish in Chicago and Miami and Portuguese in Madrid). A random sub-sampling is extracted and the calculated ratio of entropies is displayed as a function of the sample size. (c) The ratio of entropies $h_{l,c}$ as a function of the community size in number of users with a valid residence for all the communities. Every points represent a linguistic community in a city. The red vertical line marks the level of 30 users taken as a threshold. In the inset, it is shown a zoom-in with the details of the main plot. (d) We present the results concerning the ratio of entropies of a null model in which users belonging to a immigrant community is allowed to reside only in a subset N_0 of cells. These users are distributed randomly in the N_0 cells, while the local population is randomly distributed across all the gird cells. In the numerical examples, the system contains $100 \times 100 = 10000$ cells. The figure shows how the ratio of entropies changes with the number of users in the immigrant community and how the curves depend in first order on the ratio between the number of users and N_0.

https://doi.org/10.1371/journal.pone.0191612.g003

of at least 30 users to accept the data of a community. Where this value is coming from? To get a first impression of the effect that the user number has on $h_{l,c}$, we select some of the most populous migrant communities, delete a fraction of their users at random and plot in Fig 3b the value of $h_{l,c}$ as a function of the remaining users. Every random extraction produces a different value of $h_{l,c}$, so in the plot we depict the average and the error bars obtained from the standard deviation. Besides, we mark with a shadowed areas the values between which $h_{l,c}$ lies for the extractions with the largest number of users. The results depend on the particular community, but in general the values of $h_{l,c}$ enter in the shadowed areas between 10 and 100 users, 30 corresponds to the middle ground in logarithmic scale. A more systematic check can be seen in Fig 3c. There, a scatter plot with every value of $h_{l,c}$ for couples language-city is depicted as a function of the number of users associated to the particular community. After 30 users, there is no more clear dependency between $h_{l,c}$ and the number of users so it must reflect the spatial

distribution of the communities. It is also possible to perform a more detailed check in a controlled environment by introducing a null model in which the local population is randomly but uniformly distributed across the grid forming the city, while the immigrant population can only appear in a subset N_0 of cells. In those cells the immigrants are also distributed uniformly and randomly. By tuning the number of immigrant users and N_0, one can explore how the metric $h_{l,c}$ reacts to finite numbers (see Fig 3d). When the number of immigrants detected is smaller than N_0, they are indistinguishable from the local population and thus the ratio $h_{l,c}$ starts in one. As the number of immigrant users gets over N_0, the fact that their residence is restricted to a certain area of the city becomes evident and $h_{l,c}$ decays towards a fixed value. As can be seen in the inset of Fig 3d, the main control parameter of the null model is the ratio between the number of immigrant users and N_0. The curves showing $h_{l,c}$ as a function of the number of immigrants collapse by considering them as a function of such ratio. In general terms, the metric $h_{l,c}$ reaches a stable value once the number of immigrants is between 10 and 20 times larger than the cells where the community concentrates N_0. This model is a worst-case scenario for testing $h_{l,c}$, since the immigrants distribute uniformly while in more realistic applications if a ghetto exists the concentration density will not be uniform. In this latter case, lower number of users are required to measure the stable value of $h_{l,c}$.

Finally, we have been also able to run a comparison between the spatial distribution of the communities detected in three cities for which the data from census offices was available. These cities are Barcelona, London and Madrid, and for the comparison we use data from the so-called Continuous Register Statistics in Spain and the Census Office in the UK. In the Spanish case, the information is collected when people residing in a certain area must inform the municipal authorities for tax purposes and to obtain social services such as health care. The smallest spatial units for this dataset are census tracts, so Twitter data must be translated into the same geographical units (see the Supporting information for further details). We employ the Anselin Local Moran's I [55] to analyze the level of spatial correspondence of the main migrant communities. This metric provides information on the location, size and spatial coincidence 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. The details are included in the Supporting Information, but a summary with the most important results for a set of linguistic communities common to the three cities are shown in Table 1. The comparison between the location of the residence areas detected with Twitter and those registered in the census is in general good and significant, except for some of the immigrant communities such as Arabic in Barcelona and Madrid or East-Slavic in Madrid where the results lose significance and are compatible with a random distribution.

Power of integration

Once the limits of the data and the method to assess the spatial segregation levels of foreign communities have been checked, it is the moment to advance and study what can be said about the way that the cities integrate the foreign groups detected in Twitter. To this end and starting from the bipartite spatial integration network, we perform a clustering analysis based on the distribution of edge weights $h_{l,c}$. For each city c, the weights of the edges are sorted in descending order and stored into a vector $\vec{E}_c$. This vector $\vec{E}_c$ contains thus the information on how many foreign linguistic communities have been found in the city c and it quantifies how they are integrated. We can compare next the vectors $\vec{E}_c$ of pairs of cities to assess whether

Table 1. Comparison of linguistic communities detection between census and Twitter. Global Moran's *I* for a set of common languages detected in Barcelona, London and Madrid. The z-values are calculated after 99 permutations. The last column refers to the quality and significance of the spatial autocorrelations detected.

City	Language	*I*	Z-value	Autocorrelation
Barcelona	Total	0.63	236.5	Positive
	Spanish	0.62	217.0	Positive
	English	0.50	230.5	Positive
	French	0.37	151.5	Positive
	Italian	0.28	125.8	Positive
	Portuguese	0.32	151.2	Positive
	Arabic	0.08	89.9	Random
	East-Slavic	0.21	112.8	Positive
London	Total	0.71	66.5	Positive
	English	0.34	35.9	Positive
	Spanish	0.27	28.1	Positive
	French	0.25	32.1	Positive
	Italian	0.26	31.9	Positive
	Portuguese	0.15	18.5	Positive
	Arabic	0.34	48.5	Positive
Madrid	Total	0.62	268.6	Positive
	Spanish	0.62	267.3	Positive
	English	0.32	159.2	Positive
	French	0.37	151.5	Positive
	Italian	0.26	146.3	Positive
	Portuguese	0.44	204.9	Positive
	Arabic	0.07	41.5	Random
	East-Slavic	0.06	37.7	Random

https://doi.org/10.1371/journal.pone.0191612.t001

they behave in a similar way respect to the integration of external communities. Similarity metrics usually require the two vectors compared to have the same length. This difficulty can be overcome easily by adding zeros at the end of $\vec{E}_c$ until reaching the maximum length observed in the network L_{max}, namely, for London. We then perform a clustering analysis to find cities exhibiting similar distribution of edge weights by using a k-means algorithm based on Euclidean distances. The results of the analysis are confirmed by repeating the clustering detection with a Hierarchical Clustering Algorithm yielding the same results (see Fig B in the S1 File of the SI).

Fig 4a shows the three clusters (C1 in blue, C2 in red and C3 in green) obtained after applying the clustering algorithms. These three clusters are characterized by the different rhythm of decay of the entropy values in $\vec{E}_c$ as can be seen in Fig 4b. The first cluster C1 including cities like London, San Francisco, Tokyo or Los Angeles shows the slowest decay. These cities contain in general a number of communities, which are spatially distributed closely mimicking the local population. In the other extreme, the cluster C3 comprises cities with few or none migrant communities and displaying a high level of spatial segregation for the groups detected. In some cities of this club such as Guadalajara or Lima, we could only detect after applying filters the local languages. However, there are others like Toronto, Miami, Dallas, Rome or Istanbul for which the number of communities is comparable to the cities in the other clusters but the decay of the entropy is way much faster. The communities in their respective $\vec{E}_c$ are highly isolated in comparison with the local population or with similar communities in cities of C1.

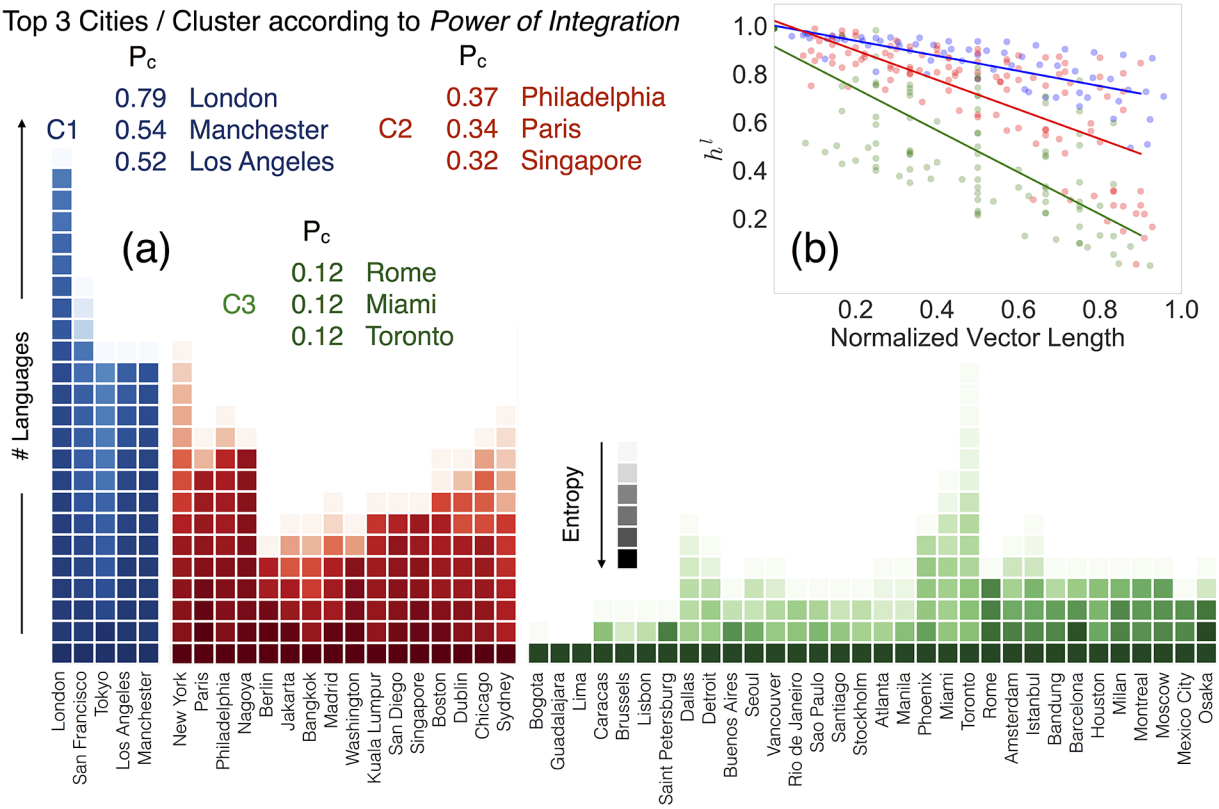

Fig 4. Clusters of cities and *Power of Integration*. In (a), three groups of cities show similar behavior in the number of communities detected and in their levels of integration. The length of the vectors represents the number of languages (communities) detected in each city; the color scale is representative of the decay of the entropy metric; the *Power of Integration* metric lead us to evaluate the potential of each city in uniformly integrating immigrant communities within its own urban area according to entropy values. In (b), decay of $h_{l,c}$ for the cities in each cluster. The points correspond to the values of the elements of $\vec{E}_c$ for each city placed in the x axis according to their index normalized to the total number of languages in c. The colors are for cities in the different clusters (C1 blue, C2 red and C3 green), and the lines are minimum square fits to the values of entropy ratios of each cluster.

https://doi.org/10.1371/journal.pone.0191612.g004

Finally, there is a middle ground in the cluster C2 containing cities as New York, Paris, Philadelphia, Chicago and Sydney. We introduce a new metric in order to summarize the distribution of entropy and to assess the city's *Power of Integration* (Table G of the S1 File of the Supporting Information). This metric is defined, for each city, as:

$$P_c = \frac{L_c}{L_{max}} Q_2 (1 - IQR), \tag{6}$$

where L_c is the number of languages spoken in city c and L_{max} is the maximum number of languages across the whole set of cities, $Q2$ is the median value of entropy and IQR the interquartile range used as a measure of dispersion. P_c is maximum when the median of the entropy ratio distribution is one or over, $IQR = 0$ and the number of languages hosted by city c is the maximum. On the other extreme, it tends to zero when there are no hosted language, the languages are spatially isolated with $Q_2 = 0$, or when the $IQR = 1$ covering the full range of values. The top three ranking cities in each cluster according to the Power of Integration are displayed in Fig 4a. According to the full ranking of cities by their *Power of Integration* (Table G in the S1 File of the SI), the metric is able to capture the contribution in the spatial integration process within each urban area: cities belonging to cluster C1 comprises values of P_c ranging from Tokyo's 0.41 to London's 0.79; the former city shows good integration of massive communities

coming from South Korea, Philippines and China. On the other side, the British capital shows almost full spatial mixing of a very large number of foreign communities. Cities belonging to cluster C2 are characterized by values of P_c ranging from Jakarta's 0.10 (characterized by mixing segregation behaviors in a scenario of spatial uniformity of most of the communities) to the 0.37 reached on the urban area of Philadelphia; here we found several communities that are uniformly spread within the city, whereas segregation appears focusing on the Arabic speaking community. The cluster as a whole mixes first segregation behaviors in a scenario of several communities involved in the process. Finally, cluster C3 is when both low number of immigrant communities are not well uniformly distributed within the urban areas, proved by the fact that P_c are very low. Brussels's 0.01 is due to the low values of entropy of the Turkish community within a scenario of few immigrant communities. Toronto, on the other side, is characterized by a very high number of immigrant communities (comparable to cities found in the cluster C2), not being well spatial integrated within the urban environment. This leads to a P_c value of 0.12. Note that the clusters are obtained directly from the similarity between vectors $\vec{E}_c$ for each city, and later their character is explained by using the decay of the ratios $h_{l,c}$ in the vectors and P_c.

Language integration network

The bipartite spatial integration network can be also be projected into the language side to gain insights on the level of integration of languages into the different countries (see Table H in the S1 File of the SI). We do the analysis at the country level because we assume that the integration of the immigrant communities is similar across the cities of the same country. When there are more than one city in the country, we take the average value of the entropy $h_{l,c}$ to build the network. The best and the worst cases of integration are displayed in Fig 5 left and right. Before proceeding to the analysis, it is important to mention that English has been excluded from the network because of its role as *lingua franca* [56]. Moreover, the role of English is dominant mainly in the worst links in terms of integration (see Fig C in the S1 File of the SI for more details). We select two thresholds of levels of integration of language in countries: in the top set (Fig 5 left) 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 in Venezuela, and the Spanish-speaking in Germany, Singapore and Turkey; the latter country shows uniformly 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 the link weights, UK leads in the role of hosting diverse communities and some other patterns emerge, such as the German presence in Japan and UK. By contrast (Fig 5 right), Arabic rises as the most common spatially segregated community followed by French-speaking communities that appear to be spatially concentrated in other European countries such as Germany and Turkey. Increasing the threshold further, 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). Note that the segregation can occur on the two extremes of the economic spectrum: poor people may need to live in ghetto-like areas but also wealthier communities may concentrate with respect to the general local population as it seems to be the case for Italian and Spanish speaking minorities in the Philippines or the English speaking community in Rome.

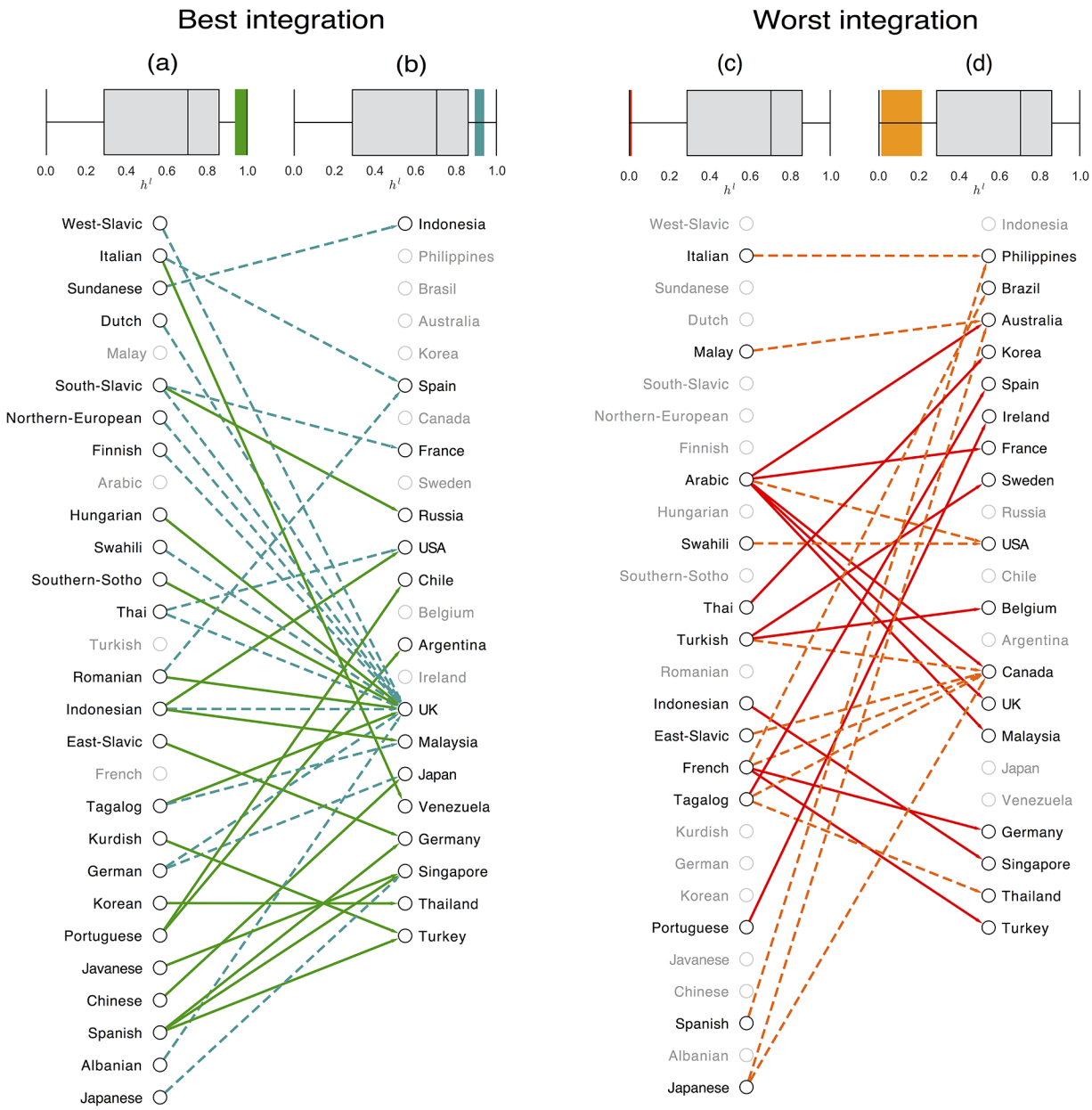

Fig 5. Language integration network. We select the sub-network representative of the best levels of spatial integration of languages in countries and display it on the left of the figure. The network is formed by the top 10% links according to the entropy distribution (the spread of the values can be seen in the boxplot (a) in comparison with all the values of $h_{l,c}$). In addition, we include an extra 10% of links (dash-lines) to the network, those between 10% and 20% best links (their spread is in the boxplot (b)). In the network only nodes that belong to the top set are highlighted. Similarly, on the right, the worst levels of spatial integration of languages in countries are shown. We filter out the bottom 10% links according to the entropy distribution (their spread of values is in the boxplot (c)), and add an extra 10% of links to the network (dash-lines), those links between the 10% and 20% worst in the ranking. Their spread is in the boxplot (d). As before, only the nodes that belong to the worst set are highlighted.

https://doi.org/10.1371/journal.pone.0191612.g005

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 studied so far by means of surveys and census data that cover

from the number of people living outside their country of birth to place of residency to features of the labor market. However, census and surveys have the disadvantage of a very high cost, geographical limitations and, typically, they have slow update frequencies. Recent works by experts in the area highlight the dare need of more agile data sources about mobility and settlement patterns of immigrant and refugee communities.

Rather than using these classical sources, in this work we explore the capability of the online social networks to provide information about the integration of immigrant communities. In particular, we use Twitter to connect users to their residence place and via a language *algebra* to determine their cultural background. This allows us to study how spatial and linguistic characteristics of people vary within the cities they are living in, and how the cities spatially integrate the diversity of languages and cultures characteristic of the global metropolises today. It is necessary to admit the potential biases of the data: the social network penetration through socio-economic hierarchies, age, generations and countries is different. This is precisely the reason why we do not detect all the possible communities in the cites under consideration. Still we have introduced a method compressing a metric that is not so sensitive to the small numbers in the users detected. As can be seen, in the validation exercise the results in the cities where we can compare with the census are significant for communities with more than 30 users. This method in general measures how well different communities are spatially integrated/segregated within urban areas. Our findings provide a new way to observe the patterns of historically immigration of people to urban areas, and any potential changes that might arise in the areas of residence. We are able to move beyond the estimation of past, current and foreshadowed global flows toward a better comprehension of the integration phenomena on a city scale. Residents' online communications can thus let us assess in an indirect way if the cultural background has been kept inside communities, although impacted on different levels by local welcoming and hosting policies. This method provides an extra alternative to the toolkit of researchers in sociology and urbanism as well as direct view in close to real time on the potential problems of integration that may appear in different areas of the cities, a knowledge that can be of great value to public managers.

Supporting information

S1 File. Pdf file containing the SI. This file includes 10 tables (Table A: Number of users and tweets in each city; Table B: Location of the city centers; C: Number of users residing in each city; D: Language aggregation process; E: Local languages in each city; F: Number of residents per language and city; G: Power of Integration of the cities; H: City-Country correspondence; I: Languages and country correspondence; J: Data validation) and 7 figures (A: Number of reliable users as a function of filtering parameters; B: Comparison of clustering methods for the cities; C: Degree and weight distribution of the bipartite networks with and without English; D: Data validation 1; E: Data validation 2; F: Relative error of the entropy as a function of the scale Δx; G: Relative error of the link weights as a function of the scale Δx).
(PDF)

S2 File. An example code in python with a query to the Twitter API.
(PDF)

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 (611307). 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). BG thanks the Moore and Sloan Foundations for support as part of the Moore-Sloan Data Science Environment at New York University.

Author Contributions

Conceptualization: Fabio Lamanna, Maxime Lenormand, Gustavo Romanillos, Bruno Gonçalves, José J. Ramasco.

Data curation: Fabio Lamanna, Maxime Lenormand, María Henar Salas-Olmedo, Gustavo Romanillos, Bruno Gonçalves, José J. Ramasco.

Formal analysis: Fabio Lamanna, Maxime Lenormand, María Henar Salas-Olmedo, Gustavo Romanillos, Bruno Gonçalves, José J. Ramasco.

Funding acquisition: Maxime Lenormand, José J. Ramasco.

Investigation: Fabio Lamanna, Maxime Lenormand, María Henar Salas-Olmedo, Gustavo Romanillos, Bruno Gonçalves, José J. Ramasco.

Methodology: Fabio Lamanna, Maxime Lenormand, María Henar Salas-Olmedo, Gustavo Romanillos, José J. Ramasco.

Software: Fabio Lamanna, Maxime Lenormand, María Henar Salas-Olmedo, Gustavo Romanillos.

Supervision: José J. Ramasco.

Validation: Fabio Lamanna, Maxime Lenormand, María Henar Salas-Olmedo, Gustavo Romanillos, José J. Ramasco.

Visualization: Fabio Lamanna, Maxime Lenormand, María Henar Salas-Olmedo, Gustavo Romanillos, Bruno Gonçalves, José J. Ramasco.

Writing – original draft: Fabio Lamanna, Maxime Lenormand, María Henar Salas-Olmedo, Gustavo Romanillos, Bruno Gonçalves, José J. Ramasco.

Writing – review & editing: Fabio Lamanna, Maxime Lenormand, María Henar Salas-Olmedo, Gustavo Romanillos, Bruno Gonçalves, José J. Ramasco.

References

1. Burgess EW, Park RE. Introduction to the Science of Sociology. University of Chicago Press; 1921.
2. Gordon MM. Assimilation in American Life: The Role of Race, Religion and National Origins. vol. 4. Oxford University Press; 1964.
3. Berry JW. Immigration, Acculturation, and Adaptation. Applied Psychology. 1997; 46(1):5–34.
4. Ager A, Strang A. Understanding integration: a conceptual framework. Journal of Refugee Studies. 2008; 21(2):166–191. https://doi.org/10.1093/jrs/fen016
5. Entzinger H, Biezeveld R. Benchmarking in Immigrant Integration. Managing Integration The European Union's Responsibilities Towards Immigrants. 2005; 1(August):123–136.
6. Gonul T. A Comparative Study of the Integration of the Turks in Germany and the Netherlands. Migration Letters. 2012; 9:25–32.
7. Massey DS, Denton NA. American apartheid: segregation and the making of the underclass. Cambridge, MA, USA: Harvard University Press; 1993.
8. Massey DS, Denton NA. Trends in the Residential Segregation of Blacks, Hispanics, and Asians: 1970-1980. American Sociological Review. 1987; 52(6):802. https://doi.org/10.2307/2095836

9. Oka M, Wong DWS. Spatializing Segregation Measures: An Approach to Better Depict Social Relationships. Cityscape: A Journal of Policy Development and Research. 2015; 17(1):97–113.
10. Abel GJ, Sander N. Quantifying global international migration flows. Science. 2014; 343(6178):1520–1522. https://doi.org/10.1126/science.1248676 PMID: 24675962
11. Butler D. What the numbers say about refugees. Nature. 2017; 543:22–23. https://doi.org/10.1038/543022a PMID: 28252091
12. Editorial article. Data on movements of refugees and migrants are flawed. Nature. 2017; 543:5–6. https://doi.org/10.1038/543005b
13. Dijstelbloem H. Migration tracking is a mess. Nature. 2017; 543:32–34. https://doi.org/10.1038/543032a PMID: 28252099
14. Beaverstock J. Lending Jobs to Global Cities: Skilled International Labour Migration, Investment Banking and the City of London. Urban Studies. 1996; 33(8):1377–1394. https://doi.org/10.1080/0042098966709
15. Sassen S. The global city: introducing a concept. The Brown Journal of World Affairs. 2005; XI(2):27–40.
16. Friedmann J. The World City Hypothesis. Development and Change. 1986; 17(1):69–83. https://doi.org/10.1111/j.1467-7660.1986.tb00231.x
17. Samers M. Immigration and the global city hypothesis: Towards an alternative research agenda. International Journal of Urban and Regional Research. 2002; 26(2):389. https://doi.org/10.1111/1468-2427.00386
18. Hamnett C. Social Polarisation in Global Cities: Theory and Evidence. Urban Studies. 1994; 31(3):401–424. https://doi.org/10.1080/00420989420080401
19. Musterd S. Social and ethnic segregation in Europe: Levels, causes, and effects. Journal of Urban Affairs. 2005; 27(3):331–348. https://doi.org/10.1111/j.0735-2166.2005.00239.x
20. Bean FD, Stevens G. America's Newcomers and the Dyamics of Diversity. Russell Sage Foundation; 2003.
21. Phalet K, Swyngedouw M. Measuring immigrant integration: The case of Belgium. Studi Emigrazione. 2003; 1(152):773–804.
22. Reades J, Calabrese F, Sevtsuk A, Ratti C. Cellular Census: Explorations in Urban Data Collection. Pervasive Computing, IEEE. 2007; 6(3):30–38. https://doi.org/10.1109/MPRV.2007.53
23. González MC, Hidalgo CA, Barabási AL. Understanding individual human mobility patterns. Nature. 2008; 453:779–782. https://doi.org/10.1038/nature06958 PMID: 18528393
24. Reades J, Calabrese F, Ratti C. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. Environment and Planning B: Planning and Design. 2009; 36(5):824–836. https://doi.org/10.1068/b34133t
25. Soto V, Frías-Martínez E. Automated land use identification using cell-phone records. In: Proceedings of the 3rd ACM international workshop on MobiArch. HotPlanet'11. New York, NY, USA: ACM; 2011. p. 17–22. Available from: http://doi.acm.org/10.1145/2000172.2000179.
26. Toole JL, Ulm M, González MC, Bauer D. Inferring Land Use from Mobile Phone Activity. In: Proceedings of the ACM SIGKDD International Workshop on Urban Computing. UrbComp'12; 2012. p. 1–8.
27. Pei T, Sobolevsky S, Ratti C, Shaw SL, Zhou C. A new insight into land use classification based on aggregated mobile phone data. International Journal of Geographical Information Science. 2014; 28:1988–2007. https://doi.org/10.1080/13658816.2014.913794
28. Louail T, Lenormand M, Garcia Cantú O, Picornell M, Herranz R, Frías-Martínez E, et al. From mobile phone data to the spatial structure of cities. Scientific Reports. 2014; 4:5276. https://doi.org/10.1038/srep05276 PMID: 24923248
29. Amini A, Kung K, Kang C, Sobolevsky S, Ratti C. The impact of social segregation on human mobility in developing and industrialized regions. EPJ Data Science. 2014; 3(1):1–20. https://doi.org/10.1140/epjds31
30. Tizzoni M, Bajardi P, Decuyper A, Kon Kam King G, Schneider CM, Blondel V, et al. On the Use of Human Mobility Proxies for Modeling Epidemics. PLoS Computational Biology. 2014; 10:e1003716. https://doi.org/10.1371/journal.pcbi.1003716 PMID: 25010676
31. Deville P, Linard C, Martin S, Gilbert M, Stevens FR, Gaughan AE, et al. Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences (USA). 2014; 111:15888–15893. https://doi.org/10.1073/pnas.1408439111
32. Grauwin S, Sobolevsky S, Moritz S, Gódor I, Ratti C. Towards a Comparative Science of Cities: Using Mobile Traffic Records in New York, London, and Hong Kong. In: Helbich M, Jokar Arsanjani J, Leitner

M, editors. Computational Approaches for Urban Environments. Cham, Switzerland: Springer International Publishing; 2015. p. 363–387.

33. Blondel V, Decuyper A, Krings G. A survey of results on mobile phone datasets analysis. EPJ Data Science. 2015; 4:10. https://doi.org/10.1140/epjds/s13688-015-0046-0
34. Louail T, Lenormand M, Picornell M, Garcia Cantú O, Herranz R, Frías-Martínez E, et al. Uncovering the spatial structure of mobility networks. Nature Communications. 2015; 6:6007. https://doi.org/10.1038/ncomms7007 PMID: 25607690
35. Lenormand M, Louail T, Cantú-Ros OG, Picornell M, Herranz R, Arias JM, et al. Influence of sociodemographics on human mobility. Scientific Reports. 2015; 5:10075. https://doi.org/10.1038/srep10075 PMID: 25993055
36. Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C. Geo-located Twitter as a proxy for global mobility patterns. Cartography and Geographic Information Science. 2014; 41:260–271. https://doi.org/10.1080/15230406.2014.890072 PMID: 27019645
37. Lenormand M, Picornell M, Garcia Cantú O, Tugores A, Louail T, Herranz R, et al. Cross-checking different source of mobility information. PLoS ONE. 2014; 9(8):e105184. https://doi.org/10.1371/journal.pone.0105184 PMID: 25133549
38. Lenormand M, Tugores A, Colet P, Ramasco JJ. Tweets on the road. PLoS ONE. 2014; 9(8):e105407. https://doi.org/10.1371/journal.pone.0105407 PMID: 25141161
39. Lenormand M, Gonçalves B, Tugores A, Ramasco JJ. Human diffusion and city influence. Journal of The Royal Society Interface. 2015; 12:20150473. https://doi.org/10.1098/rsif.2015.0473
40. Magdy A, Ghanem TM, Musleh M, Mokbel MF. Exploiting geo-tagged Tweets to Understand Localized Language Diversity. In: Proceedings of Workshop on Managing and Mining Enriched Geo-Spatial Data—GeoRich'14; 2014. p. 1–6.
41. Mocanu D, Baronchelli A, Perra N, Gonçalves B, Zhang Q, Vespignani A. The Twitter of Babel: Mapping World Languages through Microblogging Platforms. PLoS ONE. 2013; 8(4):e61981. https://doi.org/10.1371/journal.pone.0061981 PMID: 23637940
42. Jurdak R, Zhao K, Liu J, AbouJaoude M, Cameron M, Newth D. Understanding human mobility from Twitter. PLoS ONE. 2015; 10(7):35. https://doi.org/10.1371/journal.pone.0131469
43. Gonçalves B, Sanchez D. Crowdsourcing dialect characterization through twitter. PLoS ONE. 2014; 9(11):1–10.
44. Doyle G. Mapping dialectal variation by querying social media. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics; 2014. p. 98–106.
45. Mislove A, Lehmann S, Ahn Yy, Onnela Jp, Rosenquist JN. Understanding the Demographics of Twitter Users. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media; 2011. p. 554–557. Available from: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2816/3234.
46. Bokányi E, Kondor D, Dobos L, Sebők T, Stéger J, Csabai I, et al. Race, religion and the city: twitter word frequency patterns reveal dominant demographic dimensions in the United States. Palgrave Communications. 2016; 2:16010. https://doi.org/10.1057/palcomms.2016.10
47. Sloan L. Who tweets in the United Kingdom? Profiling the Twitter population using the British social attitudes survey 2015. Social Media + Society. 2017; 3(1):2056305117698981. https://doi.org/10.1177/2056305117698981
48. Arribas-Bel D. The spoken postcodes. Regional Studies, Regional Science. 2015; 2(1):458–461. https://doi.org/10.1080/21681376.2015.1067151
49. Bajardi P, Delfino M, Panisson A, Petri G, Tizzoni M. Unveiling patterns of international communities in a global city using mobile phone data. EPJ Data Science. 2015; 4(1):3. https://doi.org/10.1140/epjds/s13688-015-0041-5
50. Herdağdelen A, State B, Adamic L, Mason W. The Social Ties of Immigrant Communities in the United States. In: Proceedings of the 8th ACM Conference on Web Science. WebSci'16. New York, NY, USA: ACM; 2016. p. 78–84. Available from: http://doi.acm.org/10.1145/2908131.2908163.
51. Vigdor JL. Measuring Immigrant Assimilation in the United States. Civic Report No.53. Manhattan Institute for Policy Research. 2008;.
52. Chu Z, Gianvecchio S, Wang H, Jajodia S. Who is Tweeting on Twitter: Human, Bot, or Cyborg? In: Acsac 2010; 2010. p. 21.
53. White MJ. Segregation and Diversity Measures in Population Distribution. Population Index. 1986; 52(2):198–221. https://doi.org/10.2307/3644339 PMID: 12340704
54. Batty M. Spatial Entropy. Geographical Analysis. 1974; 6:1–31. https://doi.org/10.1111/j.1538-4632.1974.tb01014.x

55. Anselin L. Local indicators of spatial association LISA. Geographical analysis. 1995; 27(2):93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

56. Ronen S, Gonçalves B, Hu KZ, Vespignani A, Pinker S, Hidalgo Ca. Links that speak: The global language network and its association with global fame. Proceedings of the National Academy of Sciences (USA). 2014; 111(111):E5616. https://doi.org/10.1073/pnas.1410931111