

Studying Multicultural Diversity of Cities and Neighborhoods through Social Media Language Detection

Michela Arnaboldi¹, Marco Brambilla², Beatrice Cassottana¹, Paolo Ciuccarelli³,
Davide Ripamonti², Simone Vantini⁴, Riccardo Volonterio²

¹DIG, ²DEIB, ³DensityDesign, ⁴MOX. Politecnico di Milano. Piazza L. Da Vinci, 32. 20133 Milano, Italy.
{firstname.lastname}@polimi.it

Abstract

Cities are growing as melting pots of people with different culture, religion, and language. In this paper, through multilingual analysis of Twitter contents shared within a city, we analyze the prevalent language in the different neighborhoods of the city and we compare the results with census data, in order to highlight any parallelisms or discrepancies between the two data sources. We show that the officially identified neighborhoods are actually representing significantly different communities and that the use of the social media as a data source helps to detect those weak signals that are not captured from traditional data.

Introduction

Due to the increasing urban population and the exponential number of interactions among urban inhabitants, cities are the places where decisions are taken, innovation is fostered and changes are most visible. At the same time, cities are becoming a cultural melting pot, resulting from the global flows of immigrants toward cities (Sanderson, 2015; Deville, 2014). Therefore, the language mix, or the multilingualism, deriving from these population movements is a relevant feature that needs to be further analyzed. Multilingualism can be seen as intangible cultural good, to which an economic value can be attributed (Cenoz, 2015).

Within this regard, traditional data (such as census), although precious, are insufficient and slow to provide actual information on population type and densities, and needs to be integrated with alternative datasets (Deville, 2014). In order to fill the gap left by traditional data and to capture those weak signals that the census is unable to measure, social media offer a powerful alternative data source.

Big data deriving from social media has two main advantages. On the one hand it is measured at individual level (Jiang, 2013), on the other hand it is cheaper and faster compared to traditional surveys while enabling to continuously monitor certain events or features of the city (Ceron, 2014).

Specifically, the aim of this paper is to study the multilingualism in cities answering the following questions: (1) Are there neighborhoods that are characterized by the predominance of a language spoken on the social media? (2) If yes, can we assume that the predominance of a certain language on the social media reveals a group of people living in that neighborhood? (3) Can we monitor the dynamics of linguistic diversity, by detecting anomalies and possibly reporting them back to actual events?

The paper describes both the mathematical methodology used for characterizing the neighborhoods of Milan in terms of languages spoken on Twitter, including the data collection and data visualization problems.

Related Work

Our work is among the first ones that combine the multilingualism analysis of Twitter and the development of a city dashboard. Here multilingualism is intended at a societal level rather than at an individual level, and refers to the use of two or more languages in a community. Many papers study the language distributions on Twitter and focus on related topics, such as the connections and the usage differences existing between language communities. It is well known that English is the prevalent language used on Twitter (Eleta, 2014; Hong, 2011; Mocanu, 2013; Takhteyev, 2012). In all the analyzed samples of tweets, a

definitely predominant percentage, about 50% (Hong, 2011), is written in English. Mocanu et al. (2013) perform a comprehensive survey of the worldwide linguistic landscape in 191 countries. They find that English is six times more popular than the second ranked language, Spanish, while Italian ranks 12th.

Some papers analyze the different language communities in a network perspective. Eleta et al. (2014) demonstrate through logistic regression that multilingual users stand between language groups. Kim et al. (2014) state that monolingual users cluster together, while bilingual users do not and that their language-mixing ratio closely mirrors the mix of their followership, thus confirming the role of “bridge-building” of bilingual users in the Twitter network. Takhteyev et al. (2012) describes the Twitter network in terms of number of geographical ties existing between users and shows that 39% of ties are shorter than 100 km, while frequency of air travel explains non-local ties. This finding goes in the opposite direction of the commonly accepted belief that social networks entail the creation of more geographically spread communities (Cairncross, 2001).

Finally, some researches deal with behavioral differences related to language on Twitter. Hong et al. (2011) find that different language communities use Twitter for different purposes. A similar distinction is made by Kim et al. (2014) with regards to the language chosen by multilingual speakers. They state that users communicate in the local language for informative purposes and in English to talk about events, tours or enjoyment. Mocanu et al. (2013) affirm that even in countries with a homogeneous linguistic identity, more than 20% of tweets are written in English.

Some studies offer historical perspectives to multilingualism of the most multicultural cities (see, for example, (Leimgruber, 2013) for Singapore; (García, 2001) for New York; (Extra, 2004) for a cross linguistic perspective on Goteborg, Hamburg, The Hague, Brussels, Lyon and Madrid) and of their evolution (Quercia et al., 2014).

In another research field, several projects, often termed “city dashboards” (Kitchin, 2015), exist that collect, analyze and visualize the vast amount of information being created by urban services (Lee, 2015) and by the citizens themselves through social media.

A seminal group of city dashboards have a focus on crime: CompStat for the city of New York (Mattern, 2015), CitiStat for Baltimore, Maryland (Gullino, 2009); further extensions like in the case of Charlotte, North Carolina (Mattern, 2015), the AMS Institute dashboard in Amsterdam, or CityEye dashboard of Barcelona and Santander (Lee, 2015) have broadened the analysis to other topics. The London Dashboard is a visualization site that tracks the real-time information about weather, pollution, public transport, river level, electricity demand and so on.

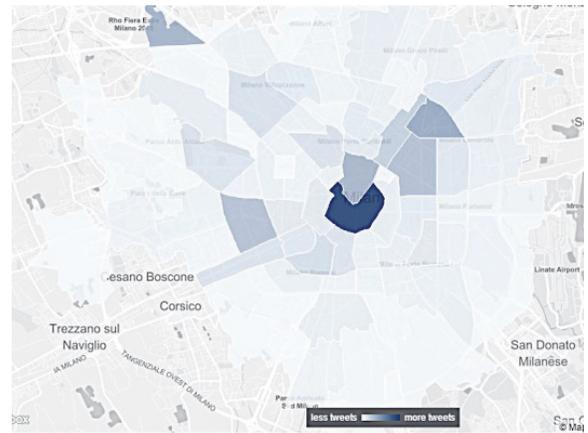


Figure 1. NILs of Milan represented according to the number of georeferenced tweets for the entire experiment period.

The Prefeitura of Rio de Janeiro in partnership with the IBM Intelligent Operations Center for Smart Cities (IOS) created a platform aggregating datasets from thirty agencies, including public transport, utility services and weather feeds (Kitchin, 2014). The demographic element appears sometimes in a second type of dashboards, which are rooted in balanced scorecard frameworks. Some examples are Dublin’s, Atlanta’s, and the Open Michigan dashboards.

Generally speaking, limited attention has been devoted to demographic and linguistic data in these dashboards. The aim of this paper is to illustrate a dashboard that addresses this gap by analyzing the demographic characteristics of citizens, and in particular the language distribution over the different zones of the city.

Experimental Scenario

In this paper we report our experience in monitoring multilingual aspects of the city of Milan, in Italy. At this purpose, we devised a solution, called Urbanscope, that takes inspiration from the concept of Macroscope (de Rosnay, 1979), i.e., a tool conceived to study complex systems. Urbanscope divides the city of Milan into 88 zones or neighborhoods, called NILs (Local Identity Nucleus) in Milan, created by the municipality to define an identifiable unit of granularity for the city. Indeed, demographic data about the population, as well as data concerning services such as transportation, education, healthcare and sport facilities are recorded for each NIL on the data portal of the municipality. Urbanscope receives data from different big data sources, which are typically initiated by private companies, such as transit authorities, mobile phone companies, and event organizers. This is also integrated with data collected from social media, specifically Twitter, Foursquare and Instagram. The project is inherently

multidisciplinary and engages researchers from fields such as management and administration, mathematics, software engineering, data visualization and data science.

In the analysis reported here, we focus only on Twitter content. Data is collected through the Twitter public API. For this study, we considered only the geo-tagged tweets within the boundaries of the municipality of Milan for a period of 18 months (the analysis is now continuously running and results are available at the live web site <http://urbanscope.polimi.it>). Thanks to the geographical metadata of Twitter, which can provide exact geographic coordinates (latitude and longitude) in case the user chooses to do so, tweets can be assigned to the NILs of Milan. Noticed however that not all the tweets are tagged with such information. A large majority of tweets do not contain geographical positioning metadata; these tweets are not considered in our analysis. For our city analysis we rely on the language detection algorithm of Twitter calculates the language automatically. Tweets containing words written in different languages are classified as “undefined language” and are excluded from our dataset.

The entire dataset analyzed here contains 1,109,69 tweets. The ones with a defined language associated are 1,007,314. The ones with geotag precisely associated with the Milan municipality area, and thus assigned to one of the 88 NILs of Milan, are 793,838. No further data cleaning and filtering is applied on the retrieved tweets before the analysis.

Figure 1 represents the map of Milan, where NILs are colored based on the number of tweets posted in that NIL for the entire period of analysis (Duomo is the NIL with the highest number of tweets).

Language Analysis

In order to provide an appropriate and statistically significant representation of the city from a linguistic perspective, our analysis focus is three-fold:

- A. Defining the proper level of data aggregation with respect to time, space, and language;
- B. Removing from the analysis the NILs which do not allow reliable estimate of language use due to low tweet counting (i.e., transparent NILs);
- C. Characterizing the NILs in terms of the most popular spoken language in Twitter (i.e., colored NILs).

With respect to task A, the proper level of aggregation is a trade-off between: (i) a level of aggregation large enough to provide a sufficiently large signal-to-noise ratio to have reliable estimates from tweet counting; and (ii) a level of aggregation small enough to allow monitoring of language changes along time and across space. In detail, to be consistent with census data we decided to keep fixed at the NIL level the space aggregation and working exclusively on the time and language dimensions. Based on quantitative

assessment, we fixed as minimal levels of aggregation: months, NILs, and the single languages.

More technically, for increasing levels of aggregation we performed a family of exact Fisher tests for stochastic independence testing (e.g., Agresti 2007) for each candidate time unit (i.e., months, two-month periods, quarters, ...) the stochastic independence between NIL and language. In this first analysis each tweet is considered as an instance of two categorical random variables: the NIL and the language of the tweet. Simultaneously, still using exact Fisher tests for stochastic independence, we tested for each candidate language the stochastic independence between NIL and time units. In this second analysis each tweet is considered instead as an instance of two categorical random categorical variables: the NIL and the corresponding time unit. An optimal setting of aggregation for the statistical analysis was then found in the identification of quarters as aggregated time units and Italian, English, and Other-languages as aggregated macro-languages.

Tasks B and C are strongly connected: the former being related to the identification of NILs whose tweet counting does not allow reliable estimates of the use of each macro-language in each quarter; and the latter being related to the identification of NILs presenting anomalously large tweet counting in at least one of the three macro-languages in a given quarter (with respect to other NILs in the same quarter). In detail, with respect to task B, for each quarter data are considered as instances of a 3d multinomial random variable defining the number of tweets in Italian, English, and Other-languages that NIL in that quarter. Assuming the total number of tweets in a given NIL in a given quarter to be ancillary with respect to the three macro-language counting, the number of tweets in one of the three macro-languages in a given NIL and quarter can be modelled as a binomial random variable $B(n,p)$ with p being the probability that the macro-language of a tweet in that NIL in that quarter is the selected one and n being instead the total number of tweets in the same NIL and quarter. In this framework, the observed macro-language percentage plays the role of an estimate of p (i.e., the ideal macro-language percentage that one could have observed if an infinite number of tweets were available). Its standard deviation is upper-bounded by $1/(2\sqrt{n})$. This bound allows to determine the minimal number of tweets required for each NIL in each quarter to have a maximal standard deviation lower than a desired threshold. In the current maps we required for each NIL a maximal standard deviation of 5%, which corresponds to at least 100 tweets for each quarter. Consequently, a NIL will become transparent in the visualization and ignored in the execution of task C if it does not count at least 100 tweets for each quarter.

With respect to task C, we decided to define three macro-languages, i.e., Italian, English, and Other-languages (representing the aggregate number of tweet in any other

languages). We give a tri-fold graphical representation of the city map for each quarter, as shown in Figure 2. For each quarter, we group non-transparent NILs into three categories: the ones presenting an extremely large value of the macro-language percentage counting with respect to the other NILs in the same quarter (i.e., dark-colored NILs in the visualizations); the ones presenting a moderately large value of the macro-language percentage counting with respect to the other NILs in the same quarter (i.e., light-colored NILs in the visualizations); and, finally, the ones not presenting a large value of the macro-language percentage counting with respect to the other NILs in the same quarter (i.e., gray NILs in the visualizations). To formally define the three categories above, we rely on a standard outlier detection method: the Box-and-Whisker Plot (e.g., Murrell 2005). According to this method, in each quarter for each one of the macro-languages, 75% of the non-transparent NILs will be gray, and 25% light- or dark-colored. In detail, the light and dark-colored NILs correspond to the ones presenting the 25% largest macro-language percentage counting among the non-transparent NILs (i.e., the ones exceeding the empirical third quartile Q_3 of the non-transparent NIL macro-language percentage counting measurements in the quarter). The remaining non-transparent NILs (i.e., the ones not exceeding the empirical third quartile Q_3) are instead gray-colored. As mentioned above, among the colored NILs we distinguish between the ones presenting an extremely large value for the macro-language percentage counting (i.e., dark-colored NILs) and the ones presenting a moderately large value for the macro-language percentage counting (i.e., light-colored NILs). In detail, dark-colored NILs are the ones for which the corresponding macro-language percentage counting exceeds the threshold $Q_3 + 1.5(Q_3 - Q_1)$ (i.e., superior outliers) where Q_1 is the empirical first quartile of the macro-language percentage counting measurements (i.e., that value of the macro-language percentage counting separating the bottom 25% and the top 75% of the non-transparent NILs). The threshold $Q_3 + 1.5(Q_3 - Q_1)$ is commonly used in robust statistics to detect superior-outliers (i.e., extremely large data that are not likely to be observed under the Normal model). Consequently, light-colored NILs are the remaining ones, namely, the non-transparent NILs exceeding Q_3 but not $Q_3 + 1.5(Q_3 - Q_1)$.

Results

While the prevalence of Italian tweets can be easily explained by the predominance of Italian speakers in a certain NIL, we are mostly interested in understanding the prevalence of English or other languages. Not surprisingly the second most numerous NILs are those for which English is the prevalent language. In accordance with literature, indeed, English is widely adopted because it is perceived as

having a greater influence in the Twitter community (Kim, 2014) and it is therefore difficult to associate a nationality to those users posting in English. More interesting are the NILs in which prevails a language other than Italian or English. Assuming that the language used on Twitter reflects the respective country of origin, the last group of NILs is probably mostly frequented by non-Italian native citizens.

Figure 2 shows the qualitative results of the language analysis, with separated Italian, English and Others contributions for two exemplary quarters, one before the Expo 2015 exhibition and one during the event.



Figure 2. Italian, English and Other language predominance: (a) January 2015 - March 2015; (b) April 2015 - June 2015.

Notice that NILs characterized by languages different from Italian or English are more often peripheral than central, while Italian frequently distinguishes the southern NILs and English the central NILs. While it is difficult to characterize the central NILs (e.g., Duomo is never colored) because of the high population and the homogeneity of this NILs (Italian residents as well as visitors and non-Italian citizens go to Duomo), some NILs might present peculiarities, which enable the characterization of the NILs because of language spoken on the Twitter. An example is Cascina Triulza – Expo, the NIL where the 2015 universal exposition took place from May until October. Indeed, Cascina Triulza – Expo becomes heavily colored in yellow in the trimester April 2015 - June 2015, showing the wave of visitors from all around the world.

In order to detect international communities within the city of Milan, NILs have been further explored by disaggregating their language composition (excluding Italian and English) and considering the whole period of analysis (August 2014-December 2015).

An example map with the Other Languages distribution in each NIL is shown in Figure 3. Table 2 reports the relevant results in summary. Interestingly, several NILs are characterized by a prevalence of tweets written in Arabic.

Spanish tweets represent a consistent percentage of non-Italian and non-English tweets for many NILs (above 20% for most of the NILs), indicating a more spread nature of the Spanish Twitter community within the city of Milan.

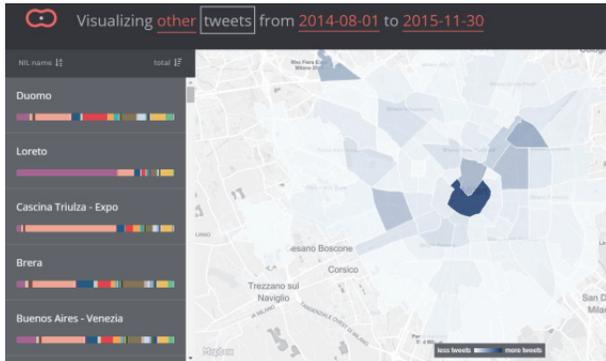


Figure 3 Explore Tweets, Other languages

Table 1 Twitter Predominant Languages for Some Relevant NILs

NIL	% of predominant language
Trenno	72% Arabic
Loreto, Umbria – Molise, Ortomercato	64% Arabic
Parco Forlanini - Ortica	63% Arabic
Quintosole	100% Spanish
Parco dei Navigli	69% Spanish
Gallaratese	66% Spanish
Cascina Triulza- Expo	57% Spanish
Ex Om – Morivione	56% Spanish
Mecenate	51% Spanish
Padova	59% Tagalog
Giambellino	57% Tagalog
Villapizzone	56% Tagalog
Bruzzano	59% Portuguese
Parco Nord	55% Dutch
Chiaravalle	75% Norwegian

Nevertheless, some NILs are characterized by a strong Spanish influence, such as Quintosole, for which 100% of the tweets are written in Spanish. Other common languages among the NILs are Tagalog (from Philippines), Portuguese, Indonesian, and Turkish. Some NILs in particular differ from the other in being the only ones characterized by a certain language, meaning that tweets written in that language overpass the 50% of the tweets written in languages other than Italian and English.

Comparison with Census Data

According to the census data of the municipality of Milan, the foreign population residing in Milan as recorded at December 31st, 2014 amounts to about 18.9% of the total population. Among them, the largest community is the Filipino one (41,237 people), followed by Egyptian (35,597 people), Chinese (25,928) and Peruvian (20,462).

By assigning to each NIL the respective language according to the census data of the municipality of Milan, we are able to compare Twitter results with traditional data. Columns of Table 2 report the percentages of non-Italian residents (grouped by spoken language) for those languages that are spoken by at least 20% of the non-Italian residents in at least one NIL. Into rows only NILs in which at least one of the languages above is spoken by at least 20% of the non-Italian residents. In the first column the characterizing language detected through Twitter is reported.

As one can observe, only in some cases the language community detected though Twitter corresponds to the majority of non-Italian residents in a certain NIL. This correspondence happens for some of the Arabic and Spanish communities. Interesting fact, Sarpi, that is the NIL typically recognized as the Chinese neighborhood, does not present a prevalence of tweets written in Chinese. This is due to absence of Twitter in China (Mocanu, 2013) replaced by Sina Weibo (www.sina.com.cn) (Crampton, 2011).

However, generally speaking the correspondence between Twitter and census is rather shallow. This is due to the fact that the two sources are actually describing very different phenomena (residents vs. any kind of social media authors, including visitors and tourists), including the fact that the users of Twitter represent a very biased subset of the population due to age and technology adoption attitude.

Table 2 Relevant Results

	Twitter top language	Percentages of non-Italian residents (Census data)					
		Arabic	Spanish	Tagalog	Ukrainian	Romanian	Albanese
Trenno	Arabic	29%	19%	7%	6%	5%	1%
Villapizzone	Tagalog	14%	23%	6%	4%	5%	4%
Umbria Molise	Arabic	22%	19%	7%	3%	5%	3%
Quintosole	Spanish	37%	3%	6%	0%	13%	21%
Parco Nord	Dutch	21%	0%	11%	21%	4%	7%
Forlanini Ortica	Arabic	23%	21%	8%	7%	15%	1%
Navigli	Spanish	9%	2%	0%	9%	28%	21%
Padova	Tagalog	14%	26%	6%	4%	5%	4%
Ortomercato	Arabic	25%	24%	14%	2%	7%	1%
Mecenate	Spanish	18%	24%	8%	5%	6%	3%
Loreto	Arabic	15%	22%	5%	3%	4%	4%
Giambellino	Tagalog	19%	20%	6%	3%	5%	4%
Gallaratese	Spanish	17%	21%	7%	5%	7%	4%
Morivione	Spanish	15%	24%	13%	5%	5%	3%
Chiaravalle	Norwegian	20%	19%	13%	4%	9%	1%

Implementation

We implemented this analysis in an online service (<http://urbanscope.polimi.it/>, section “Cities into Cities”), composed by two distinct components: the backend services retrieving the data from social network, storing it and executing the analysis, implemented with node.js and MongoDB; and a front-end portal that accesses the analysis results through REST API calls and allows visitors to visualize and explore the results. For the map, we use mapbox.js library and data from OpenstreetMap. Regarding the data/GUI binding, the Angular.js, HTML5, and Bootstrap frameworks have been adopted.

Conclusions

The aim of this work was to reveal language communities in correspondence to the neighborhoods of a large city like Milan, using Twitter and comparing the results with official census data. The discovered divergence between the prevalent Twitter language and the census nationality in a NIL clearly points out weaknesses and strengths of the two data sources supporting the need for their fruitful integration. For example, Twitter is able to intercept part of tourists, business visitors, commuters and residents. Twitter data are time-varying while census data are static (at least within each year). Twitter data are biased by the different adoption rates of different communities and population segments (age, sex, job, ...) and by the tendency of using widely spread languages in place of the native one. Vice-versa, census data are biased by the presence of not recorded residents (e.g., students).

Threats to validity. The results obtained in this work are affected by all the inherent limitations of Twitter, such as the different adoption rates of different populations, together with possible limited access to the data due to the use of the public API. Our study does not differentiate between residents and tourists/visitors and the results might be biased toward the most common language in the case that users want to reach a wider audience. Some NILs remained out of the results of our analysis because they were identified as “transparent” (not enough data available). While we accept the commonly used division into NILs applied to the city of Milan, a “bottom-up” approach, such as the head/tail breaks described by Jiang (2013), could be used in order to underline “natural cities” within the city that emerge by looking at the language distribution.

References

Cairncross, F. 2001. *The death of distance: How the communications revolution is changing our lives.* Harvard Business School Press.

Cenoz, J. N. Benefits of Linguistic Diversity and Multilingualism. Sustainable Development in a Diverse World (SUS. DIV) position paper, Research Task, 1.

Ceron, A. C. 2014. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society* 16(2): 340-358.

Crampton, T. 2011. Social Media in China: The Same, but Different. <http://www.chinabusinessreview.com/>.

Deville, P. L. 2014. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences* 111(45): 15888-15893.

Edwards, D. &. 2005. Developing a Municipal Performance-Measurement System: Reflections on the Atlanta Dashboard. *Public Administration Review* 65(3): 369-376.

Extra, G. &. 2004. *Urban multilingualism in Europe. Immigrant minority languages at home and school.* Clevedon: Multilingual Matters.

García, O. &. 2001. *The multilingual apple: languages in New York City.* Walter de Gruyter.

Gullino, S. 2009. Urban regeneration and democratization of information access: CitiStat experience in Baltimore. *Journal of Environmental Management* 90(6): 2012-2019.

Hong, L. C. 2011. Language Matters In Twitter: A Large Scale Study. In AAAI ICWSM.

Jiang, B. 2013. Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. *The Professional Geographer* 65(3): 482-494.

Kaplan, R. S. 1992. The Balanced Scorecard: Measures That Drive Performance. *Harvard Business Review* 70(1): 71-79.

Kim, S. W. 2014. Sociolinguistic analysis of twitter in multilingual societies. In 25th AMC conf. on Hypertext and Social Media, 243-248.

Kitchin, R. 2014. The real-time city? Big data and smart urbanism. *GeoJournal* 79(1): 1-14.

Lee, D. F. 2015. CityEye: Real-time Visual Dashboard for Managing Urban Services and Citizen Feedback Loops.

Leimgruber, J. 2013. The management of multilingualism in a city-state. *Multilingualism and Language Diversity in Urban Areas: Acquisition, identities, space, education*, 227-256.

Mattern, S. 2015. History of the Urban Dashboard. *Places Journal*.

Mocanu, D. B. 2013. The twitter of babel: Mapping world languages through microblogging platforms. *PloS one*, 8(4): e61981.

Quercia, D., Sáez-Trumper, D. 2014. Mining Urban Deprivation from Foursquare: Implicit Crowdsourcing of City Land Use. *IEEE Pervasive Computing* 13(2): 30-36

Sanderson, M. R. 2015. Are world cities also world immigrant cities? An international, cross-city analysis of global centrality and immigration. *International Journal of Comparative Sociology* 56(3-4): 173-197.

Takhteyev, Y. G. (2012). Geography of Twitter networks. *Social networks* 34(1): 73-81.