



Unveiling patterns of international communities in a global city using mobile phone data

Paolo Bajardi¹, Matteo Delfino², André Panisson², Giovanni Petri² and Michele Tizzoni^{2*}

*Correspondence:
michele.tizzoni@isi.it
²ISI Foundation, via Alassio 11/C,
Torino, 10126, Italy
Full list of author information is
available at the end of the article

Abstract

We analyse a large mobile phone activity dataset provided by Telecom Italia for the *Telecom Big Data Challenge* contest. The dataset reports the international country codes of every call/SMS made and received by mobile phone users in Milan, Italy, between November and December 2013, with a spatial resolution of about 200 meters. We first show that the observed spatial distribution of international codes well matches the distribution of international communities reported by official statistics, confirming the value of mobile phone data for demographic research. Next, we define an entropy function to measure the heterogeneity of the international phone activity in space and time. By comparing the entropy function to empirical data, we show that it can be used to identify the city's hotspots, defined by the presence of points of interests. Eventually, we use the entropy function to characterize the spatial distribution of international communities in the city. Adopting a topological data analysis approach, we find that international mobile phone users exhibit some robust clustering patterns that correlate with basic socio-economic variables. Our results suggest that mobile phone records can be used in conjunction with topological data analysis tools to study the geography of migrant communities in a global city.

Keywords: mobile phone data; entropy; urban geography

1 Introduction

A city is a complex system shaped by the continuously evolving interactions among its inhabitants and visitors. The recent availability of a profusion of data generated by human behavior and collected through pervasive sensors has motivated a strong and renewed interest for the study of urban spaces, building upon an established research tradition [1, 2]. Indeed, the new possibilities opened up by the current technological advances in data gathering and processing have led to the definition of a new age for the *Science of Cities* [3].

Nowadays, through the analysis of large amounts of real-time user-generated data it is possible to obtain new insights into the living conditions of a city and understand its transformations. Recently, such research has gained great popularity leading to a number of studies that leverage on urban data to monitor a city's life and structure. In particular, a major finding in this field has been the discovery that many diverse socio-economic properties of cities, from personal income to gasoline consumption, follow a power law

function of population size with scaling exponents that fall into distinct universality classes [4–8]. Analogously, similar scaling relations have been shown to hold for the social connectivity of individuals in a city [9]. Such empirical evidence has further motivated the development of several theoretical attempts to explain the observed regularities of cities [10–13] paving the way to a - yet to come - scientific theory of cities [14].

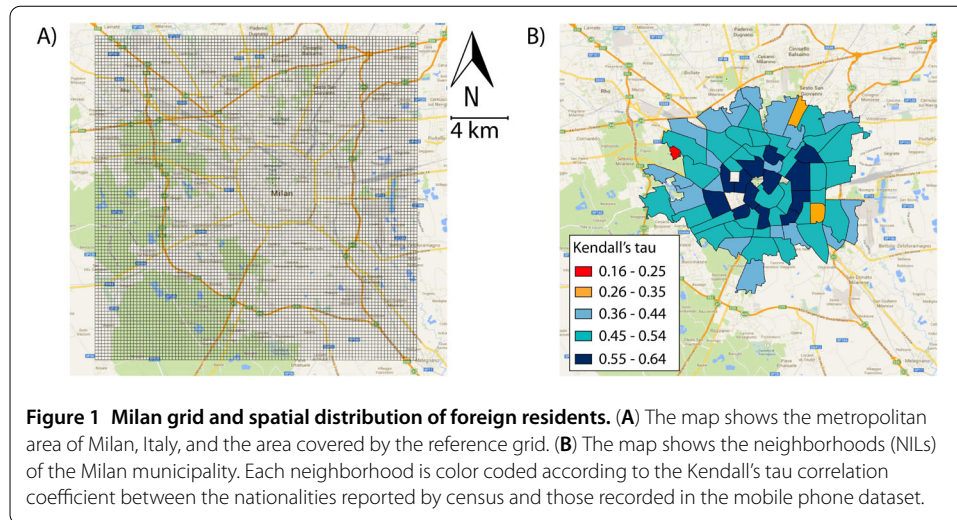
Within this context, studies of human behavior in the urban space have been strongly promoted by the availability of mobile phones call detail records (CDRs) of millions of users in different countries. Mobile phone data have been shown to be an invaluable source of information to extract individual mobility patterns at a high spatial resolution [15, 16]. Using mobile phone data, it has been possible to study migratory and commuting movements, both in industrialized [17] and developing countries [18], with applications ranging from disaster management [19, 20] to infectious disease epidemiology [21–23]. Moving beyond the analysis of human movements, recent works have shown that mobile phone data can be useful to study metropolitan land use [24–28] and, more in general, can provide some quantitative measures to classify cities in terms of their polycentrism, by identifying a city's hotspots and monitoring their stability over time [29].

In this paper, we deal with the study of urban spaces through the analysis of mobile phone records by focusing on a specific subset of mobile phone users, that is those with strong international links, which typically characterise short and long term migrant communities and those travelling to a city as tourists or for work. More precisely, we analyse a high-resolution dataset of mobile phone records collected in Milan, which has the most populous metropolitan area in Italy, that reports the international code of all calls and text messages sent and received by a user over a two-month period. Based on such call records, collected at a very high spatial and temporal resolution, we define an entropy function that measures the level of heterogeneity in the number of distinct countries - identified by their international phone code - that are present in the activity record of a city's neighborhood. By analysing the spatial and temporal features of such entropy function, we aim to identify locations that attract the most visitors and characterize how international communities occupy and make use of the urban spaces.

In particular, here we show and discuss how an entropy measure of the phone activity can be used to: (i) recognize the most attractive locations for tourists and visitors in the city, as indicated by the presence of point of interests; (ii) characterize the spatial distribution of diverse communities, recognizing - in conjunction with information obtained from topological data analysis - the different use of the metropolitan area made by visitors and residents. Additionally, we discuss how our proposed method can be advantageous with respect to only using the volume of mobile phone activity for similar analyses.

The paper is structured as follows: in Section 2 we describe the main features of the dataset under study; in Section 3 we link the diversity arising from mobile phones activity aggregates to the population diversity of Milan; in Section 4 we introduce an entropy measure to quantify the diversity of mobile phones activity and present the results obtained by analysing its properties; in Section 5 we describe how a topological data analysis based on persistent homology is able to discriminate between resident and visiting foreigners in the city, once a spatial entropy-activity distribution has been properly defined. Finally, we conclude and discuss our results.

The Python code that has been used to perform the data analysis of this work is available at the following repository: <http://github.com/micheletizzoni/Bajardi-et-al-EPJDS-2015>.



2 Dataset description

The dataset has been provided by Telecom Italia in the context of the ‘Telecom Big Data Challenge’ [30] and contains the mobile phone activity records - including both SMS and calls made or received - of all users in the Milan area (Italy) during a time span of two months, from November 1 to December 31, 2013. Data is aggregated over a 10 minutes interval and is geolocalized on a grid of 100×100 squares covering the city metropolitan area with a resolution of about 235 meters (map shown in Figure 1A). Traffic volume data has been provided with the same arbitrary units for text messages and calls. In this way, the provider assured both quantities to be directly comparable.

In our study, we focused on a specific information provided with the dataset, that is, the list of international country phone codes that are present in the activity record of each cell, with their corresponding traffic volume, for every 10-minutes time slice. As expected, the vast majority (97% of the total traffic) of the mobile phone activity is exchanged between Italian phone numbers; there is however a relatively small but very heterogeneous fraction of calls involving foreign phone numbers. During the three-month period, 221 different country codes appeared in the activity record indicating the presence of many foreigners in the city. Indeed, Milan is a diverse city with a long immigration history and a number of communities of foreigners originating from several developing countries. Moreover, Milan is a popular destination for tourist and business reasons, with important cultural attractions and hosts the headquarters of several multinational companies (the Province of Milan alone accounts for about 10% of the national GDP).

3 Mapping the diversity of mobile phones activity

Following the above considerations, it is reasonable to argue that mobile phones data could be used as a proxy to measure the diversity of the resident population living in the city without having to resort to census data. It is clear, however, that mobile phones data accounts also for individuals who are not permanent residents and, therefore, cannot be tracked by census. Furthermore, the estimation of the population diversity using mobile phones data will be biased by different usage patterns between different demographic groups [31].

Table 1 Ranked nationalities by census and by calling activity volume

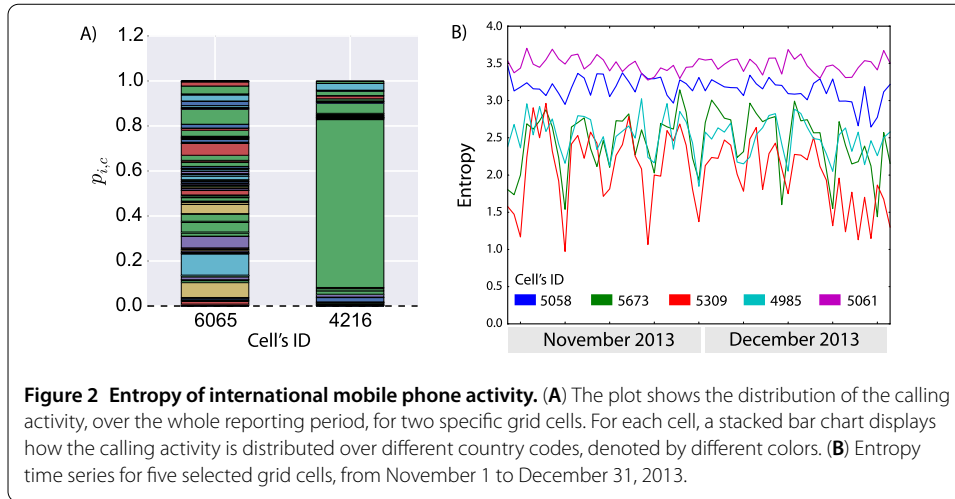
Rank	Census population	Calling activity
1	Philippines	Egypt
2	Egypt	Bangladesh
3	China	Switzerland
4	Peru	China
5	Sri Lanka	Ukraine
6	Ecuador	France
7	Romania	United Kingdom
8	Morocco	Sri Lanka
9	Ukraine	Philippines
10	Bangladesh	Romania
11	Albania	Russia
12	El Salvador	Germany
13	France	Spain
14	Brazil	Senegal
15	Moldova	Ecuador

To quantify the extent to which mobile phones data can be used as a proxy for estimating the diversity of the city population, we compared the list of countries ranked by their cumulative calling activity records with the nationalities of residents ranked by their prevalence as reported in the local census in the year 2012 [32]. As shown in Table 1, 9 out of the top 15 countries ranked by phone traffic volume are also present in the top nationalities according to the census data. The Kendall tau correlation between the ranked vectors of nationalities equals 0.69 ($p < 0.01$), showing a strong and significant correlation between the two rankings. To further investigate the correlation between mobile phones and census, we also compared the rankings of the different nationalities emerging from the two datasets at a finer spatial resolution.

Aggregating the cells at the NIL level (namely, *Nuclei di identità locale* - a subdivision of the urban area comprising 88 neighborhoods) and summing the whole traffic over the entire period of observation, we built for every NIL a vector of country-specific activities. Neighborhoods for which no census data was available (mostly parks and peripheral neighborhoods) were dropped, leaving a total of 66 NILs. We then evaluated for each NIL the Kendall's tau rank correlation coefficient between the sorted list of resident foreigner population and the sorted vector of mobile phones traffic from/to foreign countries. Each test led to statistically significant correlations ($p < 0.01$), with the exception of one NIL (the least populous one) for which no significant correlation was found between mobile phones and census rankings. For each of the remaining neighborhoods, as shown in Figure 1B, the Kendall's tau value was larger than 0.3. More specifically, it was larger than 0.4 for 55 (83%) of the NILs with median value 0.48 and 95% reference range [0.32-0.69], showing that mobile phones data can be used as a first order proxy to study the distribution of the resident foreign population even at a finer spatial resolution.

4 Entropy of mobile phone activity

Raw activity aggregates, although useful as shown in the previous section, suffer from noise and biases originating from different kinds of overlapping activities from residential, tourist and working centers. Furthermore, activity aggregates fail to convey information about the specificity and the diversity of the activity at the finest spatial and temporal levels.



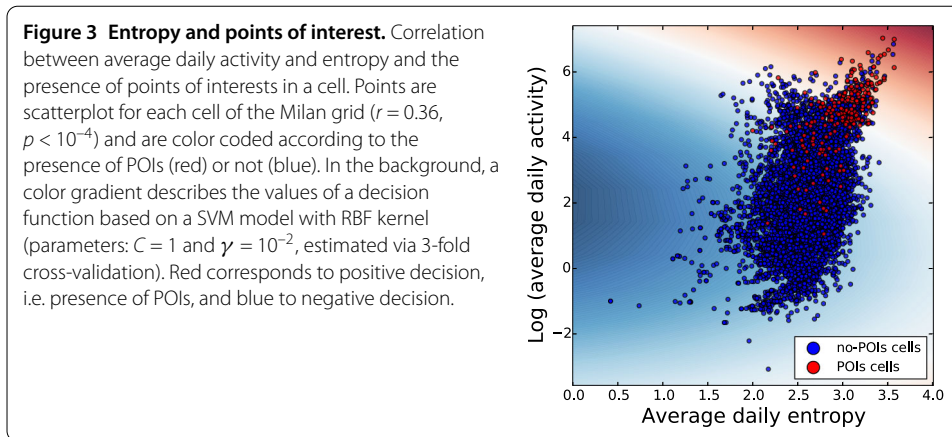
For these reasons, following previous approaches [33] we defined an entropy measure to quantify the heterogeneity of a single cell activity pattern and infer specific spatial and temporal characteristics of the corresponding urban area.

4.1 Definition

We define an entropy function on every cell of the Milan grid to measure the heterogeneity of the mobile phone activity in terms of the country codes that are present in the cell's temporal activity record. As a matter of notation, we represent the raw activity values as a time-dependent multi-dimensional matrix $A_{i,c}(t)$ where i runs over the grid ids, c refers to the country where the traffic was directed to or originating from, and t is the temporal dimension that can be tuned to different granularities (hourly, daily, etc.). Omitting the temporal dependency, we define the probability of observing a mobile phone activity related to a given country c in a cell i as $p_{i,c} = A_{i,c} / \sum_c A_{i,c}$ and thus the entropy as:

$$S_i = - \sum_c p_{i,c} \cdot \log(p_{i,c}). \quad (1)$$

As sketched in Figure 2A, a cell where the activity is evenly distributed among many countries will have a large entropy value suggesting that the phone traffic origin or destination is highly unpredictable, while if most of the activity is related to one or few countries only, the entropy will assume lower values. In the following we will use different temporal aggregations, considering both the activity over the whole period or on a daily basis, resulting in an overall entropy value for a given cell or in an entropy time series, respectively. Figure 2B shows some examples of daily entropy time series. From the plot it is possible to distinguish between two main qualitative behaviors of the entropy function: on the one hand, some cells display a high and steady entropy value over the course of the observed period (e.g. cells 5058 and 5061), on the other hand, some cells display a lower entropy value with distinct weekly pattern, where entropy gets smaller during weekends (e.g. cells 5309 and 4985). The first type of behavior is usually found in busy or central areas of the city, such as the central railway station, the Linate airport or in proximity of tourist attractions. The latter type of behavior is more typical of strictly residential and working areas, such as the University of Milan and the city outskirts.



4.2 Entropy and point of interests

A way to corroborate the hypothesis that high entropy areas can identify the busiest and most attractive locations is to compare the entropy distributions of different locations against an independent measure of their importance. To this aim, we considered the presence of point of interests (POIs) according to the popular travel website TripAdvisor (<http://www.tripadvisor.com>) as a proxy to measure a location's attractiveness. More precisely, we mapped all the locations that are listed as 'Attractions in Milan' by TripAdvisor [34] on the georeferenced grid and measured the average daily entropy of the cells where at least 1 point of interest was located. We found that the two quantities, number of POIs in a cell and its average daily entropy, are positively correlated (Spearman's correlation coefficient $r = 0.36$, $p < 10^{-4}$). Furthermore, the average daily entropy distribution of the cells without points of interest and the entropy distribution of the cells with at least one point of interest are significantly different (2-sample KS test at 1% level), with the latter being skewed towards higher entropy values (see Figure S1 in the Additional file 1).

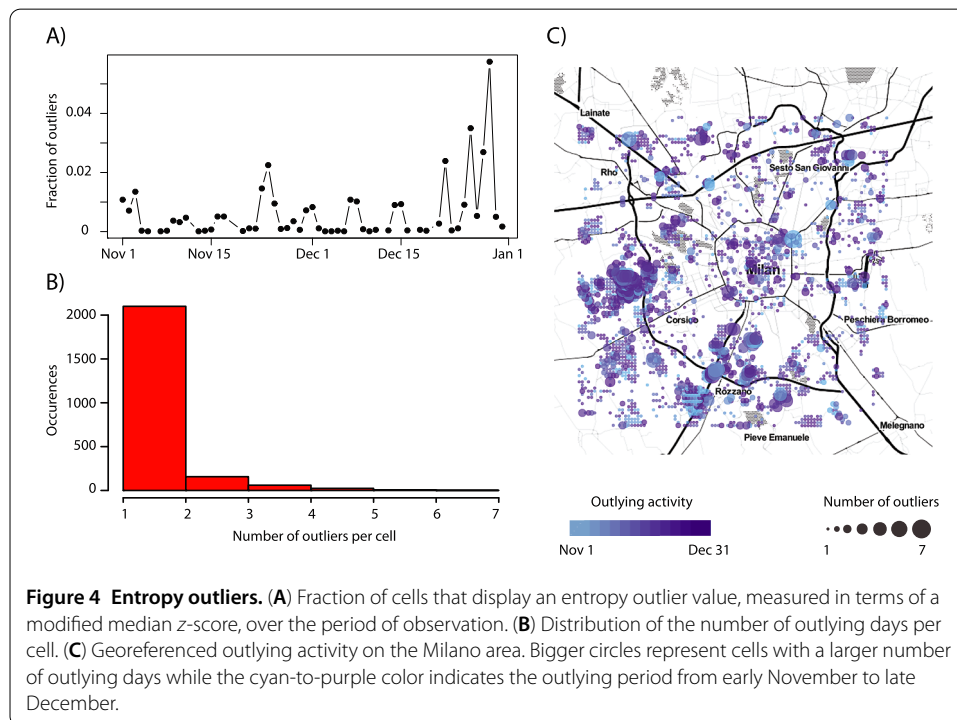
It is immediate to notice, however, that the cells having at least one point of interest are only 219 out of 10,000, and only 6 cells have 6 or more points of interest in their area (Figure S1A in the Additional file 1). Moreover, the average daily activity is also correlated with the presence of POIs in a cell (Spearman's correlation coefficient $r = 0.49$, $p < 10^{-4}$), indicating that attractive locations are also trivially characterized by a high calling volume. To examine how the entropy measure can add some useful information to this picture, we analysed the distributions of both daily activity and daily entropy in the presence or absence of points of interests. As shown in Figure 3, we found that the correlation between phone activity volume and entropy is not trivial and, more remarkably, most of the locations with points of interests (in red) are characterized by both high entropy and high activity and clustered in a relatively small region of the activity-entropy space. To provide a quantitative measure of the information added by the entropy, we trained a classification model based on a support vector machine (SVM) to automatically identify a cell with POIs. The values of the decision function are shown in Figure 3 as a background blue-to-red gradient. The decision boundary, determined by the line that separates the positive values (in red) from the negative values (in blue) of the decision function, is a nonlinear combination of both activity and entropy, and is able to classify with good accuracy the cells with attractive locations. Although the model's scores are not very relevant in this context since most of the false positives (blue points in the red area) are in fact cells adja-

cent to other attractive cells, we report that the results of a model trained with both activity and entropy are better (ROC AUC 0.93) than the results obtained by a model trained with just activity (ROC AUC 0.79). In the end, our results indicate that the entropy function of Equation (1) not only represents a reliable measure of a location attractiveness but could be effectively used to refine those methods that identify the hotspots of a city, based only on the raw activity volume.

4.3 Entropy outliers

Given the definition of entropy provided by Equation (1), for every cell i and for every time step t (in the following, a daily resolution is chosen) we are able to measure a certain entropy value and starting from the original dataset we recover 10,000 entropy time series. By calculating a modified z -score where mean and standard deviation are replaced by the median and the median absolute deviation values [35], we are able to systematically detect the days when the entropy profile shows a significantly different behavior than usual. An entropy outlier may be triggered by an unexpected increase or decrease in the mobile phone activity inwards/towards foreign locations. The rationale behind the automatic detection of outliers is to use the entropy level as a fingerprint to characterize an urban area, where the outlying points indicate some exceptional events occurring in that location.

In Figure 4A the fraction of detected outliers is shown for every day of observation. An overall pattern emerges: during the weekends and Christmas holidays, the number of cells with an exceptional activity profile increases. In Figure 4B we observe that less than one fourth of the cells has an unexpected activity during the two months period of observation, and among them, more than 70% shows only one outlier. On the other hand, there are 32 cells with 5 or more outlying days in terms of entropy profiles. Such cells are located either



in low-activity areas, which are more prone to entropy fluctuations, or close to important crossroads probably representative of a genuine special activity.

Beyond the overall patterns, we georeferenced the outlying activity cells and we mapped in Figure 4C both the median outlying day per cell and the number of outlying days per cell. The majority of darker cells - corresponding to late December - remarks the abundance of outlying activity close to the Christmas holidays, probably generated by the increasing presence of tourists and by the disruption of the typical workday patterns.

4.4 Monitoring large-scale events

Here, we demonstrate how it is also possible to leverage on the raw mobile phone's activity to monitor large-scale events in real-time. On December 11, 2013 an important football match between AC Milan and the Dutch team of AFC Ajax was hosted in Milan. Analysing the mobile phone data, we found some entropy outliers (see Figure S2 in the Additional file 1) close to the airport, the stadium and few other places around the city. In order to investigate the crowd displacements at a finer temporal scale, we focused on the raw mobile phones activity (sum of calls and sms) related to traffic from/to the Netherlands aggregated over 1-hour intervals. For every time interval, the most active 50 cells are shown on the map (Figure 5), color coded according to the recorded activity with an orange-to-blue gradient associated to low-to-high traffic. Since the mobile phone activity clearly follows strong circadian patterns, every snapshot is normalized on the hourly activity and therefore the maps describe the geolocated predominant presence of Dutch communication for every time window and the magnitude of such activity can not be directly compared among different maps.

Assuming that the majority of mobile phones activity from/to the Netherlands is linked to the presence of Dutch supporters, it is possible to monitor the geo-temporal movements of the crowd in the urban area. Before 6 am the recorded activity accounts for less than 4% of the total daily activity and the cells are quite scattered over the city with some hotspots close to the railway stations. At 6 am, the most active cells are those close to the Linate airport, Porta Genova and Central railway stations, thus registering the arrival of the first Ajax's supporters. At 7 am a large activity is also registered on the top-left corner of the Milano grid, suggesting that a group of supporters is probably reaching the city from the Malpensa airport that is located on that side of the region, as well as a sustained activity still registered at Linate and at the several railway stations. During the central hours of the day, most of the supporters were probably visiting the city center and indeed the activity is scattered over that area. Approaching the match at 6-7 pm, a strong activity is observed at the underground stations close to the stadium and the stadium itself. During the hour preceding the match, most of the activity is located at the stadium while a very localized hotspot is observed at the San Carlo Borromeo Hospital and another one close to the Niguarda Hospital: we learned from the news that six Ajax supporters had been stabbed in a fight with AC Milan hooligans [36]. During the football match (9-11 pm) the majority of the activity is observed at the stadium, as expected, and soon after the end of the match the supporters left the area and came back to the city center.

5 Topological data analysis of international communities

Previous work exploited individual geolocated digital traces [37] to study the different prominence and distribution of international communities within an English-speaking environment. The most intuitive proxy for migrant communities is the aggregated call and

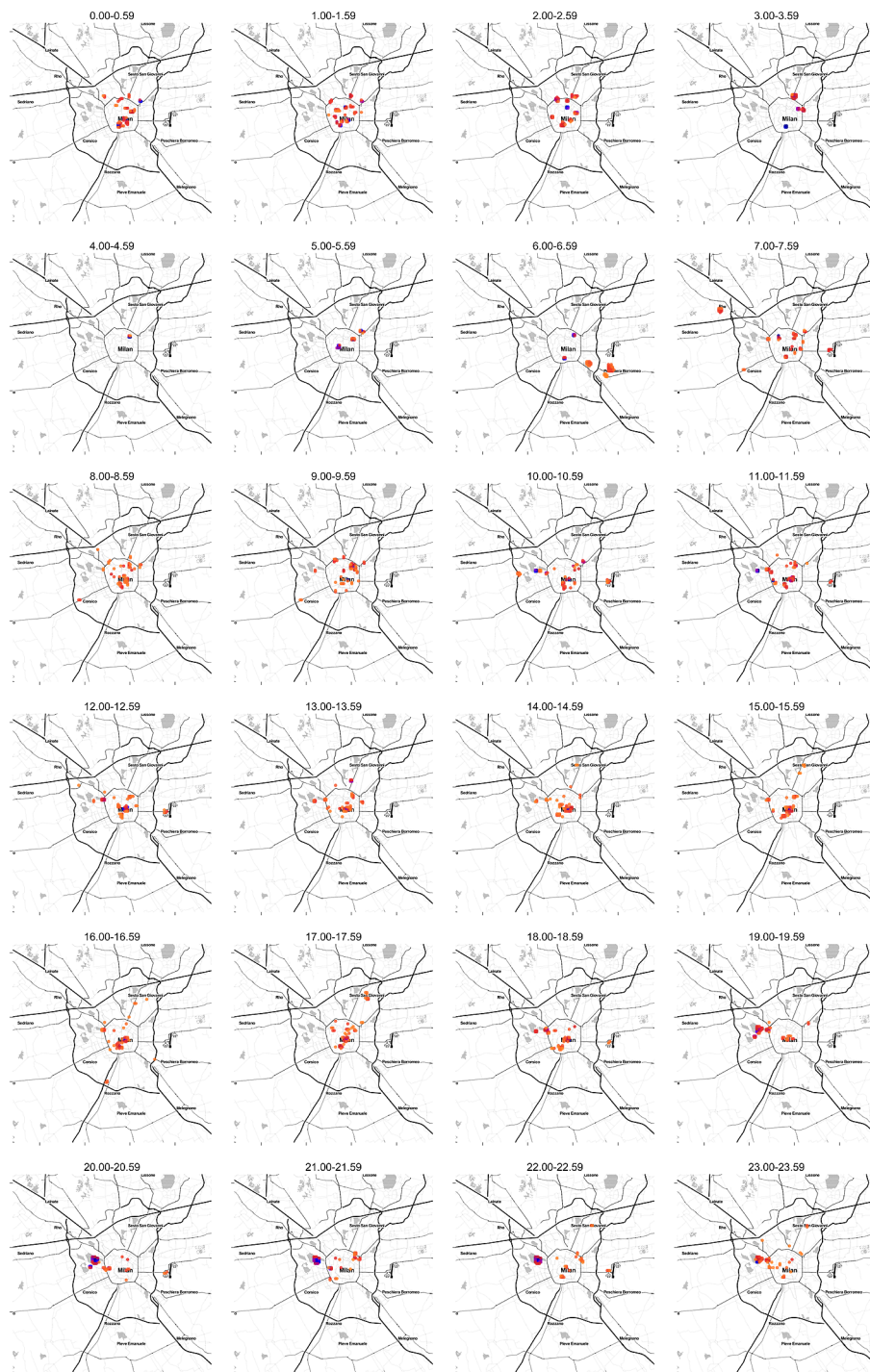
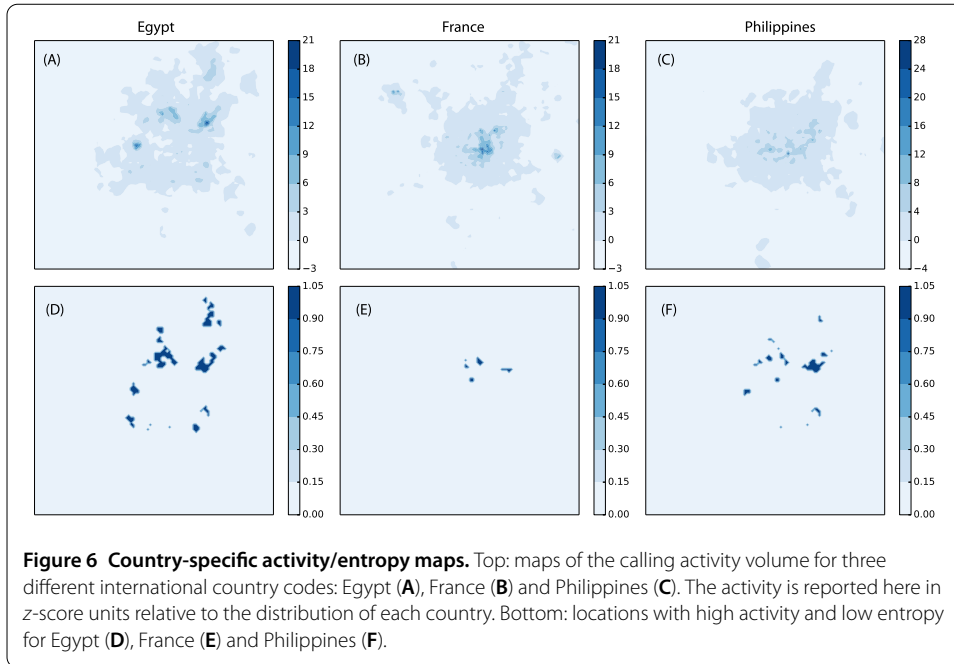


Figure 5 Phone activity during an international football match. Most active 50 cells of mobile phones traffic from/to the Netherlands on December 11, 2013. Orange-to-blue color gradient corresponds to low-to-high activity. Every hourly map is normalized by the corresponding total hourly activity.



sms activity of a cell to/from another country, the natural implication being that a large activity toward a country c highlights the presence of a substantial population of nationals of c . However, activity alone hardly produces informative results. In fact, the scale of activity can vary wildly across different cells and therefore a large total activity in a cell might drown relevant signals from others. Figure 6A helps illustrating this point. Despite some variability driven by the number of country-specific foreigners, we find a common pattern that highlights the importance of the city center. Maps of activity are indeed strongly biased by the busiest places in terms of tourist attraction, business, etc. This in turn underlines the second shortcoming of using activity alone: activity is not able to discern how specific a cell's activity is to a certain country. Entropy, on the contrary, quantifies exactly this disparity. A cell that can be reliably associated with a country should thus display a low entropy, which implies that most of its activity belongs to a small number of countries, and large activity toward the selected country.

For each country c , we considered only the cells i that had activity $A_{i,c} > \mu_{A_c} + \sigma_{A_c}$ and entropy $z_{S_i} < \bar{z}$, where μ_{A_c} and σ_{A_c} are the first and second moment of the distribution of activity restricted to country c over all cells, and z_{S_i} is the entropy z-score of the i th cell. In the rest of the paper we used a threshold value $\bar{z} = -2$, corresponding to an entropy that is 2 standard deviations smaller than the average entropy. We performed sensitivity testing by varying $\bar{z} \in [-3, -0.5]$ and we found that the patterns are robust under changes in \bar{z} .

Figure 6B shows that imposing these two constraints peculiar country-specific patterns emerge. Direct observation points to two main types of configurations:

- (single or multiple) spatially coherent clusters;
- spatially dispersed and annular clusters, with possible branching structures.

It is tempting to try to characterize quantitatively the similarity between configurations of different countries and possibly use them to cluster countries according to them. However, the intrinsic noise in the dataset and the variability in the configurations' shapes require robust shape descriptors that allow for stretching, rotations, thinning and translations.

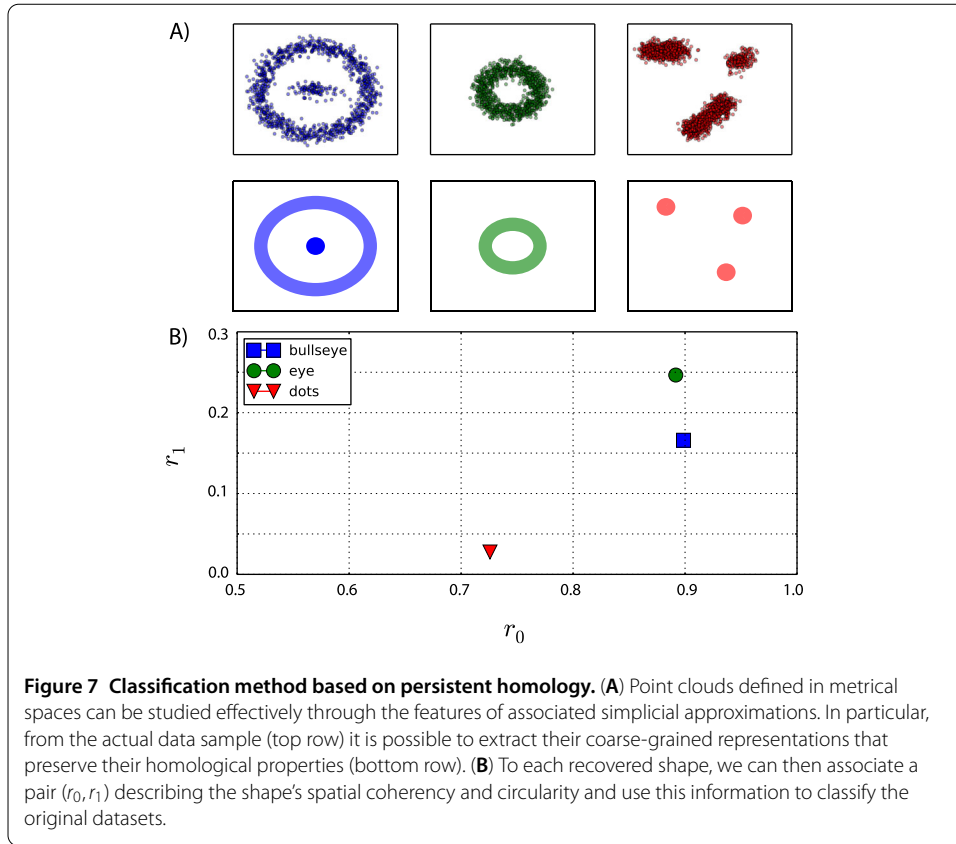


Figure 7 Classification method based on persistent homology. (A) Point clouds defined in metrical spaces can be studied effectively through the features of associated simplicial approximations. In particular, from the actual data sample (top row) it is possible to extract their coarse-grained representations that preserve their homological properties (bottom row). (B) To each recovered shape, we can then associate a pair (r_0, r_1) describing the shape's spatial coherency and circularity and use this information to classify the original datasets.

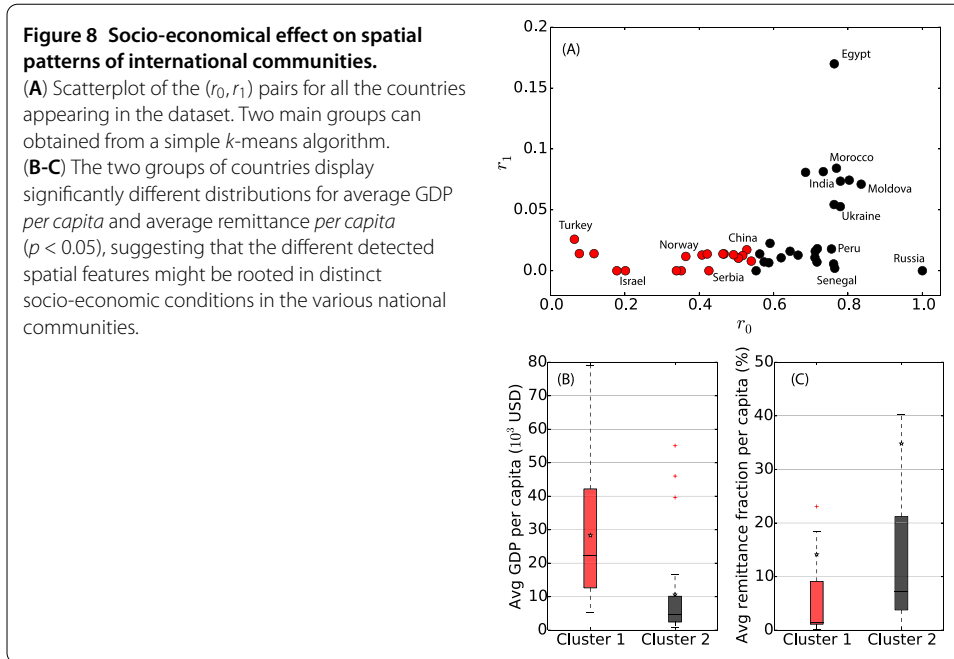
Topological features are one example of this kind of descriptors. Various recently appeared methods are able to characterize in a robust way the mesoscopic features of data spaces. They are based on ideas coming from computational topology and were introduced in the context of data analysis by [38–40]. They have been used in biology [41–44], brain functional networks [45], shape recognition [46–48], sensor network coverage [49] and complex networks [50, 51].

Figure 7A provides an ideal example of this type of shape classification: given a point cloud (Figure 7A top row), the features that are topologically detectable in two dimensions are its connectedness and circularity. In this way, it is possible to discern structures that are qualitatively different such as the three examples shown (Figure 7A bottom row). Persistent homology [38] - a parametrised version of homology for data spaces - provides the main tool to detect and quantify these differences between point clouds.

For a given data space, persistent homology yields a description of its homological structure in terms of the properties of the homology generators along a series of successive approximations of the real dataset. For our purposes however, it will be sufficient to focus on two simple derived quantities, r_0 and r_1 .

The former, r_0 , measures the spatial coherency of the point cloud, in terms of the difference between the point cloud's diameter and the radius of the point neighborhoods necessary to obtain a single connected component. We normalize r_0 by the diameter in order to obtain a value in $(0, 1)$.

The latter, r_1 , similarly conveys the large scale circularity of the point cloud by considering the ratio of the two largest circles obtained by gluing neighborhoods together. Since



the formal definition these quantities requires some mathematical technicalities, we refer the interested reader to Section 3 in the Additional file 1. However, it is easy to see that the (r_0, r_1) pairs calculated for the three examples are clearly distinct (Figure 7B) and correspond to intuitive description: the *bullseye* and the *eye* examples have larger spatial decoherence and circularity than the *dots* example, with the *eye* having the largest circularity, due to the *bullseye's* central cluster marring its symmetry.

We repeated the same analysis on entropy-activity filtered maps for all the countries, obtaining the complete list of (r_0, r_1) .

In line with the naive observations of Figure 6, we used a simple k -means clustering ($k = 2$) to single out the two emerging clusters: one containing countries characterised by low values of both ratios, the second containing those with large values of r_0 and varying values of r_1 . It is interesting to note that the two groups also appear to correspond to two different socio-economic demographics. We found indeed that countries corresponding to small r_0 and r_1 have significantly larger average GDP *per capita* (2-sample KS = 0.7, p -value < 0.05) and lower remittances ratios [52] (2-sample KS = 0.46, p -value < 0.05) as compared to countries corresponding to larger r_0 and r_1 values (Figure 8B and 8C), suggesting an interpretation of the detected spatial patterns in terms of the different types of migrations to Milan (e.g. low income labour migration versus cultural migration). Moreover, the results do not seem to be sensitive to different choice of the entropy threshold, since we found that the results are robust for a range of choices of z -threshold value ($\bar{z} \in [-3, -0.5]$) as shown in Figures S3 and S4 of the Additional file 1. Our approach provides a basis for further data-driven and time-resolved analyses of the factors that shape the spatial properties of migrant communities in a city, in addition to the standard macroscopic push-pull description [53].

6 Discussion

In the past few years, the availability of detailed user-generated data has unveiled a tremendous potential for the development of computational social sciences [54]. A fundamental contribution to this research field has derived from the analysis of mobile phones data, which has been proved to be an important source of information to study how people move [15, 55] and interact [56, 57], especially in a urban environment [24, 58–60].

In this work, we have studied a highly resolved dataset containing the international country codes of millions of anonymized and geo-referenced records of calls that have been made or received during a two-month period in the city of Milan, Italy. We focused our attention on the heterogeneity of the international codes appearing in the dataset in order to extract meaningful information about the spatial distribution and the land use related to the presence of migrant communities or tourist hubs in the city.

First, by comparing the frequency and the spatial aggregation of the international country codes to official census data, we have shown that mobile phones data can be used as a proxy for estimating the ranking of foreign residents of Milan up to the NILs level. This result generally confirms the value of mobile phones as a tool for demographic research [61, 62] and suggests the possibility to analyse mobile phone traffic to infer the population distribution of foreigners with a higher temporal and spatial resolutions than those typical of census surveys. In this respect, the advantage of using mobile phone data would be the relatively low cost compared to traditional expensive and time consuming surveys. On the other hand, mobile phone data can not be assumed to be fully representative of the population because there are differences in the use of mobile phones across demographic groups [31]. Similar biases arise from the fact that mobile phones include individuals who are not permanent residents. A solution to overcome the biases related to such issues could be mixing the analysis of large-scale mobile phone datasets with the collection of small-sample surveys [61].

The use of an entropy measure to quantify the diversity of behavioral patterns and social interactions and the analysis of its relations with socio-demographic attributes has been the subject of previous studies [33, 63]. Here, we used it as a complementary measure to characterize the metropolitan spaces based on the heterogeneity of the country codes that appear in the calling records. Previous works showed that a larger diversity of social interactions was linked to high socio-economic levels [33]. Here, we found that both high entropy and high calling activity identify the most attractive locations of a city. In particular, through a machine learning algorithm, we showed how the entropy function can be used to distinguish between those places that have a large calling volume but are not points of interests and those that are attractive for international visitors. This result suggests that the entropy function could be used in conjunction with more sophisticated methods [29] to quantify how much foreigners contribute to the importance of a city's hotspot, distinguishing between locations that attract the natives and those that attract visitors. Moreover, our findings emphasize the fact that phone activity alone is not completely informative in the case of hotspots recognition. On the other hand, while the entropy is able to provide insights into the use of urban spaces, we also showed that mobile phone activity can be effectively used to monitor in real-time city-scale mass events. In this context, our results confirm previous works where mobile phones have been used to infer the crowd density during mass gatherings [64]. In our case, we demonstrated how mobile phone data can provide very specific spatial and temporal trajectories

of visitors from a given country (the Netherlands, in our case) during a mass gathering event. Similar situations are typical of large sport, artistic and religious events, whose management requires large coordinated efforts. The analysis of the football match also hinted at the possible use of entropy outliers to automatically detect unusual events. This is, however, only an hypothesis that needs further testing on a larger dataset to be verified.

The proposed characterization of different urban areas by means of an entropy function provides a helpful description of the locations where a large diversity of people tend to gather together. In our work, we outlined two distinct behaviors that can identify either tourist places, where many point of interests attract different foreigners, or residential areas where migrant communities are well consolidated. By using a topological classification based on persistent homology, we uncovered some significant regularities in the shape of foreign communities, as identified by a combination of low-entropy and high-activity cells. When described in terms of their topological indices, the nationalities associated to the call's destinations and sources outside Italy are clustered into two main groups. The first comprises low-income countries, whose topological spatial pattern displays a strong cyclic spatial distribution. The second group is formed by higher income countries, whose spatial distribution is scattered in small areas over the city. These results indicate that migrant communities from low income labour countries tend to aggregate in cohesive spatial structures that populate the city's residential areas, mainly scattered around the city centre. On the other hand, communities associated with higher income countries tend to reflect mostly tourist movements or highly specialized labour associated with central, high-entropy urban areas. In both cases, their spatial patterns are expected to be sparser and less structured. Our findings are in the direction predicted by the spatial assimilation theory [65] and confirm the empirical observation that different socio-economic migrant conditions can show distinct spatial clustering patterns [66]. In this context, our work opens the way to a completely new approach to the geography of immigrant communities within a city, based on topological data analysis, which could help to uncover a more refined classification for such communities.

Further research is needed to fully explore the potential applications of the proposed method and validate it. First of all, our work is limited by the fact that the dataset is restricted to one city. Thus, an extension of our analysis to different cities, both in developed and developing countries, is needed. Even if strong regularities are observed across cities [5], each city has its own peculiar structure, such as being monocentric or polycentric [67], and we may expect such features to affect the spatial distribution of the entropy function. Also, the attractiveness for foreigners may significantly vary across cities in the same country, leading to a more or less important contribution of foreign country codes to the mobile phone activity. In general, we expect our results to hold for large and diverse metropolitan areas, as in the case of Milan. Finally, our work is affected by the limited period over which data has been collected, because a two-month record is not sufficient to uncover long-term trends in the activity of users. In particular, a longer time frame is necessary to better define the presence of outliers in the entropy distribution and possibly link them with the recognition of unusual events.

Additional material

Additional file 1: Supplementary information. Correlation between entropy and points of interest. Entropy outliers during the football match Ajax-Milan. Mathematical details of the persistent homology. Sensitivity analysis on the clustering of international communities.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PB, MD, AP, GP and MT conceived and designed the study, carried out the data analysis and drafted the manuscript. All authors read and approved the final manuscript.

Author details

¹Aizoon Technology Consulting, via Po 14, Torino, 10123, Italy. ²ISI Foundation, via Alassio 11/C, Torino, 10126, Italy.

Acknowledgements

The authors acknowledge Telecom Italia for providing the original dataset. The work has been partially supported by the EC FET-Proactive Project MULTIPLEX (Grant No. 317532) to MT and the EC FET-Proactive Project TOPDRIM (Grant No. 318121) to GP. We gratefully acknowledge Marco Quaggiotto for help with GIS mapping and Joanna Cordero for helpful comments.

Received: 2 October 2014 Accepted: 13 April 2015 Published online: 29 April 2015

References

- Batty M, Longley P (1994) *Fractal cities*. Academic Press, San Diego
- Fujita M, Krugman PR, Venables AJ (1999) *The spatial economy: cities, regions and international trade*. MIT Press, Cambridge
- Batty M (2013) *The new science of cities*. MIT Press, Cambridge
- Newman PW, Kenworthy JR (1989) Gasoline consumption and cities: a comparison of US cities with a global survey. *J Am Plan Assoc* 55:24-37
- Bettencourt LM, Lobo L, Helbing D, Kühn C, West GB (2007) Growth, innovation, scaling, and the pace of life in cities. *Proc Natl Acad Sci USA* 104(17):7301-7306
- Samaniego H, Moses ME (2008) Cities as organisms: allometric scaling of urban road networks. *J Trans Land Use* 1:21-39
- Rozenfeld HD, Rybski D, Andrade JS, Batty M, Stanley HE, Makse HA (2008) Laws of population growth. *Proc Natl Acad Sci USA* 105(48):18702-18707
- Bettencourt LM, Lobo L, Strumsky D, West GB (2010) Urban scaling and its deviations: revealing the structure of wealth, innovation and crime across cities. *PLoS ONE* 5(11):13541
- Schläpfer M, Bettencourt LM, Grauwin S, Raschke M, Claxton R, Smoreda Z, West GB, Ratti C (2014) The scaling of human interactions with city size. *J R Soc Interface* 11:20130789
- Makse HA, Havlin S, Staneley HE (1995) Modelling urban growth. *Nature* 377:608-612
- Bettencourt LM, West GB (2010) A unified theory of urban living. *Nature* 467:912-913
- Pan W, Ghoshal G, Krumme C, Cebrian M, Pentland A (2013) Urban characteristics attributable to density-driven tie formation. *Nat Commun* 4:1961
- Bettencourt LM (2013) The origin of scaling in cities. *Science* 340(6139):1438-1441
- Louf R, Barthélemy M (2014) How congestion shapes cities: from mobility patterns to scaling. *Sci Rep* 4:5561
- González MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. *Nature* 453(7196):779-782
- Song C, Koren T, Wang P, Barabási A-L (2010) Modelling the scaling properties of human mobility. *Nat Phys* 6(10):818-823
- Schneider CM, Belik V, Couronné T, Smoreda Z, González MC (2013) Unraveling daily human mobility motifs. *J R Soc Interface* 10:20130246
- Amini A, Kung K, Kang C, Sobolevsky S, Ratti C (2014) The impact of social segregation on human mobility in developing and industrialized regions. *EPJ Data Sci* 3:6
- Lu X, Bengtsson L, Holme P (2012) Predictability of population displacement after the 2010 Haiti earthquake. *Proc Natl Acad Sci USA* 109(29):11576-11581
- Bengtsson L, Lu X, Garfield R, von Schreeb J (2011) Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. *PLoS Med* 8:1001083
- Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, Buckee CO (2012) Quantifying the impact of human mobility on malaria. *Science* 338(6104):267-270
- Le Menach A, Tatem AJ, Cohen JM, Hay SI, Randell H, Patil AP, Smith DL (2011) Travel risk, malaria importation and malaria transmission in Zanzibar. *Sci Rep* 1:93
- Tizzoni M, Bajardi P, Decuyper A, Kon Kam King G, Schneider CM, Blondel V, Smoreda Z, Gonzalez M, Colizza V (2014) On the use of human mobility proxies for modeling epidemics. *PLoS Comput Biol* 10(7):1003716
- Ratti C, Williams S, Frenchman D, Pulselli R (2006) Mobile landscapes: using location data from cell phones for urban analysis. *Environ Plan B, Plan Des* 33(5):727-748
- Toole JL, Ulm M, González MC, Bauer D (2012) Inferring land use from mobile phone activity. In: *Proceedings of the ACM SIGKDD international workshop on urban computing*. ACM, New York, pp 1-8

26. Pei T, Sobolevsky S, Ratti C, Shaw S, Li T, Zhou C (2014) A new insight into land use classification based on aggregated mobile phone data. *Int J Geogr Inf Sci* 28:1-20
27. Becker RA, Caceres R, Hanson K, Loh JM, Urbanek S, Varshavsky A, Volinsky C (2011) A tale of one city: using cellular network data for urban planning. *IEEE Pervasive Comput* 10:18-26
28. Grauwlin S, Sobolevsky S, Moritz S, Gódor I, Ratti C (2014) Towards a comparative science of cities: using mobile traffic records in New York, London and Hong Kong. arXiv:1406.4400
29. Louail T, Lenormand M, Cantu Ros OG, Picornell M, Herranz R, Frias-Martinez E, Ramasco JJ, Barthélemy M (2014) From mobile phone data to the spatial structure of cities. *Sci Rep* 4:5276
30. Open Data Institute Trento: Telecom Big Data Challenge. <http://theodi.fbkc.eu/openbigdata/>
31. Wesolowski A, Eagle N, Noor AM, Snow RW, Buckee CO (2012) Heterogeneous mobile phone ownership and usage patterns in Kenya. *PLoS ONE* 7(4):35319
32. Comune di Milano: Milan municipality open data. <http://dati.comune.milano.it/dato/item/27-27-popolazione-residente-in-anagrafe-per-cittadinanza-e-quartiere-anni-1999-2011.html>
33. Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. *Science* 328(5981):1029-1031
34. TripAdvisor: Things to do in Milan. http://www.tripadvisor.com/Attractions-g187849-Activities-Milan_Lombardy.html
35. Iglewicz B, Hoaglin D (1993) How to detect and handle outliers. *The ASQC basic references in quality Control. Statist Techn* 16
36. Jones G (2013) Three Ajax fans stabbed during fighting before Milan game. <http://uk.reuters.com/article/2013/12/11/uk-soccer-champions-milan-violence-idUKBRE9BA13I20131211>
37. Mocanu D, Baronchelli A, Perra N, Gonçalves B, Zhang Q, Vespignani A (2013) The twitter of babel: mapping world languages through microblogging platforms. *PLoS ONE* 8(4):61981
38. Edelsbrunner H, Harer J (2008) Persistent homology - a survey. *Contemp Math* 453:257-282
39. Carlsson G (2009) Topology and data. *Bull Am Math Soc* 46:255-308
40. Tausz A, Carlsson G (2011) Applications of zigzag persistence to topological data analysis. Arxiv preprint. arXiv:1108.3545
41. Yao Y, Sun J, Huang X, Bowman GR, Singh G, Lesnick M, Guibas LJ, Pande VS, Carlsson G (2009) Topological methods for exploring low-density states in biomolecular folding pathways. *J Chem Phys* 130(14):144115
42. Chan JM, Carlsson G, Rabadan R (2013) Topology of viral evolution. *Proc Natl Acad Sci USA* 110(46):18566-18571. doi:10.1073/pnas.1313480110
43. Singh G, Memoli F, Ishkhanov T, Sapiro G, Carlsson G, Ringach DL (2008) Topological analysis of population activity in visual cortex. *Journal of vision* 8(8):11
44. Nicolau M, Levine AJ, Carlsson G (2011) Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc Natl Acad Sci USA* 108(17):7265-7270. doi:10.1073/pnas.1102826108
45. Lee H, Chung MK, Kang H, Kim BN, Lee DS (2011) Discriminative persistent homology of brain networks. *IEEE international symposium on biomedical imaging: from nano to macro, 2011*, pp 841-844
46. Singh G, Memoli F, Carlsson GE (2007) Topological methods for the analysis of high dimensional data sets and 3D object recognition. *SPBG*
47. Lum PY, Singh G, Lehman A, Ishkhanov T, Vejdemo-Johansson M, Alagappan M, Carlsson J, Carlsson G (2013) Extracting insights from the shape of complex data using topology. *Sci Rep* 3:1236
48. Silva V, Morozov D, Vejdemo-Johansson M (2011) Persistent cohomology and circular coordinates. *Discrete Comput Geom* 45(4):737-759
49. De Silva V, Ghrist R (2007) Coverage in sensor networks via persistent homology. *Algebr Geom Topol* 7:339-358
50. Horak D, Maletič S, Rajkovič M (2009) Persistent homology of complex networks. *J Stat Mech* 2009(03):03034
51. Petri G, Scolamiero M, Donato I, Vaccarino F (2013) Topological strata of weighted complex networks. *PLoS ONE* 8(6):66506
52. Personal remittances, received (% of GDP). Available at <http://data.worldbank.org/indicator/BX.TRF.PWKR.DT.GD.ZS>
53. Arango J (2000) Explaining migration: a critical view. *Int Soc Sci J* 52(165):283-296
54. Lazer D, Pentland A, Adamic LA, Aral S, Barabási A-L, Brewer D, Christakis NA, Contractor N, Fowler JH, Gutmann M, Jebara T, King G, Macy M, Roy D, Van Alstyne M (2009) Life in the network: the coming age of computational social science. *Science* 323(5915):721-723
55. Isaacman S, Becker R, Caceres R, Kobourov S, Martonosi M, Rowland J, Varshavsky A (2011) Identifying important places in people's lives from cellular network data. In: Lyons K, Hightower J, Huang EM (eds) *Pervasive computing, 9th international conference, pervasive 2011, San Francisco, USA, June 12-15, 2011. Proceedings. Lecture notes in computer science*. Springer, Berlin, pp 133-151
56. Eagle N, Pentland A, Lazer D (2009) Inferring friendship network structure by using mobile phone data. *Proc Natl Acad Sci USA* 106:15274-15278
57. Calabrese F, Smoreda Z, Blondel VD, Ratti C (2011) Interplay between telecommunications and face-to-face interactions: a study using mobile phone data. *PLoS ONE* 6(7):20814
58. Calabrese F, Colonna M, Lovisollo P, Parata D, Ratti C (2011) Real-time urban monitoring using cell phones: a case study in Rome. *IEEE Trans Intell Transp Syst* 1(12):141-151
59. Reades J, Calabrese F, Sevtsuk A, Ratti C (2007) Cellular census: exploration in urban data collection. *IEEE Pervasive Comput* 7(3):30-38
60. Reades J, Calabrese F, Ratti C (2009) Eigenplaces: analysing cities using the time-space structure of the mobile phone network. *Environ Plan B, Plan Des* 36:824-836
61. Palmer JRB, Espenshade TJ, Bartumeus F, Chung CY, Ozgencil NE, Li K (2013) New approaches to human mobility: using mobile phones for demographic research. *Demography* 50(3):1105-1128
62. Deville P, Linard C, Martin S, Gilbert M, Stevens FR, Gaughan AE, Blondel VD, Tatem AJ (2014) Dynamic population mapping using mobile phone data. *Proc Natl Acad Sci USA* 111(45):15888-15893
63. Smith C, Quercia D, Capra L (2013) Finger on the pulse: identifying deprivation using transit flow analysis. In: *Proceedings of the 2013 conference on computer supported cooperative work*. ACM, New York, pp 683-692

64. Wirz M, Franke T, Roggen D, Mitleton-Kelly E, Lukowicz P, Tröster G (2013) Probing crowd density through smartphones in city-scale mass gatherings. *EPJ Data Sci* 2:5
65. Massey DS (1985) Ethnic residential segregation: a theoretical synthesis and empirical review. *Sociol Soc Res* 69(3):315-350
66. Pamuk A (2004) Geography of immigrant clusters in global cities: a case study of San Francisco. *Int J Urban Reg Res* 28(2):287-307
67. Louf R, Barthelemy M (2013) Modeling the polycentric transition of cities. *Phys Rev Lett* 111:198702

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Immediate publication on acceptance
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ springeropen.com
