A Comparative Evaluation of Urban Fabric Detection Techniques Based on Mobile Traffic Data

Angelo Furno, Razvan Stanica, Marco Fiore

A Comparative Evaluation of Urban Fabric Detection Techniques Based on Mobile Traffic Data

Angelo Furno[1], Razvan Stanica[1] and Marco Fiore[2]

[1] *Université de Lyon, INRIA, INSA-Lyon, CITI-INRIA*
F-69621, Villeurbanne, France
Email: name.surname@insa-lyon.fr
[2] *CNR – IEIIT*
Corso Duca degli Abruzzi 24, 10129 Torino, Italy
Email: marco.fiore@ieiit.cnr.it

Mobile traffic data has been recently used to characterize the urban environment in terms of urban fabric profiles. While showing promising results, the existing urban fabric detection solutions are built without a clear understanding of the detection process chain. In this paper, we distinguish and analyze the different steps common to all urban profiling techniques. By evaluating the impact of each step of the process, we are able to propose a new solution that outperforms the state of the art techniques. Our approach uses the weekly periodicity of human activities, as well as a median-based filtering technique, resulting in a better clustering in terms of both coverage and entropy, as shown by results obtained on two large scale mobile traffic datasets covering the urban areas of Milan and Turin, in Italy.

Keywords: mobile traffic data, time series analysis, urban fabric detection, land use classification

1 Introduction

The surge of mobile traffic, at a staggering 146% compound annual growth rate during the past decade [1], is a clear evidence of the success of mobile communication technologies, with similar trends in developed and developing countries [2].

The popularity of mobile devices makes the customer base of individual mobile operators a statistically reliable sample group for macroscopic studies across different research disciplines. This has been repeatedly proven in the literature, where data captured by mobile network probes and containing fine-grained information on the activities of millions of individuals have been leveraged for many and varied purposes. Significant examples include the estimation of country-wide population distribution [3] and mobility patterns [4, 18], the identification of ethnolinguistic groups [5, 6], or the inference of economic level of geographical regions [7]. We refer the interested reader to the recent surveys in [8, 9] for a thorough discussion of these and more topics.

In this paper, we focus on one specific application of mobile traffic analysis, i.e., the detection of *urban fabrics*. By the term urban fabric we denote the type of infrastructures (e.g., roads, transportations, sports facilities, industrial plants) and human activities (e.g., work, residential, commercial, leisure) that characterize a given metropolitan area: it has thus a larger meaning than traditional land use.

Our work builds on previous studies on related subjects. The existence of a relationship between urban fabrics and the associated mobile traffic was first unveiled by a study in Lisbon, Portugal, which showed that the temporal dynamics of mobile traffic are similar in residential and suburban areas, whereas they are unique in areas including major transport arteries [10]. Subsequent analyses proved that base stations

that serve higher mobile traffic demands at noon, evening or night, respectively, reside in urban areas of diverse nature [11]. Similar conclusions were drawn for mobile traffic hotspots during working days and weekends [12].

The findings above have fostered attempts at detecting urban fabrics from mobile traffic directly. Indeed, an efficient inference technique would have important practical applications. As a matter of fact, traditional approaches to land use detection are expensive and time consuming, which results in imprecise and often outdated databases; an automated detection from mobile traffic data could significantly mitigate the problem.

Two such detection techniques have been proposed to date, in [13] and [14], respectively. In this work, we aim at providing a comprehensive evaluation of these approaches, as well as at comparing them with several novel methodologies we propose. To that end, we first present the different urban fabric detection techniques we consider, in Sec. 2. We then introduce in Sec. 3 the two reference scenarios we employ in our performance assessment, each consisting of mobile traffic and ground truth datasets. The results of the comparative evaluation are provided and discussed in Sec. 4, and conclusions are drawn in Sec. 5. Ultimately, our study identifies the solution that yields the most accurate detection of urban fabrics in heterogeneous urban scenarios.

2 Traffic clustering for urban fabric detection

We first report on the two state-of-the art approaches to cluster mobile traffic time series for land-use detection, which we refer to as Soto [13] and Cici [14]. We then introduce an original technique for urban fabric detection, based on the computation of the *median week signature (MWS)*. We also detail several variations of the basic MWS approach, which allow further exploring the design space of detection solutions.

In the following, we consider a generic dataset $\mathcal{D}$, describing the communication activity of a set of mobile subscribers for a set of days $\mathbf{d} = \{d\}$. The mobile activity in $\mathcal{D}$ is described as the aggregate of the traffic generated by all users in a same area during a given time interval; the size of the area and duration of the interval determine the spatial and temporal granularity of the dataset, respectively. More precisely, we name *unit area* the spatial aggregation level: the geographical region $\mathbf{a} = \{a\}$ is thus tessellated into unit areas a, which can map to, e.g. cell sectors, base transceiver stations, or grid elements. The time granularity is instead characterized by the duration of a *time slot*, i.e., the interval during which user activity is aggregated in each unit area. Each day $d \in \mathbf{d}$ is thus divided into a set $\mathbf{t} = \{t\}$ of time slots t.

Not only the spatio-temporal granularity, but also the nature of subscriber activity in such a mobile traffic dataset $\mathcal{D}$ can vary. It depends on the type of activity that is recorded by the operator, and representative options include, e.g., the number or duration of calls, the number of short text messages (SMS), the volume of Internet traffic, the kind of mobile services consumed by the users. For the specific purpose of land use or urban fabrics detection, voice and texting volumes yield the most significant information, and are thus the type of activity used by solutions in the literature [13, 14]. Indeed, they are a better indicator of human occupations than Internet traffic, which is much more homogeneous over time due to the applications that continuously run in background. Therefore, in this paper we adhere to the common practice, and consider that the dataset $\mathcal{D} = \{v_a(d,t)\}$ contains elements $v_a(d,t)$ that describe the total volume of voice and texting activity incoming in and outgoing from each unit area a at time slot t of day d.

As a final remark, we stress that all approaches for urban fabric detection from mobile traffic data rely on the same processing chain, consisting of the following steps.

1. **Mobile traffic signature.** First, one defines some representation of the typical mobile traffic observed at a unit area, named a *signature*. This can map to the complete time series of traffic, or to a summary of this, e.g., a statistical measure or compressed representation. Also, since signatures need to be comparable, they are normalized according to a normalization rule.

2. **Distance between signatures.** Second, one computes how similar, or different, each signature is from all other signatures in the dataset. To that end, a signature distance measure is defined.

3. **Clustering of signatures.** Third, having computed the distances among all signatures, a clustering

algorithm is run so as to separate groups of signatures that have similar shapes, i.e., that are representative of equivalent subscriber activities.

Next, we will review existing and novel approaches to urban fabric detection, decomposing them according to the three steps above.

2.1 Soto

The Soto approach [13] considers mobile traffic signatures that correspond to the average mobile traffic volume observed during (i) a working day, and (ii) a weekend day. We refer to these as *average weekday-weekend signatures*. Formally, the set of days $\mathbf{d}$ is split into two sets $\mathbf{d}^{wd}$ and $\mathbf{d}^{we}$, which contain all Mondays-to-Fridays, and all Saturdays and Sundays, respectively. Then, the element associated to t in the signature of a unit area a is

$$s_a(wd,t) = \frac{1}{|\mathbf{d}^{wd}|} \sum_{\substack{v_a(d,t)\in\mathcal{D} \\ d\in\mathbf{d}^{wd}}} v_a(d,t) \tag{1}$$

for time slots t during working days, and

$$s_a(we,t) = \frac{1}{|\mathbf{d}^{we}|} \sum_{\substack{v_a(d,t)\in\mathcal{D} \\ d\in\mathbf{d}^{we}}} v_a(d,t) \tag{2}$$

for time slots t during weekends. The signature of a is then

$$\mathbf{s}_a = \Big\|_{d\in\mathbf{d}'} \left(\Big\|_{t\in\mathbf{t}} s_a(d,t) \right). \tag{3}$$

In (3), $\mathbf{d}'$ is the condensed set of days, which, in the case of Soto approach, is $\mathbf{d}' = \{wd, we\}$. Also, $\|$ indicates the concatenation of all elements in a set: $\mathbf{s}_a$ is thus the concatenation of all elements referring to the average working day and to the average weekend day. Signatures then go through a *standard score normalization* phase, where each time slot signature obtained in (1) and (2) is normalized with respect to the mean and standard deviation of all those referring to the same unit area. Formally, for the signature element of unit area a at time slot t

$$\hat{s}_a(d,t) = \frac{s_a(d,t) - \mu(\mathbf{s}_a)}{\sigma(\mathbf{s}_a)}, \tag{4}$$

where $d \in \mathbf{d}' = \{wd, we\}$, whereas $\mu(\mathbf{s}_a)$ and $\sigma(\mathbf{s}_a)$ denote the mean and standard deviation of the set of elements concatenated in the signature $\mathbf{s}_a$. Then, the normalized signature $\hat{\mathbf{s}}_a$ is simply obtained by concatenation of $\hat{s}_a(d,t)$ for all $d \in \mathbf{d}'$ and $t \in \mathbf{t}$, as in (3).

As far as distances between signatures are concerned, Soto considers the *Euclidean distance* between the corresponding ordered vectors. Given the signatures of two unit areas a and b, their distance is

$$d_{ab} = \sqrt{\sum_{d\in\mathbf{d}'} \sum_{t\in\mathbf{t}} (\hat{s}_a(d,t) - \hat{s}_b(d,t))^2}, \tag{5}$$

where $\mathbf{d}'$ is always the same condensed day set $\{wd, we\}$.

Finally, the clustering of signatures is performed in Soto by running a *k-means algorithm* over the set of $\hat{\mathbf{s}}_a, a \in \mathbf{a}$, using (5) as the k-means distance measure. The algorithm requires the parametrization of k, i.e., the desired number of clusters: in Soto, k is selected according to the validity index proposed in [15]. In their considered dataset, the best results are obtained with k=5.

2.2 Cici

The Cici solution [14] consider a *whole-time-series signature* for each unit area. In other words, the signature of area a is

$$\mathbf{s}_a = \Big\|_{d \in \mathbf{d}} \left(\Big\|_{t \in \mathbf{t}} s_a(d,t) \right), \tag{6}$$

and the number of elements that compose it is not bounded, but depends on the timespan of the dataset $\mathcal{D}$. In addition, Cici applies a Fast Fourier Transform (FFT) to the signature above, so as to clean it from irregular patterns. Specifically, once converted to the frequency domain with FFT, only the highest power frequencies are kept, and the time signal is reconstructed with and Inverse FFT (IFFT) from the selected frequencies. The filtering returns the *Seasonal Communication Series (SCS)* of the original signature. Normalization of whole-time-series SCS-filtered signatures is then performed using the standard-score approach in (4). However, in the case of Cici, $d \in \mathbf{d}$, since signatures do not condense days, but include the full time series.

The similarity between two such signatures is computed in Cici using the *Pearson correlation coefficient*, which, for two unit areas a and b, is computed as

$$p_{ab} = \frac{\sum_{d \in \mathbf{d}'} \sum_{t \in \mathbf{t}} (s_a(d,t) - \mu(\mathbf{s}_a))(s_b(d,t) - \mu(\mathbf{s}_b))}{\sqrt{\sum_{d \in \mathbf{d}'} \sum_{t \in \mathbf{t}} (s_a(d,t) - \mu(\mathbf{s}_a))^2} \cdot \sqrt{\sum_{d \in \mathbf{d}'} \sum_{t \in \mathbf{t}} (s_b(d,t) - \mu(\mathbf{s}_b))^2}}. \tag{7}$$

In the case of Cici, there is no day condensation and all days are considered separately in the signature, thus $\mathbf{d}' = \mathbf{d}$ in (7). The distance measure is based on (7) and defined as

$$d_{ab} = 1 - p_{ab} \tag{8}$$

Concerning clustering of signatures, Cici adopts an agglomerative hierarchical clustering, i.e., the *linkage clustering with average distance criterion*. This hierarchical clustering outputs a whole family of solutions that can be expressed through a dendrogram: it thus returns a richer information than a single clustering solution, as in the case of, e.g., k-means. However, this also implies that some criterion must be adopted to select the best clustering among all those in the family. To that end, Cici evaluates the skewness of the cluster sizes at the different levels of the dendrogram build by the hierarchical clustering: selecting the level with minimum skewness allows grouping unit areas into clusters of relatively comparable sizes.

It is important to note that, by using the lowest-skewness criterion, the number of generated clusters can be high, in the order of hundreds. Since this makes the analysis cumbersome, Cici limits the analysis to the 10 largest clusters, which are expected to represent the most relevant urban fabrics in the considered region.

2.3 MWS: clustering the median week signatures

The Soto approach is based on a signature definition that is very compact but omits too much information when confronted to the original data [14]. On the other hand, the signature employed in Cici has a number of elements equal to that of the original time series, and, intuitively, is a loss-less representation of the same. However, when considering months of mobile traffic activity, clustering Cici signatures incurs into the well-known curse of dimensionality [16], and is in all cases very expensive from a computational standpoint.

We thus present a novel signature model that aims at combining the advantages of Soto and Cici signatures, while overcoming their limitations. Our *median week signature (MWS)* is based on two considerations.

- First, it has been repeatedly shown that there exists a strong weekly periodicity in human occupations [17, 18], which implies that most of the diversity in mobile traffic activity occurs within a one-week period. We thus speculate that a signature describing the typical weekly behavior of the mobile demand at one unit area contains the vast majority of the significant information about the nature of that area. This allows defining a compact, week-long signature that avoids the dimensionality problems of the Cici model, and does not lose important knowledge as in Soto.

Tab. 1: Summary of the solutions for urban fabric detection from mobile traffic data.

Name	Signature	Filtering	Normalization	Distance	Clustering
Soto-5	average weekday-weekend	–	standard score	Euclidean	k-means, k=5
Soto-10	average weekday-weekend	–	standard score	Euclidean	k-means, k=10
Cici	whole time series	SCS	standard score	Pearson correlation	linkage
MWS-stdscr-pearson	median week	–	standard score	Pearson correlation	linkage
MWS-stdscr-euclidean	median week	–	standard score	Euclidean	linkage
MWS-daily-pearson	median week	–	daily	Pearson correlation	linkage
MWS-daily-euclidean	median week	–	daily	Euclidean	linkage
MWS-scs-stdscr-pearson	median week	SCS	standard score	Pearson correlation	linkage
MWS-scs-daily-euclidean	median week	SCS	daily	Euclidean	linkage

- Second, we deem the median to be a more reliable statistical measure than others used in Soto or Cici (e.g., the average or the absolute values), when it comes to assessing the typical activity in mobile traffic. As a matter of fact, the median is much more robust to outliers, which are frequent in mobile traffic due to special events of social, political, sports, or cultural nature [19–21].

The MWS is computed according to these guidelines, as follows. The whole set of days $\mathbf{d}$ is divided into seven sets, each containing elements of the dataset $\mathcal{D}$ that refer to one day of the week, from Monday to Sunday. In other words, $\mathbf{d}^{mon} \cup \mathbf{d}^{tue} \cup \mathbf{d}^{wed} \cup \mathbf{d}^{thu} \cup \mathbf{d}^{fri} \cup \mathbf{d}^{sat} \cup \mathbf{d}^{sun} = \mathbf{d}$. Then, the element associated to time slot t in the signature of unit area a is

$$s_a(mon,t) = \mu_{1/2}\left(\{v_a(d,t) \,|\, d \in \mathbf{d}^{mon}\}\right), \tag{9}$$

for time slots t corresponding to Mondays, and equivalently for all other days. In (9), $\mu_{1/2}(\cdot)$ represents the median of the set within parenthesis.

Then, the MWS is defined as the concatenation in (3), where $\mathbf{d}' = \{mon, tue, wed, thu, fri, sat, sun\}$ is the condensed set of days. Taking the MWS model as a pivot, we explore the design space of a complete solution for urban fabrics detection, as discussed next.

First, we assess the impact of SCS filtering, proposed in [14] by considering both the case where the MWS is passed through the FFT/IFFT procedure described in Sec. 2.2, and the case where MWS is used as is.

Second, we evaluate two different techniques to normalize MWS. One option is the standard score normalization introduced above; in this case, MWS are normalized according to (4), where $\mathbf{d}' = \{mon, tue, wed, thu, fri, sat, sun\}$. The other option is *daily normalization*, where the signature element of unit area a at time slot t

$$\hat{s}_a(d,t) = \frac{s_a(d,t)}{\sum_{t \in \mathbf{t}} s_a(d,t)}, \tag{10}$$

where again $d \in \mathbf{d}' = \{mon, tue, wed, thu, fri, sat, sun\}$. Thus, daily normalization normalizes each element with respect to the total activity during the weekday the element belongs to.

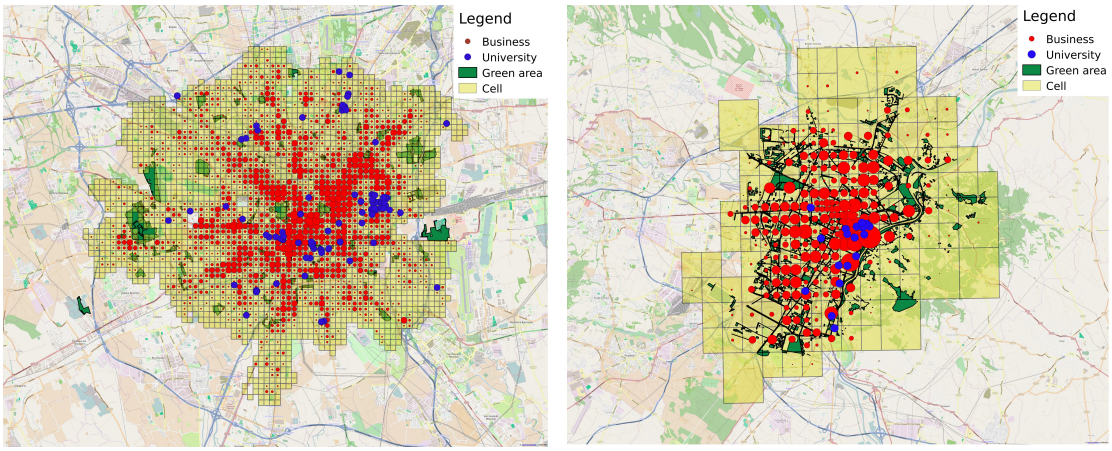
Third, we combine MWS with both distance measures used in Soto and Cici, i.e., the Euclidean distance in (5) and the distance in (8) based on the Pearson correlation coefficient; in both cases $d \in \mathbf{d}' = \{mon, tue, wed, thu, fri, sat, sun\}$.

Finally, signature clustering is performed as in Cici, using the agglomerative hierarchical algorithm described in Sec. 2.2.

A summary of the different solutions we test in Sec. 4 is provided in Tab. 1. We stress that we test Soto with both k=5 and k=10 since the former value is that originally used in [13] and the second is aligned with the number of large clusters considered by all of the other solutions.

3 Datasets

In order to evaluate the different approaches described in Sec. 2 we consider two citywide case studies, employing datasets provided by Telecom Italia Mobile (TIM) as part of its Big Data Challenge initiative, in



(a) **Milan-2013** (b) **Turin-2015**

Fig. 1: Spatial tessellation into unit areas (i.e., *cells* in the legend), and ground-truth data for the **Milan-2013** (a) and **Turin-2015** (b) datasets.

the two editions of 2014 [22] and 2015 [23].

3.1 Milan, 2013

The first dataset is referred as **Milan-2013**, and describes the communication activity of TIM subscribers in the conurbation of Milan, Italy, for a two-month period in November and December 2013. The dataset differentiates among incoming and outgoing calls, providing information about their number and duration. The dataset also contains the number of received and sent SMS and amount of Internet traffic.

Traffic load information are provided every 10 minutes and reported with respect to a tessellation of the surface of the city of Milan in cells. Each cell has a 235×235 m^2 size, i.e., an area of 0.055 Km^2, and maps to a unit area in our analysis. We consider an area of approximately 150 Km^2 containing 2726 cells of the tessellation in the following. The **Milan-2013** dataset is the same used in [14] for the evaluation of the Cici approach with respect to Soto. In order to perform a fair comparison, we focused on the same subset of cell-phone activity as in [14]: this relates to a 4-week period ranging from November 4$^{\text{th}}$, 2013 to December 1$^{\text{st}}$, 2013, where the available data related to incoming/outgoing calls and incoming/outgoing SMS is summed up, so as to avoid data sparsity, and according to the fact that they have the same scale and are comparable.

We also considered the same ground-truth information[†] used in [14], which was retrieved from publicly available databases [24]. Such ground-truth data conveys information on urban infrastructures and land uses that can be associated to different kinds of human activities. Specifically, it reports, for each unit area, the number of business activities, sport centers, universities, schools and bus stops, as well as the percentage of green area covering the area and the number of people living within it. Fig. 1a shows a map of Milan, including the set of analyzed unit areas, together with markers for university, green areas and business presence.

3.2 Turin, 2015

The second dataset, named **Turin-2015**, describes the mobile traffic activity in March and April 2015 for the city of Turin, Italy. Also in this case, a tessellation of the city is provided by the operator in order to describe communication activity per cell. Differently from the Milan dataset, traffic load is provided every 15 minutes and cells present heterogeneous sizes. In our analysis, we considered an area of approximately 150 Km^2, as for Milan, containing 261 cells, i.e., unit areas, with variable size, ranging from a minimum of 255×325 m^2 to a maximum of 2×2.5 Km^2. As a traffic load measure we considered the sum of incoming

[†] `http://odysseas.calit2.uci.edu/doku.php?id=public:urbanecology`.

and outgoing calls and SMS, as in the Milan scenario. For the sake of consistency with the Milan case, we limited our analysis to one month of communication data, by considering a 4-week period ranging from the 1st to 28th of March 2015.

We built ground truth data for each of unit area in **Turin-2015** using the open data published by the local municipality [25]. We selected information related to the latitude-longitude coordinates of schools, universities, and business activities and, as in [14], we associated them to individual unit areas. We also leveraged open data on green areas and population distribution in order to determine their presence in each unit area[‡]. Fig. 1b shows the considered unit areas for the city of Turin, together with a representation of some ground truth data, related to universities, business activities and green areas.

4 Results

In this section, we provide a complete comparative evaluation of the techniques described in Sec. 2 on the reference datasets in Sec. 3. First, we present the metrics we consider in order to assess the quality of the urban fabric detection, and then show the results we obtain.

4.1 Metrics

We evaluate the quality of the unit area classification with respect to the available ground-truth data according to the following metrics, borrowed in part from [14].

Density. The *density* $D_{\mathcal{G}}(\mathbf{c},c)$ is a measure of the frequency of ground-truth elements of a given class $\mathcal{G}$ within a cluster $c \in \mathbf{c}$, where $\mathbf{c}$ is the set of clusters determined by the current urban fabrics detection approach. Let us define as $\mathbf{k}_{\mathcal{G}}$ the set of elements of class $\mathcal{G}$ (e.g., the set of universities) in the ground-truth data; also, $\mathbb{1}_c(k)$ is an indicator function that is one if a ground-truth element $k \in \mathbf{k}_{\mathcal{G}}$ ends up in unit areas belonging to cluster c, and zero otherwise. Formally, the density is then defined, for a given clustering $\mathbf{c}$ as

$$D_{\mathcal{G}}(\mathbf{c},c) = \frac{1}{|c|} \sum_{k \in \mathbf{k}_{\mathcal{G}}} \mathbb{1}_c(k), \tag{11}$$

where $|c|$ denotes the size of the cluster $c \in \mathbf{c}$, i.e., the number of unit areas it includes. The density allows comparing different clusters for a same class $\mathcal{G}$, so as to understand in which clusters elements of $\mathcal{G}$ are more frequent.

Entropy. The *entropy* $H_{\mathcal{G}}(\mathbf{c})$ associated to a ground-truth class $\mathcal{G}$ for a given clustering $\mathbf{c}$ allows estimating the dispersion of $\mathcal{G}$ across the clusters defined by $\mathbf{c}$. It is defined as [26]

$$H_{\mathcal{G}}(\mathbf{c}) = -\sum_{c \in \mathbf{c}} P_{\mathcal{G}}(c) \log P_{\mathcal{G}}(c). \tag{12}$$

In (12), $P_{\mathcal{G}}(c)$ is the probability that a ground-truth element of class $\mathcal{G}$ falls into cluster c, i.e.,

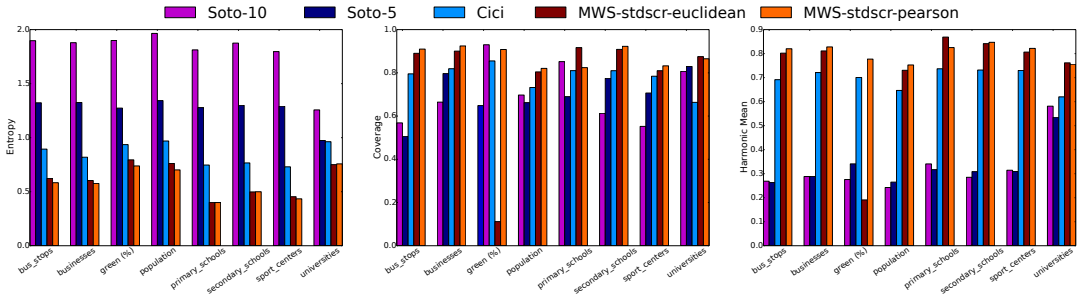
$$P_{\mathcal{G}}(c) = \frac{1}{|\mathbf{k}_{\mathcal{G}}|} \sum_{k \in \mathbf{k}_{\mathcal{G}}} \mathbb{1}_c(k). \tag{13}$$

Lower entropy is thus an indicator of a less random, i.e., more precise, assignment of ground-truth data of a given class to clusters defined by the detection strategy.

Coverage. The *coverage* $C_{\mathcal{G}}(\mathbf{c})$ of $\mathcal{G}$-class elements for a clustering $\mathbf{c}$ is defined as the percentage of ground-truth elements of class $\mathcal{G}$ included within those clusters of $\mathbf{c}$ that are the most relevant to $\mathcal{G}$. Specifically, let us define a subset of clusters $\mathbf{c}_{\mathcal{G}} \subseteq \mathbf{c}$ that have higher-than-average density for ground-truth class $\mathcal{G}$, i.e., $\mathbf{c}_{\mathcal{G}} = \left\{ c \in \mathbf{c} \; s.t. \; D_{\mathcal{G}}(\mathbf{c},c) > |\mathbf{k}_{\mathcal{G}}| / \sum_{c \in \mathbf{c}} |c| \right\}$. Then, the coverage is defined as

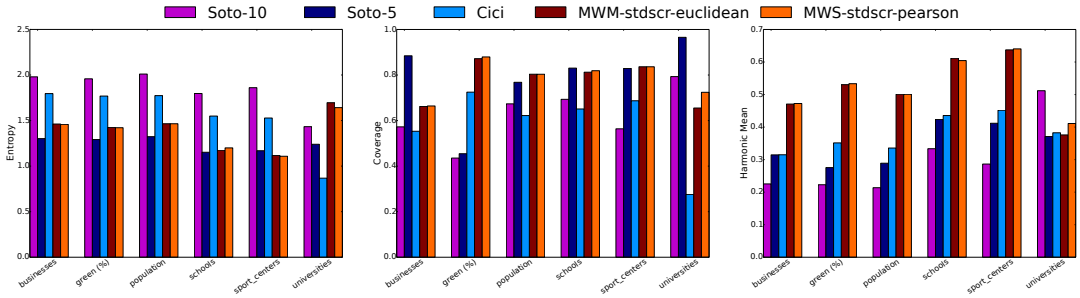
$$C_{\mathcal{G}}(\mathbf{c}) = \sum_{c \in \mathbf{c}_{\mathcal{G}}} P_{\mathcal{G}}(c). \tag{14}$$

[‡] Green area and demographic data are associated to polygon shapes. This first required the computation of the intersections between the polygons and the unit area cells, and then the assignment to each unit area of an amount of green or population percentage proportional to the intersecting surface.



(a) Entropy (b) Coverage (c) F-measure

Fig. 2: Milan-2013. Performance comparison among **Soto-5**, **Soto-10**, **Cici** (lowest skewness at 92 clusters), **MWS-stdscr-pearson** (lowest skewness at 63 clusters) and **MWS-stdscr-euclidean** (lowest skewness at 77 clusters).



(a) Entropy (b) Coverage (c) F-measure

Fig. 3: Turin-2015. Performance comparison among **Soto-5**, **Soto-10**, **Cici** (lowest skewness at 60 clusters), **MWS-stdscr-pearson** (lowest skewness at 33 clusters) and **MWS-stdscr-euclidean** (lowest skewness at 33 clusters).

Higher coverage indicates that the ground-truth data for a class $\mathcal{G}$ is better matched by those clusters that are deemed considered meaningful for $\mathcal{G}$.

F-score. The *F-score* index allows determining a single, final score to each detection technique, by combining entropy and coverage for each class $\mathcal{G}$, as follows

$$F_{\mathcal{G}}(\mathbf{c}) = 2 \cdot \frac{(1-\hat{H}_{\mathcal{G}}(\mathbf{c})) \cdot C_{\mathcal{G}}(\mathbf{c})}{(1-\hat{H}_{\mathcal{G}}(\mathbf{c})) + C_{\mathcal{G}}(\mathbf{c})}, \tag{15}$$

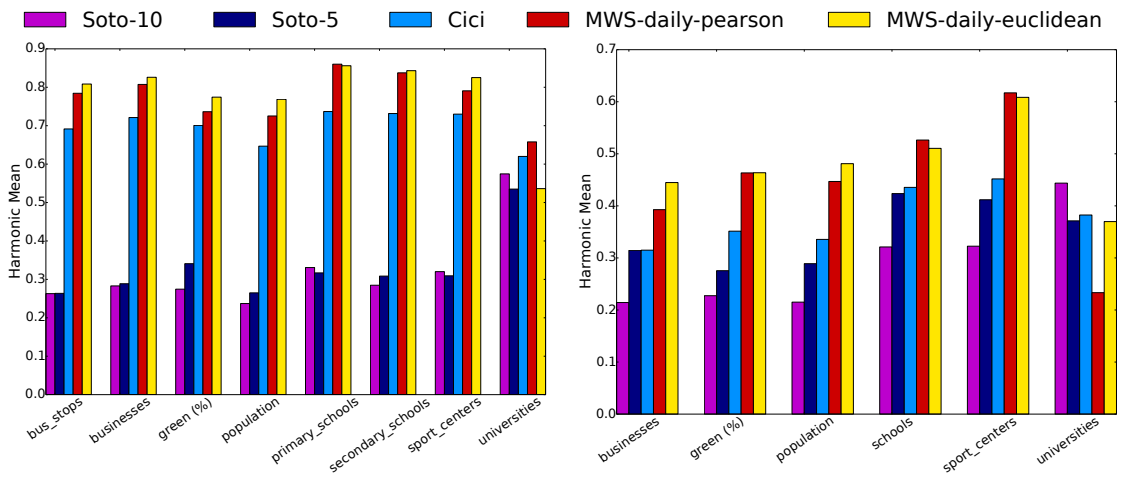
where $\hat{H}_{\mathcal{G}}(\mathbf{c}) = \dfrac{H_{\mathcal{G}}(\mathbf{c})}{\log(|\mathbf{c}|)}$ is the normalized Shannon entropy. The F-score index ranges in [0, 1], with 1 indicating the best performance achievable by the given cluster set, with respect to ground truth class $\mathcal{G}$.

4.2 Performance evaluation

We begin by comparing the first five urban fabrics detection solutions reported in Tab. 1. Fig. 2 and Fig. 3 thus provide results in terms of entropy, coverage and F-score for the state-of-the-art approaches of **Soto-5**, **Soto-10**, and **Cici** as well as for our proposed **MWS-stdscr-pearson** and **MWS-stdscr-euclidean** approaches. The two groups of plots refer to the cases of **Milan-2013** and **Turin-2015** datasets, respectively.

As shown in Fig. 2a, regardless of the specific distance measure used, solutions based on the MWS model attain a significantly lower entropy than solutions based on the other signatures proposed by **Soto** and **Cici**. As mentioned in Sec. 4.1, this indicates a reduced level of randomness, and thus a more precise classification of unit areas with respect to the ground-truth data for **Milan-2013**. Also, the increased accuracy does not come at a cost in terms of coverage, in Fig. 2b. In fact, the entropy gain granted by MWS is associated to an increase in coverage, thus proving the higher effectiveness of a median-week representation of mobile traffic signatures of unit areas.

The result above is summarized in Fig. 2c, which depicts the F-score that considers both entropy and coverage. As one can expect, the F-score further evidences that solutions based on MWS improve current



(a) **Milan-2013**. (b) **Turin-2015**.

Fig. 4: **Milan-2013** (a) and **Turin-2015** (b). F-score comparison among **Soto-5**, **Soto-10**, **Cici** (lowest skewness at 92 and 60 clusters), **MWS-daily-pearson** (lowest skewness at 120 and 60 clusters) and **MWS-daily-euclidean** (lowest skewness at 103 and 80 clusters).

state-of-the-art techniques in urban fabrics detection: the advantage in terms of F-score constantly ranges between 10% and 15%.

The conclusions above are not specific to the **Milan-2013** case study. Fig 3 shows that they hold also for the **Turin-2015** dataset. The most notable difference emerging when comparing the two scenarios is observed in Fig. 3a: there **Soto-5** outperforms all other approaches in terms of entropy. However, such entropy values are related to the lower number of classes obtained by this solution compared with all the other approaches (5 instead of 10): this introduces a bias in favor of **Soto-5**, since the calculation in (12) is a sum over all clusters. Indeed, when entropy values are normalized with respect to the number of clusters, as for the F-score in (15), approaches based on MWS still emerge as clear winners, with the only exception of the universities class, where **Soto-10** gives the best results.

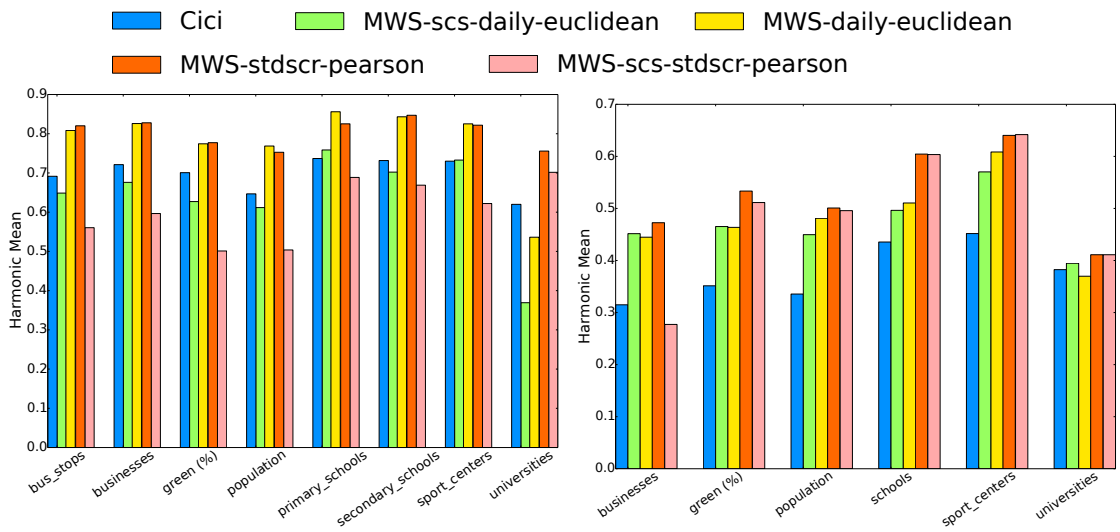
Having substantiated the advantage brought by MWS, an important result from Fig. 2 and Fig. 3 is the performance comparison between **MWS-stdscr-pearson** and **MWS-stdscr-euclidean**. Both solutions employ MWS, but they use different signature distance measures, i.e., Pearson correlation and Euclidean, respectively. Results show that the Pearson correlation distance achieves slightly better and more stable (see, in particular, the *green* ground-truth class in the **Milan-2013** case) results with respect to Euclidean distance. This holds in both mobile traffic scenarios.

We further explore the design space of urban fabrics detection solutions, by testing the impact of a different normalization approach. Fig. 4a and Fig. 4b show the performance of the current state-of-the-art approaches of **Soto-5**, **Soto-10**, and **Cici** against those attained by the MWS model combined with *daily normalization* instead of the *standard score* used before. The two figures refer to the **Milan-2013** and **Turin-2015** scenarios, respectively. For the sake of brevity, we limit results to the F-score, which is a more comprehensive metric according to the discussion of previous results.

We still remark that MWS-based solutions tend to outperform techniques based on other signature models. In this case, however, slightly better performance is achieved by the **MWS-daily-euclidean** approach, and thus *Euclidean* distance appears to work better than *Pearson correlation* in combination with *daily normalization.* Again, results are consistent through different urban scenarios.

An ultimate comparison between the schemes that provide the best performance in the previous tests in provided in Fig. 5. There, we include the F-score attained by **MWS-stdscr-pearson** and **MWS-daily-euclidean**, as well as that of **Cici**, which yielded a much more accurate detection than **Soto** in general. When confronting the three solutions above, **MWS-stdscr-pearson** emerges as the preferred approach, although the difference with respect to the other MWS-based technique is not large.

In addition, Fig. 5 evaluates the *SCS filtering* used by **Soto** on the MWS-based solutions. It can be noted



(a) **Milan-2013**. (b) **Turin-2015**.

Fig. 5: Milan-2013 (a) and **Turin-2015** (b). F-score comparison among **Cici**, (lowest skewness at 92 and 60 clusters), **MWS-stdscr-pearson** (lowest skewness at 63 and 33 clusters), **MWS-scs-stdscr-pearson** (lowest skewness at 110 and 29 clusters), **MWS-daily-euclidean** (lowest skewness at 103 and 80 clusters), and **MWS-scs-daily-euclidean** (lowest skewness at 120 and 100 clusters).

that using the FFT/IFFT processing imposed by SCS does not improve, but instead degrades the performance of both **MWS-stdscr-pearson** and **MWS-daily-euclidean**. This suggests that the median week already provides a sufficiently de-noised representation of typical mobile traffic observed at a unit area: thus, further attempts at removing noise, e.g., using SCS, only risk to disrupt the embedded information.

In conclusion, we deem **MWS-stdscr-pearson** to be the best performing solution among those evaluated, and we consider that most of its gain is due to the MWS model it uses to represent unit area signatures. We stress that MWS is an original contribution of our work.

We provide additional results for the **MWS-stdscr-pearson** approach in Fig. 6, which shows an intuitive idea of the clustering achieved by this technique. We observe how the clusters, denoted by different colors, are spatially localized, i.e., tend to bring together (groups of) unit areas that are nearby. The accuracy is better in the **Milan-2013** case, since the finer spatial granularity of cells allows for a more detailed view of mobile traffic activity.

The nature of the clusters is explained by Fig. 7. The figure portrays, for each urban scenario and ground-truth class, one plot showing the density – computed as in (11) – of corresponding elements across the 10 largest clusters. For example, the bottom-right plot in Fig. 7a shows that university campuses are almost all located in the orange cluster.

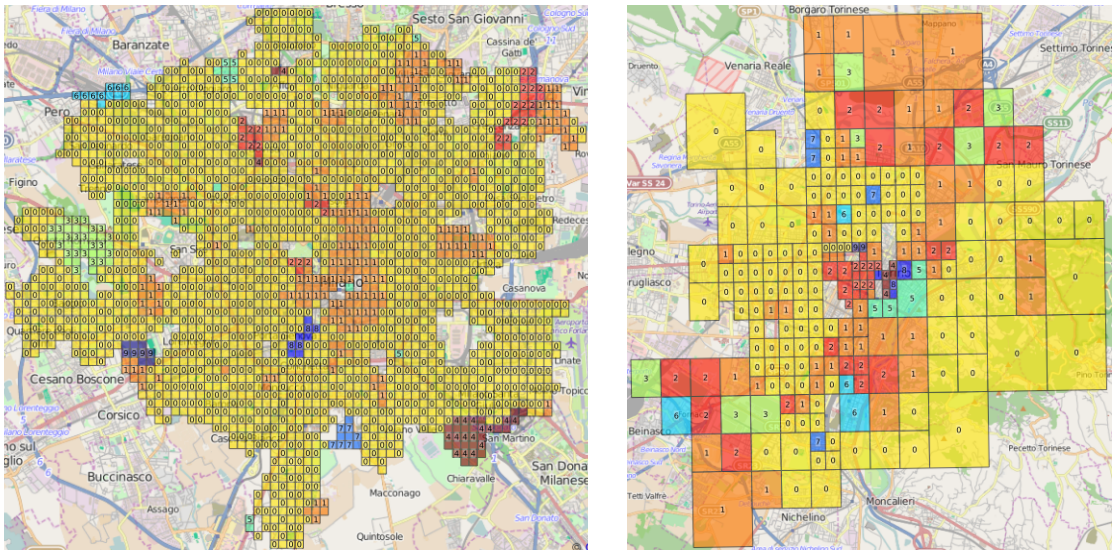
A detailed analysis of the urban fabrics identified by **MWS-stdscr-pearson** is however out of the scope of this work, which aimed at identifying the best detection techniques among a wide range of options. This analysis will be part of our future works.

5 Conclusion

Automatic detection of urban fabric profiles can be of significant interest when studying the evolution of our urban society and planning the shape of future smart cities.

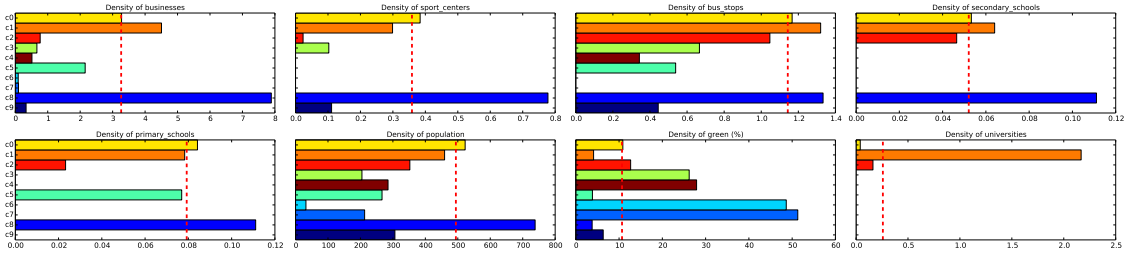
In this paper, we study the usage of mobile traffic datasets to infer this information about the urban landscape. Starting from two state of the art solutions, we distinguish the major algorithmic steps of urban fabric detection techniques: building the mobile traffic signatures of different urban areas, computing the signature distance between these areas and, finally, clustering them to find activity classes.

Our detailed analysis of these solutions underlines their drawbacks, and pushes us to propose a novel

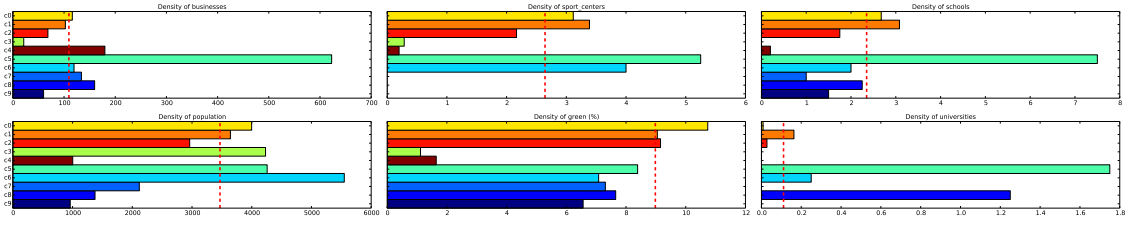


(a) **Milan-2013** (b) **Turin-2015**

Fig. 6: Geographical representation of the 10 largest clusters identified through **MWS-stdscr-pearson**.



(a) **Milan-2013**



(b) **Turin-2015**

Fig. 7: Per-ground-truth class densities of the 10 largest clusters identified through **MWS-stdscr-pearson**.

approach to overcome these limitations, while still exploiting the strengths of these related studies. Our urban fabric detection technique uses the weekly periodicity of human activities, combined with a robust median-value based filtering, to obtain results much closer to the ground truth data.

This work opens a number of perspectives related to the analysis of urban fabric profiles. We are planning as future work a detailed analysis of the geographical clusters obtained by the proposed solution, using more datasets, possibly from different countries. This would allow us to compare the structure of different cities, and possibly find general patterns, i.e. areas showing a similar mobile usage throughout the world, as well as city- or country-specific patterns, related to cultural and social differences.

References

[1] Cisco, "Visual Networking Index – Forecast and Methodology, 2007-2012," 2008.

[2] Pew Research Center, "Emerging Nations Embrace Internet, Mobile Technology," 2014.

[3] R.W. Douglass, D.A. Meyer, M. Ram, D. Rideout, D. Song, "High Resolution Population Estimates from Telecommunications Data," *EPJ Data Science*, 4(4), 2015.

[4] M.C. Gonzalez, C.A. Hidalgo, A.-L. Barabasi, "Understanding Individual Human Mobility Patterns," *Nature*, 453(7196):779–782, 2008.

[5] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, "Fast Unfolding of Communities in Large Networks", *Journal of Statistical Mechanics: Theory and Experiment*, 10, 2008.

[6] J. Blumenstock, O. Toomet, R. Ahas, E. Saluveer, "Neighborhood and Network Segregation: Ethnic Homophily in a Silently Separate Society," *Proc. NetMob 2015*, Cambridge, MA, USA, Apr. 2015.

[7] C. Smith, A. Mashhadi, L. Capra, "Ubiquitous Sensing for Mapping Poverty in Developing Countries", *Proc. NetMob 2013*, Cambridge, MA, USA, May 2013.

[8] V.D. Blondel, A. Decuyper, G. Krings, "A Survey of Results on Mobile Phone Datasets Analysis," *arXiv:1502.03406 [physics.soc-ph]*, 2015.

[9] D. Naboulsi, M. Fiore, S. Ribot, R. Stanica, "Large-scale Mobile Traffic Analysis: A Survey," *hal-01132385*, 2015.

[10] S. Almeida, J. Queijo, L. M. Correia, "Spatial and Temporal Traffic Distribution Models for GSM," *Proc. IEEE VTC Fall 1999*, Amsterdam, Netherlands, Sep.1999.

[11] I. Trestian, S. Ranjan, A. Kuzmanovic, A. Nucci, "Measuring Serendipity: Connecting People, Locations and Interests in a Mobile 3G Network," *Proc. ACM IMC 2009*, Chicago, IL, USA, Nov. 2009.

[12] M. R. Vieira, V. Frias-Martinez, N. Oliver, E. Frias-Martinez, "Characterizing Dense Urban Areas from Mobile Phone-Call Data: Discovery and Social Dynamics," *Proc. IEEE SocialCom 2010*, Minneapolis, MN, USA, Aug. 2010.

[13] V. Soto, E. Frias-Martinez, "Automated Land Use Identification using Cell-Phone Records," *Proc. ACM HotPlanet 2011*, Washington, DC, USA, Jun. 2011.

[14] B. Cici, M. Gjoka, A. Markopoulou, C.T. Butts, "On the Decomposition of Cell Phone Activity Patterns and their Connection with Urban Ecology," *Proc. ACM MobiHoc 2015*, Hangzhou, China, Jun. 2015.

[15] S. Ray, R. H. Turi, "Determination of Number of Clusters in k-means Clustering and Application in Colour Image Segmentation," *Proc. ICAPRDT 1999*, Calcutta, India, Dec. 1999.

[16] A. Hinneburg, D.A. Keim, "Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering," *Proc. VLDB 1999*, Edinburgh, Scotland, UK, Sep. 1999.

[17] H. Zang, J. Bolot, "Mining Call and Mobility Data to Improve Paging Efficiency in Cellular Networks," *Proc. ACM MobiCom 2007*, Montreal, QC, Canada, Sep. 2007.

[18] F. Calabrese, G. Di Lorenzo, L. Liu, C. Ratti, "Estimating Origin- Destination Flows using Mobile Phone Location Data," *IEEE Pervasive Computing*, 10(4):3644, Oct. 2011.

[19] M.Z. Shafiq, L. Ji, A.X. Liu, J. Pang, S. Venkataraman, J. Wang, "A First Look at Cellular Network Performance during Crowded Events," *Proc. ACM SIGMETRICS 2013*, Pittsburgh, PA, USA, Jun.2013.

[20] P. Paraskevopoulos, T.-C. Dinh, Z. Dashdorj, T. Palpanas, L. Serafini, "Identification and Characterization of Human Behavior Patterns from Mobile Phone Data," *Proc. NetMob 2013*, Cambridge, MA, USA, May 2013.

[21] D. Naboulsi, R. Stanica, M. Fiore, "Classifying Call Profiles in Large-scale Mobile Traffic Datasets," *Proc. IEEE Infocom 2014*, Toronto, ON, Canada, Apr. 2014.

[22] Telecom Italia Big Data Challenge '14. [Online]. `https://www.telecomitalia.com/tit/en/innovazione/tutte-le-news/big-data-challenge.html`.

[23] Telecom Italia Big Data Challenge '15. [Online]. `http://www.telecomitalia.com/tit/en/innovazione/big-data-challenge-2015.html`.

[24] Milan open data. [Online]. `http://dati.comune.milano.it`.

[25] Turin open data. [Online]. `http://aperto.comune.torino.it/`.

[26] C. D. Manning, P. Raghavan and H. Schtze, "Introduction to Information Retrieval,", Chapter 8, p. 358. Cambridge University Press. 2008.