Characterizing Dense Urban Areas from Mobile Phone-Call Data: Discovery and Social Dynamics

Marcos R. Vieira [1#], Vanessa Frías-Martínez [‡], Nuria Oliver [‡] Enrique Frías-Martínez [‡]

[#] *Department of Computer Science, University of California, Riverside, CA – USA*
[‡] *Data Mining and User Modeling Group, Telefónica Research, Madrid – Spain*
mvieira@cs.ucr.edu {vanessa,nuria,efm}@tid.es

***Abstract*— The recent adoption of ubiquitous computing technologies (*e.g.* GPS, WLAN networks) has enabled capturing large amounts of spatio-temporal data about human motion. The digital footprints computed from these datasets provide complementary information for the study of social and human dynamics, with applications ranging from urban planning to transportation and epidemiology. A common problem for all these applications is the detection of dense areas, *i.e.* areas where individuals concentrate within a specific geographical region and time period. Nevertheless, the techniques used so far face an important limitation: they tend to identify as dense areas regions that do not respect the natural tessellation of the underlying space. In this paper, we propose a novel technique, called DAD-MST, to detect dense areas based on the Maximum Spanning Tree (MST) algorithm applied over the communication antennas of a cell phone infrastructure. We evaluate and validate our approach with a real dataset containing the Call Detail Records (CDR) of over one million individuals, and apply the methodology to study social dynamics in an urban environment.**

I. Introduction

While the mobility of animals has already been quantitatively studied, *e.g.* marine predators [1], our understanding of individual human mobility is somewhat limited, mostly due to the lack of large scale quantitative mobility data. However, the recent adoption of ubiquitous computing technologies by very large portions of the population (*e.g.* GPS devices, ubiquitous cellular networks) has enabled the capture of large scale quantitative data about human motion [2], [3], [4]. Some of the areas that directly benefit from this new source of information are urban computing and smart cities [5], [6]. These areas focus on improving the quality of life of an urban environment by understanding the city dynamics through the data provided by ubiquitous technologies.

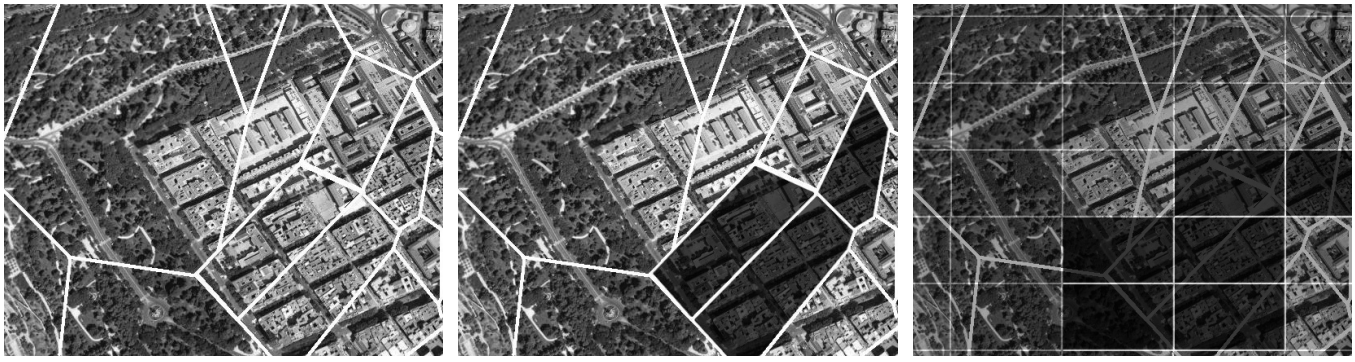
A city is an inherently self-organized human-driven organization where individuals and their behavior play an important role in defining the pulse and the dynamics of the city. This implies that in order to efficiently model human mobility, individual information is necessary in order to reflect that location is, at least in part, each individuals decision [6]. The datasets captured by ubiquitous computing technologies inherently reflect individual information relating to mobility and social dynamics. This fact represents a huge improvement when compared to how mobility data has been typically collected: using questionnaires and surveys, and in more advanced studies, using proxies such as bills, public transport, etc [7].

[1]Work done while author was an intern at Telefónica Research, Madrid.

Some of the applications of smart cities and the study of social dynamics include traffic forecasting [8], modeling of the spread of biological viruses [9], urban and transportation design [8] and location-based services [10]. A challenging and interesting problem related to social and human dynamics is of identifying areas of high density of individuals and their evolution over time. This information is of paramount importance for, among many others, urban and transport planners, emergency relief and public health officials, as it provides key insights on *where* and *when* there are areas of high density of individuals in an urban environment. Urban planners can use this information to improve the public transport system by identifying dense areas that are not well covered by the current infrastructure, and determine at which specific times the service is more needed. On the other hand, public health officials can use the information to identify the geographical areas in which epidemics can spread faster and, thus, prioritize preventive and relief plans accordingly.

The problem of dense area detection was initially presented in the data mining community as the identification of the set(s) of regions, from spatio-temporal data, that satisfy a minimum density value. This problem was initially solved for spatial and multidimensional domains [11], and later for spatio-temporal domain [12], [13], [14]. In the former proposals, no time dimension is considered, while in the later ones only moving objects, typically represented by GPS sensors that continuously report their locations, are considered. Common to all of the above methods is that a fixed-size non-overlapping grid or circle employed to aggregate the values over the spatial dimensions are considered. Therefore, these methods "constrain" the shape of the detected areas and, generally, identify dense areas that are a *superset/subset* of the desirable dense areas. Ideally, we seek a technique that is able to detect dense areas whose shape is as similar as possible to the underlying dense geographical areas.

In this paper we propose the *Dense Area Discovery (DAD-MST)* algorithm to automatically detect dense areas in cell phone networks. Our approach, unlike the previous approaches, is not based on fixed-size grids, but on the natural tessellation of the spatial domain, thus overcoming the limitations of all the previous approaches. The DAD-MST is especially suited to work with human mobility data from cell phones. Nevertheless, the type of information used by the DAD-MST is not only available to telecommunication companies but also to an increasingly large number of companies



(a) Original Tessellation (b) Desired dense area (c) Grid dense area

Fig. 1. (a) the original tessellation for an urban area, where each polygon defines the coverage of a cell phone tower; (b) the ideal dense area (highlighted) based on the tessellation of the data; (c) common techniques based on fixed-size grids fails on the identification of dense area.

that provide location-based services and mobile services which also collect (or are able to collect) human mobility data using the cell phone network infrastructure. Moreover, although the DAD-MST algorithm has been designed considering the infrastructure of a wireless phone network, it can also be applied to any problem where the data is represented in a domain that has a *natural* tessellation (*e.g.* zip codes). We evaluate the proposed algorithm with a very large, real-world Call Detail Record (CDR) dataset, and then validate it with a study of the social dynamics in a urban environment.

Note that the focus of this paper is on the detection and study of dense areas, not *hotspots* [15]. *Hotspots*, as defined by scan statistics, are the *largest discrepancy areas in which an independent variable has statistically different count values from the rest of the geographical areas* [16]. Conversely, *dense areas* are defined as the *(global or local) maxima of the distribution of the function under study* [17]. Thus, the information provided by both approaches is different, while *hotspots* can be used to identify events, dense areas identify regions in space with a minimum critical mass of individuals.

The remainder of this paper is organized as follows: Section II discusses the related work; Section III formally defines the problem of discovering dense areas; Section IV describes the proposed approach and its evaluation and validation appear in Section V; A case study of the dynamics of a city from a dense area perspective is described in Section VI; and Section VII concludes the paper.

II. Related Work

In the GIS, Urban Planning, Transportation and Virus Spreading communities there has been (for a long time) a variety of models to study human and city dynamics. Traditional approaches divide the geographical region under study in zones which exchange population among themselves. Each zone is characterized by a vector of socio-economic indicators [5], typically collected and generated using surveys. Also, this information can be completed with proxy sources for human mobility such as transport infrastructures, air connections, etc [7], [9]. These approaches provide information about human behavior in a geographical environment but they are very difficult to update and limit the results to a moment in time [18]. In any case, these models substitute humans with derivatives of their activities, ignoring the self-driven nature of human mobility. The use of data originating in pervasive infrastructures captures each individual mobility and is ideal to represent the self-driven nature of the problem, complementing traditional approaches. In [18], [19], the authors discuss initial guidelines on how mobile phone data can be relevant for urban planning and transportation communities.

Previous works on the identification of dense areas, not necessarily for the study of social dynamics, have been carried out following three main approaches: **(1)** density-based clustering techniques; **(2)** detecting dense fixed-size grids in spatio-temporal data; and **(3)** spatial-based techniques to detect local *maxima* areas.

Clustering algorithms for spatial, multidimensional and spatio-temporal data have been the focus of a variety of studies (*e.g.* [20], [21], [22], [23]). Common to all of the above methods is that clusters with high numbers of objects in a specific geographical area are associated, using spatial properties of the data, to denser regions. Furthermore, all of these methods require choosing some number of clusters or making underlying distributional assumptions of the data, which is not always easy to estimate.

There are a variety of solutions for detecting dense areas in spatial [11] and spatio-temporal [12], [13], [14] domains. The STING method [11] is a fixed-size grid-based approach to generate hierarchical statistical information from spatial data. Hadjieleftheriou *et al.* [14] present another method based on fixed-size grids where the main goal is to detect areas with a number of trajectories higher than a predefined threshold. Algorithms using a fixed-size window are proposed in [12], [13] to scan the spatial domain in order to find fixed-size dense regions. All of these approaches are specifically designed to work for trajectory data where the exact location and speed direction of a trajectory are used in order to aggregate values in each grid for the spatial domain. Unfortunately, these methods cannot be applied to our domain since in the majority of mobile phone databases mobile users are not continuous tracked. Furthermore, all the works described here detect dense areas of fixed-size above a threshold using a predefined grid.

Some solutions to detect dense areas are based on the iden-

tification of local *maxima*, typically using techniques inherited from computer vision (*e.g. mean-shift* [16], [15]). Mean shift is a non-parametric feature-space analysis technique that identifies the modes of a density function given a discrete dataset sampled from that function. As in previous approaches, the geographical space under study is divided into a grid, hence ignoring other original (natural) tessellations. Crandall *et al.* [17] use *mean-shift* to identify geographical landmarks from geo-tagged images.

In summary, previous works, among other limitations, typically identifies dense areas by overlaying a fixed grid on the geographical region, which might not correspond to the real shape of the underlying dense area. Although this problem can be tackled to some extent with the creation of a grid with enough granularity to linearly approximate the natural tessellation of the area under study, the exponential increase in complexity makes this solution computationally unfeasible.

III. Preliminaries

In order to study the social dynamics of a geographical area, we propose a new technique to identify dense areas and study their evolution over time using the ubiquitous infrastructure provided by a cell phone network. Cell phone networks are built using a set of cell towers, also called Base Transceiver Stations (BTS), that are in charge of communicating cell phones with the network. Each BTS has a latitude and a longitude, indicating its location, and gives cellular coverage to an area called a *cell*. We assume that the *cell* of each BTS is a 2-dimensional non-overlapping polygon, and we use a Voronoi tessellation to define its coverage area. Neighboring towers can thus be identified using the Delaunay triangulation. Ideally, an algorithm to detect dense areas in this context should respect the tessellation produced by the Voronoi.

An example of the problems that arise when using the traditional techniques presented in the previous section is illustrated in Figure 1: Figure 1(a) depicts the natural tessellation of a city given by cell towers, with each polygon representing its coverage; Figure 1(b) highlights the desired dense area (in this case an area of high mobile-phone activity) based on the original tessellation; and Figure 1(c) represents the dense area identified by a *state-of-the-art* grid technique (with cell size similar to the size of the cells in the tessellation). Note that in the latter case, the dense area found by the algorithm includes geographical regions with low density of activity (*e.g.* parks) due to the grid structure employed. Moreover, due to the nature of the grid structure, a larger area than what it is in reality is returned by the grid-based algorithms.

Call Detailed Record (CDR) databases are populated whenever a mobile phone makes/receives a phone call or uses a service (*e.g.* SMS, MMS). Hence, there is an entry in the CDR database for each phone call/SMS/MMS sent/received, with its associated *timestamp* and the BTS that handled it, which gives an *indication* of the geographical location of the mobile phone at a given moment in time. Note that no information about the position of a user within a cell is known.

We characterize the information handled by each BTS of two types: *activities* and *users*. The *activities*, $A(\delta t)_i$, at bts_i correspond to the number of different calls that were handled by bts_i during the time period δt. Likewise, $U(\delta t)_i$ measures the number of *unique individuals* whose calls where handled by bts_i during the time period δt.

In order to study the social dynamics of a region using cell phone networks, we propose the DAD-MST algorithm to automatically discover dense areas of *activities* or *unique users* in a specific geographical region and during a determined period of time δt such that: **(1)** it respects the original tessellation of the space defined by the cell phone network; **(2)** it does not need as input the number of dense areas (*e.g.* the number of clusters) to be identified; and **(3)** it guarantees that *all* dense areas are identified, covering up to a maximum percentage λ of the total region under consideration. In our scenario, the geographical region corresponds to the total area where the dense spots are to be identified, and we carry out our analysis at three levels: *urban*, *regional* and *national*.

IV. Discovering Dense Areas from CDR Data

Given an initial set of *BTS*=$\{bts_1, bts_2, ..., bts_n\}$ that gives coverage to a geographical region R characterized by its Voronoi tessellation $R = \{V_1 \cup V_2 \cup ... \cup V_n\}$, we seek to discover the optimal disjoint subsets of *BTS* that cover areas within R where either the number of *activities* or *unique users* reaches a *maximum* in a specific time period δt. An exhaustive exploration of all possible disjoint subsets of *BTS* becomes a daunting task as the number of *BTS* increases. Thus, we propose a *greedy algorithm* based on the Maximum Spanning Tree (MST) algorithm [24] that selects, at each step, the best subsets of *BTS*. In order to smooth noisy data, the minimum number of *BTS* that define a dense area is set to 2.

The algorithm computes the dense areas in a geographical region R given two parameters: coverage λ and granularity ξ. The coverage λ corresponds to the maximum percentage of the geographical area R that can be covered by the dense areas identified by the algorithm. Typical values for λ are in the range 0.05 to 0.5 (5% and 50%, respectively). Smaller λ values may risk not identifying dense areas, and larger values are considered not relevant as the areas identified would cover most of the region R under study.

The granularity ξ represents the maximum distance between two BTS in order to consider them to be part of the same dense area and to be joined to form a potential subset. Hence, the parameter ξ sets the spatial granularity at which dense areas are identified (*e.g. urban*, *regional* or *national* levels) and it is similar to the *scale of observation* parameters employed in the *mean-shift* approach in [17]. When seeking an adequate value for ξ, the distribution of *BTS* is a key factor. In urban areas this distribution is typically very dense and homogeneous such that each cell covers similar extension of areas. However, in sparsely populated areas (*e.g.* rural area), the distribution of *BTS* is scarce. For example, the average distance between two neighboring urban BTS is around 1km, while in rural environments this value may increase up to 11km. Therefore, suitable values for ξ could be 1km, 10km and 100km to detect dense areas in *urban*, *regional*, and *national* level, respectively.

The proposed Dense Area Discovery via MST algorithm (DAD-MST) consists of three phases (explained below): **(A)** Graph Construction; **(B)** Computation of Dense Areas; and **(C)** Post-processing. It receives as inputs: the geographical region R, the time period δt for which the dense areas need to be computed, the set of *BTS* in the region R, the coverage λ and the granularity ξ. It generates as outputs the subsets of *BTS* that correspond to the dense areas in region R with coverage λ and granularity ξ.

A. Graph Construction

First, a graph G=(V,E) is built using Delaunay triangulation, where each vertex $v_i \in V$ corresponds to $bts_i \in$ BTS in the geographical region R, and each edge $e_{i,j} \in E$ represents the connection between bts_i and bts_j. The Delaunay triangulation is implemented following the *Divide and Conquer* approach [25], with an approximate complexity of $O(V \log V)$. Next, all the edges in E with a distance between the two connecting BTS larger than ξ are eliminated from the graph, in order to ensure the desired spatial granularity given by ξ. The distance between two BTS is computed by translating their geographical coordinates into Cartesian coordinates and then computing their Euclidean distance.

After that, a weight $w_{i,j}$ is associated to each edge $e_{i,j} \in E$ that has not been eliminated. The weight represents the average density of the area covered by bts_i and bts_j during the time period δt. The density is given by two types: the total activity ($A(\delta t)_i$+$A(\delta t)_j$) or the total number of (unique) users ($U(\delta t)_i$+$U(\delta t)_j$) observed at bts_i and bts_j during δt, divided by the geographical area (in km^2) covered by bts_i and bts_j. Both values are computed from the CDR database using a query system (see subsection CDR Query System). The details of the algorithm are presented in Algorithm 1.

Algorithm 1 *GraphConstruction*($type$,*BTS*,ξ,δt)

```
1: G(V, E) ← Delaunay(bts_1, ..., bts_n)
2: for each edge e_{i,j} ∈ E do
3:     if distance(bts_i, bts_j) > ξ then
4:         E ← E \ e_{i,j}
5:     else
6:         w_{i,j} ← QueryDB(type, bts_i, bts_j, δt)
```

B. Computation of Dense Areas

A variation of the *Maximum Spanning Tree* algorithm is used to detect dense areas given by $G(V,E)$ and the associated weights W (see Algorithm 2). The edges in E are first sorted by decreasing weight W. At each step the edge $e_{i,j} \in E$ with the highest weight $w_{i,j}$ is removed from E and added to the list L of edges that represent dense areas if and only if the edge connects vertices that belong to two different subsets (trees) of *BTS*. In Algorithm 2, *MakeSet* creates a potential tree for each vertex, *FindSet* identifies the tree in which a vertex is included in L, and *Union* joins two trees. A detailed description of *MakeSet*, *FindSet* and *Union* is given in the formal definition of the MST algorithm [24].

This process selects, in a greedy manner, the subsets of vertices (*BTS*) that are associated to high values of either *activities* or *unique users*. Edges are added to L until the total geographic area covered by the *BTS* that are connected by the edges in L is equal or larger than $\lambda * |R|$, where λ is the coverage and $|R|$ is the size of the area under study. Note that the coverage of bts_i is approximated by the area of its associated Voronoi cell *Voronoi*(bts_i), such that the algorithm computes the tree until $\sum$*Voronoi*(bts_i) $\forall i \in$ unique bts_i of $|L| > \lambda * |R|$. Additionally, every time an edge $e_{i,j}$ is added to L, the edges in E where either i or j are one of the vertices, are re-weighted in order to avoid double counting of *activities* or *unique users*. Once the stopping condition is satisfied, the list L contains all the edges (and associated pairs of BTS) that correspond to the dense areas in the graph. The complexity of *ComputeDenseAreas* is $O(E \log E)$.

Algorithm 2 *ComputeDenseAreas*($G(V,E)$,R,λ)

```
1: sort E by decreasing weight W
2: L ← ∅
3: for each v_i ∈ V do
4:     MakeSet(v_i)
5: while ∑ Voronoi(bts_i) ∀i ∈ unique bts_i of |L| < λ * |R| do
6:     e_{i,j} ← E.top()
7:     E ← E \ e_{i,j}
8:     if FindSet(i) ≠ FindSet(j) then
9:         L ← L ⋃ e_{i,j}
10:        Union(i, j)
11:        re-weight ∀e_{i',j'} ∈ E affected by e_{i,j}
12:        sort E by decreasing weight W
```

C. Post-processing

The post-processing phase computes all the connected components from the final list L in order to visualize the dense areas on a map. Each subset of connected edges in L represents a subset of *BTS* associated to a dense area. Specifically, we use the *Shiloach-Vishkin* [26] algorithm to compute the connected components of the graph (with a complexity of $O(E \log V)$). Once the connected edges are obtained, the final density of *activities* or *unique users* associated to each dense area are computed as the sum of the weights of all of its edges divided by the geographical area (in km^2) covered by all the *BTS* in the dense area. Finally, a color is assigned to each dense area (subset of *BTS* based on its level of *activities* or *users*: *warm* (red, orange, ...) and *cold* (blue, grey, ...) colors are used to represent areas with high and low, respectively, dense levels.

D. CDR Query System

The most computationally expensive part of the proposed algorithm is the calculation of the weights W associated to the edges E. Since processing a very large CDR database containing several millions of records for a specific period of time δt can be computationally expensive (especially when long periods of time are considered), in this work we make use of a spatio-temporal query system designed specifically for CDR databases [4] that guarantees a timely retrieval of information associated to any BTS. Basically, for each bts_i, two index structures are built: one B$^+$-tree to organize entries by the temporal attribute *timestamp*; and one *inverted-index*

where entries are ordered by ($phone_{id}$,*timestamp*). This index-based structure allows us to compute the weights of the edges by querying the system with the time period δt, the bts_i and bts_j, and the type of query (*activities*/*users*) under study.

V. Experimental Evaluation

We collected cell phone data in the form of CDR from a single carrier of a state with an approximate area of 80,000 km^2. The state contains two large metropolitan areas (of approximately 4,000,000 and 400,000 residents respectively) and other smaller urban areas. Here we use a sample of this dataset containing the calls of over one million anonymized unique customers over a period of four months, with around 50 million CDR entries collected with 5,000 BTS towers[2].

A. Quantitative Evaluation

First, we experimentally analyze the effect that ξ (granularity) and λ (coverage) have in the number of dense areas identified and by extension in the granularity of the social dynamics modelled. For that purpose, we consider two different settings for the geographical region R: *urban*, R_u, defined by a rectangle (R_u=30km*35km) that covers the main metropolitan area ($4,000,000$ residents); and *regional*, R_r, defined by a rectangle (R_r=400km*200km) that approximately covers all the geographical area of the state. Regarding the temporal range δt, we consider "weekdays" and "weekends" separately and within each type of day, we identify four time slots: "mornings" (6am–10am), "afternoons" (10am–2pm), "evenings" (2pm–6pm) and "nights" (6pm–10pm). Note that we present here results for δt="weekdays in the morning" and for the type of query *number of users* due to space constraints, but we obtained similar results with the other temporal ranges and the *number of activities*.

Figure 2 (top) shows the number of dense areas (Y axis) obtained by the DAD-MST algorithm with different values of coverage λ from 5% to 50% (X axis) for the *urban* geographical region R_u=30km*35km. Each line in the plot represents a different value for ξ: 1km (urban), 10km (regional) and 100km (national). Considering ξ=1km, we observe that as λ increases, the number of dense areas identified by the algorithm increases linearly. The algorithm successfully identifies different dense areas due to the fact that ξ=1km does not allow for many BTS connected by Delaunay to be merged together as the area of coverage increases. However, we observe that for larger values of ξ, the increase in coverage results in a reduction in the number of dense areas. This is due to the fact that larger ξs merge dense areas together as the coverage area is increased. In fact, we observe that for ξ=10km and ξ=100km, when the area of coverage is larger than 10%, only one dense area is identified as all of the edge subsets are joined. In sum, this analysis highlights the importance of selecting an appropriate value for ξ that will adequately identify areas within the region under study. In the case of an urban environment, the value of ξ=1km successfully achieves this result.

[2]Company policy does not allow us to reveal the geographical origin of the data.

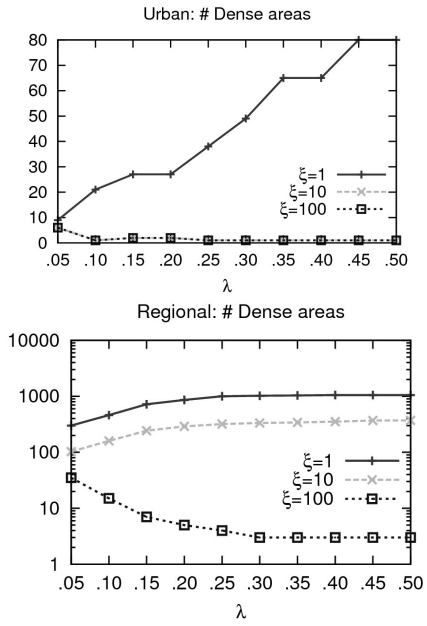


Fig. 2. Total number of dense areas identified (Y axis) for values of λ (X axis) ranging from 5% to 50% and for ξ=1, 10, 100km. R represents an urban (top) and a regional (bottom - with Y in log scale) area.

Figure 2 (bottom) shows a similar analysis for a regional geographical area R_r=400km*200km (Y axis in log scale). In this case, ξ=1km generates a large number of dense areas of small size as the coverage area λ increases. In fact, we are capturing the dense areas within the neighborhoods of metropolitan areas located in the region under study. Higher values of ξ allow us to merge and thus reduce the number of dense areas identified as λ increases. We observe that ξ=10km stabilizes after identifying approximately 200 dense areas, which correspond to small suburban areas (generally near metropolitan areas) and/or big neighborhoods within metropolitan areas. Finally, for ξ=100km, we observe that values of λ higher that 20% only allow for the detection of two dense areas that correspond to the two big metropolitan areas in the region.

Hence, a value of ξ=100km yields the identification of big metropolitan areas whereas ξ=10km leads to detecting small suburban areas within the region under study. From a computational efficiency perspective, the scale of the experiment deeply affect the computation time. At an urban scale R_u, a reduced number of BTS is analyzed (around 500), yielding a processing time of less than 30 seconds per evaluation tuple (λ, ξ). At a regional scale R_r, the number of BTS is one order of magnitude larger (around 5,000). Hence, the processing time is of the order of 30 minutes per evaluation tuple. All experiments were run on a Dual Intel Xeon E5540 2.53GHz running Linux 2.6.22 with 32GBytes memory.

B. Qualitative Validation

In order to assess the quality of the areas identified by the proposed algorithm, Figure 3 shows some landmarks of the city under study, having as a reference the subway system.

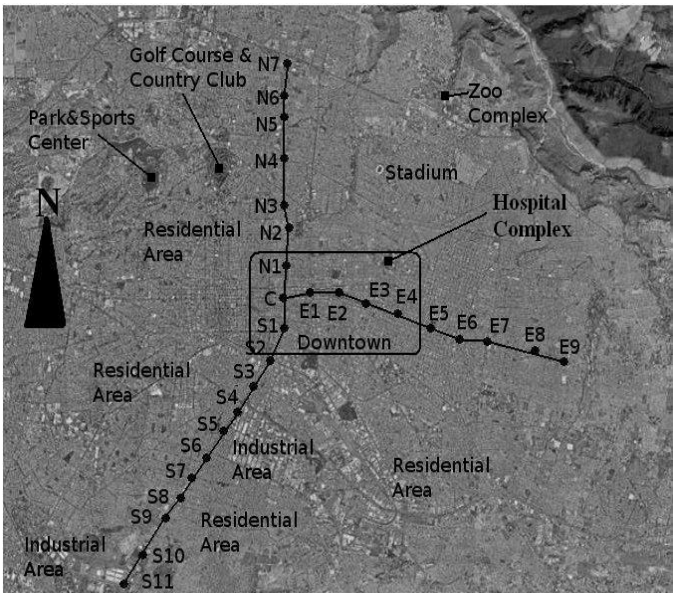


Fig. 3. General description of the city under study with the subway system as the main reference landmark.

The two subway lines in the city (represented by dotted-black lines), run East-West (L1) and North-South (L2) with one central station in common. For reference purposes the central station is denoted by C, with stops north of C denoted as N1 to N7, stops south of C denoted S1 to S11 and stops east of C denoted E1 to E9. The downtown area is geographically located around C, E1, E2, E3 and E4. Near C we find university buildings, government offices and parks. The vicinities of E1 and E2 form the commercial part of the city with markets, commercial streets and hotels. E3 and E4 have more university buildings and night life area. The rest of L1 services mainly residential areas. Regarding L2, around S3 to S11 there are mainly residential neighborhoods with light industrial areas. N2 to N7 serve residential areas with some commercial and entertainment places. The map also indicates other places such as a Stadium complex (S) and the city Zoo (Z), the main zoo of the country. For the areas not commented, as a general rule, there is a mixture of residential areas (with different densities) and light industrial areas, with the north and north-west having more affluent areas than the south.

Figures 4 and 5 depict the graphical representation of the dense areas detected at an urban level (R_u=30km*35km, ξ=1km, λ=15%) during mornings, afternoons, evenings and nights in term of *number of users* for weekdays and weekends respectively. R covers the city and its metropolitan area as to reflect the social dynamics between the metropolitan belt and the city. The area presented in the figures is a smaller rectangle of 20km*15km centered in the city in order to appreciate downtown dense areas. Note that the same color in different time frameworks does not imply the same density, just the same relative importance.

The dense areas identified by DAD-MST align well with the two subway lines, highlighting the intricate relationship between public transportation and city dynamics. This alignment is also an indirect validation of our algorithm, as it is expected that people will concentrate around the areas served by the public transport system. Finally, we repeated the analysis for the variable *number of activities* with similar results.

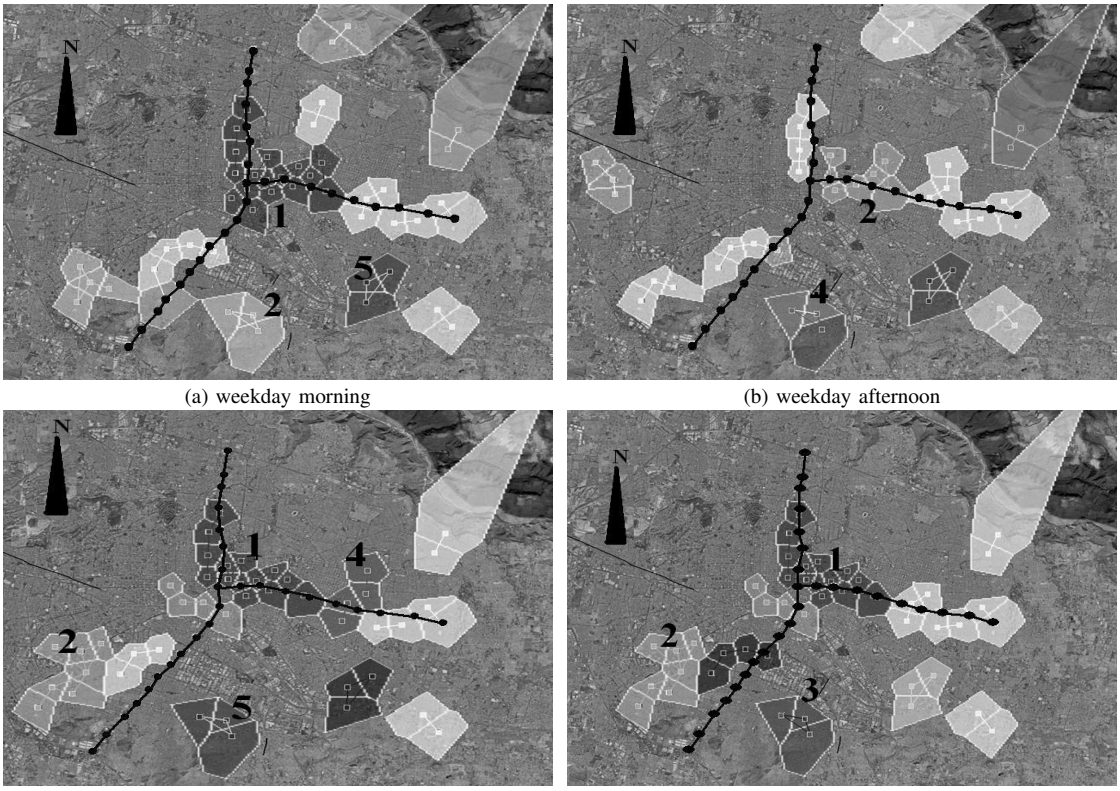
VI. City Dynamics

The combination of Figures 4 and 5 shows the evolution of dense areas during weekdays and weekends and thus an indication of how people live and move in the city. It has to be noted that the dynamics reflected by the evolution of dense areas does not necessarily imply *flocks of individuals* [27] moving from one area to another, just density of individuals changing over time. The top five dense areas presented in each map of Figures 4 and 5 have been numbered from 1 to 5 (being dense area #1 the one with the highest density) to facilitate the reading of the relevance[3]. If one number is not present is because it is not located in the area of 20km*15km showed in the Figures 4 and 5 but is present somewhere in R, *i.e.* in those cases the dense areas are located outside the city somewhere in the metropolitan belt.

Following the temporal sequence of the evolution of dense areas during weekdays (see Figure 4) it can be observed that downtown is covered by a big dense area in the morning, focussing on the university, the commercial and the governmental district. Also in the morning, the second and fifth dense areas are located in residential zones and dense areas #3 and #4 are outside the city. This indicates that in those hours although downtown is the top dense area, the metropolitan belt of the city also has important density of individuals. In the afternoon, the top dense area that appeared in the morning disappears, and a new dense area, ranked #2, located around E1 to E4 and focussing on the commercial area appears. Also in the afternoon, 3 out of the top 5 dense areas identified are in the metropolitan belt. Note that because of the nature of the data, we are strictly representing the number of people that have used their cell phone in that period of time. It can be the case that once people get to the working place or to the university in downtown, cell phones are not used with the same frequency. This would motivate the disappearance of the dense area in the government and university districts. On weekday evenings and nights the activity is concentrated in downtown, with the top dense area around the commercial and business districts and aligning with the subway lines. Both in the evening and at night there is a shift, when compared to mornings and afternoons, in the localization of the top dense areas from the metropolitan belt to downtown. To represent that shift, in the evening 4 out of the 5 top dense areas are in the city, while at night the 3 top ones are in the city.

Following the temporal sequence of the evolution of dense areas during weekends (see Figure 5) it can be observed that in the morning and in the afternoon the top dense area (#1) is outside the city. In both cases there are dense areas in the commercial and business districts in downtown, although they are not in the top 3. Two dense areas appear north of downtown, which include the stadium complex (marked with

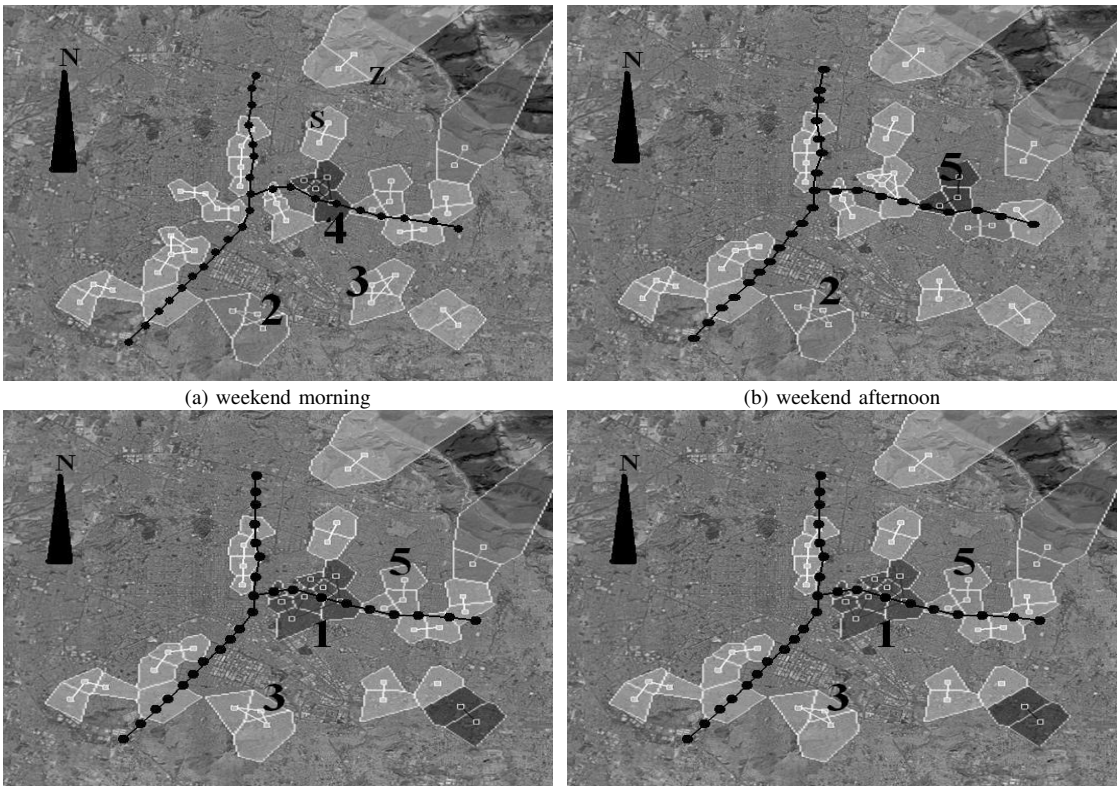
[3]We have ranked the dense areas because the color scheme used in our graphic interface does not translate into interpretable grey levels.



(a) weekday morning (b) weekday afternoon

(c) weekday evening (d) weekday night

Fig. 4. Area of 20km*15km of the city under study and the dense areas detected by DAD-MST in the morning, afternoon, evening and night during weekdays with ξ=1km and λ=.15. The black line with black dots represents the subway system and the corresponding stations.



(a) weekend morning (b) weekend afternoon

(c) weekend evening (d) weekend night

Fig. 5. Area of 20km*15km of the city under study and the dense areas detected by DAD-MST in the the morning, afternoon, evening and night during weekends with ξ=1km and λ=.15.The black line with black dots represents the subway system and the corresponding stations.

S) and the zoo (marked with Z) respectively. Both dense areas are present in the four range hours during weekends (note that both complexes cover a small geographical area of the dense areas identified). During the evenings and at night the top dense area is in downtown around the subway stops E1 to E4, *i.e.* in the commercial and nigh life area. Both evening and nights during weekends have the same dense areas, indicating that there is no change in the dynamics of the city. As it happened during weekdays, in the evening and at night there is a shift in the top dense areas from the metropolitan belt to the city.

Both during weekdays and weekends residential areas are identified south of downtown. The identification of dense areas in residential neighborhoods is tightly related to the density of housing, where lower income neighborhoods tend to have a higher density than more affluent neighborhoods. This is probably one of the reasons why residential dense areas are identified mainly in the south and none in the north.

A direct application of these knowledge regarding the social dynamics of the city is to help in the decision process of the design of the public transport infrastructure. In general, as mentioned in the previous section, there is an alignment between the subway lines (especially N1-N11 and E1 to E7) and the dense areas, indicating that the main dense areas are covered. The dense areas of the south-west are not directly covered by S1-S11, although they are in the vicinity. Note that our algorithm has also identified dense areas in the south-east corner of the city, where there is no subway service right now. From an urban planning perspective, this information could be used to propose line extensions to public transport officials. Also it is relevant that dense areas are very relevant in the metropolitan belt, and typically in the top 5 most important ones, so enough means of communication (buses, trains, etc.) have to communicate those areas with the city, specially between afternoon and evening when dense areas shift from the metropolitan belt to the city.

VII. Conclusion and Future Work

Ubiquitous computing infrastructures are opening new doors for the study of social dynamics, specially in the field of urban computing. The application of these new techniques can be used as a complement to traditional approaches in areas such as urban planning and transportation design. The identification of dense areas is a topic that is key to a variety of social dynamic studies, and as such has received attention from a variety of research fields. Nevertheless, the techniques used so far suffer from limitations – particularly when using large scale human activity data sets, such as mobile phone-call records – including poor spatial resolution due to the use of grids. In this paper, we have proposed the DAD-MST algorithm to identify dense areas from Call Detail Records that is able to process large scale datasets and respects the original tessellation of the space. The The DAD-MST algorithm has been tested and validated using a real CDR dataset of almost 50 million entries for over 1 million unique users over a four-month period. The dense areas identified have been qualitatively validated using the subway system of the city under study. The dynamics of the dense areas identified revealed the use that the citizens make of their city, indicating differences between different hour ranges and weekdays and weekends. We consider that although the results presented can only be applied to the city under study, the framework can be generalized to study social dynamics not only at an urban level, but also at a regional and national levels using the proper parameters of the DAD-MST algorithm. As future work we plan to combine our technique with data originating from other urban sensors, such as traffic, and analyze the applicability of our work for the recommendation of urban planning decisions.

References

[1] D. Sims and et al., "Scaling laws of marine predator search behaviour," *Nature*, vol. 451, 2008.

[2] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, 2008.

[3] N. Eagle, Y.-A. de Montjoye, and L. Bettencourt, "Community computing: Comparisons between rural and urban societies using mobile phone data," in *SocialCom*, 2009.

[4] M. Vieira, E. Frias-Martinez, P. Bakalov, V. Frias-Martinez, and V. Tsotras, "Querying Spatio-Temporal Patterns in Mobile Phone-Call Databases," *IEEE Mobile Data Management Conf.*, 2010.

[5] I. Benenson, "Modeling population dynamics in the city: from a regional to a multi-agent approach," *Discrete Dynamics in Nat. and Soc.*, 1999.

[6] I. Benenson and E. Hatna, "Human choice behavior makes city dynamics robust and, thus, predictable," in *GeoComputation*, 2003.

[7] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, 2006.

[8] L. Liao, D. J. Patterson, D. Fox, and H. Kautz, "Learning and inferring transportation routines," *Artificial Intelligence*, vol. 171, 2007.

[9] D. Brockmann, "Human mobility and spatial disease dynamics," *Review of Nonlinear Dynamics and Complexity - Wiley*, 2009.

[10] J. Schiller and A. Voisard, *Location-Based Services*. Morgan Kaufmann, 2004.

[11] W. Wang and R. Muntz, "Sting: A statistical information grid approach to spatial data mining," in *VLDB Conf.*, 1997.

[12] J. Ni and C. Ravishankar, "Pointwise-dense region queries in spatio-temporal databases," in *ICDE*, 2007.

[13] C. Jensen, D. Lin, B. C. Ooi, and R. Zhang, "Effective density queries on continuously moving objects," in *ICDE*, 2006.

[14] M. Hadjieleftheriou, G. Kollios, D. Gunopulos, and V. Tsotras, "On-line discovery of dense areas in spatio-temporal databases," in *SSTD*, 2003.

[15] D. Agarwal, A. McGregor, J. Phillips, S. Venkatasubramanian, and Z. Zhu, "Spatial scan statistics: Approximations and performance study," in *ACM SIGKDD*, 2006.

[16] M. Kulldorff, "A spatial scan statistic," *Communications in Statistics-Theory and methods*, 1997.

[17] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *WWW Conf.*, 2009.

[18] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, "Cellular census: Explorations in urban data collection," *Pervasive Computing*, 2007.

[19] C. Ratti, S. Williams, D. Frenchman, and R. M. Pulselli, "Mobile landscapes: using location data from cell phones for urban analysis," *Environment and Planning B: Planning and Design*, 2006.

[20] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A local-density based spatial clustering algorithm with noise," in *ACM SIGKDD*, 1996.

[21] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: An efficient data clustering method for very large databases," in *SIGMOD Conf.*, 1996.

[22] P. Kalnis, N. Mamoulis, and S. Bakiras, "On discovering moving clusters in spatio-temporal data," in *SSTD*, 2005, pp. 364–381.

[23] J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory clustering: a partition-and-group framework," in *SIGMOD Conf.*, 2007.

[24] J. B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proc. of the American Math. Soc.*, 1956.

[25] L. Guibas and J. Stolfi, "Primitives for the manipulation of general subdivisions and the computation of Voronoi," *Trans. on Graphics*, 1985.

[26] Y. Shiloach and U. Vishkin, "An O($log\ n$) parallel connectivity algorithm," *Journal of Algorithms*, vol. 3, 1982.

[27] M. Vieira, P. Bakalov, and V. Tsotras, "On-line discovery of flock patterns in spatio-temporal data," in *ACM SIGSPATIAL Conf.*, 2009.