

# From Routine to Network Deployment for Data Offloading in Metropolitan Areas

Eduardo Mucelli Rezende Oliveira  
Ecole Polytechnique and Inria

Aline Carneiro Viana  
Inria

**Abstract**—This paper tackles the WiFi hotspot deployment problem in a metropolitan area by leveraging mobile users' context and content, i.e., their trajectories, scenario interactions, and traffic demands. The careful deployment of hotspots in such areas allow to maximize WiFi offloading, a viable solution to the recent boost up of mobile data consumption. Our proposed strategy considers the restrictions imposed by transportation modes to people trajectories and the space-time interaction between people and urban locations, key points for an efficient network planning. Using a real-life metropolitan trace, we show our routine-based strategy guarantees higher offload ratio than the current approach in the literature while using a realistic traffic model.

## I. INTRODUCTION

With the steady growth of smart-phones sales [1], the demand for services that generate mobile data traffic has grown tremendously. In the last four years, the used amount mobile data has grown more than twice in relation to the previous year. The average generation of mobile data traffic grew 81% in 2012, and the global data traffic will increase 8-fold until 2017 [2]. This recent boost up of mobile data consumption are struggling the 3G cellular networks, which are not always prepared to receive such demand [3]. From 2007 to 2009, AT&T faced 5000 percent growth in mobile data traffic on its network [3]. In 2009 AT&T saw its network overloaded by mobile data traffic, mainly generated by iPhone users and, as a result, an unresponsive system with delayed text messages, dropped calls, and slow download speed [4].

WiFi offloading seems to be promising solution to the recent boost up of mobile data consumption that is making excessive demands on 3G cellular networks in metropolitan areas. [5], [6]. The idea consists in shifting the traffic off of cellular networks to WiFi networks. Carefully deploying WiFi hotspots can both be cheaper than upgrade the current cellular network structure and can concede significant improvement in the network capacity [5]. Nevertheless, one question remains: *how WiFi hotspots should be deployed?* The following factors make the answer to this question a challenge task.

The expansion of metropolitan areas increased the possibility of moving around [7]. This fact together with the increase of smartphone use results in highly dynamic links, which may significantly affect the performance of the network [8]. Moreover, people may use different transportation modes,

which significantly impacts their trajectories: e.g., a person riding a bike or walking can decide the path to follow contrarily to someone inside a bus. Finally, it is also important to take into account the space-time interaction between people and urban locations, a key point for an efficient network planning. Such considerations can reveal fundamental insights in terms of network usability. Popular sites for instance, are the source of the most of the traffic on the network [9].

In order to consider these issues, *this work tackles the WiFi hotspot deployment problem in a metropolitan area by leveraging mobile users' context and content, i.e., their trajectories, scenario interaction and traffic demands.* To the best of our knowledge, this is the first work leveraging users' three dimensional information (time, space and interest) in data offloading services. Our objective is to define what are the best places to receive WiFi hotspots in order to maximize the offloaded traffic on an urban scenario. This is a convenient solution for both cellular operators and users: The former can see the traffic being shifted to inexpensive networks while the latter can take advantage of higher data rates and less monetary costs than using cellular networks. We claim that unplanned deployment of hotspots may lead to both under-utilized and over-utilized network areas.

Related to our work, [10], [11] proposes the current approaches in the literature to deploy hotspots that considers user mobility characteristics only. Section II presents a deep comparison between those two works and the one herein proposed.

To accomplish our objective, we study the mobility context of people in a metropolitan area of a major city and identify a set of locations to well deploy WiFi hotspots (cf. Section III). Our strategy (cf. Section IV) is methodologically structured as follows. First, we create a time dependent graph to represent the interaction between people's mobility and locations suitable to receive a hotspot. Then, we measure how much data offloading a location can contribute to. For this, we use a metric herein proposed to rank which locations are suitable to support more data offloading: Better positioned hotspots are likely able to offload more data (cf. Section V).

Through experiments on a real-life trace, we evaluate the performance of our routine-based network deployment in terms of offloaded traffic, by varying the number of deployed hotspots. We also compare our solution with the current work on the literature. The results reveal that when using a realistic synthetic traffic model, our strategy provides up to 12% more offloaded traffic than the current solution on the literature (cf. Section

This work was supported by the EU FP7 ERANET program under grant CHIST-ERA-2012 MACACO.

VI). Finally, Section VIII concludes this work.

## II. RELATED WORK

To the best of our knowledge, [10], [11] are the most prominent approaches to deploy WiFi that takes in consideration user mobility characteristics.

In [10] authors create a mobility graph in which edges correspond to the road section traveled by the user and nodes are either builds or road intersections. Weight on the edges represent the number of Access Points (APs) required to completely cover the route. Based on a trace of which user connected on which AP, authors define the probability of a path being taken by a person. The probabilities guide the algorithms on choosing which APs are going to be deployed. Differently, in [11] no graph is built. Instead, the city is divided in a grid in which one hotspot will be positioned on each cell center. Based on the mobility of taxis, a data request is made every 5 seconds. Hotspots are ranked by their offload of data requests, i.e., better hotspots are those positioned in cells with more data requests.

Besides, they present some significant differences from the work herein proposed which are summarized in Table I.

First, our objective is not to provide continuous coverage as considered in [10]: In an urban scenario, this is prohibitively expensive since it would require the deployment of hotspots over the whole area, where most of them may be under-utilized. Similar to [11], our goal is to maximize the offload ratio. Second, we consider an urban scenario, which presents significantly higher complexity than the campus scenario considered in [10]: i.e., higher densities, many popular areas, diverse types of mobility (imposed by a variety of transportation modes), bigger area, etc. [11] employs a metropolitan scenario but considers taxi as the only transportation mode. Third, our approach is not restricted to the consideration of only one popular spot as done in [10]: In an urban scenario, diverse popular areas may exist and their features may vary according to space-time issues. Fourth, [10] does not apply a traffic model while [11] employs a non realistic traffic model in order to measure the offloaded traffic. Contrarily, our work uses a realistic synthetic traffic model that is strongly based on parameters and real measurements from the literature.

Some other related works provide solutions for data offloading but not related to hotspot deployment e.g., delegating the data offloading for people’s devices [6].

## III. RATIONALE

Before presenting our strategy for hotspot placement, we provide in this section the insights considered in our solution and discuss our system model.

### A. General view

People are routinary semi-rational entities, they have regular circles of actions guided by their decisions but unexpected situations may interfere on their directions [12]. A person may change their itinerary due to a traffic jam, problems on the public transportation, etc. When choosing an itinerary, people tend to use the shortest-path to reach their destination, also known as “*desire line*”. The desire line is the shortest line

TABLE I  
RELATED WORK COMPARISON

Work	[10]	[11]	Our proposal
Objective	Maximize continuous coverage	Maximize offloaded traffic	Maximize offloaded traffic
Scenario	Campus	City	City
Transportation modes	Walking	Taxi	10 different
Points of Interest	Real	Grid	Real
Traffic	Not considered	One request every 5 seconds	Based in the literature

between origin and destination, and expresses the way a person would like to go, if such a way were available [13]. Furthermore, the people’s itinerary is characterized by its *confinement*, i.e., despite of choosing the shortest-paths, people will roam close by their main physical address [14].

Inspired by these observations, our hotspot deployment proposal considers people context described by how they move. For this, we consider a real dataset describing movements in a large metropolitan area of almost two hundred people and several transportation modes.

In general terms, we want to investigate the possibility of providing a better networking service for people based on their *context* and *content*: i.e., their trajectories, scenario interactions and traffic demands. Specifically, in the work herein proposed, we are going to instantiate such service as a data offloading architecture. Therefore, our main question is: *how to provide a context- and content-aware data offloading architecture based on the people’s daily trajectories?*

### B. System model

Our system model represents a fairly real urban scenario composed of 182 people and their routes. Besides, we use data describing more than two hundred thousand real locations in a metropolitan area.

1) *Urban scenario*: In this work, we are using the latest version of GeoLife dataset [15]. GeoLife is considered to be unique in the literature. This is due to the fact that it provides a rich view of people mobility using 11 different transportation modes in an urban area for a long period of time. It provides geolocalized and timestamped points from 182 people during a 4 year span, from 2007 to 2011, mostly in Beijing. For each person, the dataset provides a set of geolocalized points ascendingly sorted by timestamp, i.e., a GPS trajectory. All components are based on geolocalized information, i.e., latitude and longitude coordinates within a 1960 km<sup>2</sup> central area in Beijing. Moreover, to better understand specific behaviors inherent from different periods of the day, every day is divided into four periods of time with 6 hours, from 00:00 to 05:59, from 06:00 to 11:59, from 12:00 to 17:59, and from 18:00 to 23:59.

2) *Trajectories*: A trajectory represents how people moves around and is described as a set of points representing GPS

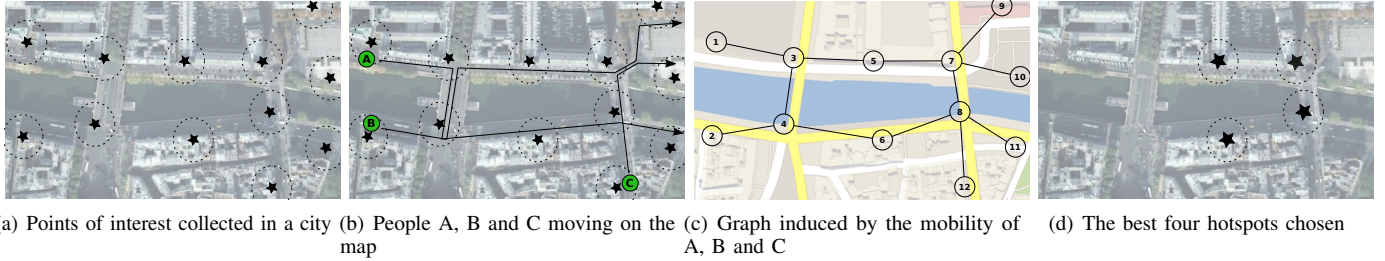


Fig. 1. A general view of our proposed methodology.

coordinates periodically collected. In GeoLife dataset, people may move around building their trajectories using at most ten different transportation modes such as taxi, bike, run, bus, walk, train, subway, car, boat, and motorcycle. Therefore, to capture this urban behavior, a trajectory is divided into *legs*. A leg is a subset of points from a trajectory with an unique transportation mode. For example, a trajectory composed by two legs, “car” and “walk” representing a situation in which a person went by car to his work, parked the car and went walking until his office.

3) *People and Points of Interest*: People move, build their trajectories, and carry mobile devices capable of WiFi communication and able to receive GPS information. While walking by, people may “interact” with points of interest, e.g., bar, bus station, supermarket, etc. Those points of interest describe more than mere locations in the map but they reflect a social aspect: e.g., students are frequently going to meet their colleagues in a coffeehouse close to the university they attend to. Indeed, this represents a routinary behavior that involves not only people but also their interaction with points of interest in a city. In our scenario, points of interest are geolocalized physical venues, e.g., bank, cafe, school, stadium, train station, university, etc.

4) *Communication*: While moving around a person passes by many points of interest and sometimes may stop by. We consider that the interaction between a person and a point of interest lasts as long as the former is inside the interaction range of the latter. The interaction range is defined as the WiFi range on a urban scenario. Taking into account the interferences caused by buildings, vehicles, or any other obstacle, the effective WiFi range in outdoor environment can vary from 5m to 75m [16]. We consider a interaction range of 50 meters in our experiments: if a person is, at most, 50 meters from a point of interest they are able to wirelessly communicate and consequently, the latter covers the former. It is important to enhance that initially a point of interest is not compulsory considered as a hotspot but as potential place to receive a WiFi hotspot structure.

## IV. PROPOSAL

In order to understand the relation between people and urban scenario during time, we construct a time dependent graph that describes interaction between people’s trajectories (edges) and points of interest (vertices). This section presents how mobility

is mapped into a time dependent graph and introduces our hotspot placement strategy.

### A. Graph creation

Let  $p \in P = \{1, 2, \dots, q\}$  be a geographic coordinate on the trajectory of a person. Let  $s \in S = \{1, 2, \dots, n\}$  be a point of interest. Both  $P$  and  $S$  are uniquely identified by their geographical position:  $\forall s(x,y) \in S, \forall p(x,y) \in P$ , where  $(x, y)$  is a pair of geographical coordinates. Since we are using real data traces containing GPS-based positioning,  $x$  and  $y$  are respectively, latitude and longitude. Fig. 1(b) illustrates a set of spots of interest presented as stars on map. Initially, we assign all points of interest with a hypothetical range illustrated as dashed circles<sup>1</sup>.

Our graph  $G(V, E)$  represents the interaction between people and the urban scenario. The graph should provide a simple and easy-to-see representation of people’s mobility. While walking, people create trajectories and their interactions with the urban scenario may point out important places to start planning the deployment of a network infra-structure. Fig. 1(b) illustrates the three trajectories created by three hypothetical people (A, B, and C) and their respective interactions with points of interest. A vertex  $v \in V = \{1, 2, \dots, w\}$  where  $w \leq q$  is created on the same coordinates of the point of interest if the latter covers at least one person passing by. A vertex coverage of a person  $a$  by point of interest  $s$  with a range  $r$  is expressed as  $\forall p \in P_a, \exists p | (p_x - s_x)^2 + (p_y - s_y)^2 \leq r^2$ , i.e.,  $a$  is inside  $s$ ’s circular range. Note that a point of interest that does not respect this condition is not considered as vertex in the graph.

An edge  $e \in E$  is created between two vertices if their corresponding points of interest sequentially cover a person during its trajectory. Fig. 1(c) shows the graph when three people are covered by points of interest.

### B. Routine characterization

The features present on people’s routinary mobility follow three intuitions: *Desire lines*, *confinement* and *repetitiveness* of people’s mobility. While repetitiveness is computed using a metric introduced hereafter, Desire lines and Confinement are described by Stress and Closeness Centrality calculated on the vertices of the graph. Following, related details are provided.

<sup>1</sup>Unit disk graph model.

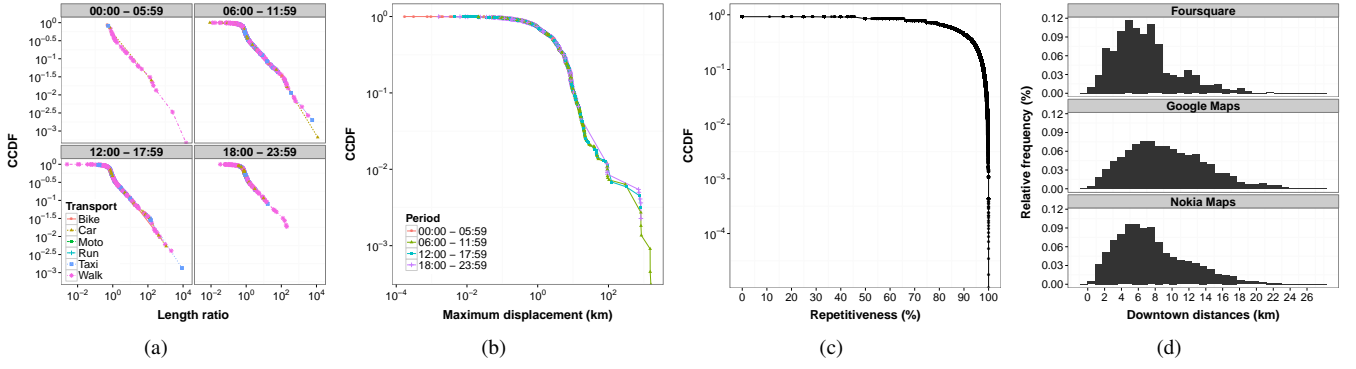


Fig. 2. (a) CCDF of the length ratio. (b) Maximum displacement. (c) Repetitiveness of points of interest. (d) Distance from Points of Interest to the downtown.

1) *Desire lines*: The concept of “desire lines” states that people tend to choose the shortest-paths to arrive on their destinations. In order to verify that, we have compared the length of each travelled leg against the length of the corresponding shortest path considering the same initial and final points of the original leg. Dividing the length of the original leg by the length of the shortest path allows us measuring how longer the path made by a person is from the shortest path. We have used Google Directions API<sup>2</sup> to compute the shortest path. The API receives the coordinates of both initial and final points and a *travel mode*, i.e., transportation mode. Then, it returns the shortest path considering the restrictions imposed by the existing routes and obstacles in the city for a specific transportation mode. Note that, we only have considered transportation modes where people have decision control of their paths. This excludes for example, buses, boats or trains.

Legs traveled by walking, running, and biking had their lengths divided by their respective shortest paths computed while using the API in *walking* mode. Google Directions API indeed has a *bicycling* travel mode, but at the moment, it does not contain routes in Beijing. Legs travelled by taxi, car, and motorcycle had their lengths divided by results of the API in the *driving* mode. Fig. 2(a) shows the CCDF of the ratio between the original legs length and the Google shortest path, by transportation mode and period. It shows that the periods from 00:00 to 05:59, from 06:00 to 11:59, from 12:00 to 17:59, and from 18:00 to 23:59 present, respectively, 46%, 70%, 62% and 83% of the legs measuring, at most, half longer than the shortest path. That results show that on late night people tend to walk around not directly going to their destination. Indeed, on late night people tend to go for bars, night clubs and are more susceptible to create routes that are way longer than the shortest ones. On the other hand, periods representing early morning and early night show high percentage of legs closest to the shortest one and describing how people go directly to their destinations, e.g., work, home, etc. The period containing the early afternoon hours present an intermediate percentage of legs close to the shortest path. Indeed, this period mixes people walking around careless about shortest paths (e.g, someone shopping, or looking

for restaurants), and people more concerned about their being on time (e.g, people coming back from the lunchtime towards the work). Moreover, it is possible to see that the length ratio does not change in function of the transportation mode. It shows that regardless of the transportation mode, people tend to be oriented by the shortest paths.

**Stress Centrality**: Considering such feature, vertices are ranked with the *stress centrality metric* [17]. Consider a graph  $G(V, E)$ , the Stress Centrality ( $st$ ) for a certain vertex  $v \in V$  is defined as follow:

$$st(v_i) = \sum_{j=1}^N \sum_{k=1}^{j-1} \sigma_{jk}(v_i) \quad (1)$$

where  $N$  is the number of vertices on the graph,  $\sigma_{jk}(v_i)$  is the number of shortest-paths that contains  $v_i$  between the pair of vertices  $v_j$  and  $v_k$ . Vertices with high stress are those that lay on most of people’s shortest routes and may become well positioned hotspots. Vertices with high stress values are likely to be places for hotspot deployment. Indeed, centrality metrics are widely used in the literature, but not yet explored in the problem we are considering.

2) *Confinement*: People’s mobility is generally confined. Even if people are not using the shortest routes, they are at least not going far from their home location. To check how that premise occurs on our scenario, we have measured how confined the trajectories are by their maximum displacement. Maximum displacement is the distance between a trajectory’s initial and farthest point (not necessarily the last point). Fig. 2(b) presents the CCDF of the maximum displacement for all trajectories grouped by period of the day. It shows that 90% of the trajectories have, at most, 10 km maximum displacement and the period of the day does not affect this behavior. This result shows that generally people do not move far away from their starting point, and presents high confinement.

**Closeness Centrality**: On our proposal, we consider *Closeness Centrality metric* [18] as a way to measure how a place is close to each others. Closeness is calculated based on the geodesic distance between all pairs of vertices on the network. Consider a graph  $G(V, E)$ , the Closeness Centrality ( $cl$ ) for a certain

<sup>2</sup><https://developers.google.com/maps/documentation/directions>

vertex  $v_i \in V$  is defined as follow:

$$cl(v_i) = \sum_{k=1}^N d(v_i, v_k) \quad (2)$$

where  $N$  is the number of vertices on the graph,  $i \neq k$  and  $d(v_i, v_k)$  is the shortest-path distance between  $v_i$  and  $v_k$ . It assigns higher values for vertices closer to the rest of the network, that is, on a city those are probably hospitals, and markets, i.e., places planned to be close to most of the people's trajectories.

3) *Repetitiveness*: People with confined mobility tend to repetitively visit the same areas. To quantitatively express how repetitively visited is a point of interest, a metric called *Repetitiveness* was conceived. The repetitiveness of a point of interest  $v_i$  is based on the number of unique users ( $NUU$ ) and total users ( $NTU$ ) that visited it as following:

$$re(v_i) = \frac{NTU_{v_i} - NUU_{v_i}}{NTU_{v_i}} * 100 \quad (3)$$

Fig. 2(c) shows the results of repetitiveness for all points of interest during the whole evaluated period. This result shows that the majority of the points of interest present high repetitiveness. Indeed, 67% of the points of interest present repetitiveness higher than 80%. Contrarily, only 2% of the points of interest have less than 50% of repetitiveness. Besides, 10% of the points of interest are highly repetitively visited and show 98% of repetitiveness.

### C. Synthetic traffic model

To the best of our knowledge, there is no freely available dataset describing both people's mobility and their traffic demands on an urban scenario. Therefore, we have created a realistic synthetic traffic model that takes into consideration traffic parameters and measurements from the literature [19], [20], [2], [21]. In addition, it takes in consideration different traffic demands required by people using different transportation modes. The traffic model herein discussed is responsible for simulating the content generation and offload as if it was done by people participating on GeoLife while performing their trajectories. Table II shows the parameters used on the synthetic traffic model.

1) *Content generation*: Considering projections made in [2], by 2017 two-thirds of the mobile data traffic will be used to transfer video. Therefore, our synthetic traffic model will generate data assuming users generating video-sized files. File size is Gamma distributed based on workload characterization study made on 250 thousand popular videos crawled from Youtube on a five month period [20]. Besides, its shape was taken from study made with more than 2.6 million videos crawled from Youtube during a three month period [19]. Furthermore, it has been shown that the Internet access times of mobile users are exponentially distributed [5].

The amount of data per day was estimated from monthly volumes projected in [2] and the transfer rate was projected in [2].

The synthetic traffic model simulates content creation by generating *loads*. A load contains an initial timestamp

TABLE II  
SYNTHETIC TRAFFIC MODEL PARAMETERS

Parameter	Value
File size	Gamma (shape=2 [19], scale=8.5 MB [20])
Inter-arrival time (IAT)	Exponential [5]
Amount per day	94.54 MB [2]
Transfer rate	3.9 Mbps [2]
Connection time	1.83 second [21]

indicating when it was generated, the file size, and the final timestamp that is the sum of its initial timestamp and transfer time. We consider a constant transfer rate during the connection between an user's device and a hotspot. Therefore, transfer time is the file size of the load divided by the transfer rate. For each leg of a person's trajectory, our model generates a load and divides it throughout the points of this leg as defined in Section III-B (Figure 3(a)). The division respects the duration of the leg and the available transfer rate, i.e., no content can be generated after the leg was travelled and the load is divided to points limited to the maximum transfer rate.

2) *Content offload*: When a hotspot covers a trajectory point, a content offloading is likely to happen (Figure 3(b)). Connection duration is taken into consideration on the data offloading, to do so, there is a 1.83 second gap between the start of a coverage and the availability of the data offloading. This was made aiming to simulate the IP acquisition time [21]. Therefore, after the initial connection gap, the traffic generated will be offloaded from trajectories points to the hotspot (Figure 3(b)).

Moreover, the synthetic traffic model was designed to couple with specific demands inherent in a scenario that contains people moving using different kinds of transportation modes. The objective is to create a scenario that is closer to the reality and, to do so, synthetic traffic model should consider only reasonable situations in which someone would be generating traffic depending on how this person is moving around. When a person is covered by a point of interest, and the transportation mode being used is "taxi", "bus", "walk", "train", "subway", "car" or "boat" the traffic generated by our model is normally made available to be offloaded. Correspondingly, when the person gets covered and uses other transportation modes, e.g., "run", "motorcycle" and "bike", no content will be available for offloading.

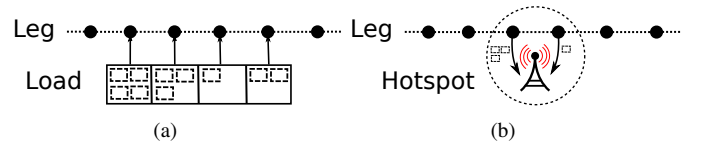


Fig. 3. (a) Content generation. (b) Content offload.

### D. Objective formalization

Consider a graph  $G_t(V_t, E_t)$  constructed as described in Section IV-A for a period of time  $t \in T = \{1, 2, \dots, u\}$ . Let  $\phi$  be the number of period of times per day  $d \in D$ , which represent the set of days in which we analyze mobility data.

Therefore,  $|T| = |D| * \phi$ .  $G_t$  represents all the interactions between people and urban scenario for the hours contained in  $t$ . Therefore,  $V_t \subseteq V$  is the the set of vertices that covered at least one person during  $t$  and  $E_t \subseteq E$  is the subset of edges that represents the mobility of people during  $t$ . Our objective is to maximize data offloading while limiting the number of deployed hotspots. We advocate that in order to deploy a WiFi network for data offloading on an urban scenario, it is important to take into consideration both traffic and mobility characteristics. For instance, consider two hotspots  $h_i$  and  $h_j$  that receive a certain amount  $\theta$  of traffic to be offloaded during a period of time  $t$ . If we consider only offloaded traffic, both would be equally selected with no priority. Our intuition is that if we also consider people’s mobility characteristics it is possible to provide a better network deployment for data offloading envisioning future traffic. Popular regions in the cities tend to be more popular and more traffic likely will be generated.

Consider the mobility characteristics metrics Stress ( $st$ ), Closeness ( $cl$ ) and Repetitiveness ( $re$ ) as described in Section IV-B. Consider the traffic ( $tr$ ) as the content offloaded as described in Section IV-C2. Consider  $U_{i,t,m}$  as the value of the metric  $m \in M = \{st, cl, re, tr\}$  on the period  $t$  for the vertex  $i$ . The traffic on a period  $t$  for the vertex  $i$  is calculated as the sum of all  $b$  bytes of content offloaded from all people covered by  $i$  during  $t$ , i.e.  $U_{i,t,tr} = \sum_b b_{i,t}$ . Due to the different magnitudes of the metrics, we have normalized  $U_{i,t,m}$  between 0 and 1 by the maximum value of  $m$  on the period  $t$ .

Let  $\vec{R}_{i,t}$  be a vector of metrics calculated for the vertex  $i \in V_t$ . Due to the mobility, a vertex  $i$  may appear on the graph on a certain period  $t = k$ , but do not appear on a period  $t = k + 1$ . Therefore,  $[R_{i,t,m} = U_{i,t,m} \cdot z_{i,t}]$ , where  $z_{i,t} = 1$  if  $i$  appears on  $t$  and  $z_{i,t} = 0$  if  $i$  does not appear in  $t$ . The result of metrics attributed to the same vertex on different periods is aggregated through a sum, i.e.,  $\sum_{i,t} R_{i,t,m} = Total_m \forall m \in M$ . Then, our problem is to find the set of vertices that maximizes the metrics with a limited number of hotspots  $\lambda$ , formally:

$$\begin{aligned} & \text{Maximize } \sum_{i,t} R_{i,t,m} \\ & \text{Subject to } \sum z_{i,t} < \lambda \end{aligned} \quad (4)$$

Besides, our model can be tuned in order to prioritize either mobility characteristics or traffic:  $\alpha * (\sum_{i,t} R_{i,t,st} + \sum_{i,t} R_{i,t,cl} + \sum_{i,t} R_{i,t,re}) + \beta * \sum_{i,t} R_{i,t,tr}$  by choosing weights  $\alpha$  and  $\beta$  depending on operators’ demands. Considering  $\alpha \rightarrow 0$  and  $\beta \rightarrow 1$  results in a deployment that prioritizes traffic demands. On the other hand, when  $\alpha \rightarrow 1$  and  $\beta \rightarrow 0$  hotspots will likely be placed on more important areas of the city regardless their traffic offload. For this first analyses, we are going to consider both traffic and mobility characteristics with the same importance.

1) *Spots selection*: The metrics are calculated on a offline centralized manner, i.e., the solution herein proposed may be applied on data previously collected by the interested entity, e.g., a Telecom operator. This premise is reasonable since network deployment planning generally uses historical data

TABLE III  
AVERAGE SPEED RANGES BY TRANSPORTATION MODE

Transportation mode	Average speed range (m/s)
Walk	$\leq 1.5$
Bus	$> 1.5$ and $\leq 4$ [22]
Bike (99.7%) or Run (0.3%)	$> 4$ and $\leq 4.4$ [23]
Taxi (39%), Motorcycle (0.1%), or Car (60%)	$> 4.4$ and $\leq 11.5$ [24]
Subway	$> 11.5$ and $\leq 28$ [25]
Train	$> 28$ and $\leq 250$
Airplane	$> 250$

as input. The result is a set of points of interest ranked by how good they are in providing traffic offloading. Besides, the points of interest are greedily selected based on our objective function. From that point, the operator needs to decide how many hotspots  $\lambda$  they are willing to deploy and the trade-off between the monetary cost and the predicted amount of WiFi offloaded traffic.

## V. EXPERIMENTAL SETUP

This section describes technical details from three components of our experimental setup: dataset, trajectories, points of interest.

### A. Dataset

All evaluations were made based on a subset of Geolife’s dataset. Since our work bases its premises on routinary behavior present on the mobility of people, a subset of Geolife data is already enough to show how we can explore routine to provide a better user experience. Therefore, the data of the two most active months are selected. This subset spans from 1st November to 31th December of 2008 and contains 39 users and 2203 trajectories.

### B. Trajectory labeling

A trajectory is a set of geolocalized points describing a route traveled by a person using at least one transportation mode. However, not every trajectory from the dataset herein considered are labeled with transportation modes. That is due to the fact that labeling was not mandatory for people participating on the GeoLife experiment. In order to label all the trajectories, we have created a model that labels legs by their speed compared against known average speeds for transportation modes in Beijing.

In particular, we calculate the average speed of a leg. The result is matched against known speed ranges in Beijing shown in Table III. Lets consider we have an unlabeled leg  $l$  traveled with average speed of  $5$  m/s. It is possible to see that  $l$  falls on a range in which three transportation modes present overlapping average speed ranges, “taxi”, “motorcycle”, and “car”. In order to keep the proportion of legs that were originally labeled by the users in the experiment, we have calculated the percentage of legs (shown between parentheses) on each overlapping range. Therefore, in this example,  $l$  will be labeled either as “taxi”, “motorcycle”, or “car” with 39%, 0.1% and 60% of chance, respectively.

### C. Points of Interest

The proposed graph is built based on a set of points of interest in Beijing. Points of interest are places generally present in most of major cities, e.g., bars, universities, supermarkets, etc. In order to study the deployment with real points of interest in Beijing, we have collected information from different databases of places (e.g., Google Places<sup>3</sup>). Such databases are growing and are the most accurate source of public information about points of interest. To avoid to be biased by the characteristics of one unique database, we have collected data from multiple sources namely Google Places, Nokia Maps, and Foursquare, counting more than 202 thousand real unique points of interest with their respective IDs, latitudes and longitudes. While Google Places and Nokia Maps databases provide information about points of interest collected from city hall, owner of venues, etc, Foursquare provides only information from places where its users checked in, generally places related with leisure and social relationships. For each set of places collected from a source, repeated ones were removed by keeping an unique occurrence of each ID. The set of collected points of interest contains 77919 from Google Places, 119346 from Nokia Maps, and 5059 from Foursquare totalizing 202324 places distributed in 98 categories, e.g., market, library, school, etc.

Figure 2(d) shows the distribution of points of interest by their distance to Beijing downtown. We have considered the downtown as being centered at the Forbidden city. This result shows that, regardless the source, points of interest are not evenly distributed in Beijing. Indeed, 70% of the points of interest are located at most 10km from the downtown.

## VI. PERFORMANCE EVALUATION

To better understand the results of our proposal, hereafter, we have first described the comparison approach then assessed our main results.

### A. Comparison

As described in Section II, [11] is the most related work with the one herein proposed. Therefore, we compare our proposal against it which we name “grid-based”. In order to create a comparable scenario considering the differences presented on Table I, we have adapted some aspects of the original scenario presented by the grid-based. It is important to enhance that both the original algorithm and the metric called “data requests” proposed in the grid-based work were applied as originally. The data requests metric ranks the cells based on the number of data requests generated inside the respective cell. Cells in which more data requests were generated are considered to be more capable on providing data offloading and, consequently, are better ranked.

Two are the main changes we have applied to the grid-based scenario in order to create a fair comparison scenario.

First, grid-based traffic model was solely a constant generation of data requests every 5 seconds. The simplicity of this content generation model does not employ any realistic characteristic. Consequently, the offload measurement made on the grid-based work likely does not represent a reliable

measurement. Therefore, we have chosen to apply our synthetic traffic model on the grid-based approach in order to have a more precise offload measurement.

Second, grid-based work divides the city on square-shaped cells and restrict one hotspot per cell positioned on its center, resulting on a grid fashion throughout the city. That is not realistic by any means, grid-shaped deployment does not respect the constraints of the city, e.g., a hotspot in the middle of a road may have its deployment hardly feasible. Therefore, in one of the compared approaches, we have used a subset of our collected point of interest, limiting one per cell selected as being the one closest to the center of the respective cell. For this purpose, Beijing map was divided into cells of 50 square meters, in a grid shape. To summarize, the approaches we compare against our solution are:

**Optimal:** The spots are deployed on points of interest selected as being those that maximize only the offloaded traffic. Points of interest and the synthetic traffic model are the ones provided in our work.

**Grid-based:** The spots selection algorithm, the data request metric and the spot positioning (i.e., center of each cell) are kept as presented in [11]. The synthetic traffic model is the one proposed in Section IV-C. Finally, the spots are chosen based on the data requests and their respective offloaded traffic is calculated based on the synthetic traffic.

**Grid-based (RS):** Differentiates from the Grid-based by its spots positioning, which are a subset of the real collected spots with one per cell selected to be the one closer to the cell center. This approach was created in order to use an approximate grid-shaped deployment proposed in [11], while using points of interest that are feasible to receive a hotspot.

### B. Offloaded traffic

Figure 4(a) shows the offloaded traffic by the different approaches presented on Section VI. It is important to note that, to be counted on the graph, a hotspot must have covered at least one person in the simulation scenario. Considering that Grid-based and Grid-based (RS) are limited by one hotspot per cell, their superior limit in number of hotspots is the number of cells in which the hotspots have covered someone. The number of cells which presents both trajectory points and hotspots is 19226. In addition, for the graphic, we consider only hotspots that offloaded traffic. Based in this criteria, Grid-based and Grid-based (RS) presents 1843 hotspots. Differently, our approach does not limit the number of points of interest per cell. Thus, our approach presents 2109 hotspots.

Regarding the offloaded traffic, the routine-based approach shows better results. For instance, when deploying 210, 506 and 1012 hotspots, 51%, 70% and 85% of data is offloaded, respectively. For the same amount of hotspots, the Grid-based approach presents 31%, 52% and 73% of offloaded traffic. This results shows that taking into consideration routine characteristics indeed provides better results than using the data requests approach proposed by the Grid-based. Considering the same amount of hotspots, the optimal approach presents 57%, 77% and 91% of offloaded data, a difference of 6%, 7% and 6%, respectively when compared to our approach. When comparing Grid-based and Grid-based (RS) no significant difference

<sup>3</sup><https://developers.google.com/maps/documentation/places>

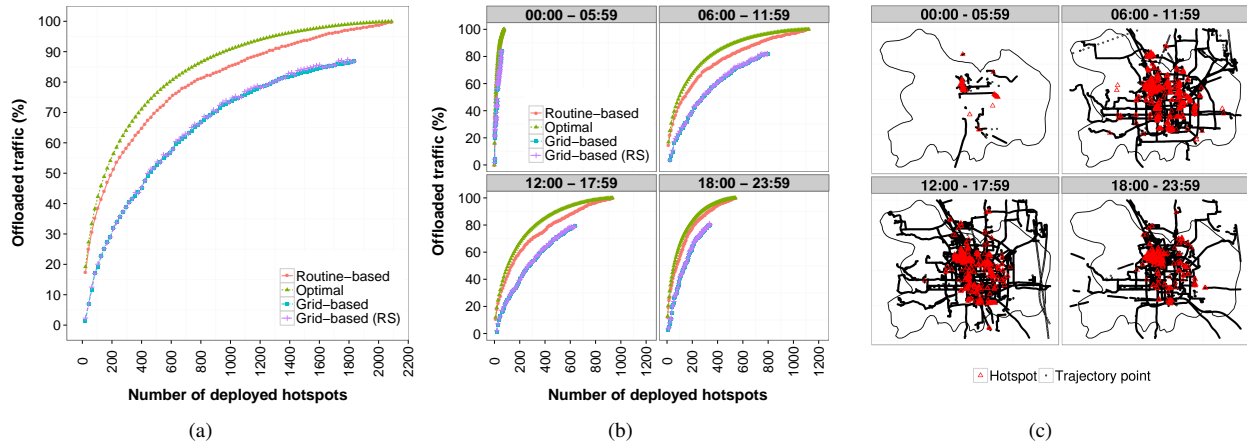


Fig. 4. (a) Offloaded traffic as a function of the number of deployed hotspots. (b) Offloaded traffic as a function of the number of deployed hotspots by period of the day. (c) The map of hotspots by period created with the routine-based approach.

appears. The usage of one real spot per cell positioned on an almost grid-shaped configuration does not increase significantly the efficiency of the Grid-based approach. Indeed, at most there is 1% of difference when comparing those two approaches regarding offloaded data. However, the difference should not be in the amount of offloaded traffic, but in reality, if the location can receive a hotspot.

Results presented on Figure 4(a) describe the offloaded traffic for a hotspot deployment configuration that merges result from different periods of the day. Due to the specific characteristics of the mobility on different periods of the day (e.g., period from 00:00 to 05:59 tends to have less trajectories than the period from 12:00 to 17:59), it might be important to assess the deployment per period. Indeed, individual mobile users consume resources in different ways depending on the time and the location that they access the network. Comprehending the differences inherent from different periods of the day allow operators to better allocate network resources.

Figure 4(b) shows the offloaded traffic given by the different approaches presented on Section VI, according to periods of the day. Each period of the day has its own number of trajectories which in turn generates a respective amount of traffic. If it is necessary to guarantee an offload ratio per period, different number of hotspots are needed to offload the traffic on each period of the day. Indeed, in order to achieve 80% of offloaded traffic on each period of the day, the routine-based approach employs, respectively, 36, 492, 457 and 216 hotspots for periods from 00:00 to 05:59, from 06:00 to 11:59, from 12:00 to 17:59, and from 18:00 to 23:59. More trajectories likely implies more areas being explored in the city. Consequently, more hotspots are needed to couple with the traffic demands. For the same coverage the grid-based approach needs 53, 739 and 340 hotspots for periods from 00:00 to 05:59, from 06:00 to 11:59, and from 18:00 to 23:59, respectively. The same approach was not able to reach 80% of data offloading for the period from 12:00 to 17:59.

Consider a requirement to deploy hotspots and guarantee a certain offload ratio per period. It is important to note that it is likely to happen an intersection between the sets

TABLE IV  
NUMBER OF COMMON HOTSPOTS BETWEEN PERIODS CONSIDERING 80% OF OFFLOADED TRAFFIC BY APPROACH

Approach	06:00 - 11:59		12:00 - 17:59		18:00 - 23:59	
	Routine	Grid	Routine	Grid	Routine	Grid
00:00 - 05:59	46	39	33	28	36	33
06:00 - 11:59	-	-	253	189	166	125
12:00 - 17:59	-	-	-	-	184	132

of hotspots from different periods. Table IV shows the number of common hotspots between periods by approach to guarantee 80% of offloaded traffic on each period. We refer to “Routine” and “Grid” the routine-based and grid-based approaches, respectively. More intersection means that more hotspots can be reused in a scenario that is capable of switch on and off certain hotspots during some periods of the day. On all intersections between pairs of periods the routine-based approach presented more intersection values. The intersection between all periods results in a few common hotspots due to the different mobility characteristics in each period. Indeed, the number of common hotspots for routine and grid-based approaches is 18 and 16 respectively.

Finally, Figure 4(c) (better visualized in colors) shows the map of the deployed hotspots using the routine-based approach over Beijing by period of the day. Regardless of the period of the day, a northwest region in the central area consistently shows a concentration of deployed hotspots, which is due to the fact that Microsoft Research Asia headquarter is located there. GeoLife experiment was conducted mostly with Microsoft members that were constantly walking nearby the working place. Therefore, many hotspots were deployed in that region in order to offload the traffic generated there.

## VII. DISCUSSION

It is important to note that as a first study using real points of interest, we do not consider the monetary cost variation per hotspot based on its location. On a completely real scenario the cost might vary depending on the location,



because the availability of power and backhaul capacity are widely different at different locations. In this context, it may be economically prohibitive to install hotspots at some of the locations determined to be most useful by the proposed algorithm. On the other hand, it may be very cheap to set up hotspots at other locations, for example, co-located with a cellular base station, fixed-line pay phones, or the homes of DSL customers. However, our solution could be completely applied with an improved version of our database of points of interest which adds their respective hotspots' installation cost. Then, our problem would be to find the set of vertices that maximizes the herein proposed metrics with a limited deployment cost  $\omega$ , i.e., maximize  $\sum_{i,t} R_{i,t,m}$  subject to  $\sum z_{i,t} < \omega$ , where  $z_{i,t} = \omega_i$  if  $i$  appears on the period  $t$  and  $z_{i,t} = 0$  if  $i$  does not appear in  $t$ .

Related to the cost issue, a partially connected topology may present some specific characteristics. For instance, consider a parking area that is not provided with network connection a-priori. The cost of linking the parking area to the operator's network may not pay off the offloaded traffic gains it will produce. A possible solution would be to place other hotspots connected to the network bridging the gap between the parking area and the operator's network. It is important to enhance that our resulting hotspots topology does not guarantee a connected topology, and we think that it would require a prohibitive amount of hotspots to create a completely connected graph covering an urban scenario. However, it may be interesting to assess the creation of connected components in order to provide offload on certain isolated areas.

## VIII. CONCLUSIONS AND NEXT STEPS

In this work, we have presented to the best of our knowledge, the first analyses of a metropolitan-wide hotspot deployment which employs a realistic traffic model. We have proposed a graph model to represent the relationship between people and the city infra-structure. Furthermore, we have created a realistic synthetic traffic model to deal with the traffic generation, which takes into consideration transportation modes. Besides, based on common behaviors presented on people's real routines and their traffic demands, we have proposed a methodology to select points of interest more suitable to receive hotspot placement data offloading. Our results show that with a small quantity of hotspots it is possible to provide high offload ratio and our routine-based approach provides higher offload ratio than the current solution in the literature.

As future work, we intend to investigate how good in terms of offload ratio is our deployment configuration for a different period of the dataset. Additionally, we are going to further investigate the monetary trade-off of a deployment in which the cost of the hotspots is not unique and the cost of offloaded traffic is considered. Finally, we envision to study the impact of changes on the mobility and traffic parameters  $\alpha$  and  $\beta$  on the deployed configuration.

## REFERENCES

[1] Gartner. (2011, April 7) Gartner says worldwide smartphone sales soared in fourth quarter of 2011 with 47 percent growth. Gartner. [Online]. Available: <http://www.gartner.com/it/page.jsp?id=1924314>

[2] Cisco. (2012, Feb.) Cisco visual networking index: Global mobile data traffic forecast update, 2011–2017. Cisco. [Online]. Available: [http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white\\_paper\\_c11-520862.html](http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html)

[3] AT&T. (2010, May) AT&T Launches Pilot Wi-Fi Project in Times Square. AT&T. [Online]. Available: <http://www.att.com/gen/press-room?pid=17961&cdv=news&newsarticleid=30838>

[4] J. Wortham. (2009, September 2) Customers Angered as iPhones Overload AT&T. The New York Times. [Online]. Available: <http://www.nytimes.com/2009/09/03/technology/companies/03att.html>

[5] K. Lee, I. Rhee, J. Lee, Y. Yi, and S. Chong, "Mobile data offloading: how much can wifi deliver?" *SIGCOMM Comp. Comm. Rev.*, vol. 41, no. 4, Aug. 2010.

[6] K. Thilakarathna, H. Petander, and A. Seneviratne, "Performance of content replication in mobitribe: A distributed architecture for mobile ugc sharing," in *Local Computer Networks (LCN), 2011 IEEE 36th Conference on*, October 2011, pp. 554–562.

[7] K. C. Seto, M. Fragkias, B. Gneralp, and M. K. Reilly, "A meta-analysis of global urban land expansion," *PLoS ONE*, vol. 6, no. 8, p. e23777, 08 2011.

[8] G. Judd, X. Wang, and P. Steenkiste, "Efficient channel-aware rate adaptation in dynamic environments," in *Proceedings of the 6th international conference on Mobile systems, applications, and services*, ser. MobiSys '08. New York, NY, USA: ACM, 2008, pp. 118–131.

[9] M. Kim, D. Kotz, and S. Kim, "Extracting a mobility model from real user traces," in *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, April 2006, pp. 1–13.

[10] T. Wang, G. Xing, M. Li, and W. Jia, "Efficient wifi deployment algorithms based on realistic mobility characteristics," in *IEEE MASS*, Nov. 2010, pp. 422–431.

[11] E. Bulut and B. K. Szymanski, "Wifi access point deployment for efficient mobile data offloading," in *Proceedings of the First ACM International Workshop on Practical Issues and Applications in Next Generation Wireless Networks*, ser. PINGEN '12. New York, NY, USA: ACM, 2012, pp. 45–50.

[12] D. S. Hamermesh, "Routine," National Bureau of Economic Research, Working Paper 9440, January 2003. [Online]. Available: <http://www.nber.org/papers/w9440>

[13] J. A. Throgmorton and B. Eckstein. (2000, Nov.) Desire lines: The Chicago area transportation study and the paradox of self in post-war America. [Online]. Available: <http://www.nottingham.ac.uk/3cities/throgeck.htm>

[14] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, Jan. 2006.

[15] Y. Zheng, X. Xie, and W.-Y. Ma, "Geolife: A collaborative social networking service among user, location and trajectory," *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–39, 2010.

[16] S. Dimatteo, P. Hui, B. Han, and V. Li, "Cellular traffic offloading through wifi networks," in *Mobile Adhoc and Sensor Systems (MASS), 2011 IEEE 8th International Conference on*, 2011, pp. 192–201.

[17] A. Shimbel, "Structural parameters of communication networks," *The bulletin of mathematical biophysics*, vol. 15, no. 4, pp. 501–507, 1953.

[18] G. Sabidussi, "The centrality index of a graph," *Psychometrika*, vol. 31, no. 4, pp. 581–603, 1966.

[19] A. Abhari and M. Soraya, "Workload generation for youtube," *Multimedia Tools Appl.*, vol. 46, no. 1, pp. 91–118, Jan. 2010.

[20] X. Cheng, C. Dale, and J. Liu, "Understanding the characteristics of internet short video sharing: Youtube as a case study," *CoRR*, vol. abs/0707.3670, 2007.

[21] D. Hadaller, S. Keshav, T. Brecht, and S. Agarwal, "Vehicular opportunistic communication under the microscope," in *Proceedings of the 5th international conference on Mobile systems, applications and services*, ser. MobiSys '07. New York, NY, USA: ACM, 2007, pp. 206–219.

[22] Z.-R. Peng, "Urban transportation strategies in chinese cities and their impacts on the urban poor," *Transportation Research Board 85th Annual Meeting*, p. 14, 2005.

[23] J. Lin, "Bicycles in beijing," project: Transportation (Green Design and the City).

[24] S. Wang, J. Sun, C. Shao, and F. Wang, *Evaluation on Vehicle Restriction Measure in Beijing*, ch. 39, pp. 433–443.

[25] Z. Xin. (2012, November) Subway line 6 to start running in december. China Daily. [Online]. Available: [http://www.chinadaily.com.cn/beijing/2012-11/26/content\\_15998073.htm](http://www.chinadaily.com.cn/beijing/2012-11/26/content_15998073.htm)