Impacts of Human Mobility in Mobile Data Offloading

Emanuel Lima, A. Aguiar

Instituto de Telecomunicações,
Department of Electrical and
Computer Engineering, Faculty of
Engineering of the University of
Porto
{emanuellima;anaa}@fe.up.pt

Paulo Carvalho

Centro Algoritmi, Departamento
de Informática, Universidade do
Minho
pmc@di.uminho.pt

Aline C. Viana

INRIA Saclay, France
aline.viana@inria.fr

ABSTRACT

Due to the limited coverage of WiFi APs, users' mobility has a severe impact on the performance of mobile offloading systems. The present study is a contribution in this context as offloading zones are identified and characterized from individual GPS trajectories when small offloading time windows are considered. The results show that (i) attending to users mobility, ten seconds is the minimum offloading time window that can be considered; (ii) offloading predictive methods can have variable performance according to the period of the day; and (iii) per-user opportunistic decision models can determine offloading system design and performance.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; • **Networks** → *Cloud computing*; *Location based services*;

KEYWORDS

Offloading; Human Mobility; Systems Design; EDA

1 INTRODUCTION

According to Cisco forecasts [6], mobile data traffic will grow at a compound annual growth rate of 47% from 2016 to 2021 with smartphones surpassing four-fifths of mobile data traffic. It is known that mobile network operators are struggling to keep up with such traffic demand, and part of the solution is to offload communications to WiFi networks [4]. In this sense, mobile data offloading systems can assist mobile devices in the decision making of when and what to offload to WiFi networks. However, due to the limited coverage of a WiFi AP (Access Point), the expected offloading performance of such a system is linked with the users mobility.

Several studies on the analysis of human mobility patterns have been carried out focusing on the identification and characterization of important locations in users' life in general. We intend to extend these works by studying human mobility from the perspective of mobile data offloading. This brings two major differences compared to the related work. First, high temporal resolution of positioning datasets is needed. In the majority of the related work, important locations have a temporal dimension representing the time spent by a user in that location, which confers its degree of importance. This time is usually in the order of several minutes which is suitable for the case of detecting important locations but not for a mobile data offloading scenario. Here, according to the amount of data traffic that needs to be offloaded, locations with a visiting temporal resolution of few seconds may be enough for data offloading. Thus, we expect to discover additional offloading opportunities, which were not visible with a coarser temporal resolution. Second, while important locations are usually limited in size, offloading locations can have arbitrary shape and size.

In this work, offloading regions are defined as spatially aggregated locations where users have mobility suitable to offload. The main contribution of this work are:

- The identification of offloading regions (ORs) on an individual basis through unsupervised learning;
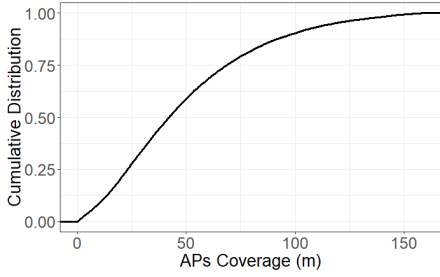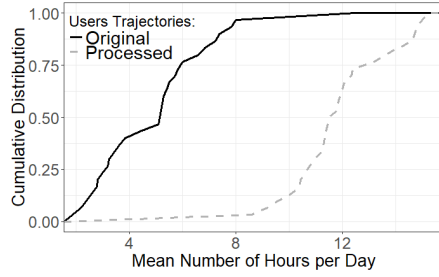
Figure 1: APs coverage (over 40k APs).

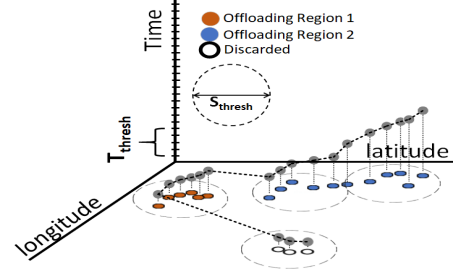**Figure 2: Daily hours with GPS per user.**

Figure 3: ORs extraction methodology from a user trajectory.

- The characterization of these regions in terms of availability, sojourn, and transition time based on their relevance;
- The discussion of the users mobility impacts on the design of mobile offloading systems.

The remainder of this paper is organized as follows. The methodology to identify and extract ORs from trajectory traces is explained in Section 2. The relevance categories to be used are defined in Section 3. The results are presented in Section 4 and the related work on human mobility patterns is discussed in Section 5. Finally, the conclusions are included in Section 6.

2 OFFLOADING REGIONS

Due to the limited coverage of WiFi APs, users' mobility dictates the time window available for offloading. Locations where a user is stopped or moving at lower speeds would be preferred for offloading, since the time window is maximized. In this context, we define an OR as a geographical area where a user exhibits a mobility suitable for offloading. In the next sections, we will introduce the methodology to define and extract ORs from trajectory traces.

2.1 Mobility Constraints for Offloading

In this work, the dataset studied in [9] is explored in order to select the mobility trajectories from 29 users with data from one to three weeks. During this period, GPS information such as users' positions and speed was being periodically sampled at 1 Hz using smartphones as probes.

An OR can be identified by applying space and time constraints to the users mobility. Space and time are interrelated as a user should stay in a geographical area for a certain amount of time. Therefore, the space and time thresholds - S_{thresh} and T_{thresh} - are defined in order to identify ORs.

Considering an offloading scenario based on WiFi networks, we use the coverage of APs within an urban environment as the space threshold to define an OR. Figure 1 shows the APs coverage for more than 40k APs, extracted from

the same dataset as [9]. The AP's coverage was computed as the euclidean distance traveled by the users while listening to beacons frames received from an AP. Only beacons received with good RSSI (higher than -85 dBm) were considered. Ground truth shows that more than 50% of the APs have a coverage range higher than 44m. Thus, the median coverage of an AP in a urban scenario is used as a threshold to the space constraint (i.e., $S_{thresh} = 44m$).

The T_{thresh} is the time spent by a user in a geographical area similar to an AP coverage range. Implicitly, this threshold represents the time window for offloading. Different applications may require different time windows for offloading depending on the offloading task and the network connection. In order to discover additional opportunities of using small time windows, different time thresholds are considered in this study ranging from 40 to 5 s ($T_{thresh} = \{40, 20, 10, 5\}$).

2.2 Trajectories Pre-Processing

Given the relationship between time and space constraints, the users' speed is limited as $max\_speed = \frac{S_{thresh}}{T_{thresh}}$. This allows to apply a high-speed filter with the goal to reduce the dataset size eliminating data samples which do not fulfill this relationship.

The users' trajectories may present gaps corresponding to periods where no GPS information was collected due to the lack of GPS signal in indoor environments or mobile devices/sensing applications being turned off. Moreover, some sensing applications used in crowdsensing, as in our case, use automatic start/stop strategies that minimize energy cost by collecting data only when the user is moving, and therefore, creating trajectories that favor movement phases.

In order to avoid the underestimation of ORs and capture the locations where users stopped or were inside buildings, we first identify the gaps corresponding to these cases (cf. Definition 2.1). Through the whole gap duration (in sec), we then use linear interpolation to fill the gaps and add pseudo locations to it.

Table 1: Activity periods in a day.

Definition	Day Period	Usual Activities
DP1	01:00h to 06:59h	Sleeping
DP2	07:00h to 09:59h	Commuting
DP3	10:00h to 11:59h	Working
DP4	12:00h to 13:59h	Lunch
DP5	14:00h to 16:59h	Working
DP6	17:00h to 19:59h	Commuting
DP7	20:00h to 00:59h	Social

Definition 2.1. For a user trajectory $Traj^u = (P_1^u, P_2^u, ..., P_n^u)$, a gap G has a beginning and an ending position corresponding to (P_i, P_{i+1}). We then define G as the spatio-temporal break satisfying the condition $(distance(P_{i+1}, P_i) < max(G_{dist})) \wedge (duration(P_{i+1}, P_i) > min(G_{dur}))$, for $min(G_{dur}) = \frac{max(G_{dist})}{max\_speed}$ and where max_speed depends on the T_{thresh} in use.

Hereafter, we define $max(G_{dist}) = 100m$ for capturing indoor locations. Figure 2 shows the mean number of hours where the location of users is known. As shown, after processing the trajectories based on the dataset in use, 50% of users are over 12h per day in known locations, 7h more than in the original trajectories. This shows the importance of filling the gaps when dealing with crowdsensing applications that favor movement phases during data collection.

2.3 Offloading Regions Extraction

In order to extract ORs from the users trajectory traces, the notion of offloading location candidate (OLC) is used. An OLC is a geographical area defined by S_{thresh} where a user stays during T_{thresh}. Then, an OR is defined as the aggregation of contiguous OLCs, as represented in Figure 3, reflecting aggregated areas where the user has a mobility suitable for offloading, e.g. walking path from home to bus station. When no aggregation is performed an OR corresponds to an OLC.

As mentioned before, we use $S_{thresh} = 44m$ and $T_{thresh} = \{40, 20, 10, 5\}$ s to define an OR. If a constant sampling rate is assumed (e.g., 1 Hz), OLCs will be represented by a geographical area with higher concentration of recorded location samples. Here, the number of samples correspond to the user sojourn time in that location. This advocates the use of an unsupervised density-based clustering algorithm to extract the ORs. Contrarily to partition-based approaches, density-based clustering allows to obtain clusters with arbitrary shape focusing only on the locations/areas with high concentration of points. The DBSCAN [5] clustering algorithm was used as it simplifies the process of identification and aggregation of OLCs.

There are two parameters that need to be defined in DBSCAN before the clustering process: *MinPts* and ϵ. A data point is a core point if it has at least *MinPts* in its neighborhood ϵ. We use the notion of core point to define an OLC. As
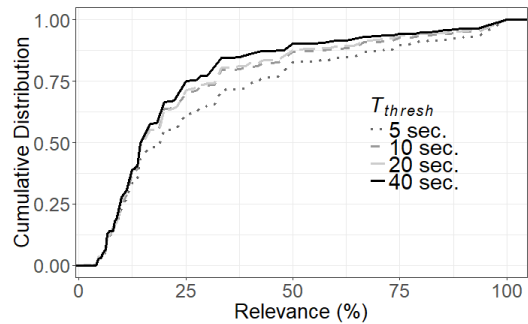

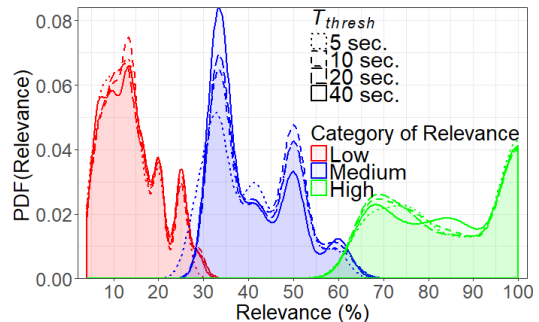**Figure 4: Offloading regions relevance.**

Figure 5: PDF of the ORs relevance per category.

in our dataset the location samples are produced at a fixed rate of 1 Hz, we define *MinPts* as T_{thresh} and the neighborhood ϵ as S_{thresh}. Then, ORs are the clusters formed by OLCs that are density connected. Clusters that do not meet the OLC criterion will be discarded, while adjacent OLCs will be aggregated in the same cluster forming ORs. Section 4 brings a detailed characterization of ORs extracted from the studied dataset.

The analysis will be focused on weekdays considering several day periods as representing established daily routines in peoples' life (Table 1). Therefore, for each user, the clustering process described above is applied to the trajectories traces of each day period.

3 OFFLOADING REGIONS RELEVANCE

Users may have several ORs, with distinct relevance. We define *relevance* based on the frequency a user visits an OR. For each user, we compute the conditional probability of visiting an OR in a given day period (Table 1). We choose the day as temporal metric to represent visits as it represents the reference time window when considering people's routine. The relevance is a simple but powerful metric since it captures the likelihood of a user visiting a specific OR attending to its mobility pattern (ORs visited every day will have a relevance value of 100%).

Figure 4 shows the relevance values of all ORs and users when considering different T_{thresh}. The majority of the ORs are visited occasionally (45% of them have a relevance lower than 15%), while a small minority is visited almost daily (10% of the ORs have a relevance higher than 77%). These results reflect the fact that users tend to visit many ORs, but very few on a regular basis. However, 20% of the ORs have a relevance higher than 50%. indicating that there are other ORs apart from home and work that can be used for offloading. ORs for $T_{thresh} = \{20, 10, 5\}$ s exhibit similar visiting patterns as a result of the similarity between their relevance distributions.

3.1 Categories of Relevance

Users visit different ORs per day with different roles in their lives. In order to better understand and characterize the ORs we classify them into different categories of relevance. The classification process encompass two technical challenges: i) the definition of the number of categories to use; and ii) the mapping of an OR to the right category of relevance. To define the number of categories that best suits all users, an unsupervised learning algorithm (namely k-means) is applied on a per user basis. Here, the relevance values from the ORs are clustered using a different number of clusters (k) that represents the number of categories to be used. Then, k is determined using the elbow method along with the total WSS (Within Sum of Squares) where k is chosen as being the number of clusters for which the WSS curve stabilizes [8]. For the vast majority of users, ground truth led to k=3 as the best number of clusters. Therefore, three categories of relevance were defined: *low, medium* and *high*.

Mapping the ORs to the categories mentioned above implies the definition of the relevance values to be used as bounds for each category. However, these bounds are dependent of the users' mobility patterns and so, they cannot be predefined. To solve this problem, the ORs' relevance values of each user are clustered using k-means with k=3, and then, each cluster is classified as *low, medium* or *high* category according to the distribution of relevance values.

Figure 5 shows the probability density function of the ORs' relevance for each relevance category obtained using the kernel density estimation (KDE). The good separation between the categories' distributions validates the choice of the number of categories of relevance (k=3) and the small overlapping area between distributions reveals that some users have similar bounds for each relevance category. Moreover, the results evince that ORs within the high-relevance category are visited by users almost daily with the distribution peaking at 100%. This category contains locations such as home, work place and commute paths. ORs within the medium-relevance category may be occasionally visited by users (favorite restaurant, gym, etc.) with the distribution
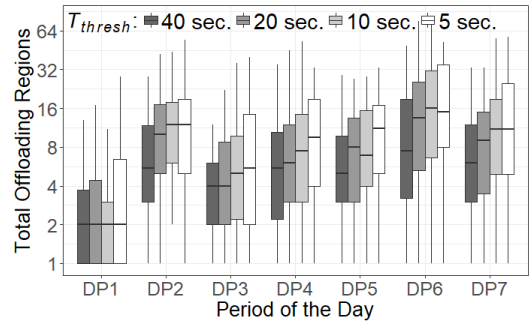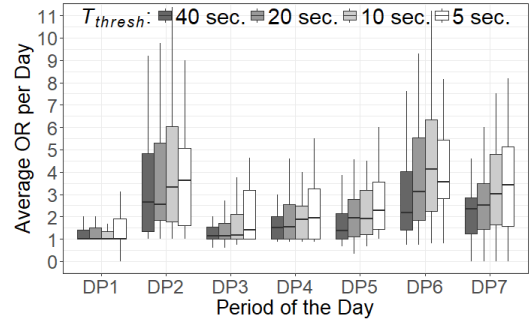

Figure 6: Total number of ORs per user.

Figure 7: Average number of ORs visited per day per user.

peaking at 32.5%. Finally, ORs within the low-relevance category are visited sporadically (distribution peaking at 13%).

4 CHARACTERIZING OFFLOADING REGIONS

Attending to the users mobility, ORs represent geographical areas where offloading tasks could be performed. In this section, these regions are characterized in order to better understand the impact of the human mobility on mobile data offloading.

4.1 Availability of Offloading Regions

The total number of ORs extracted from each user trajectory traces is presented in Figure 6. As shown, users have more ORs in the DP2, DP6 and DP7 periods. These periods are usually associated with higher mobility, which increases the probability of a user visiting new locations that fit the criteria of an OR. As expected, the night period (DP1) is the period with less ORs, where 50% of the users have less than 2 ORs.

Figure 7 depicts the users' average number of ORs visited per day. It is shown that for $T_{thresh} = \{40, 20, 10, 5\}$ s, 50% of the users visit, on average, more than 2.6, 2.5, 3.3 and 3.6 ORs in the DP2; 2.1, 3.1. 4.1 and 3.5 in DP6; and 2.3, 2.5, 3
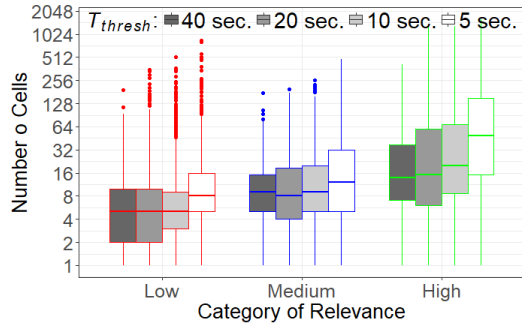
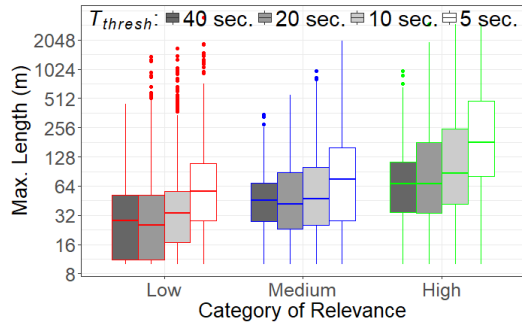**Figure 8: ORs' number of cells (11x11m).**

Figure 9: ORs' maximum length.

and 3.4 in DP7, respectively. However, for the other periods (DP3, DP4 and DP5), 75% of the users visit less than 3 ORs.

When evaluating the impact of considering different T_{thresh}, the results show that for $T_{thresh} = \{10, 5\}$ s the total and the average number of ORs visited per day is similar.

4.2 Offloading Sites Characteristics

ORs can be used as offloading sites and therefore their spatial characteristics represent the areas where offloading could be performed attending to the users' mobility. Figure 8 shows the distribution of the number of cells (11x11m grid squares) in the ORs for each category of relevance. As illustrated, the size of the ORs increases with the relevance category. However it is importance to notice that ORs extracted for $T_{thresh} = 5$ s are the largest. Defining small time windows for the OR extraction increases the number of OLCs that can be aggregated, forming larger ORs. However, it is interesting to notice that 5 s is the critical value since other T_{thresh} has a minimum impact on the ORs size.

ORs were extracted from the trajectory traces using a density-based clustering approach (section 2.3) and therefore, they can have arbitrary shapes. The number of cells reflects the total space but not the maximum distance that can be traveled inside an OR. For that, we computed the maximum length of the ORs as the maximum distance between two

cells belonging to the same OR. As shown in Figure 9, the maximum length of the ORs presents similar tendencies when compared to the number of cells regarding the impact of the T_{thresh} and the relevance categories.

The ORs in the high-relevance category are the largest regarding the maximum number of cells and the maximum length. In this category, ORs are visited almost daily, which increases the number of dataset samples in the surroundings of these locations contributing to the increasing of size. On the other hand, ORs in the low-relevance category counts with considerably less data samples reflecting the real size of the offloading location.

4.3 Sojourn and Transition Time

The sojourn time of a user in an OR represents the time available to perform an offloading task, while the transition time between ORs indicates for how long users will not be able to offload. Due to the reduced number of ORs during the night period (DP1), this period was excluded from the sojourn and transition time analysis.

Figure 10 shows the daily median sojourn time in the ORs for all users per category of relevance. The results show a positive correlation between sojourn time and relevance, where users stay longer in ORs with high relevance. This category includes ORs such as home and work, typically long stay locations. The outliers in this category correspond to OR than lie along the commute path.

Decreasing T_{thresh} reduces the global users' sojourn time in the ORs since more ORs with smaller sojourn time are considered due to the relaxation of the time constrain. This effect is different for each category of relevance. Whereas a gradual decrease of the users sojourn time occurs for the low-relevance category, for the medium-relevance categories and for $T_{thresh} = \{40, 20\}$ s, the users sojourn time is similar. The same occurs for $T_{thresh} = \{20, 10\}$ s. This shows that for this category of relevance, $T_{thresh} = \{20, 10\}$ s are the critical time windows to be used. Finally, regarding the high-relevance category, users usually spend more than 40 s in these ORs, therefore, any T_{thresh} can be used to extract these regions. The different sojourn times are related to the different ORs size (Figure 8) where larger ORs lead to slightly higher sojourn time.

Figure 11 shows the median time that users spend moving between ORs for each day period. The results show that, excepting DP3 and DP5, 75% of users stay less than 200 s in the transitions between ORs for $T_{thresh} = 40$ s. This time decreases with T_{thresh}, reaching its minimum when $T_{thresh} = 5$ s. Reducing T_{thresh} increases the number of ORs and their sizes, reducing consequently, the transition time.
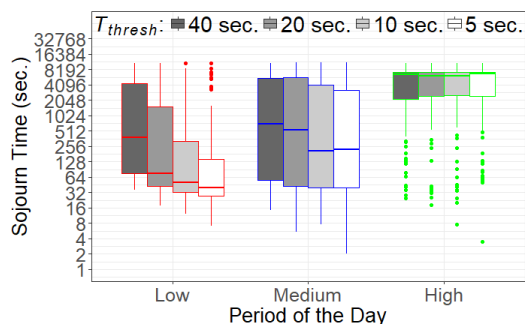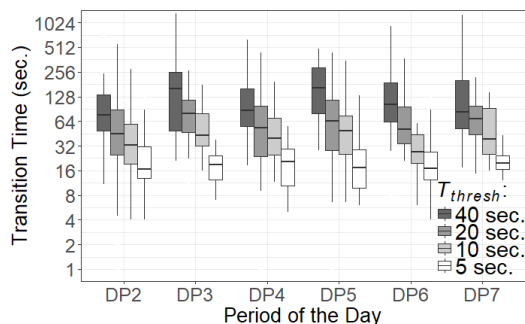
Figure 10: Sojourn time in the ORs per user.

Figure 11: Median transition time between ORs per user.

4.4 Discussion

In this section, we discuss the characterization results from the perspective of mobile data offloading systems. The discussion is focused on the impact of using different T_{thresh} when extracting ORs and the impacts of human mobility on the design of mobile data offloading systems.

4.4.1 Offloading Opportunities. T_{thresh} expresses the minimum time window to offload when extracting ORs (Section 2.3). Reducing T_{thresh} increases the number of OLCs, as users have to stay less time in a geographical area for it to be considered an OLC. This has different impacts on the availability and characteristics of the ORs. The *results show that for areas that are visited regularly, these OLCs are aggregated forming large ORs* (see Section 4.2). This effect is more significant for T_{thresh} lower than 10 s. OLCs that are not aggregated form small ORs, belonging to the low-relevance category. These ORs also present lower sojourn time evincing that *users tend to stay longer in ORs that they visit often.*

Similar results regarding the total and daily average number of ORs available to the users were obtained for a T_{thresh} of 5 and 10 s. This shows that, attending to the users' mobility, *partitioning an offloading task into data transfers lower*

than 10 s will not increase considerably the amount of offloading opportunities provided to users. However, the results show that T_{thresh} = 5 s halves the transition time when compared to T_{thresh} = 10 s.

4.4.2 Offloading Decision Models. Offloading decision models can combine mobility patterns with predictive methods to decide where and when to offload. To better understand the impact of the users' mobility on offloading decision models, we evaluate the likelihood of an OR belonging to a relevance category for each day period and different T_{thresh}. ORs belonging to the high and medium-relevance categories are frequently visited by users and, therefore, are easier to predict.

For each user, we compute the percentage of ORs in each relevance category for a specific day period. We then use the average as the probability of an OR to belong to a relevance category in a specific day period, see Figure 12. Results show that *ORs are more likely to belong to the low-relevance category for all periods of the day, excepting DP1 (night).* During this period, users are usually at home and, therefore, an OR is likely to have high relevance. After DP1, DP7 is the day period with higher probability for users to find ORs from the medium and high-relevance categories showing that in this period users present a more predictable mobility pattern, which can be leveraged by mobile offloading systems.

The commuting periods (DP2 and DP6) present the highest total number of ORs due to the higher mobility associated with these periods. However, interestingly, in the morning commuting period, users present a more predictable mobility than in the afternoon. In fact, DP6 is the most challenging day period to predictive methods, as the ORs have the highest probability to belong to the low-relevance category.

The high probability of users having ORs occasionally visited evinces the need for offloading decision models based on opportunistic strategies, as these regions may be used for offloading: 50% of the users present a median sojourn time higher than 40 s for T_{thresh} = 5 s.

When evaluating the availability and the characteristics of the ORs, the *results show high variability among users.* Therefore, to improve the performance of mobile offloading systems, *we advocate a per-user basis offloading decision model taking into consideration each individual mobility pattern.*

5 RELATED WORK

Works such as [2, 7, 10], use CDR (Call Detail Record) datasets to study human mobility patterns. CDRs data contains the user identifier and cell tower location once a user initiates or receives a voice call. However, this data is often sparse in space and time which is not adequate for a mobile data offloading scenario.
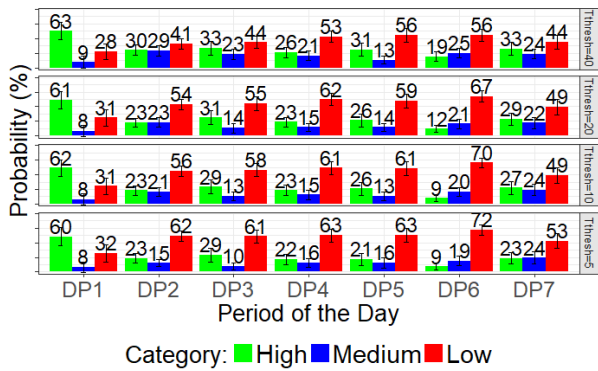
Figure 12: Probability of an OR belonging to a relevance category per period of the day.

The lower cost of GPS devices increased the possibility of collecting high-sampling mobility data. In [3], techniques capable of extracting semantic locations from GPS data are presented. In [1], the authors propose a two-step method to infer the significant locations through the loss of GPS signal and a variation of the k-means algorithm. As the loss of GPS signal is used as the main indicator to identify significant locations, places where the signal is available, such as outdoor places, may be lost. Moreover, rather than detecting locations with an arbitrary shape, they retrieve only circular locations. To overcome the k-means limitations, a series of density-based approaches have been proposed. In [13], the authors propose a density and join-based clustering algorithm called DJ-Cluster to infer significant locations. In [11, 12] similar approaches are used to extract significant locations in order to infer transportation modes and to predict the users' preferred locations.

The present work is inspired by the above methods. Nevertheless, our analysis is focused on a mobile data offloading scenario rather than on the identification of important locations to users.

6 CONCLUSION

In this work, we extracted ORs from individual trajectories of 29 users using a density-based clustering approach. These are zones where users exhibit a mobility suitable for offloading. We assessed the impact of the human mobility on mobile data offloading by inspecting the characteristics of these zones concerning their relevance, availability sojourn and transition time. Results showed that offloading time windows smaller that 10 seconds do not provide more offloading opportunities to users.

Regarding the offloading decision models, the results showed the need for per-user basis strategies and that predictive methods can be more effective for periods associated with social activities. However, due to the high percentage of ORs occasionally visited by users, opportunistic decision models can determine a better offloading system performance.

REFERENCES

[1] Daniel Ashbrook and Thad Starner. 2003. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous computing* 7, 5 (2003), 275–286.

[2] Francesco Calabrese, Massimo Colonna, Piero Lovisolo, Dario Parata, and Carlo Ratti. 2011. Real-time urban monitoring using cell phones: A case study in Rome. *IEEE Transactions on Intelligent Transportation Systems* 12, 1 (2011), 141–151.

[3] Xin Cao, Gao Cong, and Christian S Jensen. 2010. Mining significant semantic locations from GPS data. *Proceedings of the VLDB Endowment* 3, 1-2 (2010), 1009–1020.

[4] Man Hon Cheung and Jianwei Huang. 2015. DAWN: Delay-aware Wi-Fi offloading and network selection. *IEEE Journal on Selected Areas in Communications* 33, 6 (2015), 1214–1223.

[5] Sabhia Firdaus and Md Ashraf Uddin. 2015. A Survey on Clustering Algorithms and Complexity Analysis. *International Journal of Computer Science Issues (IJCSI)* 12, 2 (2015), 62.

[6] Cisco Visual Networking Index. 2017. Global Mobile Data Traffic Forecast Update, 2016–2021 White Paper. (2017). https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html (accessed 01-06-2018).

[7] Shan Jiang, Joseph Ferreira, and Marta C Gonzalez. 2017. Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. *IEEE Transactions on Big Data* 3, 2 (2017), 208–219.

[8] Trupti M Kodinariya and Prashant R Makwana. 2013. Review on determining number of Cluster in K-Means Clustering. *International Journal* 1, 6 (2013), 90–95.

[9] Emanuel Lima, Ana Aguiar, and Paulo Carvalho. 2017. Offloading Surrogates Characterization via Mobile Crowdsensing. In *Proceedings of the First ACM Workshop on Mobile Crowdsensing Systems and Applications*. ACM, 7–12.

[10] Michela Papandrea, Karim Keramat Jahromi, Matteo Zignani, Sabrina Gaito, Silvia Giordano, and Gian Paolo Rossi. 2016. On the properties of human mobility. *Computer Communications* 87 (2016), 19–36.

[11] Yu Zheng, Like Liu, Longhao Wang, and Xing Xie. 2008. Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 247–256.

[12] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. 2009. Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th international conference on World wide web*. ACM, 791–800.

[13] Changqing Zhou, Nupur Bhatnagar, Shashi Shekhar, and Loren Terveen. 2007. Mining personally important places from GPS tracks. In *Data Engineering Workshop, 2007 IEEE 23rd International Conference on*. IEEE, 517–526.