Poster Abstract: Using Unlabeled Wi-Fi Scan Data to Discover Occupancy Patterns of Private Households

Wilhelm Kleiminger, Christian Beckel
Institute for Pervasive Computing
ETH Zurich, Switzerland
{kleiminger,beckel}@inf.ethz.ch

Anind Dey
HCI Institute
Carnegie Mellon University, USA
anind@cs.cmu.edu

Silvia Santini
WSN Lab
TU Darmstadt, Germany
santinis@wsn.tu-darmstadt.de

ABSTRACT

This paper introduces the *homeset* algorithm, a novel approach to estimate occupancy schedules of private households from sensor data. The algorithm relies on unlabeled Wi-Fi scans and anonymized GPS traces collected by the mobile phones of household occupants and is able to autonomously determine the reliability of the computed schedules. We validate our approach using a data set from the Nokia Lausanne Data Collection Campaign that contains mobile phone traces of 38 participants over more than a year.

1. INTRODUCTION

Recent studies show that the analysis of human mobility traces allows to determine hot spots of social activities in a city, identify places of interest in the daily lives of individuals, or predict the places they will most likely be visiting next [1, 5, 6]. In this work, we focus on the problem of determining the *occupancy schedules* of users' households from their mobility traces. We present the *homeset* algorithm, which relies on Wi-Fi scans recorded by the mobile phones of households' occupants to determine such schedules.

Occupancy schedules are typically used to, e.g., develop and evaluate algorithms that perform smart heating control [3]. Actual ground truth occupancy data is however very cumbersome and time-consuming to collect and large, public data sets of occupancy data are not available yet. We show that the homeset algorithm is able to reliably retrieve occupancy schedules from raw Wi-Fi or GPS traces. The homeset algorithm thus enables extracting occupancy schedules from available data sets of human mobility traces. We validate our approach using a data set from the Nokia Lausanne Data Collection Campaign (MDC data set) that contains mobility traces of 38 users over more than one year [4].

SenSys '13, Nov 11-15 2013, Roma, Italy
ACM 978-1-4503-2027-6/13/11.

2. THE HOMESET ALGORITHM

The goal of the homeset algorithm is to compute the *occupancy schedule* of a household. To this end, the homeset algorithm relies on logs of Wi-Fi scans collected using the mobile phones of household's occupants. Each time a mobile phone detects the presence of a Wi-Fi access point (AP) it stores several pieces of information. Among these, the homeset algorithm only uses the timestamp of the scan and the MAC addresses of the visible APs. A single Wi-Fi scan is a tuple $< ts, AP_0, AP_1, \ldots, AP_{m-1} >$ where m is the total number of APs seen in a particular scan and AP_i is the MAC address of, and thus uniquely identifies, a specific AP. The homeset algorithm uses these scans to identify a set of APs that are located within, or in the immediate proximity of, the household of a mobile phone user. We call this set the *homeset* (HS) and assume it contains n APs, so that $HS = \{AP_0^{HS}, AP_1^{HS}, ..., AP_{n-1}^{HS}\}$.

For a single week, a *occupancy schedule* is represented as a matrix P with 7 columns and N_s rows. N_s is the number of temporal *slots* within a day. We set $N_s = 15$ minutes.[1]. Given a Wi-Fi scan $< ts, AP_0, AP_1, \ldots, AP_{m-1} >$ the homeset algorithm tests whether $\{AP_0, AP_1, AP_2, ..., AP_{m-1}\} \cap HS \neq \emptyset$. In the affirmative case, the algorithm assumes the household to be occupied in the slot i of day j corresponding to the timestamp of the scan. If no scan is available for a given time slot, heuristic methods can be applied to reconstruct the missing information [2].

To initialize the homeset algorithm the AP AP_0^{HS} is determined as described in [2]. Once AP_0^{HS} has been identified, the homeset is constructed by including in HS any other APs that appear in a Wi-Fi scan together with AP_0^{HS}. Relying on several access points instead of only on the "dominant" one (AP_0^{HS}) increases the reliability of the homeset algorithm. We quantify this increase in reliability using a metric called *stability*. We compute the stability π_x of an AP x over a certain time interval T_π, as the ratio of two quantities. The numerator is the total number of scans in which the access point x appears in the period T_π. The denominator is the total number of scans in the period T_π, whereby the scans are counted only if the access point x is seen at least once in the period T_π. In this study, we set T_π to be the interval between $3am$ and $4am$. A value of π_x equal to 1 thus means that if the access point is seen on any given night, it is going to be seen in all other scans

[1] In the data set, the interval between consecutive Wi-Fi scans is less than 15 minutes in 95% of the cases.

between 3am and 4am, and thus that it is a stable indicator of household occupancy. The rationale behind the fact that we consider a set of APs instead of a single one, is that a set of APs has a higher stability than a single one, even if this one is the private AP of the household. For instance, for user 009 in our data set using the HS instead of AP_0^{HS} only increases stability from 0.477 to 0.954. More extensive results are presented in [2].

3. VALIDATION

To evaluate the performance of the homeset algorithm ground-truth data about the absence from, or presence in, the household of the mobile phone owners is needed. As this information is not available in the MDC data set, we set out to validate our findings using an indirect approach. To this end, we leverage the fact that the GPS data available in the MDC data set has been partially anonymized in order to protect users' privacy. In particular, the latitude and longitude coordinates of selected places (e.g., users' home or workplace) have been occasionally truncated to the 3rd decimal digit. As the coordinates are reported along with a timestamp, it is possible to retrieve statistics about *when* participants were in such "sensitive" places.

We extract all the truncated instances of the GPS data from the data set and assign each unique pair of truncated latitude and longitude coordinates to a symbolic location k. For each location, we create a frequency count vector $\vec{CV}_k = (c_0, c_1, \dots, c_{23})$ with 24 elements, one for each hour of the day. Over the whole data set, we count the number of occurrences of a location k in a given hour of the day and store this value in the corresponding element of the vector CV_k. We thus count how many times a specific symbolic location has been "anonymized". Figure 1 shows the results of this analysis for participant 002 (the figure shows the 6 most relevant symbolic locations). As visible in this figure, location 1 is often anonymized between 1pm and 5pm and is never anonymized before 8am or after 9pm. We thus conjecture that this location corresponds to the workplace of the participant, as it is likely that between 1pm and 5pm the participant is at work and thus there is a higher need to truncate coordinates that correspond to this sensitive location. On the other side, location 5 is the one that is anonymized most frequently and consistently over the whole course of the day. Therefore, we conjecture that this is the location of the home of the participant.

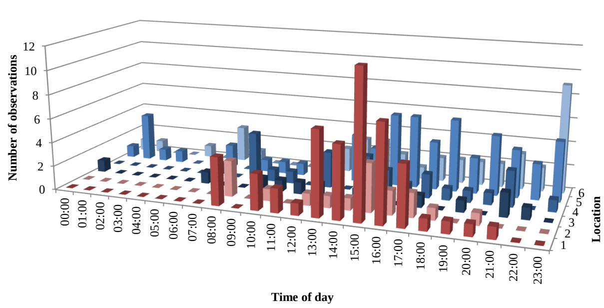

Figure 1: Time-frequency analysis of the anonymized locations for participant 002.

In order to automatically assess if a particular set of coordinates could identify a home location, we compute a score for each location. To make results comparable, we round CV_k to binary values and multiply it with a weighting vector $\vec{w} = (w_0, w_1, \dots, w_{23})$. Times between 9 and 17 (i.e., w_9 to w_{17}) are set to $\frac{2}{7}$ while all other times are set to 1. We chose this weighting assuming a normal "nine to five" schedule with little presence during the day except on weekends. A set of coordinates can score a maximum of 18.3 points under this metric. We have chosen a threshold of 10 for a location to be accepted as a possible home location.

After having retrieved the (truncated and thus anonymized) location of the home of each participant using the method described above, we compare the symbolic location with the GPS coordinates of the Wi-Fi APs. To this end, we compute the locations of the APs using temporal matching between the Wi-Fi and anonymized GPS data. For 20 out of the 38 participants included in the dataset, a match was found. Of the remaining cases, 13 times the score of the candidate locations was below 10 and in 5 cases no anonymized coordinates could be found for the homeset APs.

4. CONCLUSIONS

We described the homeset algorithm, a method to extract households' occupancy schedules from Wi-Fi scan traces. We evaluated the proposed approach using actual data from 38 users collected over more than one year. For this evaluation, we developed a technique that leverages anonymized GPS data to identify the home of mobile phone users. The derived occupancy schedules can be used to evaluate, e.g., algorithms for smart heating control. For more details about this work please refer to [2].

5. REFERENCES

[1] P. Baumann, W. Kleiminger, and S. Santini. The Influence of Temporal and Spatial Features on the Performance of Next-place Prediction Algorithms. In *Proc. of the 2013 ACM Intl. Joint Conf. on Pervasive and Ubiquitous Computing (UbiComp'13)*, Sept. 2013.

[2] W. Kleiminger, C. Beckel, A. Dey, and S. Santini. Inferring Household Occupancy Patterns from Unlabelled Sensor Data. Technical Report 795, ETH Zurich, Department of Computer Science, Sept. 2013.

[3] J. Krumm and A. J. B. Brush. Learning Time-based Presence Probabilities. In *Proc. of the 9th Intl. Conf. on Pervasive Computing (Pervasive'11)*, June 2011.

[4] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. Do, O. Dousse, J. Eberle, and M. Miettinen. The Mobile Data Challenge: Big Data for Mobile Computing Research. In *Proc. of the Mobile Data Challenge by Nokia Workshop (co-located with Pervasive'12)*, June 2012.

[5] R. Montoliu, J. Blom, and D. Gatica-Perez. Discovering Places of Interest in Everyday Life from Smartphone Data. *Multimedia Tools and Applications*, 62(1):179–207, Jan. 2013.

[6] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell. NextPlace: A Spatio-Temporal Prediction Framework for Pervasive Systems. In *Proc. of the 9th Intl. Conference on Pervasive Computing (Pervasive'11)*, June 2011.