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

1 Introduction

Recent studies show that using human mobility traces collected through mobile phones it is possible to, e.g., determine hot spots of social activities within a city [4], identify places of interest in the daily lives of individuals [5, 1], or predict the places that will most likely be visited by the mobile phone users [6]. In this work, we focus on the problem of determining the *occupancy schedules* of users' households. We present the *homeset* algorithm, a simple but effective approach to retrieve households' occupancy schedules. The algorithm relies on Wi-Fi scans recorded by the mobile phones of households' occupants and is able to autonomously determine the reliability of the computed schedules.

The availability of occupancy schedules is usually necessary to develop (and evaluate) algorithms that, e.g., perform smart heating control [2]. Actual ground truth occupancy data is however very cumbersome and time-consuming to collect and there exist no publicly available large data sets of occupancy data. We show that the homeset algorithm is able to reliably retrieve occupancy schedules from raw Wi-Fi or GPS traces thus enabling the extraction of occupancy schedules from existing human mobility traces. We validate our approach using a data set from the Nokia Lausanne Data Collection Campaign that contains mobility traces of 38 users over more than a year [3].

2 The *homeset* algorithm

The goal of the homeset algorithm is to compute the *occupancy schedule* of a mobile phone user. For a single week, a schedule is represented as a matrix P with 7 columns, one for each day of the week, and N_s rows. N_s is the number of temporal *slots* within a day. N_s can be set to an arbitrary value, depending on the time granularity of the schedules. In the context of this work we consider 15-minute slots[1].

To compute the occupancy schedules, the homeset algorithm relies on logs of Wi-Fi scans only. 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 thus 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 $HS = \{AP_0^{HS}, AP_1^{HS}, ..., AP_{n-1}^{HS}\}$. 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. The choice of a specific method to perform this data cleaning depends on the intended use of the data.

To initialize the homeset algorithm the AP AP_0^{HS} must be determined. If the household has a private AP, this should ideally be set to be AP_0^{HS}. As information about private APs is not available in the MDC data, we resort to an empirical procedure to obtain it. To this end, we compute the empirical probability ω_x of seeing AP x at least once between $3am$ and $4am$ on any particular night. This procedure relies on the assumption that people spend most of their nights at home. The AP with the highest value for ω_x is set to be AP_0^{HS}.

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 access point 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 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. Table 1 shows evidence of this observation for

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

ID	$\pi_{AP_0^{HS}}$	$\omega_{AP_0^{HS}}$	π_{HS}	ω_{HS}
009	0.477	0.78	0.954	0.962
026	0.588	0.875	0.963	0.971
117	0.375	0.825	0.964	0.997

Table 1. Empirical probability ω and stability π of $AP_0^H S$ and HS for users 009, 026, 117.

selected users included in our MDC data set. For user 009, for instance, using the HS instead of AP_0^{HS} only, increased stability from 0.477 to 0.954.

3 Validating the homeset algorithm

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 necessary. 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 GPS data available in the MDC data set and the fact this data has been partially anonymized in order to protect the privacy of the users. In particular, the latitude and longitude coordinates of sensitive 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 sensitive places, even though the "semantic" of the places is unknown.

We thus 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 then create a frequency count vector $\vec{CV}_k = (c_0, c_1, \ldots, c_{23})$ 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 "anonymised". Figure 1 shows the results of this analysis exemplary for participant 002, whereby we only display the 6 most relevant symbolic locations. As visible in this figure, location 1 is anonymised most of the times between $1pm$ and $5pm$ and is never anonymised 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 anonymised 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.

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 $\vec{w} = (w_0, w_1, \ldots, w_{23})$. 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.

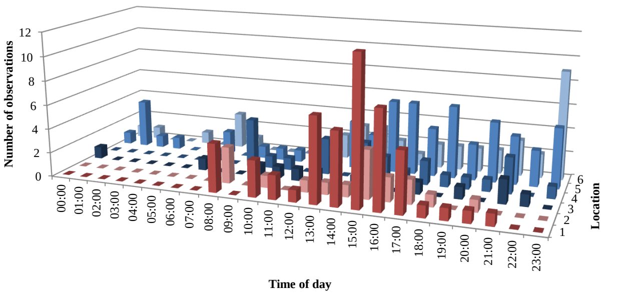

Figure 1. Time-frequency analysis of the anonymised locations for participant 002. The plot shows only locations with more than 10 observations in total.

Once we retrieved the (truncated and thus anonymised) 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 anonymised 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 anonymised coordinates could be found for the homeset APs.

4 Conclusions

This poster abstract describes a heuristic method – called the *homeset* algorithm – to extract occupancy schedules of private households from Wi-Fi scan traces. The effectiveness of the proposed approach is evaluated using actual data from 38 users collected over more than one year. For this evaluation, we developed a technique that leverages anonymised GPS data to identify the home of mobile phone users.

Using the occupancy schedules derived with the homeset algorithm we plan to develop and evaluate methods to predict when people arrive at home based on past behaviour.

5 References

[1] D. H. Kim, J. Hightower, R. Govindan, and D. Estrin. Discovering Semantically Meaningful Places from Pervasive RF-beacons. In *Proc. of the 11th Intl. Conf. on Ubiquitous Computing (UbiComp'09)*, Sept. 2009.

[2] 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.

[3] 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, in conjunction with the 10th Intl. Conf. on Pervasive Computing (Pervasive '12)*, June 2012.

[4] M. Loecher and T. Jebara. CitySense: Multiscale Space Time Clustering of GPS Points and Trajectories. In *Proc. of the Joint Statistical Meeting (JSM'09)*, August 2009.

[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, January 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.

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

Wilhelm Kleiminger[a], Christian Beckel[a], Anind Dey[b], Silvia Santini[c]
wilhelmk@ethz.ch, beckelc@ethz.ch, anind@cs.cmu.edu, santinis@wsn.tu-darmstadt.de
[a] Institute for Pervasive Computing, ETH Zurich, Switzerland
[b] HCI Institute, Carnegie Mellon University, USA
[c] Wireless Sensor Networks Lab, TU Darmstadt, Germany

The 11th ACM Conference on Embedded Networked Sensor Systems (SenSys 2013),
November 11th to 15th, Rome, Italy

Motivation: Obtaining reliable schedules for occupancy prediction algorithms

Building automation tasks such as smart heating require **occupancy schedules** [1,2]. Figure 1 shows such a (probabilistic) occupancy schedule, which determines the probability of a household being occupied at any point in time during the week.

Recently released datasets [3] **do not contain semantic place information** (labels), which would allow us to build occupancy schedules from Wi-Fi and GPS data.

We therefore propose to **extract occupancy schedules** from Wi-Fi scans using **temporal inference** and **implicit semantic information** introduced through anonymisation.

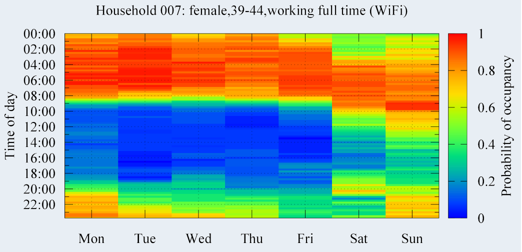

Figure 1: Probabilistic occupancy schedule for participant 007. The participant is likely not to be at home between 8am and 7pm.

The homeset: Initialisation and usage

Initialising the homeset:

The homeset algorithm uses Wi-Fi scans to identify the **homeset** HS, a set of access points located in close proximity of the household of a mobile phone owner (figure 2).

The homeset is built by first identifying AP_0^{HS}, the access point x with the highest empirical probability ω_x of being seen least once between $3am$ and $4am$.

HS then contains n access points, s.t. $HS = \{AP_0^{HS}, AP_1^{HS}, ..., AP_{n-1}^{HS}\}$ and $AP_1^{HS}, ..., AP_{n-1}^{HS}$ are access points which have appeared in a scan with AP_0^{HS}.

Building occupancy schedules:

Figure 3 shows given a Wi-Fi scan $< ts, AP_0, AP_1, \ldots, AP_{m-1} >$ the homeset algorithm tests whether $\{AP_0, AP_1, AP_2, ..., AP_{m-1}\} \cap HS \neq \emptyset$.

If the test returns a non-empty set at least once during a 15-minute interval, this interval is classified as *home*, otherwise the occupants are assumed to be *away*.

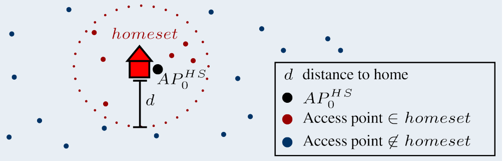
homeset
AP_0^{HS}
d
d distance to home
● AP_0^{HS}
● Access point $\in$ *homeset*
● Access point $\notin$ *homeset*

Figure 2: The homeset is the set of access points in the vicinity of the home access point AP_0^{HS}.

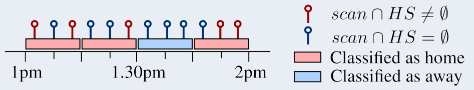

Figure 3: Interval classification based on multiple scans and homeset.

Homeset validation using implicit labels

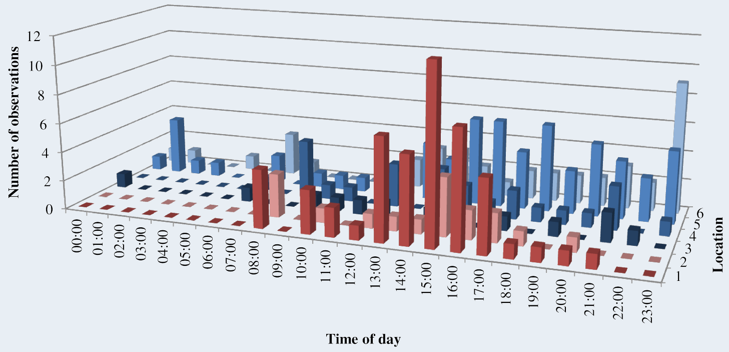

Figure 4: Time-frequency analysis of the anonymised locations for participant 002. Locations with less than 10 observations are excluded.

Over the whole data set, we count the number of occurrences of unique truncated (latitude, longitude) tuples k (i.e. anonymised coordinate pairs) in a given hour of the day. Figure 4 shows that location 5 is most probably the **home of the participant** as it is anonymised most frequently and consistently over the whole course of the day. As location 5 is anonymised most of the times between $1pm$ and $5pm$ and is never anonymised before $8am$ or after $9pm$, we conjecture that this is the **workplace of the participant**.

Dataset

Dataset obtained during Nokia Research's Lausanne Data Collection Campaign (LDCC) [3].

Data includes **GPS locations**, **Wi-Fi scans**, **accelerometer readings** and **call records**.

For privacy reasons, GPS coordinates are **truncated** and Wi-Fi SSIDs and BSSIDs are **hashed** when the participants were at sensitive places (home, office).

Conclusions

We build an **heuristic algorithm** to extract occupancy schedules from incomplete, anonymised mobile phone datasets.

The proposed approach was evaluated using actual data from 38 users collected over more than one year

Using **implicit semantic information** introduced through anonymisation we could validate the home location for 20 participants.

References

[1] J. Krumm, A. J. Brush, Learning time-based presence probabilities, in: Proc. Pervasive'11, San Francisco, CA, USA, IEEE, 2011, pp. 79-96.
[2] J. Lu, T. Sookoor, V. Srinivasan, G. Gao, B. Holben, J. Stankovic, E. Field, K. Whitehouse, The Smart Thermostat: Using occupancy sensors to save energy in homes, in: Proc. SenSys'10, Zurich, Switzerland, ACM, 2010, pp. 211-224.
[3] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, J. Laurila, Towards rich mobile phone datasets: Lausanne data collection campaign, in: Proc. ICPS'10, Berlin, Germany, ACM, 2010.