Detecting Dwelling in Urban Environments Using GPS, WiFi, and Geolocation Measurements

Niels Brouwers
Delft University of Technology
n.brouwers@tudelft.nl

Matthias Woehrle
Delft University of Technology
m.woehrle@tudelft.nl

ABSTRACT

A fundamental part of studying human mobility is to detect dwelling. When we dwell at places such as our homes, the supermarket or the local park, we are not necessarily still or stationary, but move around in a confined area. Most, but not all, of our significant places are indoors, which hampers the detection using GPS. In this work, we discuss three different sensor sources and their idiosyncrasies when used for dwelling detection. In particular, we use geolocation services based on collected WiFi fingerprints and show that they may be used complementary to GPS information. Our study is based on data collected on mobile phones in cities of various sizes in four different European countries.

1. INTRODUCTION

Understanding human mobility is crucial for many different application areas such as traffic prediction, city planning, and for determining social interactions. Therefore human mobility has been widely empirically studied in the social sciences, e. g., [2, 16]. Note that understanding mobility has two components: (i) Understanding how we move, i. e., determining transportation modes [13]. (ii) Understanding where we stop, i. e., determining the (important) *points of interests* (POIs) in our life [16]. For determining whether we stop at a POI, we need to distinguish whether we are *dwelling* at a location, e. g., at home or at our local supermarket, or if we are *mobile*. The focus of this paper is to determine whether users are dwelling based on traces collected on their mobile phone.

In previous work, empirical data has often been very coarse-granular, e. g., GSM cell tower information [15] or merely cellphone call data [2, 16]. However, modern smart phones provide a wealth of data including GPS and WiFi connectivity. Social sciences can benefit from these additional "sensors"by increasing the fidelity of models of human mobility. As with any sensor technology, the information measured is subject to noise, uncertainty and availability issues. Additionally, sensing, e. g., using the GPS chip, consumes energy.

PhoneSense 2011, Seattle, WA

This may negatively impact user experience by draining the battery; hence, the use of sensors needs to be carefully examined.

This work presents a comparative study of three different sensors and their quality w. r. t. determining whether users were dwelling. The sensors we consider are GPS, the "raw" WiFi scan information about surrounding access points (APs), and a geolocation service based on WiFi data. The contributions of this paper can be summarized as:

- We introduce an off-the-shelf geolocation service as a beneficial sensor type for detecting dwelling.
- We compare the data from three sensors for determining whether users are mobile or dwelling and the corresponding POIs based on offline analysis of mobile phone measurement traces.
- We present idiosyncrasies of the sensor types and identify that the information obtained by different sensors may be used complementary.

In the following we show how our work relates to previous work in Section 2. We present the sensors and their corresponding features that we utilize in Section 3. Section 4 discusses the data we collected. In Section 5 we present an evaluation of dwelling and POI detection on the collected data. We conclude in Section 6 with a summary.

2. RELATED WORK

Several researchers have proposed systems to detect mobility by monitoring the signal strength of beacons received from fixed network infrastructure such as GSM and WiFi APs. As a user moves around, the set of stations her mobile phone can overhear and the received signal strength (RSSI) of the corresponding beacons of these station, change over time. Moreover, the RSSI of individual base stations tends to fluctuate more when the receiving device is in motion.

Sohn et al. [15] apply these principles to GSM cell tower information and propose a classifier based on seven features. The classifier distinguishes between three mobility states (stationary, walking, and driving), and achieves an overall accuracy of 85%. Muthukrishnan et al. [10] show how to determine whether a user is in motion using similar features, however on information from WiFi scans. In [9] a combination of both GSM and WiFi was used to differentiate between the same three states as [15], achieving a classification rate of 88%.

Other sensors commonly found on mobile phones are accelerometers and GPS. These have been used in [13] to determine a classifier with an average accuracy of 93.6%. Using speed as indicated by the GPS receiver in combination with features extracted from the accelerometer data, the proposed classifier can additionally recognize biking and running.

Users move from one location to another; in between, they spend a certain amount of time at each location. Extracting these significant locations, i. e., POIs, can be done by analysis of time-annotated location traces. For example, [1] uses the fact that a GPS signal is lost indoors, and detects these cut-off points in the trace. Clustering is then used to gather such locations into POIs. However, relying on GPS signal loss will miss many important landmarks such as outdoor locations, indoor locations where GPS is still available, and will generate false-positives in urban canyons. Other approaches have focused on finding spatially and temporally constrained clusters in GPS traces [5, 6], but these assume clusters to be circular point clouds, and we found them not to be robust against noisy location estimates (cf. Figure 4). The DJ clustering algorithm [17] has fewer assumptions on clusters and performed the best in our experiments (cf. Section 5). In this paper we focus on simple decision tree models based on selected features and spatial clusters. However there may be a benefit for temporal modeling, e. g., using hidden Markov models [13] and conditional random fields [8]; we leave this as future work. Nurmi et al. [11] explore statistical methods in particular Markov chain monte carlo methods for detecting spatial clusters based on location information. Clusters are modeled as multivariate normal distributions. The described approach suffers from false positives as any data point no matter if part of a location or from travel between locations must be part of a cluster. Moreover, this approach only uses location information, it cannot directly use sequence or temporal information and cannot benefit from the wealth of information that mobile phones can collect.

Finally, SensLoc [7] integrates motion- and place detection using accelerometers, WiFi access point scanning, and GPS, to deliver a highly accurate system for location recognition and path tracking. However, WiFi scan results are used only for recognizing previously visited POIs and detecting entrance- and departure events, while localization relies entirely on GPS. In contrast, we perform a comparative study of different sensor sources, features and algorithms for the actual localization of POIs.

3. SENSORS

In the following we describe the three different sensors that we consider and the features that we extract from each individual sensor.

3.1 GPS

Most, if not all, modern smart phones come equipped with GPS sensors. These provide accurate measurements of both position and speed in outdoor locations, but signal quality is reduced or completely lost in indoor environments. Moreover, phone users tend to keep GPS turned off when not in use to avoid battery drain. When the GPS signal is available however, it tends to be a very good candidate for differentiating between dwelling and mobility [13].

We extract the following features for GPS: (*i*) Measured speed provided directly by the GPS, (*ii*) speed calculated from the distance of GPS locations, (*iii*) the difference between calculated and measured speed, (*iv*) a boolean that indicates whether the GPS had a fix, and (*v*) the number of location samples around the current location within a specific radius r_{loc}.

3.2 WiFi

Continuous scanning for WiFi APs has been used in context-aware computing to detect user mobility. This method is attractive because it can be performed on-line and in real-time, both desirable qualities for this class of applications. Several features have been discussed in literature, of which we have selected the following: (*i*) The Euclidean distance of RSSI measurements, (*ii*) the number of stations that are in the fingerprint (scan result) and (*iii*) the Jaccard index as a measure of difference between consecutive fingerprints.

3.3 Geolocation

An alternative to studying WiFi scan results directly is to pass them into a localization service such as Google's geolocation API [3] or Skyhook Wireless' localization service [14]. These services use large databases of location-annotated access point scans to compute a user location based on WiFi scan results. In this way the WiFi chip can act as a "poor man's" GPS, providing estimates of user location. Google trains its database using a background service built into Android devices that reports GPS coordinates and WiFi scan results to their servers at regular intervals. This results in good accuracy and broad coverage. Because of this and the open nature of the API, we chose this service for our experiments. We extract the following features: (*i*) the speed calculated from location distances, and (*ii*) the number of location samples around the current location within a specific radius r_{loc}. Note that these features are corresponding to those based on GPS locations.

Combinations.

In order to investigate whether individual sensor sources can be combined for an additional benefit in detection accuracy, we also look at the combinations of sensors.

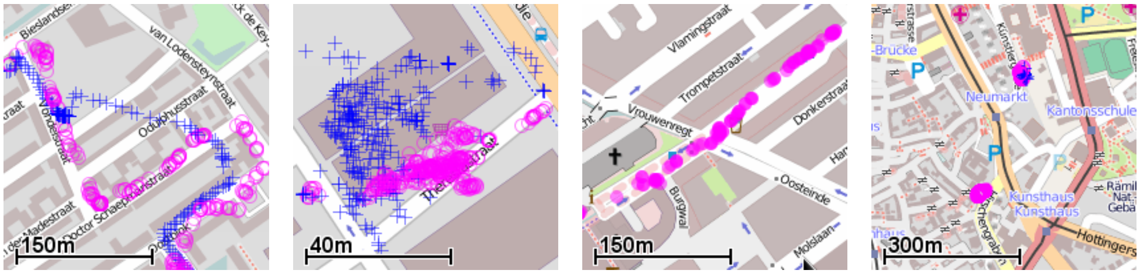
4. DATA COLLECTION

We collected data traces from various locations in seven cities in The Netherlands, Germany, Denmark, and Switzerland using a custom Android application on ZTE Blade and Sony Experia phones. The users are knowledge workers at a university and were asked to log and annotate parts of their day where they traveled by foot or bike to some place of interest in their lives. During the collection process users manually annotated traces with their state (walking, dwelling, ...) using the Android application on their mobile phones.[1] Since our emphasis is on detecting dwelling and POIs, we sought out suitable locations such as supermarkets, bars, bus stops, university buildings, and homes. In the following we discuss our data collection approach and the sensor data idiosyncrasies we identified in the collected data traces.

4.1 Data collection

WiFi scans were performed at 2 second intervals; for each scan we recorded the returned list of APs and their signal strengths, along with the most recent GPS measurement. The raw traces were sanitized by removing GPS outliers as well as WiFi beacons from locally administered APs [7] in

[1]We excluded driving in this work because it is considerably easier to distinguish from other activities.



(a) GPS Drift (b) Supermarket (c) Walking (d) University Building

Figure 1: Sensor idiosyncrasies. GPS and WiFi-Geolocation measurements shown in blue crosses and magenta circles, respectively. Note that in (d) there are two WiFi clusters.

Users	Traces	Samples GPS/WiFi	POIs (unique)
2	57	115,406 / 75,738	93 (54)

Table 1: Overview of the collected data from four western european countries. Note that we selected traces for a variety in activity and points of interests.

order to rely on fixed APs only. We then obtained geolocation data by passing the WiFi scan results to the Google geolocation service [3]. The returned result is a location and an estimate of its accuracy. Querying can be done either online from the phone if an Internet connection is available, or off-line at a central server.

Table 1 summarizes the data we collected over the course of 6 weeks. We selected 57 traces, containing a total of 115,406 GPS and 75,738 WiFi samples, collected in 8 different-sized cities (from $\approx 40,000$ to about $500,000$ inhabitants) across Western Europe. These traces include 93 dwelling locations. 54 locations are unique, i. e., some locations are visited multiple times like the users' home and a favorite café. 34 of the unique locations are indoors, while 20 are outdoors.

4.2 Sensor idiosyncrasies

GPS locations, especially from such low-end to mid-range smart phones as the ones used in this study, are not always perfect. Figure 1(a) shows the GPS trace drifting significantly through housing blocks. The geolocation estimates show the true path taken. On the other hand, we found that we were able to get a GPS lock in a surprisingly large number of indoor locations. For example, Figure 1(b) shows a trace from a supermarket. This shows that the assumption that GPS is always lost indoors simply does not hold.

The quality of Google's geolocation service depends on the accuracy of their database and how many APs are in range at a given location. We identified two interesting phenomena when inspecting the geolocation data. First, we found that even when a user is moving through a city at constant speed (i.e. walking), the returned locations are not spread out evenly, but rather tend to cluster around focal points along the route, as illustrated in Figure 1(a) and Figure 1(c). The second phenomenon was for dwelling locations: the reported location sometimes "jumped" between two points spaces several hundred meters apart when the user was dwelling, as shown in Figure 1(d).

5. EVALUATION

Based on the collected data, we performed two experiments: (i) Classification of user state, i. e., whether a user is mobile or dwelling, and (ii) detecting the POIs where a user dwells. Note that these two approaches are closely related, yet take a different approach. While the first merely relies on a single set of features at a given instant in time, the second approach focuses on spatial distribution of determined locations.[2]

5.1 Classification

We perform classification on the set of features of each individual sensor. In particular, we want to distinguish between *mobile* users and users that are *dwelling*.[3] In the following we therefore focus on supervised learning based on the two class labels. We use the weka data mining toolset [4] and in particular the J48 implementation of the C4.5 decision tree algorithm [12]. The selection of classifiers is a multi-objective problem: we want to maximize both (i) precision and (ii) recall. For each sensor (or set of sensors), we determine the power set of its features and determine a decision tree for each combination of features individually. We determine the Pareto set of classification results by maximizing for both precision and recall. Precision and recall are calculated by using 10-folds cross-validation in order to understand the generalization performance.

The extracted features have three distinct parameters: a *RSSI threshold* parameter for WiFi APs, a *window size* that determines the time intervals considered for a given feature and a *density* parameter for the clustering algorithm. We explored different parameter settings and compared the results based on the hyper-volume of the solution set. Our exploration yields the best results for a RSSI threshold of -80 dB, a window size of 50 seconds and a density radius $r_{loc} = 30m$.

Figure 2 details on the performance of the classifiers.[4] While two GPS classifiers are dominated by geolocation classifiers, the other two GPS-based classifiers have better recall than geolocation information, however with lower precision. The (pure) WiFi-based classifiers have good recall, yet a lower precision and are thus dominated by both geolocation and GPS results. Evident from the figure, the most accurate classifier can be obtained by using both GPS and geolocation data. Note that the combined classifiers do not necessarily use all available information. For the following presentation

[2] Another difference is that for classification, we perform supervised learning and rely on user annotations in the traces.

[3] Since we are focusing on dwelling in this work, for the classification we label all non-dwelling activities as mobile. We sanitize traces by removing instances without a class label.

[4] For the sake of clarity, we omit other combinations of sensor sources than gps and geolocation.

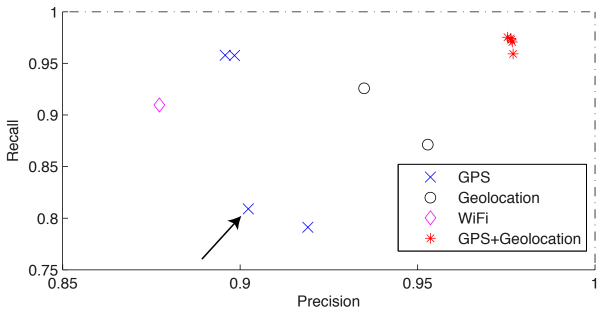


Figure 2: Plot of precision and recall results for different sensor sources. Intuitively the best classifier is near the upper right, i. e., has a precision close to 1 and a recall close to 1. Note that to show detailed differences both axis do not start at 0. The annotated classifier is a simple GPS classifier using only 2 features.

we annotate the simplest of the determined GPS classifiers in Figure 2. This GPS classifier merely uses 2 out of 5 features: (*i*) the difference between measured and calculated GPS speed and (*ii*) the information on whether GPS had a fix. For the sake of brevity we omit the full decision tree; it is fairly simple with 55 nodes and 28 leaves.

5.2 Dwelling Locations

In the following we focus on extracting POIs from GPS and geolocation localization data. Extracting POIs from sensor traces involves scanning the data for locations where the user spent a certain amount of time (i.e. more than two minutes). Such algorithms operate under the assumption that when a user is stationary, so are the location measurements. Unfortunately when this does not hold, as is common for both GPS and WiFi, it means that not all POIs will be found.

In order to determine how GPS and WiFi geolocation compare for detecting POIs we extracted sub-traces from our data set of dwelling sessions lasting at least three minutes. This resulted in a set of 93 POIs, where 21 are outside and 72 are indoors. We compute the coverage of both localization data, GPS and geolocation, at each of these POIs. Coverage is calculated as the number of samples where the sensor was able to compute an up-to-date location divided by the total number of samples. Figure 3 shows the distribution of this metric. We found, in accordance with our intuition, that WiFi is available in almost any urban setting. GPS coverage varied from site-to-site, but still was able to provide good coverage, even in many indoor locations.

We evaluate the quality of the two sensors and their applicability to the location detection problem by investigating their location point clouds. Figure 4 shows two examples, an indoor and outdoor location respectively. The GPS cloud is much more compact in the outdoor case, whereas the WiFi cloud is much more concentrated in the indoor example.

Several low-complexity clustering algorithms have been proposed in literature for POI extraction. We implemented four different algorithms: Ashbrook et al. [1], Hariharan et al. [5], the time-based clustering algorithm found in [6], and the DJ-Cluster algorithm from [17]. All four algorithms use parameters in both the spatial and time domain. The distance

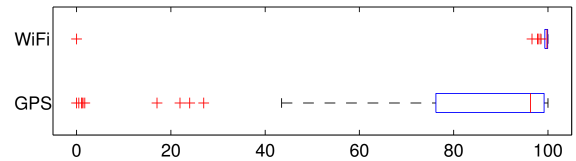
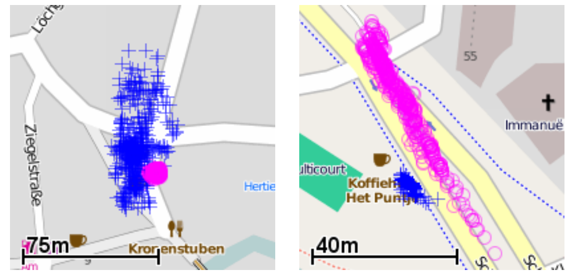


Figure 3: Coverage (%) of WiFi and GPS location services.



(a) Café (b) Outdoor Coffee Place

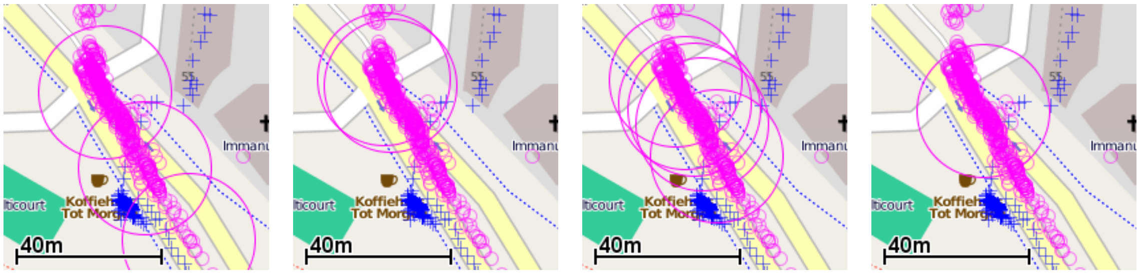
Figure 4: Dwelling locations. GPS and WiFi measurements shown in blue crosses and magenta circles respectively.

parameter is set to 20m; the time threshold for a location to be identified as a POI is set to two minutes.

From visual inspection of the results, we found that most algorithms were not able to cope with oddly-shaped point clouds. Figure 5 shows a particularly extreme case where WiFi coverage is low, and the location estimates drift over time. Only the DJ-Cluster algorithm was able to detect this cloud as a single cluster, because DJ-Cluster contains logic that merges overlapping clusters.

We compare GPS and geolocation by running the clustering algorithms on the respective point clouds of each of the traces. In the ideal case, a single cluster should be returned for each trace. When a clustering algorithm returns zero clusters, this is counted as a false negative. When more than one cluster is returned, these extra clusters are counted as false positives. We also perform a simple heuristic to combine results from both sensors. We gather the clusters extracted from the GPS and geolocation data, and then remove all geolocation clusters that overlap with GPS clusters. In this way the WiFi chip acts as a back-up for when GPS is not available.

Table 2 shows the results. As previously mentioned, the first three algorithms generate many false positives when the location estimates are noisy, resulting in low precision. DJ-Cluster provides the best results overall, so we will focus our discussion on the DJ-cluster results in the lower part of Table 2. When only using GPS, precision is high but recall is only 70.97% simply because sensor data was not available everywhere. WiFi geolocation on its own yields surprisingly good results but has a lower precision because the "jumping" problem discussed in Section 4.2 caused it to yield more false positives. When the sensors are combined, overall coverage is increased. In several cases where the geolocation was "jumping", GPS was actually available to provide a single, concentrated cluster. The combined results therefore miss fewer locations and generate fewer false positives, which results in a precision of 98.88% and a recall of 94.62%.



(a) Ashbrook (b) Hariharan (c) Time-based Clustering (d) DJ-Cluster

Figure 5: Clustering algorithms applied to a noisy WiFi point cloud.

Method	Precision	Recall
Ashbrook GPS	58.59%	80.65%
Ashbrook WiFi	73.33%	94.62%
Ashbrook GPS+WiFi	60.00%	96.77%
Hariharan GPS	63.12%	95.70%
Hariharan WiFi	51.72%	96.77%
Hariharan GPS+WiFi	55.09%	98.92%
Timebased GPS	69.70%	74.19%
Timebased WiFi	70.49%	92.47%
Timebased GPS+WiFi	70.08%	95.70%
DJ-Cluster GPS	98.51%	70.97%
DJ-Cluster WiFi	96.63%	92.47%
DJ-Cluster GPS+WiFi	98.88%	94.62%

Table 2: Clustering results

5.3 Discussion

Our comparative study of different sensors is based on empirical data. As such the selection of users, mobile phones using the Android OS and the selection of POIs may be biased; however, we specifically collected data from various different cities across several European countries. The difference in these locations varies the sensor measurements, however the discussed phenomena remain the same. Nevertheless, we identified in our traces characteristics of Western European cities: Firstly, the population density in urban areas results in urban canyons that considerably challenge GPS measurements. Secondly, WiFi coverage is excellent. The geolocation information based on WiFi scans provides mostly accurate location estimates. In turn, in these environments GPS and WiFi can be seen as complementary sensor sources. Still, larger studies are needed to validate these findings.

6. SUMMARY

In this work, we studied how GPS and WiFi information can be used in order to detect that a user dwelled at a certain POI. Firstly, we studied how the different sensors support the building of a decision tree classifier for detecting dwelling. Secondly, we investigated the use in the detection of places where the users dwelled, their points of interest. We verified that GPS is the prime candidate for determining dwelling of users. Additionally, we could identify that geolocation is an interesting sensor that is complementary to GPS information and helps dwelling and POI detection.

7. REFERENCES

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

[2] Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. Understanding individual human mobility patterns. *Nature 453*, 7196 (06 2008), 779–782.

[3] Google geolocation API. `http://code.google.com/apis/gears/api_geolocation.html`.

[4] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The WEKA data mining software: an update. *SIGKDD Explor. Newsl. 11* (2009), 10–18.

[5] Hariharan, R., and Toyama, K. Project lachesis: Parsing and modeling location histories. In *Geographic Information Science*, vol. 3234 of *LNCS*. Berlin, Heidelberg, 2004, ch. 8, pp. 106–124.

[6] Kang, J. H., Welbourne, W., Stewart, B., and Borriello, G. Extracting places from traces of locations. In *2nd ACM Int'l Workshop on Wireless mobile applications and services on WLAN hotspots* (2004), pp. 110–118.

[7] Kim, D. H., Kim, Y., Estrin, D., and Srivastava, M. B. SensLoc: sensing everyday places and paths using less energy. In *8th ACM Conf. on Embedded Networked Sensor Systems* (New York, NY, USA, 2010), pp. 43–56.

[8] Liao, L., Fox, D., and Kautz, H. Extracting places and activities from gps traces using hierarchical conditional random fields. *Int'l Journal of Robotics Research 26* (2007).

[9] Mun, M., Estrin, D., Burke, D., and Hansen, M. Parsimonious mobility classification using GSM and WiFi traces. In *5th Workshop on Embedded Networked Sensors* (2008).

[10] Muthukrishnan, K., van der Zwaag, B. J., and Havinga, P. Inferring motion and location using WLAN RSSI. In *2nd int'l Conf. on Mobile entity localization and tracking in GPS-less environments* (2009), pp. 163–182.

[11] Nurmi, P., and Bhattacharya, S. Identifying meaningful places: The non-parametric way. In *6th Int'l Conf. on Pervasive Computing* (2008), pp. 111–127.

[12] Quinlan, J. R. *C4.5: programs for machine learning.* Morgan Kaufmann Publishers Inc., 1993.

[13] Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., and Srivastava, M. Using mobile phones to determine transportation modes. *ACM Trans. Sen. Netw. 6* (March 2010), 13:1–13:27.

[14] Skyhook wireless. `http://www.skyhookwireless.com/`.

[15] Sohn, T., Varshavsky, A., LaMarca, A., Chen, M., Choudhury, T., Smith, I., Consolvo, S., Hightower, J., Griswold, W., and de Lara, E. Mobility detection using everyday GSM traces. In *UbiComp 2006: Ubiquitous Computing*, vol. 4206 of *LNCS*. 2006, pp. 212–224.

[16] Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. Limits of predictability in human mobility. *Science 327*, 5968 (2010), 1018–1021.

[17] Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., and Terveen, L. Discovering personal gazetteers: an interactive clustering approach. In *12th ACM Int'l Workshop on Geographic information systems* (New York, NY, USA, 2004), GIS '04, pp. 266–273.