ARTICLE IN PRESS

Pervasive and Mobile Computing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

journal homepage: www.elsevier.com/locate/pmc

Anonymizing mobility data using semantic cloaking

Omer Barak, Gabriella Cohen, Eran Toch *

Department of Industrial Engineering, Tel-Aviv University, Ramat Aviv 6997801, Israel

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Location privacy
Human mobility
Semantic labeling
Semantic cloaking
k-anonymity

ABSTRACT

The prevalence of mobile phones has led to an explosion in the amounts of human mobility data stored in the cloud. It has been shown that seemingly anonymized location datasets are highly susceptible for re-identification, and may not provide enough privacy protection. In this paper we quantitatively show how *semantic cloaking*, the application of semantic labeling to achieve anonymization, can improve the privacy of a mobility dataset for use cases where location coordinates can be replaced by semantic categories. We develop a semantic labeling framework, apply it and evaluate it using the dataset uniqueness (ε) measure. Our experiments show an improvement in uniqueness ranging between two- and twenty two-fold in comparison to the original, naively anonymized, dataset.

1. Introduction

Human mobility data is being collected today in unprecedented amounts, owing to the prevalence of mobile devices fitted with GPS and other positioning technologies. The increasing sophistication and decreasing cost of cloud computing resources and Big Data analysis tools [1] allow mobility data to be collected, stored and analyzed at high resolution. This data has countless applications in mobile clouds: from location sharing in social networks, through real-time transportation forecasting, and to context-aware advertising services. In academic research, mobility data is used today in various research fields, including urban planning [2], analysis of commuting patterns [3,4] and social relations analysis [5,6].

In today's mobile computing ecosystem, data-collectors are numerous: the device itself (that locally stores positioning data and may upload it to the cloud), providers of Wi-Fi positioning services, various applications installed and websites visited, mobile network operators triangulating user's location via cellular antennas and certain government authorities which are permitted to perform electronic location surveillances. Many of the data-collectors can use the data for their own needs, sell it to third parties, or even release it to the research community. All the while, they are usually obliged by law (or their own privacy policies) to protect the privacy of their users.

An important aspect of users' privacy is **anonymity**, that is, ensuring the true identity of tracked users cannot be exposed and linked back to the data collected. Anonymity provides several advantages over other privacy protection approaches, such as encryption and position sharing [7]. It can protect the data from the data collector as well as from 3rd parties, it allows analyzing the data without transgressing on users' private information, and it does not require any additional burden of the user. In many legal frameworks, anonymity is the crucial requirement that enables mobile carriers to use information about the whereabouts of their users [8]. Recent data breaches (e.g. [9,10]) underscore the importance of providing anonymity, even if the data is originally intended only for the data-collector's internal use.

A typical data record in a mobility dataset, a waypoint, is comprised of a unique user identifier (user-id), a timestamp and a *physical location*. The physical location is a pair of numbers drawn from a Geographic Coordinate System

* Corresponding author. Tel.: +972 3 640 6978.
E-mail address: erant@post.tau.ac.il (E. Toch).

http://dx.doi.org/10.1016/j.pmcj.2015.10.013
1574-1192/

Please cite this article in press as: O. Barak, et al., Anonymizing mobility data using semantic cloaking, Pervasive and Mobile Computing (2015), http://dx.doi.org/10.1016/j.pmcj.2015.10.013

(e.g. ⟨32.156, 34.69⟩, with a possible third coordinate for altitude). The set of waypoints pertaining to a single user over a time interval is called a *physical trajectory*. Providing anonymity in such a dataset is not a simple task and it has been shown [11–15] that the naive approach of replacing the user-id with an arbitrary id ("pseudo-id") is ineffective as users can be re-identified even when the dataset contains no personal information other than a user's location samples. Several privacy-enhancing measures have been offered in the literature, most of which involve data distortion or degradation of the dataset's spatial or temporal resolution, as we detail in Section 2. Recent results by de Montjoye et al. [16] show that the uniqueness of mobility traces decays only as the 1/10 power of the degradation in spatial and temporal resolution. When the uniqueness remains high it means less anonymity, so it follows that even coarse datasets cannot guarantee anonymity, unless they are degraded to a level that impairs their utility.

In this paper we present *semantic cloaking*, a different approach to anonymization of human mobility datasets. We begin by observing that many applications using location data are not interested in the location coordinates *per-se*. For example, a personal assistant application such as Apple's Siri[1] can be set to trigger a reminder when the user reaches *Home* or *Work* and is indifferent as to Home and Work's physical locations. Therefore, if the location dataset supplies *semantic locations* (describing a useful category of the location, e.g. Home, Work and Shopping Center) instead of physical locations, it can retain its usefulness while, as we shall see, improving the users' privacy. Personalized ad services and online social networks (where semantic locations can be shared instead of physical ones) are also examples of usages where semantic location suffices. Such representation is also useful in research, as can be seen in the following examples: in the social sciences, *time-use surveys* are conducted, which effectively try to build a travel diary of semantic locations for each user [17]; a recent paper evaluated the privacy preferences of users given their semantic location [18]; and one of Nokia's Mobile Data Challenge (MDC) tasks [19] invited scientists to build prediction models of semantic locations. Finally, law enforcement bodies trying to discover anomalous movement behavior within a crowd (as done today in airport security and other indoor settings [20]) can decrease the privacy impact, and perhaps even improve the effectiveness of anomaly detection, by analyzing the semantic – instead of physical – locations.

These observations lead to an approach where physical locations are abstracted away entirely and replaced with the corresponding semantic locations, a privacy enhancing approach we call *semantic cloaking*. The main challenge in implementing this approach is to derive the appropriate semantic location for each physical one. We address this challenge by developing a semantic labeling framework that accepts a user's physical trajectory as input and outputs a *semantic trajectory*, which is the set of semantic locations pertaining to that user. The complete framework we have implemented draws on existing work in the field of semantic labeling with some original contribution, and is evaluated on real-world datasets. In this paper, we will provide the general outline of that framework.

While the method of exposing only semantic locations from a physical location dataset has been used before (e.g. in the public release of Nokia's MDC dataset [19] mentioned earlier), our work is the first to empirically evaluate the method's effect on privacy using quantitative measures. To that end, we develop and implement a semantic labeling framework, train it on ground-truth (labeled) data and apply it to a sample of 100 users taken from a real-world cellular dataset. We then use the *unicity test* recently defined by de Montjoye et al. [16] to obtain the *uniqueness* (ε) metric and use it to evaluate the anonymity the dataset provides before and after semantic cloaking. The unicity test is essentially done by repeatedly simulating a brute-force privacy attack using varying amount of background knowledge the attacker is assumed to have and measuring its success. We show that our approach improves the uniqueness by twofold and, when combined with temporal obfuscation, achieves a twenty two-fold improvement. The rest of the paper is structured as follows: Section 2 provides some background and related work. Section 3 presents an outline of the semantic labeling framework we developed, Section 4 describes the method we used to evaluate the framework's application for enhancing dataset privacy and Section 5 presents the results obtained. Section 6 discusses the finding and concludes the work.

2. Background and related work

Cloud-based services collecting user mobility data are faced with the problem of guaranteeing anonymity to the multiple users in their dataset. Whether they expose the data to third-party applications through an API (Application Programming Interface) or use it internally, the threat of privacy attacks – stemming from misuse of their API, rogue employees or external cyber-attackers – is viable. While the proliferation of mobile clouds is a recent phenomenon, the issue described is in fact equivalent to a well-known problem in the field of information privacy, called public dataset release. An informal definition is given by Sweeney [21]:

> "*Consider a data holder, such as a hospital or a bank, that has a privately held collection of person-specific, field structured data. Suppose the data holder wants to share a version of the data with researchers. How can a data holder release a version of its private data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful?*"

Sweeney shows that publicly released – or otherwise leaked – datasets can be linked back to a specific person even when explicit identifiers such as name, address and telephone number (collectively referred to as *personally identifying information*

[1] http://www.apple.com/ios/siri/.

Please cite this article in press as: O. Barak, et al., Anonymizing mobility data using semantic cloaking, Pervasive and Mobile Computing (2015), http://dx.doi.org/10.1016/j.pmcj.2015.10.013

or *PII*) are removed. This is done by using *quasi-identifiers* (such as birth date and gender that may uniquely identify an individual when combined), which can be looked up in other available datasets. In an oft-quoted example, Sweeney linked a publicly available voter registration list with an allegedly anonymous Massachusetts health record to extract the patient-specific information for the governor of Massachusetts. Later work show how such privacy attacks can be used on datasets of physical trajectories [11–16], even when these contain no personal information other than a user's waypoints.

The adversarial scenario we are examining involves a physical trajectory dataset, which the collector has naively anonymized by replacing the user-id with a pseudo-id, and an attacker who managed to obtain a few timestamp-location records belonging to a specific, identified user. The attacker gains access to the dataset, either legitimately (say, if he is the developer of a location-collecting app or a researcher working on a publicly released dataset) or illegitimately, by breaking into the database system. Now, the attacker tries to utilize the few identified records he has for complete re-identification of the user in the dataset. For example, if Alice shared a post on a social network saying she is at "Joe's Café" on Sunday evening, Eve the attacker can use a public geocoding service[2] to lookup the physical location of Joe's Café and then retrieve from the location dataset all users who were around that location at that time on Sunday. If only one user fits this query, than it must be Alice. Eve will note Alice's pseudo-id and use it to re-identify her entire trajectory in the dataset. If more users fit the query, Eve can add another known location (e.g. an observation she made of Alice entering her workplace on Monday morning) and retrieve all users matching **both** locations, a query that has a higher chance to single-out Alice.

A large body of work exists in the field of location dataset privacy protection. Krumm surveyed and compared several methods in [14,22], and more recently Wernke et al. [7] provided a classification of privacy protection goals, existing attacks on those goals and the appropriate protection methods. In the following paragraphs we will position our approach and attack scenario in these categories.

Wernke et al. identify three possible *protection goals*, which are the attributes the user would like to protect: the user's identity, spatial information and temporal information. Accordingly, the protection goals in our work are the user's identity and spatial information, specifically the physical location itself, which can be utilized for identification. On their second classification, the type of attack, our attack scenario is a combination of an *observation (personal context linking) attack* and a *location tracking attack*. In observation attacks, the attacker manages to get hold of some user-identified waypoints. In location tracking attacks, the attacker may query the dataset multiple times. According to Wernke et al., these two are identified as attacks that have the least coverage by existing protection methods.

The classification divides privacy protection approaches into seven categories. Generally speaking, two of the categories (obfuscation and *k*-anonymity) report true physical locations but degrade the temporal and/or spatial resolution of the data; and three categories (coordinate transformation, position dummies and mix zones) distort the data: the first two induce false locations and the third occasionally changes the pseudo-id in a way that prevents tracking the same user for a long period of time. The other two categories, encryption and position sharing, are meant to protect the location data in scenarios where one or more of the location servers (cloud repositories) are compromised. Our approach conceptually protects privacy by guaranteeing *k*-anonymity to users. Nonetheless, the fact that our approach removes the physical locations from the dataset can be considered a distortion, though not one fitting into any of the three distortion categories identified above.

It is important to note that our approach differs from other methods for location privacy protection that rely on location semantics, such as [23–25]. In these works, the authors use the category of nearby points of interest (POI) to decide on the degree of spatial obfuscation to use: lower the resolution until each waypoints include several POI categories [23,25] or coarse the spatial resolution more near sensitive locations such as hospitals [24]. The authors assume the attacker wishes to perform *map matching*, and extract the semantic locations corresponding to physical locations reported by the user. Therefore, the user's protection goal in these cases are the semantics of visited location, and the protection methods obfuscate the physical locations in light of this goal. Our approach is different in several ways: the user's protection goal is different (the user's identity), physical locations are not exposed at all, and semantic locations are the output of the method, and not an aid to it.

One more aspect by which privacy protection approaches are characterized is the nature of the component that applies the protection method. The standard reference model is of a *mobile device* (carried by the tracked user), a *location server* (which may or may not be a trusted anonymizer) collecting data from several devices and *clients* (applications) querying the location server for users' location. Our solution does not define the component performing the anonymization (the method can be applied in any of them), but certainly the safest approach would be to apply semantic cloaking at the source (the device) or in the cloud (location server), after which the physical locations will be discarded.

The central ingredient of the semantic cloaking approach is a *semantic labeling* framework. Semantic Labeling is a vivid research topic, with several recent papers proposing different frameworks [4,26–30]. It is possible to identify a typical structure common to most frameworks published, which can be seen in Figs. 1 and 2, where the framework's input is a user's physical trajectory and the output is a semantic trajectory. A user's instantaneous position may be determined using different technologies: satellite positioning (using GPS signals), cellular positioning (by triangulating visible cell towers), Wi-Fi positioning (by using the MAC addresses and known locations of nearby Wi-Fi hotspots), etc. In the setting discussed here, it is the responsibility of the data collector (the mobile device or the cloud-based service) to properly transform these reading

[2] Geocoding is the process of retrieving the physical location of some point of interest using its street address, zip code or business name. Several free geocoding services are available on the Internet.

Please cite this article in press as: O. Barak, et al., Anonymizing mobility data using semantic cloaking, Pervasive and Mobile Computing (2015), http://dx.doi.org/10.1016/j.pmcj.2015.10.013

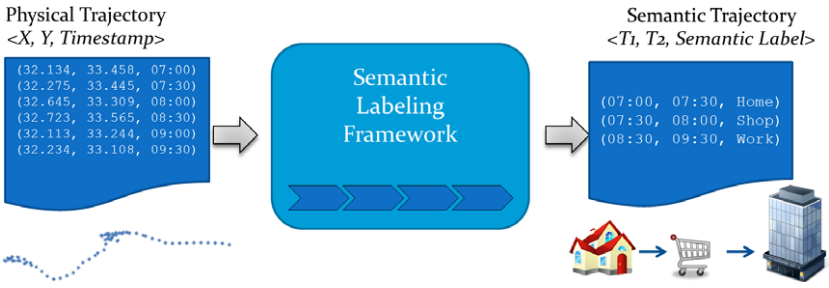

Fig. 1. Illustration of a semantic labeling framework. The different stages of semantic labeling in our framework are shown in detail in Fig. 2.

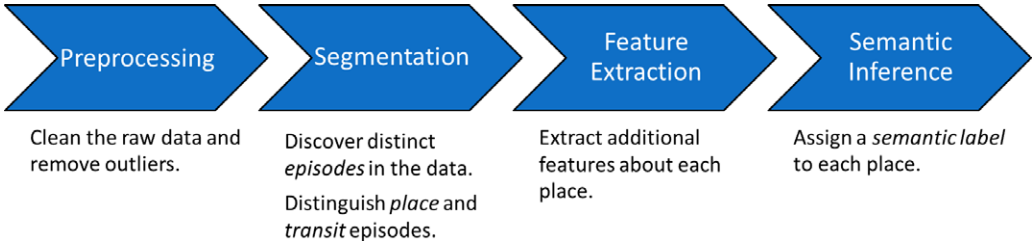

Fig. 2. The four main stages comprising our semantic labeling framework and the main purpose of each stage.

into a set of separate, time-stamped, physical locations represented using a coordinate system which together comprise the physical trajectory fed into the framework.

3. The semantic labeling framework developed

The semantic labeling framework includes four stages (Fig. 2), which are typically found, perhaps with different names, in most published frameworks.

In the **preprocessing** stage, the raw data is cleaned out of noise and outliers, which are intrinsic to real-world location recording. We have used a simple technique based on a speed threshold Δ_{speed} which is set at 40 m/s, (approximately 90 mph or 145 km/h). In the **segmentation** stage, the physical trajectory is first divided into segments called episodes. Then the episodes are classified into "places", which are locations where the user was observed staying for a predefined timespan and "non-places" such as transits where the user was moving between places. Usually these "non-places" are discarded. We have implemented a segmentation algorithm based on time, distance and density criteria—measures appropriate for a time-ordered human mobility dataset, as two waypoints close in space and time usually belong to the same semantic place. The algorithm therefore greedily merges consecutive waypoints into the same "place" episode given they satisfy all of the following:

1. The time difference between two consecutive waypoints is less than Δ_t minutes ($\mathbf{\Delta_t = 180\ min}$ was used).
2. The distance between two consecutive waypoints is less than Δ_d meters ($\mathbf{\Delta_d = 300\ m}$ was used).
3. The distance between any waypoint and the episode center is less than $\Delta_{d\_episode}$ meters ($\mathbf{\Delta_{d\_episode} = 600\ m}$ was used).
4. The total stay time within the episode is at least $\Delta_{t\_episode}$ minutes ($\mathbf{\Delta_{t\_episode} = 30\ min}$ was used).

Condition 1 relates to time and condition 2 to distance. Conditions 3 and 4 together relate to episode density: to be considered a place episode, the user must stay within a certain radius ($\Delta_{d\_episode}$) from the episode center for at least a set amount of time ($\Delta_{t\_episode}$).

The algorithm starts by treating the first waypoint as an episode and then tries to merge consecutive waypoints to that episode as explained herein. When a waypoint does not satisfy the conditions to join an episode—that existing episode is "closed" and a new episode starts, initially consisting only of that new waypoint. If the closed episode satisfied the density condition, it will be labeled a place episode, otherwise—it will be labeled a transit and discarded.

In the **feature extraction** stage, each place episode is enriched with information that will aid in inferring its semantic label. Specifically, we extract the land-use in the episode's center (e.g. is the episode located in a residential zone, commercial zone, etc.) and several temporal features, such as duration of stay and day of week, which are obtained directly from time-stamped location data. In addition, the categories of nearby points-of-interest (POI) are extracted from an online POI directory (a category can be Residence, Food, Shop and Service, etc.). We also use real-world popularity data (based on Foursquare/Swarm[3] check-in information) to determine the most popular POI category in the area.

[3] https://www.swarmapp.com/.

Please cite this article in press as: O. Barak, et al., Anonymizing mobility data using semantic cloaking, Pervasive and Mobile Computing (2015), http://dx.doi.org/10.1016/j.pmcj.2015.10.013

Table 1
The semantic labels used in D1.

ID	Semantic label
1	Home
2	Workplace
3	Studying place
4	Shopping and services
5	Social, recreation and eating out
6	Home of friends or family
7	Transport
8	Other

Table 2
Random forest performance measures.

D1—random forest classifier	
Average accuracy	76%
Avg. weighted F-measure	74.2%
AUC	94.8%

In the final stage, **semantic inference**, a machine-learning classifier is applied to the episodes containing the newly extracted features – 18 features altogether – to output the most probable semantic label for each one. The classifier must be trained in advance on a labeled (ground-truth) dataset, a process we will describe in the next paragraphs. We have used the open-source data mining software WEKA [31] to train and evaluate different classifiers and choose the best-performing one.

The dataset we use for training and evaluation is a physical trajectories dataset, denoted by D1, which includes the true (user reported) semantic label of each waypoint. D1 comprises of data collected in two user studies conducted by our research group. The user studies used a designated Android smartphone application that sampled the users' location every 15 min on average, and popped-up randomly several times a day with survey questions regarding a location the user had visited. One of the survey questions asks the user to choose the semantic label of that location. In addition, a post-experiment survey asked users for the label of locations where the user was seen to stay but no semantic label was recorded for them. Both studies were authorized by the institutional review committee and were conducted between April and July 2012. One experiment was conducted for two weeks encompassing 25 users and the second was conducted for two and a half months encompassing 50 users. Only participants who were cooperative (i.e. answered enough survey questions) were chosen, so that at least 70% of their waypoints were labeled. Eventually, D1 comprised of 42 users and 50,000 different waypoints. The final list of semantic labels used in D1 can be seen in Table 1. We note this list is based on the data available from the original user studies, and is not meant as a comprehensive mapping of all possible semantic labels.

We fed dataset D1 into the semantic labeling framework. For the semantic inference stage we have evaluated 3 different classifiers: J48 decision tree, Bayesian network and Random Forest. These classifiers were chosen because they are well-suited for the feature types at hand and widely used for semantic labeling (for example, in [4,32–34]). The following measures were used to evaluate the framework's labeling performance:

- **Accuracy**—the fraction of correctly predicted semantic labels (compared to the true semantic label for each episode).
- **Area under the ROC curve (AUC)**—The ROC (Receiver Operating Characteristic) curve is a plot that shows the performance of the classifier for a specific class vs. all-other classes as a function of the discriminative threshold used (the probability threshold above which we take the classifier prediction). The AUC is the fraction of area which is bound between the curve and the diagonal, and is a popular metric in classification tasks.
- **Per-label precision and recall**—we measure the per-label performance using the F-measure, which is the harmonic mean of the *recall rate* (from all episodes that were truly labeled "Home", how many did we predict correctly?) and the *precision rate* (from all episodes we predicted as "Home", how many were correct prediction?). The F-measure is also a popular metric in classification tasks.

The average accuracy, AUC and F-measure (weighted over all classes) were calculated using 10-fold cross validation for each of the classifiers. Random Forest classifier was shown to perform best according to all three measures and was chosen for the rest of the analysis. Its results are shown in Table 2.

4. Evaluation of semantic cloaking

Following the implementation of the framework and its training and evaluation on a ground-truth dataset, we move on to apply it to a larger dataset and test the effect of semantic cloaking on the dataset's privacy.

4.1. Preliminaries

In this experiment we use a new dataset, denoted *D2*. *D2* comprises of waypoints of mobile phone users, collected by an Israeli mobile carrier over a period of two months (November 2012 to January 2013). The dataset is naively anonymized, with

Please cite this article in press as: O. Barak, et al., Anonymizing mobility data using semantic cloaking, Pervasive and Mobile Computing (2015), http://dx.doi.org/10.1016/j.pmcj.2015.10.013

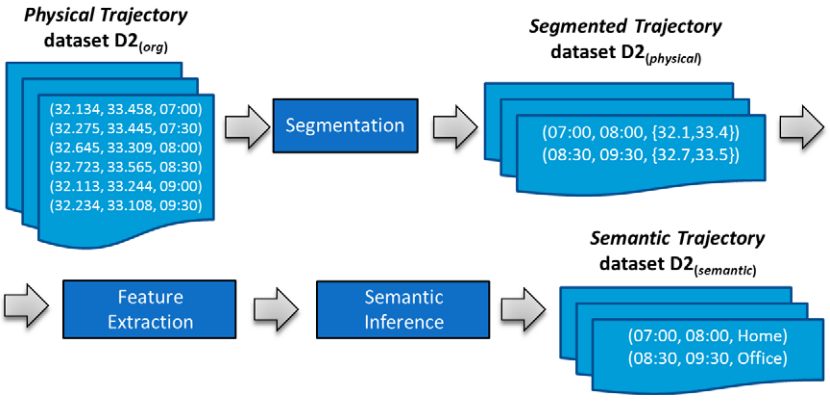

Fig. 3. Illustration of how $D2_{(physical)}$ and $D2_{(semantic)}$, the two datasets used in the privacy comparison, were obtained from $D2_{(org)}$.

the unique IMSI (International Mobile Subscriber Identity) number for each device replaced by a pseudo-id. Each data record holds the timestamp, the visible cell towers and a large amount of network auxiliary data. The records were triangulated by the carrier to extract their physical locations, yielding locations with accuracy that depends on the cell tower density in the area. The carrier states that accuracy varies between 60 and 300 m within metropolitan areas and between 1000 and 5000 m in open and less populated areas. For $D2$, the 100 most active users within a metropolitan area in Israel were extracted. Owing to the high sampling rate (a new record is collected whenever one of numerous network/in-call events occurs), the dataset contained a total of 27 million waypoints.

4.2. Method

To estimate the dataset's privacy level we use the uniqueness measure recently defined and explored by de Montjoye et al. [16]. This measure can quantify how "hard" it will be for an attacker to re-identify an average user in the reference scenario we described. Low dataset uniqueness means users' privacy is more protected and the attacker's work will be harder. The measure's definition and computation follows: *uniqueness* (ε) of a dataset D is defined with respect to some waypoint set I_p, which contains p spatio-temporal waypoints of a single user u. Note that this set I_p is equivalent to the p known locations the attacker obtained in the adversarial scenario we outlined in Section 2. The uniqueness of dataset D given a specific set I_p is then evaluated by extracting from D the subset of traces $S(I_p)$ that matches all waypoints in I_p. This extraction is called a unicity test. User's u trace is called *unique* if $\left|S(I_p)\right| = 1$, that is if only one user matches that set of waypoints and therefore u can be singled-out from D using I_p. The dataset's average uniqueness can be evaluated by repeatedly choosing random users and random sets of size p, applying the unicity test to extract the subset $S(I_p)$ and eventually calculating the fraction of unique user traces found. The higher the dataset's uniqueness, so is the dataset more vulnerable to such attacks and its privacy is lower. In the cited paper, the uniqueness of a massive dataset of mobile phone users was found to be 95% for $p = 4$. That is, for 95% of the users, the dataset allows an attacker knowing only 4 locations the user visited, to identify the user's entire physical trajectory.

Our experiment will compare the uniqueness of dataset $D2$ before and after semantic cloaking, to show the effect of this approach on privacy. To this end, let $D2_{(org)}$ be the original physical trajectory dataset and let $D2_{(semantic)}$ be the final semantic trajectory dataset obtained after applying the semantic labeling framework (with the Random Forest classifier) to $D2_{(org)}$. Finally, let $D2_{(physical)}$ be the interim dataset obtained after the segmentation (episode extraction) stage of the framework (see Fig. 3). $D2_{(physical)}$ and $D2_{(semantic)}$ consist of the same episodes but in one the episodes are given with physical locations and in the other with semantic locations. These are the two datasets to be compared.

To estimate the datasets' uniqueness, first a value of p (number of waypoints) is chosen. We have experimented with values of p between 1 and 12. Next, for each of the two datasets, 100 random sets of waypoints are sampled (with replacement), by first randomly choosing a user and then randomly choosing p waypoints for that user. Each set is in fact a set I_p for the value of p chosen beforehand and each waypoint in the set is a tuple $\langle timestamp, location\rangle$. For $D2_{(semantic)}$, *location* is the semantic location and for $D2_{(physical)}$, *location* is the physical location of the square obtained after partitioning the geographic space into squares of 1 sq. kilometer (de Montjoye et al. used instead the partitions dictated by the network cells in which the user was located). For the first experimental setting, *timestamp* includes the full date with time discretized to 5 bins (Morning, Prenoon, Afternoon, Evening, Night). The other two experimental settings use a different *timestamp* representation, as explained below.

5. Results

Let N_I be the number of I_p sets sampled for a certain value of p (in our experiment $N_I = 100$ for each of the two datasets). For each dataset, for each of the 100 sets $I_p^{(1)}, I_p^{(2)}, I_p^{(3)} \ldots I_p^{(100)}$ the dataset was queried for all users who were recorded at

Please cite this article in press as: O. Barak, et al., Anonymizing mobility data using semantic cloaking, Pervasive and Mobile Computing (2015), http://dx.doi.org/10.1016/j.pmcj.2015.10.013

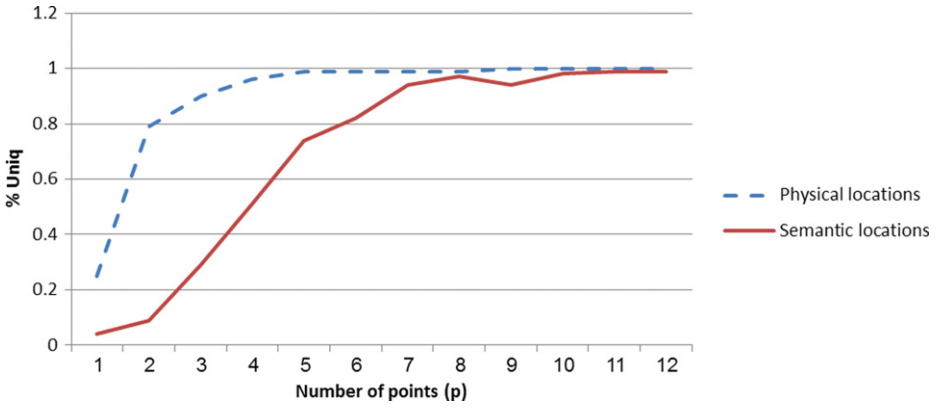

Fig. 4. Uniqueness ε for $D2_{(semantic)}$ and $D2_{(physical)}$ in the first experiment setting (timestamp includes date and time).

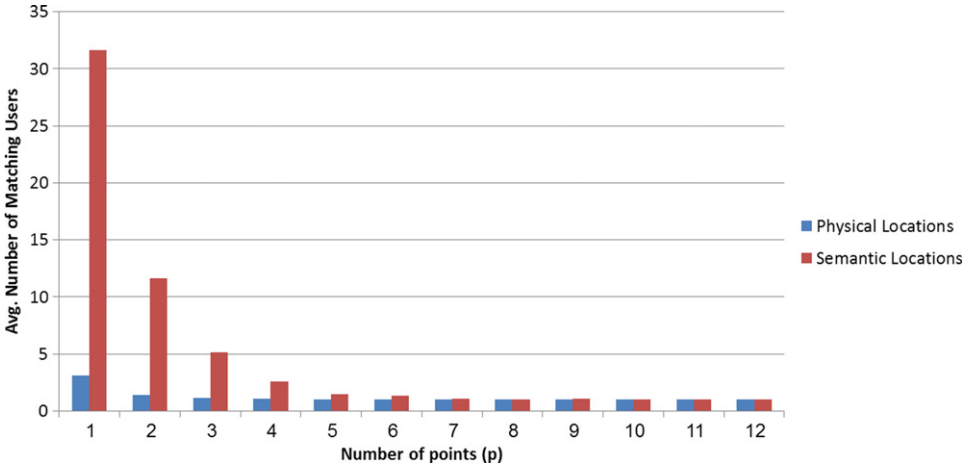

Fig. 5. Average number of matching users ($\bar{N}$) for $D2_{(semantic)}$ and $D2_{(physical)}$ in the first experiment setting (timestamp includes date and time).

all the waypoints in I_p. Let the number of matching users for the set $I_p^{(k)}$ be $N^{(k)}$. Then, for each dataset, two measures were obtained. The first, denoted $\bar{N}$ is the average number of matching users (Eq. (1)).

$$\bar{N} = \left(\frac{\sum_k N^{(k)}}{N_I} \right) = \left(\frac{\sum_k N^{(k)}}{100} \right). \tag{1}$$

The second is the fraction of sets for which only one user was returned, which is in fact the dataset uniqueness ε (Eq. (2)).

$$\varepsilon = \left(\frac{\sum_{\{k|N^{(k)}=1\}} N^{(k)}}{N_I} \right) = \left(\frac{\sum_{\{k|N^{(k)}=1\}} 1}{100} \right). \tag{2}$$

The comparison of the uniqueness for $D2_{(semantic)}$ and $D2_{(physical)}$ is shown in Fig. 4.

For $D2_{(physical)}$ the results are similar to those obtained by de Montjoye et al. In the two studies, uniqueness of 95% is seen at $p = 4$. For $D2_{(semantic)}$, the rise in uniqueness as p grows is slower, measuring $\varepsilon = 50\%$ for $p = 4$ and reaching 95% only when $p = 8$.

Fig. 5 compares the average number of matching users, $\bar{N}$, for both datasets (here a higher number indicates higher privacy). We can see that for $D2_{(physical)}$ $\bar{N}$ is quite low even for $p = 1$. For $D2_{(semantic)}$, the value drops sharply with p and seems to fix around $\bar{N} \approx 1$ (meaning $\varepsilon \approx 100\%$) for values of p larger than 10.

It can be observed that the semantic location dataset provides a higher degree of privacy in comparison to physical location dataset. However, the protection is limited: identifying half of the users in a dataset using only 4 waypoints (the interpretation of $\varepsilon = 50\%$ for $p = 4$) seems to make the dataset unsafe against privacy attacks. A further improvement in privacy can be achieved by combining semantic cloaking with lowering of the temporal resolution. If we present the

Please cite this article in press as: O. Barak, et al., Anonymizing mobility data using semantic cloaking, Pervasive and Mobile Computing (2015), http://dx.doi.org/10.1016/j.pmcj.2015.10.013

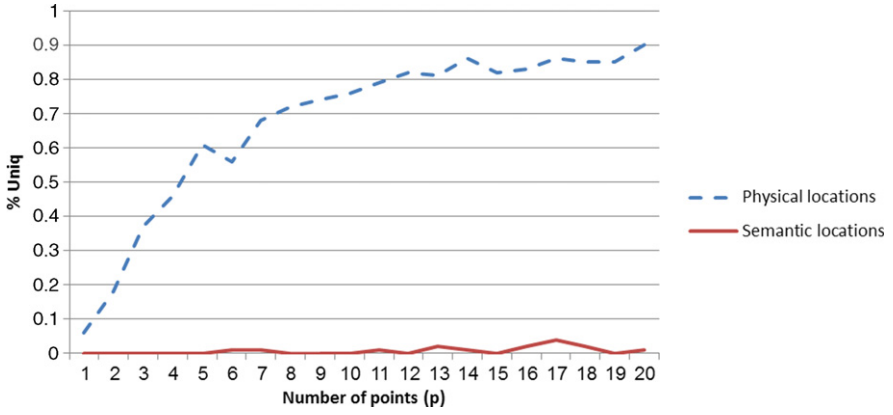

Fig. 6. Uniqueness ε for $D2_{(semantic)}$ and $D2_{(physical)}$ in the second experiment setting (timestamp includes only time).

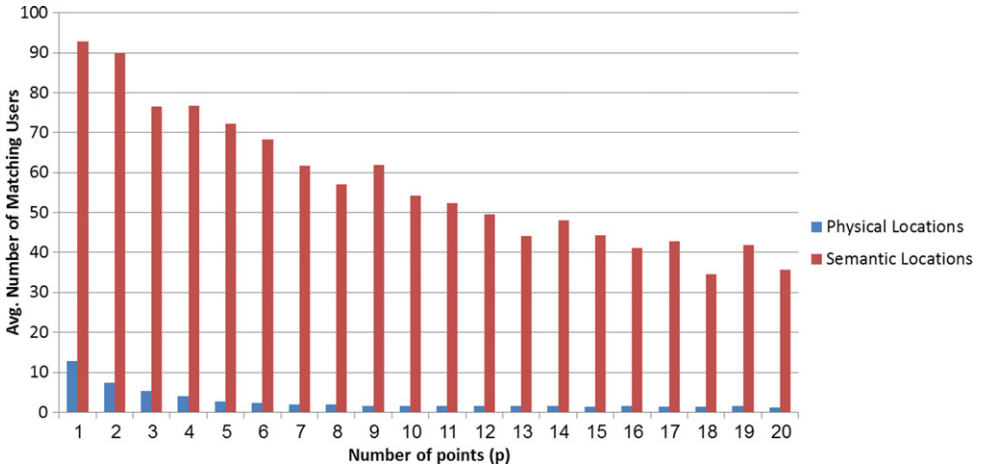

Fig. 7. Average number of matching users ($\bar{N}$) for $D2_{(semantic)}$ and $D2_{(physical)}$ in the second experiment setting (timestamp includes only time).

episode's *timestamp* using only the time bin and without the date we can make a visit to a place on Sunday morning indistinguishable from a visit on Tuesday morning. Note that if we wish to keep the order of the dataset's episodes as-is (according to the temporal order of the user movement), we must not expose the start and end date of the dataset, or else the attacker can still approximate the date of each record. This can be done either by hiding these dates (i.e. keeping them separately from the dataset itself) or by keeping the dates with lowered resolution, e.g. "data collected over Spring–Summer 2014" instead of exact dates.

We have repeated the experiment under this second setting, this time varying the values of p from 1 to 20 (Fig. 6). As expected, the anonymity improves for both datasets, but the gain of using semantic labeling is dramatically higher. Under this setting, the uniqueness of the physical location dataset increases more slowly crossing $\varepsilon = 50\%$ at $p = 5$ and hitting $\varepsilon = 90\%$ only for $p = 20$. But the semantic location dataset's uniqueness remains very low across all values of p, reaching a maximum of only $\varepsilon = 4\%$.

These dynamics are clearer when we observe the average number of matching users ($\bar{N}$) for both datasets, shown in Fig. 7. Again, the higher the number of matching users, the higher the privacy the dataset supplies.

We see $\bar{N}$ is higher for $D2_{(semantic)}$ across all values of p. Moreover, for $D2_{(semantic)}$ the decrease in $\bar{N}$ slows-down and it seems stabilized for higher p values, implying a diminishing return for the attacker trying to improve her attack by collecting more waypoints. The explanation for the results so far is simple: if a user is spotted in a café near her home, she will be non-unique in $D2_{(physical)}$ only if another user was in that exact same geographical area at the same time bin. But in $D2_{(semantic)}$ she will gain uniqueness from **all** the café-goers in the dataset, regardless of whether "their" café was near her own home. When lowering the temporal resolution (thus adding more potential user episodes to which she might be similar) the effect is strengthened.

We next turn to compare the privacy protection obtained by semantic cloaking with the protection obtained by a different approach, spatial obfuscation, which retains the physical locations but degrades their resolution. To this end, we keep the lowered temporal resolution ("only time") from the previous experiment and re-partition $D2_{(physical)}$ to further lower its spatial resolution as compared to the resolution of 1 sq. kilometer used before. We experiment with squares of 2.5, 5 and

Please cite this article in press as: O. Barak, et al., Anonymizing mobility data using semantic cloaking, Pervasive and Mobile Computing (2015), http://dx.doi.org/10.1016/j.pmcj.2015.10.013

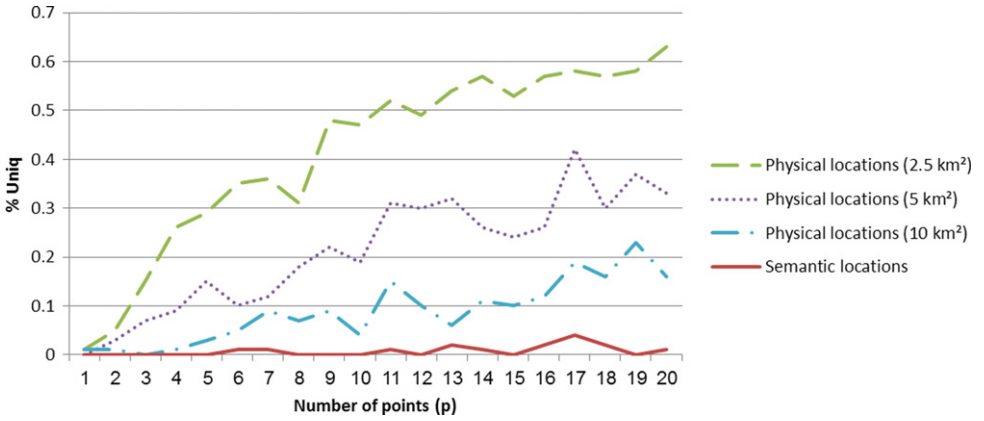

Fig. 8. Uniqueness ε for $D2_{(semantic)}$ and $D2_{(physical)}$ when applying varying degrees of spatial obfuscation to $D2_{(physical)}$.

10 sq. kilometers and estimate the dataset's uniqueness for each value (Fig. 8). It is important to note that high degrees of spatial obfuscation, like 5 and 10 sq. kilometers, may hamper the usability of the dataset for many applications, particularly in an urban setting (for perspective, the entire area of Central Park in New York is 3.4 sq. kilometers). And still, as we can see, the gain in privacy protection is not as high as one would expect, and is lower than the protection supplied by semantic cloaking.

6. Discussion and conclusions

In this work we have shown that semantic cloaking is an effective solution to location-based cloud applications in use cases that can rely solely on semantic locations. That is, removing the physical locations completely will not affect their usability, or will affect it in a way which is bearable given the increase in privacy. Using a semantic labeling framework we have developed, we showed an increase in the privacy level of a dataset by a ratio ranging between two- and twenty two-fold (as measured by users' uniqueness). The higher protection of privacy is achieved by combining semantic cloaking (which can be considered a distortion of the spatial data) with the distortion of temporal data. Varying the amount of distortion in both dimensions allows a balance between the privacy and the utility of the dataset.

Our ability to generalize these results from the dataset sample we used to other datasets depends on the required semantics and their relations to the underlying mobility datasets. The dataset size by itself is not a limitation: Our results regarding uniqueness of physical locations are similar to those obtained by de Montjoye et al. [16], with one and a half million users. Our results of the uniqueness of semantic cloaking provide an empirical upper bound on average uniqueness, as an increase in the number of users will tend to decrease semantic uniqueness. If the number of semantic categories is fixed, adding more users is expected to increase the number of users in each combination of semantic locations.

Designers of cloud services that consider semantic cloaking should take several aspects into consideration. First, a basic requirement is that the service does not require an actual physical location. If the service does require such locations, e.g., serving ads near specific physical locations, then other methods should be used, such as pre-calculated access graphs [35] or private proximity testing [11]. Our method can only be used if the locations can be mapped to some semantic vocabulary. Second, unlike other works that identify places [36] or paths [36,37], our work is based on offline segmentation of waypoints and requires large portions of the location trace for segmentation and inference. Therefore, our method would requires additional development to be used in real time. Third, as the set of semantic categories should be decided in advance, it limits the possible ways location data can be used when the application is running. However, the process of designing the particular semantics of tracking is a good example of privacy-by-design engineering method [38]. From the point of view of privacy protection, asking the designer to predefine the collected information is actually preferable to collecting all available location data. This also applies to the resolution of semantic labels, as labels which are too detailed (e.g. separating the Eating Out category to specific types of restaurants) might cause less users to fit each label and lower the privacy protection. As mentioned, in this work the labels were pre-dictated by the data we had at hand, so we did not explore this issue further. The last point to consider is the labeling accuracy and its effect on the dataset utility. The results we showed are a valid quantitative measure for the effect of semantic cloaking on privacy. Yet, the utility of the dataset can also be affected by semantic cloaking. This can range from zero-effect when the data holder has the true semantic labels (like in the Nokia dataset release [19]) to concrete effect when the inference accuracy is low. The inference method used in our experiment showed a 76% labeling accuracy, which we believe allows a fair trade-off between privacy and utility. Achieving higher accuracy levels is possible by using larger datasets for the training data, and is of course also affected by the labels used.

We conclude by mentioning that semantic labeling can improve a dataset's privacy not only by the methods described, but also as an aid to other dataset anonymization method. For example, Lee et al. [25] present a method that improves the privacy of a physical location dataset in a different reference scenario than the one described here (where a single, identified,

Please cite this article in press as: O. Barak, et al., Anonymizing mobility data using semantic cloaking, Pervasive and Mobile Computing (2015), http://dx.doi.org/10.1016/j.pmcj.2015.10.013

user wants to conceal specific places she visits) and depends greatly on the locations' semantic labels. Also, future works can use semantic locations in combination with physical location. For example, by replacing only some of the physical locations with semantic ones and releasing the rest as-is. The physical locations to be replaced are exactly those that are unique in the dataset and would pose a threat on the user's privacy if obtained by an attacker.

Acknowledgment

This work is supported by the Israeli Ministry of Science, Technology, and Space, Grant no. 3-8709: Learning and mining mobility patterns using stochastic models. We would also like to thank Irad Ben Gal, Boaz Lerner, Dan Halbersberg, Gabriella Cohen and Lior Rokach for their help and insightful comments.

References

[1] D. Agrawal, S. Das, A. El Abbadi, Big data and cloud computing: Current state and future opportunities, in: Proceedings of the 14th International Conference on Extending Database Technology, ACM, New York, NY, USA, 2011, pp. 530–533. http://dx.doi.org/10.1145/1951365.1951432.
[2] R.A. Becker, R. Caceres, K. Hanson, J.M. Loh, S. Urbanek, A. Varshavsky, et al., A tale of one city: Using cellular network data for urban planning, IEEE Pervasive Comput. 10 (2011) 18–26. http://dx.doi.org/10.1109/MPRV.2011.44.
[3] J. Candia, M.C. González, P. Wang, T. Schoenharl, G. Madey, A.-L. Barabási, Uncovering individual and collective human dynamics from mobile phone records, J. Phys. A: Math. Theor. 41 (2008) 224015. http://dx.doi.org/10.1088/1751-8113/41/22/224015.
[4] M. Lv, L. Chen, G. Chen, Discovering personally semantic places from GPS trajectories, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, 2012, pp. 1552–1556. http://dx.doi.org/10.1145/2396761.2398471.
[5] N. Eagle, A. Pentland, Reality mining: sensing complex social systems, Pers. Ubiquitous Comput. 10 (2006) 255–268. http://dx.doi.org/10.1007/s00779-005-0046-3.
[6] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, et al., Computational social science, Science 323 (2009) 721–723. http://dx.doi.org/10.1126/science.1167742.
[7] M. Wernke, P. Skvortsov, F. Dürr, K. Rothermel, A classification of location privacy attacks and approaches, Pers. Ubiquitous Comput. 18 (2014) 163–175. http://dx.doi.org/10.1007/s00779-012-0633-z.
[8] G.A. Gow, Information privacy and mobile phones, Converg.: Int. J. Res. New Media Technol. 11 (2005) 76–87. http://dx.doi.org/10.1177/135485650501100208.
[9] Adobe confirms data security breach, BBC News. (n.d.). http://www.bbc.co.uk/news/business-24392819 (accessed 9.02.15).
[10] C. Riley, Insurance giant Anthem hit by massive data breach, CNNMoney. 2015. http://money.cnn.com/2015/02/04/technology/anthem-insurance-hack-data-security/index.html (accessed 9.02.15).
[11] N.T. Arvind Narayanan, Location privacy via private proximity testing, in: Network and IT Security Conference, NDSS 2011, San Diego, California, 2011.
[12] P. Golle, K. Partridge, On the anonymity of home/work location pairs, in: Proceedings of the 7th International Conference on Pervasive Computing, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 390–397. http://dx.doi.org/10.1007/978-3-642-01516-8_26.
[13] M. Gruteser, B. Hoh, On the anonymity of periodic location samples, Secur. Pervasive Comput. 3450 (2005) 179–192. http://dx.doi.org/10.1007/978-3-540-32004-3_19.
[14] J. Krumm, Inference attacks on location tracks, IEEE Pervasive Comput. 4480 (2007) 127–143. http://dx.doi.org/10.1007/978-3-540-72037-9_8.
[15] H. Zang, J. Bolot, Anonymization of location data does not work: A large-scale measurement study, in: Proceedings of the 17th Annual International Conference on Mobile Computing and Networking, ACM, New York, NY, USA, 2011, pp. 145–156. http://dx.doi.org/10.1145/2030613.2030630.
[16] Y.-A. de Montjoye, C.A. Hidalgo, M. Verleysen, V.D. Blondel, Unique in the crowd: The privacy bounds of human mobility, Sci. Rep. 3 (2013) http://dx.doi.org/10.1038/srep01376.
[17] K. Partridge, P. Golle, On using existing time-use study data for ubiquitous computing applications, in: Proceedings of the 10th International Conference on Ubiquitous Computing, ACM, New York, NY, USA, 2008, pp. 144–153. http://dx.doi.org/10.1145/1409635.1409655.
[18] O. Barak, G. Cohen, A. Gazit, E. Toch, The price is right?: economic value of location sharing, in: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication, ACM Press, 2013, p. 891. http://dx.doi.org/10.1145/2494091.2497343.
[19] J.K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.M.T. Do, et al. The mobile data challenge: Big data for mobile computing research, in: Proc. Mobile Data Challenge by Nokia Workshop, in Conjunction with Int. Conf.. on Pervasive Computing, Newcastle, 2012.
[20] S.-I. Yu, Y. Yang, A. Hauptmann, Harry Potter's Marauder's map: Localizing and tracking multiple persons-of-interest by nonnegative discretization, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2013, pp. 3714–3720. http://dx.doi.org/10.1109/CVPR.2013.476.
[21] L. Sweeney, *k*-Anonymity: A model for protecting privacy, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10 (05) (2002) 557–570.
[22] J. Krumm, A survey of computational location privacy, Pers. Ubiquitous Comput. 13 (2009) 391–399. http://dx.doi.org/10.1007/s00779-008-0212-5.
[23] B. Bamba, L. Liu, P. Pesti, T. Wang, Supporting anonymous location queries in mobile environments with privacygrid, in: Proceedings of the 17th International Conference on World Wide Web, ACM, New York, NY, USA, 2008, pp. 237–246. http://dx.doi.org/10.1145/1367497.1367531.
[24] M.L. Damiani, E. Bertino, C. Silvestri, Protecting location privacy against spatial inferences: The PROBE approach, in: Proceedings of the 2nd SIGSPATIAL ACM GIS 2009: International Workshop on Security and Privacy in GIS and LBS, ACM, New York, NY, USA, 2009, pp. 32–41. http://dx.doi.org/10.1145/1667502.1667511.
[25] B. Lee, J. Oh, H. Yu, J. Kim, Protecting location privacy using location semantics, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2011, pp. 1289–1297. http://dx.doi.org/10.1145/2020408.2020602.
[26] Juhong Liu, O. Wolfson, Huabei Yin, Extracting semantic location from outdoor positioning systems, in: MDM 2006, 7th International Conference on Mobile Data Management, 2006, pp. 73–73.
[27] J.J.-C. Ying, W.-C. Lee, T.-C. Weng, V.S. Tseng, Semantic trajectory mining for location prediction, in: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, New York, NY, USA, 2011, pp. 34–43. http://dx.doi.org/10.1145/2093973.2093980.
[28] C. Renso, M. Baglioni, J.A.F. de Macedo, R. Trasarti, M. Wachowicz, How you move reveals who you are: understanding human behavior by analyzing trajectory data, Knowl. Inf. Syst. (2013) 1–32. http://dx.doi.org/10.1007/s10115-012-0511-z.
[29] O. Barak, Bootstrapping semantic locations from human mobility data (M.Sc. Thesis), Tel Aviv University, 2014.
[30] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, K. Aberer, Semantic trajectories: Mobility data computation and annotation, ACM Trans. Intell. Syst. Technol. 4 (2013) 49:1–49:38. http://dx.doi.org/10.1145/2483669.2483682.
[31] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: An update, SIGKDD Exp. Newsl. 11 (2009) 10–18. http://dx.doi.org/10.1145/1656274.1656278.
[32] J. Krumm, D. Rouhana, Placer: semantic place labels from diary data, in: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2013, pp. 163–172.
[33] Y. Zhu, Y. Sun, Y. Wang, Nokia mobile data challenge: Predicting semantic place and next place via mobile data, in: Proc. Mobile Data Challenge by Nokia Workshop, in Conjunction with Int. Conf. on Pervasive Computing, Newcastle, UK, 2012.

Please cite this article in press as: O. Barak, et al., Anonymizing mobility data using semantic cloaking, Pervasive and Mobile Computing (2015), http://dx.doi.org/10.1016/j.pmcj.2015.10.013

[34] Y. Zhu, E. Zhong, Z. Lu, Q. Yang, Feature engineering for place category classification, in: Proc. Mobile Data Challenge by Nokia Workshop, in Conjunction with Int. Conf. on Pervasive Computing, Newcastle, UK, 2012.
[35] U. Hengartner, P. Steenkiste, Avoiding privacy violations caused by context-sensitive services, in: Fourth Annual IEEE International Conference on Pervasive Computing and Communications, 2006. PerCom 2006, 2006, pp. 223–233. http://dx.doi.org/10.1109/PERCOM.2006.11.
[36] D.H. Kim, Y. Kim, D. Estrin, M.B. Srivastava, SensLoc: sensing everyday places and paths using less energy, in: Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems, ACM, New York, NY, USA, 2010, pp. 43–56. http://dx.doi.org/10.1145/1869983.1869989.
[37] Y. Jiang, D. Li, G. Yang, Q. Lv, Z. Liu, Deliberation for intuition: A framework for energy-efficient trip detection on cellular phones, in: Proceedings of the 13th International Conference on Ubiquitous Computing, ACM, New York, NY, USA, 2011, pp. 315–324. http://dx.doi.org/10.1145/2030112.2030156.
[38] M. Langheinrich, Privacy by design—Principles of privacy-aware ubiquitous systems, in: G.D. Abowd, B. Brumitt, S. Shafer (Eds.), Ubicomp 2001: Ubiquitous Computing, Springer, Berlin, Heidelberg, 2001, pp. 273–291.

Please cite this article in press as: O. Barak, et al., Anonymizing mobility data using semantic cloaking, Pervasive and Mobile Computing (2015), http://dx.doi.org/10.1016/j.pmcj.2015.10.013