# Protecting Location Privacy Using Location Semantics *

Byoungyoung Lee[α], Jinoh Oh[α], Hwanjo Yu[α], and Jong Kim[β]
[α]Department of Computer Science and Engineering, [β]Division of IT convergence engineering
Pohang University of Science and Technology (POSTECH)
Pohang, South Korea
{override, kurin, hwanjoyu, jkim}@postech.ac.kr

## ABSTRACT

As the use of mobile devices increases, a location-based service (LBS) becomes increasingly popular because it provides more convenient context-aware services. However, LBS introduces problematic issues for location privacy due to the nature of the service. Location privacy protection methods based on $k$-anonymity and $\ell$-diversity have been proposed to provide anonymized use of LBS. However, the $k$-anonymity and $\ell$-diversity methods still can endanger the user's privacy because location *semantic* information could easily be breached while using LBS. This paper presents a novel location privacy protection technique, which protects the location semantics from an adversary. In our scheme, location semantics are first learned from location data. Then, the trusted-anonymization server performs the anonymization using the location semantic information by cloaking with *semantically heterogeneous* locations. Thus, the location *semantic* information is kept secure as the cloaking is done with semantically heterogeneous locations and the true location information is not delivered to the LBS applications. This paper proposes algorithms for learning location semantics and achieving semantically secure cloaking.

## Categories and Subject Descriptors

H.2.0 [**Database Management**]: General—*Security, integrity, and protection*; H.2.8 [**Database Management**]: Database Applications—*Spatial databases and GIS*

## General Terms

Security, Algorithm, Experimentation

## Keywords

Location Privacy, Location Semantics, $\theta$-Secure Cloaking Area

*This research is supported by WCU(World Class University) program (R31-2008-000-10100-0) and Research Grant (KRF-2008-331-D00528), both through the National Research Foundation of Korea funded by the Korean Government.


*KDD'11,* August 21–24, 2011, San Diego, California, USA.


## 1. INTRODUCTION

The use of mobile devices has increased dramatically in the last decade. As mobile device technology has developed, context awareness services have become available and mobile devices now support more convenient and user-friendly services. The representative service of context awareness services is a location-based service (LBS). LBS is an information and entertainment service based on the geographical position of the mobile device. There are many different kinds of LBS services, such as navigation services, requesting the nearest business locations, receiving traffic alerts or notifications, and so on.

However, LBS services introduce problematic issues for location privacy due to the nature of the service. The history of certain user's locations could be accumulated, and private information could be exposed if an adversary has access to that history. For example, support for a certain political party, or the whereabouts of a user at a certain time could be handed to adversaries and abused by them. This is a critical problem because most people are reluctant to use LBS services if their location privacy is in danger in spite of its convenience.

In order to protect location privacy, previous research has been done using $k$-anonymity [23] and $\ell$-diversity [19]. A cloaking area, which is an extended area from the exact position of a mobile user, is computed by the anonymization server and the anonymization server delegates the LBS requests for a mobile user. For computing a cloaking area, $k$-anonymity based location privacy [10, 6, 20, 3, 14, 26, 27] extends a cloaking area until '$k$-1' other users are included, and $\ell$-diversity based location privacy [1, 24, 28] extends until '$\ell$-1' different locations are included. As a result, an exact position is abstracted with other users ($k$-anonymity) and other locations ($\ell$-diversity), which makes it difficult for an adversary to infer valuable information (see Section 7 for related work).

Although these previous methods guarantee some degree of location privacy, both techniques have a critical limitation. The cloaking area could breach location semantic information, which possibly endangers the user's privacy. To be specific, the cloaking area could include only semantically similar locations even if it is mixed with other users and locations, and the adversary would be able to infer semantic meanings from the extended area. For example, if the extended area only includes an elementary school, high school, and university, then the adversary could infer that a mobile user is doing work related to 'teaching' or 'studying'.

In this paper, we propose a novel location privacy protection technique, which protects the location semantics from an adversary. In our scheme, location semantics are first learned from location data. Then, the trusted-anonymization server performs the anonymization using the location semantic information by cloaking with *semantically heterogeneous* locations. Thus, the location *semantic* information is kept secure as the cloaking is done with semantically heterogeneous locations and the true location information is not delivered to the LBS applications.

Our primary novel contributions are summarized as follows.

- We propose a method for mining location semantics from the perspective of location privacy. A staying duration feature is presented to capture the location semantics from trajectory data, and such mined location semantics are stored in an abstracted graph to be efficiently used.
- We propose a method to obtain a cloaking area which protects location semantic leakages. An adversary's prior and posterior knowledge of location semantics is generally modeled and the adversary's information gain from a cloaking area is restricted below a certain degree. According to the experimental results, our method is much safer at the same cost than $k$-anonymity and $\ell$-diversity based location privacy methods in terms of a semantic heterogeneity.

The remainder of this paper is organized as follows. Section 2 introduces the background of location privacy protection and points out its limitations. Section 3 describes how to obtain location semantic information and Section 4 presents how to compute an extended area with semantically heterogeneous locations. Section 5 shows the evaluation results of our proposed methods. Section 6 discusses the limitations and future work of our method, and Section 7 surveys related work. Section 8 concludes the paper.

## 2. BACKGROUND AND LIMITATIONS

Many researchers have tried to guarantee location privacy in using LBS, but their attempts have limitations. In this section, we first give the background of location privacy protection techniques and then describe the limitations of such techniques.

### 2.1 Background: Location Privacy Protection

A primary cause of location privacy breaches in using LBS lies in the fact that the exact position of a mobile device should be used and known to LBS applications. Thus, in order to protect exact position information, location privacy protection techniques use a *cloaking area* instead of exact position information for LBS requests. A *cloaking area* is defined as an area which includes the current position of a mobile device for the purpose of hiding an exact position. Based on using a cloaking area, the adversary cannot easily breach a mobile user's privacy since the exact current position is abstracted. There are largely two approaches for computing the cloaking area, each of which is based on well known data publishing protection techniques, $k$-anonymity [23] and $\ell$-diversity [19]. Depending on which technique has been adopted for location privacy, we refer to as *location $k$-anonymity* or *location $\ell$-diversity*.

The most well known approach is location $k$-anonymity [10, 6, 20, 3, 14, 26, 27], which provides at least a '$k$' anonymity level. In location $k$-anonymity, the cloaking area is extended until '$k$-1' other users are included. This is a good starting point for protecting location privacy because the adversary has to classify each person among '$k$' people to identify who actually submitted LBS requests. Similar to $k$-anonymity, location $\ell$-diversity [1, 18, 24, 28] extends a cloaking area until '$\ell$-1' different locations are included. In location $\ell$-diversity, the adversary cannot simply tell which location a mobile user actually visited since the cloaking area includes multiple locations.

These location $k$-anonymity and $\ell$-diversity techniques are performed by a trusted-anonymization server for a mobile device and

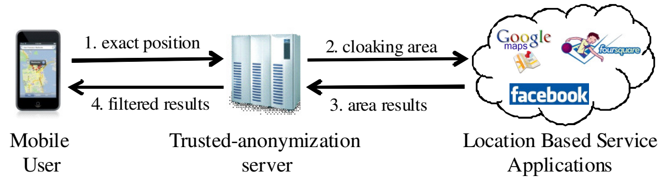
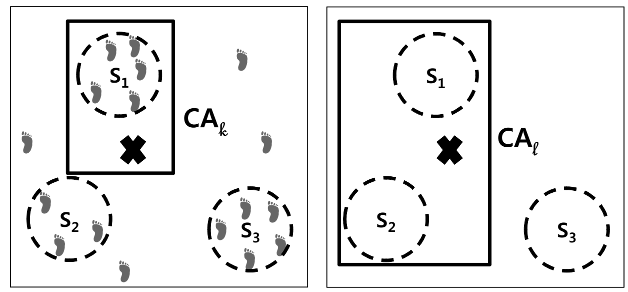



**Figure 1: Trusted-anonymization server based location privacy architecture**




(a) $k$-anonymity ($k$=5) (b) $\ell$-diversity ($\ell$=2)

**Figure 2: An example of a location similarity attack against a cloaking area**

we call this model a *trusted-anonymization server based model* hereafter. Figure 1 shows a trusted-anonymization server based model. At first, a mobile user requests a service to the trusted-anonymization server (line 1). In this request, the mobile user specifies her exact current position. Then the anonymization server computes the cloaking area using either a number of users (location $k$-anonymity) or locations (location $\ell$-diversity) nearby the user's position. The cloaking area is passed to LBS applications (line 2) and the LBS applications return all results related to the cloaking area (line 3). The anonymization server filters out unnecessary results and gives back the result corresponding to the mobile user's current location (line 4).

As a result, the exact position is not exposed to LBS applications because a cloaking area is used instead of the position. Though the delegating anonymization server knows the exact position of a mobile device, LBS applications only see an abstracted range of an area.

### 2.2 Limitations of Previous Location Privacy Protection

Though location $k$-anonymity and $\ell$-diversity approaches guarantee some degree of location privacy, both protection schemes are vulnerable to a *location similarity attack* which possibly endangers LBS user's privacy. In other words, a cloaking area $CA$, which contains $n$ locations denoted as $L(CA) = \{S_1, S_2, ..., S_n\}$, is vulnerable to a location similarity attack if all locations in $CA$ are semantically similar.

Figure 2 shows an example of a *location similarity attack*. Assume that each node from $S_1$ to $S_3$ represents a location; $S_1$ and $S_2$ are hospitals and $S_3$ is a library. Note that 'x' marked in the center represents the current location of a mobile user. Based on this setting, two rectangles, $CA_k$ and $CA_\ell$, represent the cloaking area under $k$-anonymity ($k = 5$) and $\ell$-diversity ($\ell = 2$) respectively. $CA_k$ in Figure 2-(a) is vulnerable to a location similarity attack since it only includes a single location, $S_1$. Thus, the mobile user using the cloaking area $CA_k$ for LBS requests would be highly linked to hospitals and can be suspected of having treatment. $CA_\ell$

in Figure 2-(b) is also vulnerable because two locations in $CA_\ell$ are similar in terms of the purpose of its visits. Similarly, a mobile user using $CA_\ell$ would be suspected for the same reason.

What we actually desire would be to have a cloaking area with semantically heterogeneous locations, which needs to include $S_3$ in the above example. However, both location $k$-anonymity and $\ell$-diversity methods fail to protect a mobile user from a location similarity attack. Location $k$-anonymity fails because it picks the users without considering their locations and location $l$-diversity fails because it picks the locations without considering the semantics of locations.

# 3. MINING LOCATION SEMANTICS

In order to compute secure cloaking area based on location semantics, we must be aware of *location semantics* beforehand. In this paper, location semantics are interpreted as which type of services are provided at locations. This interpretation makes sense from the perspective of location privacy, because *what people want to secure in location privacy is what they did in a location*.

Based on such an interpretation, this section describes how to obtain location semantics. Locations are identified first (Section 3.1) and features for capturing location semantics are proposed next (Section 3.2). Finally, a location semantic graph which represents semantic relations between locations is constructed (Section 3.3). The location semantic graph will be used to compute the cloaking area with semantically heterogeneous locations (Section 4).

## 3.1 Identifying Locations

There are several methods for finding a location. The first one is to utilize point of interest (POI) collections, which can be publicly available through OpenStreetMap [11], etc. Since POI collections provide information on locations that people may find useful or interesting, locations including coordinate information can be obtained. The second method, which we use in our experiments, is to analyze trajectory data. The trajectory data contains coordinate information (usually GPS points) and corresponding timestamps. If a trajectory stays in a limited area over a time threshold value, it indicates that someone stayed in that location and did some meaningful job. Thus, locations can be discovered by identifying such a limited area. For interested readers, please refer to [31] for more details.

## 3.2 Features for Location Semantics

Our proposed method for discovering location semantics is based on the following observations. People visit locations mostly with a reason. We go to restaurants to have food, schools to attend classes, or hospitals to see a doctor. Since we have reasons for a visit, we stay for a while in a location for those reasons. Moreover, we spend a different amount of time depending on these reasons. Motivated by these observations, we propose two quantitative features for extracting location semantics, which are named staying duration, and usage time context.

**Staying duration:** People spend different amount of time in a location depending on what they do there. We call this amount of time staying duration. Intuitively, having food in a restaurant generally takes one or two hours, whereas students usually stay more than six hours at school. In addition, restaurants themselves have different staying duration distributions according to what they actually serve. For example, eating at a fine dining restaurant takes much longer time than at a fastfood restaurant. Thus visiting purposes of each location can be captured by using the staying duration.

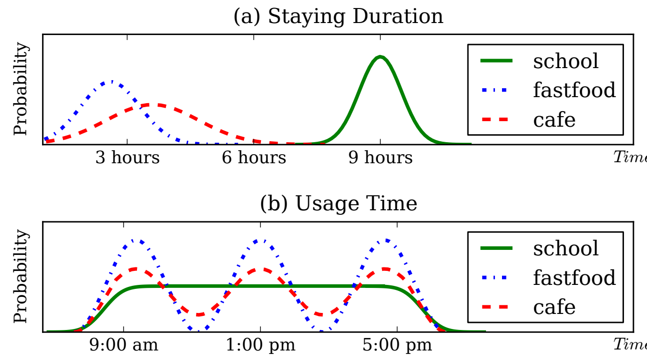



**Figure 3: PDF of features for location semantics**

**Usage time context:** It is common to see that places have their own active (popular) times. For example, most restaurants are full of customers during lunch or dinner time. On the contrary, most bars are full during the night but quiet in the morning. Relying on common sense, locations have differences in the distribution of usage times depending on services provided.

The feature data can be acquired via trajectory or survey data which contains such information. From the data, each feature's distribution is computed for each location. For example, hypothetical distributions of each feature are plotted in Figure 3. Distributions for locations such as school, fastfood restaurant, and cafe, are represented according to each location's characteristics. In Figure 3-(a), the x-axis represents staying duration and it reflects the fact that people stay longer at school than a fastfood restaurant or cafe. In Figure 3-(b), the x-axis represents usage time. A fastfood restaurant and cafe are crowded during meal times, but a school is active throughout the daytime.

Though the two features mentioned above can be used to characterize semantics of locations, we choose to use staying duration in our experiments. This is because the usage time context does not capture location semantics in some cases. For instance, a cafeteria and a dining restaurant are difficult to distinguish based on the usage time context because both places have similar crowded times. However, the staying duration feature is able to distinguish such places well because most people spend more time in a dining restaurant than in a cafeteria.

## 3.3 Constructing Location Semantic Graph

In order to represent location semantics with a simplified data structure, we present a *location semantic graph*. A distance between places (edge-weight) is computed first and a cluster of locations (node) are determined next. Finally, a location semantic graph is built with edges and nodes.

### 3.3.1 Distance measure

The distance measure should be able to capture the semantic differences between locations. Since all features can be represented in distributions, Kullback-Leibler (KL) divergence could be a reasonable choice to measure the distances between two distributions. KL divergence measures the distances of two distributions $P, Q$ with

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \left\{ \frac{p(x)}{q(x)} \right\} dx.$$

However, KL divergence is not an appropriate measure to capture semantic differences, which is explained in the following example.

Figure 4 shows three staying duration distributions P, Q, and R. Intuitively, P is similar to Q than to R. That is, our desired result

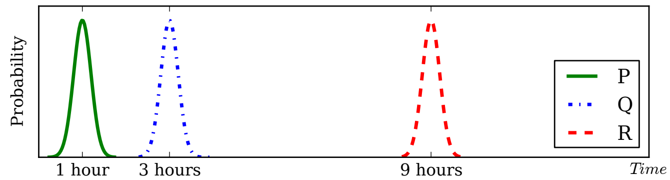



**Figure 4: Example of staying duration distribution**

is $D(P||Q) < D(P||R)$. However, the actual result from KL divergence is $D_{KL}(P||Q) \approx D_{KL}(P||R)$ because KL divergence cannot capture that the semantic distance between 1 hour and 9 hours is bigger than that between 1 hour and 3 hours is.

In order to overcome this limitation of KL divergence, we use Earth Mover's Distance which has recently been adopted for privacy protection of publishing data [15].

**Earth Mover's Distance (EMD):**[1] A distribution can be interpreted as an arbitrary arrangement of a mass of particles. In this view, a distribution can be transformed to another distribution by moving particles. EMD captures the minimum costs of transporting particles to equalize two distributions, which can be formally defined using the transportation problem. Suppose $P$ and $Q$ can be represented $P = \{p_1, p_2, \ldots, p_m\}$, $Q = \{q_1, q_2, \ldots, q_m\}$. The workload, needed to make two distribution the same, is defined as follows.

$$\text{WORK}(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{m} f_{ij} d_{ij}$$

with constraints

$$f_{ij} \geq 0, \quad 1 \leq i, j \leq m, \quad p_i - \sum_{j=1}^{m} f_{ij} + \sum_{i=1}^{m} f_{ij} = q_i,$$

$$\sum_{i=1}^{m} \sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{m} p_i = \sum_{j=1}^{m} q_j = 1,$$

where $f_{ij}$ is a flow of mass (the amount of moving particles) from $i$ to $j$, $d_{ij}$ is a ground distance from $i$ to $j$, and all flows $F = [f_{ij}]$. From this setting, EMD measures the minimum workload, defined as

$$D_{EMD}(P, Q) = \min_{F} \text{WORK}(P, Q, F).$$

The virtue of EMD is in adjusting ground distance which enables us to capture semantic differences. In our case, the ground distance ($d_{ij}$) is set to be the normalized difference between staying durations, i.e., the difference of staying duration divided by the maximum difference. Thus, $0 \leq d_{ij} \leq 1$ for all $i$ and $j$, which results in $0 \leq D_{EMD}(P, Q) \leq 1$ [15]. When revisiting the example in Figure 4, EMD performs $D_{EMD}(P, Q) < D_{EMD}(P, R)$ as desired since EMD sees (1 hour, 3 hours) pair has a much smaller ground distance than (1 hour, 9 hours) pair has. As a result, EMD is a better distance measure than KL divergence for our application and we adopt EMD as the distance measure for location semantic differences.

#### *3.3.2 Location clustering*

Based on EMD, locations are grouped into clusters before being structured in a graph. The main reason for performing clustering is that the data which would represent location semantics across all locations is too complicated and huge. Because there are numerous locations which are bases of people activities, computational complexities and exchanging costs with such data would be considerable burdens.

In this respect, $k$-means clustering is performed by grouping semantically similar locations. Pair-wise distances between locations are represented using EMD and the centroid of each cluster is updated by computing the average of the distributions in the cluster. Detailed explanations for $k$-means clustering can be found in [12].

#### *3.3.3 Location semantic graph*

After having clustered locations, all semantic information is represented by a graph based structure, which we call a *location semantic graph*. In a location semantic graph, a node represents clustered locations and an edge weight represents EMD between corresponding nodes. EMD between cluster nodes is defined as

$$D(C_i, C_j) = \sum_{l_i \in C_i} \frac{1}{|C_i|} \sum_{l_j \in C_j} \frac{D_{EMD}(l_i, l_j)}{|C_j|}$$

where $C_i, C_j$ are clusters, $|C_i|, |C_j|$ are the number of locations in a cluster, and $l_i, l_j$ are the distributions of each location. Note that $D(C_i, C_j)$ is also normalized into [0,1] because $0 \leq D_{EMD}(l_i, l_j) \leq 1$ for all $i, j$.

Figure 5 shows an example of a location semantic graph. The nodes indicate clustered locations, each of which is a group of semantically similar locations. The edge weight indicates the semantic differences between clusters computed by EMD. The number beside the cluster is the normalized number of locations in the cluster. Since all clusters have one location respectively, all clusters are 0.25.

## 4. $\theta$-SECURE CLOAKING AREA

Obtaining a privacy preserving cloaking area is difficult because the way of checking the safety of the cloaking area in terms of location semantics is unexplored. In other words, it is difficult to evaluate *how much location semantic information an adversary will gain from the cloaking area*. To handle this, we propose methods for evaluating the safety of cloaking areas (Section 4.1) and obtaining a $\theta$-secure cloaking area (Section 4.2).

### 4.1 Computing EMD of Cloaking Area

An adversary gains location semantic information from a cloaking area only if it is different from what he/she already knew. In other words, before seeing the cloaking area, the adversary's knowledge is the location semantics of an entire area since he/she has no idea where a mobile user is located (a prior belief). After seeing the cloaking area, the adversary obtains more specific location semantics corresponding to the cloaking area (a posterior belief). Thus, the information gain of the adversary from the cloaking area can be the difference between such prior beliefs and posterior beliefs, which is directly linked to the safety of the cloaking area. Motivated by this observation, we aim at evaluating the differences of the prior belief and the posterior belief.

The prior belief is the location semantic information of an arbitrary area since the adversary has no idea where a mobile user is located. Thus, the prior belief is represented as a location semantic graph in a hypothetically large area. For the posterior belief, the adversary sees a cloaking area which possibly contains more specific semantic information. Thus, the posterior belief is a more elaborated location semantic graph which is built upon the prior belief.

[1]We do not present a detailed description of EMD. Interested readers please refer to [22, 15].

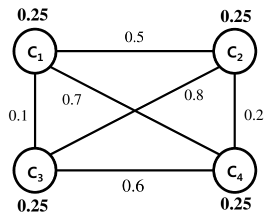
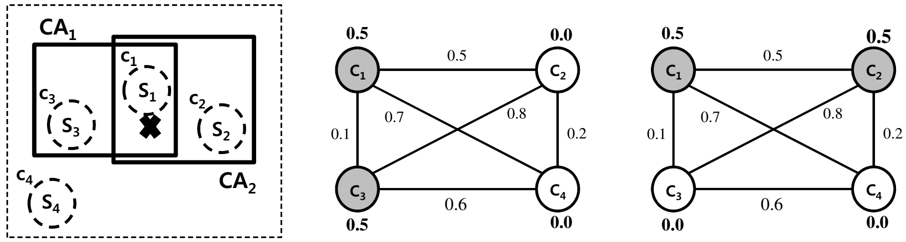



**Figure 5: Location semantic graph**




(a) Spatial snapshot of a mobile user

(b1) Location semantic graph of $CA_1$

(b2) Location semantic graph of $CA_2$

**Figure 6: A spatial snapshot of a mobile user and its location semantic graph**

Figure 6 illustrates a spatial snapshot of a mobile user (a) and its location semantic graph (b1) and (b2). Note that the location semantic graph in Figure 5 is constructed based on Figure 6-(a) to be a running example. In Figure 6, locations from $S_1$ to $S_4$ are represented with their corresponding cluster labels. A mobile user located at 'x' marked in the center has two choices for the cloaking area, $CA_1$ and $CA_2$. The cluster weight of the location semantic graph in (b1) and (b2) is changed because it only considers the cloaking area. For example, (b1) shows 0.5 on $C_1$, $C_3$, and 0 on $C_2$, $C_4$, because $C_1$, $C_3$ have one location respectively but $C_2$, $C_4$ have no location.

To measure the safety of a cloaking area by comparing the prior belief and the posterior belief, we compute the EMD of location semantic graphs between before and after seeing the cloaking area. A node (clustered locations) in a location semantic graph is converted into a discrete domain in EMD, and an edge weight (semantic differences) is converted into a ground distance $d_{ij}$. Numerical computational results of $CA_1$ and $CA_2$ are shown below as $D_{EMD}(P_{CA_1}, P_E)$ and $D_{EMD}(P_{CA_2}, P_E)$ respectively.

$$\begin{aligned} D_{EMD}(P_{CA_1}, P_E) &= \min_f \sum_i \sum_j f_{ij} d_{ij} \\ &= f^*_{12}0.5 + f^*_{14}0.7 + f^*_{32}0.8 + f^*_{34}0.6 \\ &= 0.25 \cdot 0.5 + 0 \cdot 0.7 + 0 \cdot 0.8 + 0.25 \cdot 0.6 \\ &= 0.275 \\ D_{EMD}(P_{CA_2}, P_E) &= \min_f \sum_i \sum_j f_{ij} d_{ij} = 0.075 \end{aligned}$$

where $P_{CA_k}$ indicates the posterior belief of cloaking area $CA_k$, $P_E$ indicates the prior belief. $d_{ij}$ and $f_{ij}$ is a ground distance and a flow between clusters $C_i$ and $C_j$ respectively, and $f^*_{ij}$ is an optimal flow making two beliefs the same (see Section 6 for the discussion on the time complexity of EMD computations). Note that $0 \le D_{EMD}(P_{CA}, P_E) \le 1$ is satisfied, because $0 \le d_{ij} = D(C_i, C_j) \le 1$ holds.

Based on $D_{EMD}(P_{CA_2}, P_E) < D_{EMD}(P_{CA_1}, P_E)$, we argue that $CA_2$ is more secure than $CA_1$. This interpretation makes sense since $C_1$ and $C_2$ have bigger semantic distances than $C_1$ and $C_3$ have.

## 4.2 Finding $\theta$-Secure Cloaking Area

Based on the safety measure of the cloaking area, a $\theta$-secure cloaking area is defined below.

*Definition 1.* ***θ-Secure Cloaking Area.*** If a cloaking area $CA$ satisfies $D_{EMD}(P_{CA}, P_E) \le \theta$, we denote this cloaking area as a $\theta$-secure cloaking area.

In order to obtain a $\theta$-secure cloaking area, a cloaking area is extended until it satisfies $\theta$-secure cloaking area. Algorithm 1 describes a greedy algorithm for finding a $\theta$-secure cloaking area. Note that all locations are mapped onto a $2^n$-by-$2^n$ grid for efficient computation ($n$ is selected by the user).

**Algorithm 1.** Finding $\theta$-secure cloaking area

**input** : a grid map of location Map, a location semantic graph G, a prior belief $\mathsf{P_E}$, a threshold value $\theta$, and a maximum number of iterations maxLoop.
**output**: $\theta$-secure cloaking area

1 CA = `getInitialCA` () ;
2 **for** *i=1* **to** maxLoop **do**
3   **foreach** dir $\in$ `getPossibleDirs` (CA, Map) **do**
4     $\mathsf{CA_{dir}}$ = `extendCloakingArea` (CA, Map, dir) ;
5     $\mathsf{P_{CA_{dir}}}$ = `computePosteriorBelief` (CA, G) ;
6     $\mathsf{EMD_{CA_{dir}}} = D_{EMD}(\mathsf{P_{CA_{dir}}}, \mathsf{P_E})$ ;
7   (p, $\mathsf{EMD_{min}}$) = `findMinEMD` () ;
8   CA = $\mathsf{CA_p}$;
9   **if** $\mathsf{EMD_{min}} \le \theta$ **then**
10     break ;
11 **return** CA ;

In Algorithm 1, a single cell, which includes a current mobile position, is chosen as an initial cloaking area (line 1). Next, obtain possible directions from {North, East, South, West}, which would make a rectangular form of a cloaking area (line 2). For each direction, the cloaking area is extended (line 4), and a posterior belief and EMD are computed (line 5-6). After iterating over all possible directions, select a cloaking area which has the minimum EMD value (line 7-8). Finally, if the minimum EMD value is less than the threshold $\theta$ (line 9), then return the corresponding cloaking area which is a $\theta$-secure cloaking area (line 11).

# 5. EVALUATION

## 5.1 Experimental Setting

We extended a traffic simulator [2] to consider the human mobility patterns, which enables us to evaluate the effectiveness and performance of our proposed methods.[2] As described in Section 3, people stay a while in a location according to the location semantics. To be able to reflect such characteristics, the simulator should be able to determine the staying duration of each location. However, to the best of our knowledge, all known traffic simulators do not consider location semantics but simply generate random

[2]All implementations and data sets used for the evaluation are available on the project page, http://hpc.postech.ac.kr/locpriv.

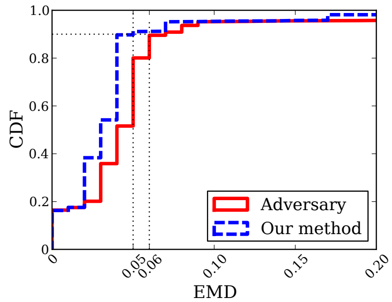



**Figure 7: The cumulative distribution function (CDF) of cloaking areas' EMD ($\theta_t$ = 0.05)**

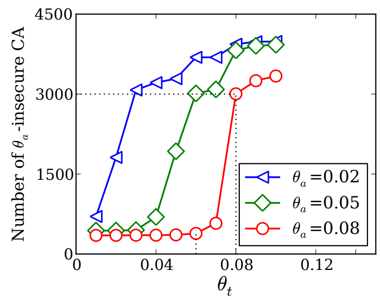



**Figure 8: $\theta_a$-insecure cloaking areas**

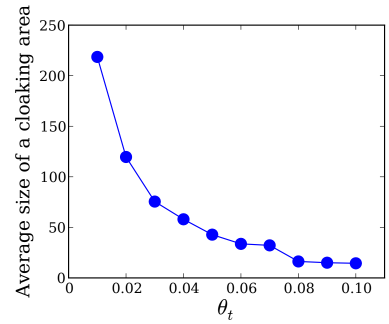



**Figure 9: The average size of a cloaking area**

**Table 1: Parameters for the traffic simulator**

| Parameters | $N_c$ | $t_{\min}$ | $t_{\max}$ | $N_l$ | $\alpha$ | $\beta$ | $N_{users}$ |
|---|---|---|---|---|---|---|---|
| Values | 4 | 50 | 400 | 2000 | 10 | 10 | 4000 |

movements. Thus, we modify the state of the art traffic simulator, Network-based Generator of Moving Objects [2], in order to reflect realistic human movement patterns.

In our modified simulator, staying duration patterns are modeled relying on the Gaussian distribution with latent variables. First, we sample $N_c$ clusters which represent the group of semantically similar locations. For a cluster $C_k$, the mean $\mu_{c_k}$ and the standard deviation $\sigma_{c_k}$ of the cluster staying duration pattern are picked from the uniform distributions, $\mu_{c_k} \sim \mathcal{U}(t_{\min}, t_{\max})$ and $\sigma_{c_k} \sim \mathcal{U}(0, \alpha)$, where $t_{\max}$ and $t_{\min}$ represent the maximum and minimum staying duration and $\alpha$ controls the variance of the cluster. Next, $N_l$ locations are sampled from the clusters. For a location $l_{ki}$ from the cluster $C_k$, the mean $\mu_{l_{ki}}$ and the standard deviation $\sigma_{l_{ki}}$ of a location staying duration pattern are picked from the Gaussian distributions, $\mu_{l_{ki}} \sim \mathcal{N}(\mu_{c_k}, \beta)$ and $\sigma_{l_{ki}} \sim \mathcal{N}(\sigma_{c_k}, \beta)$, where $\beta$ controls the variance among the locations in each cluster. Once a moving object reaches the location, an actual staying duration $t$ is determined from the Gaussian distribution $t \sim \mathcal{N}(\mu_{l_{ki}}, \sigma_{l_{ki}})$. In addition, a moving object chooses its next destination based on a lévy-flight process that is known to be followed by human mobility patterns according to a recent study [9]. Relying on the lévy-flight process, the probability of visiting nearby locations is higher than those of far away locations.

We generated trajectory data using our modified simulator. $N_{users}$ objects were moved over a real road map of Oldenburg, Germany. The road map contains 6,105 nodes and 7,035 edges, a city about 15×15 km$^2$, which is presented in a $2^7$-by-$2^7$ grid. Table 1 lists the parameters used for simulating moving objects. It is assumed that a LBS request is sent with a 2% probability from the reported positions. A staying duration pattern for each location is obtained from the trajectory and a location semantic graph is learned from the staying duration patterns. Using the location semantic graph, a $\theta$-secure cloaking area is computed for each LBS request. In order to compare with the proposed method, a cloaking area is also computed based on $k$-anonymity and $\ell$-diversity techniques as described in Section 2.1.

## 5.2 Experimental Results

### 5.2.1 *Evaluation on location semantic learning*

From 2000 locations used for generating trajectory data, 1948 locations were identified. Some locations were missing because a small number of moving objects passed by them. Among the identified locations, $k$-means clustering was performed with a parameter $k = 4$, and the clustering result was close to the perfect; i.e. $F_1 = 0.997$. In order to measure the correctness of the location semantic graph, we measured the normalized edge-weight differences between the modeled location semantic graph ($G_M$) and learned location semantic graph ($G_L$) as

$$D(G_M, G_L) = \frac{2}{N_C(N_C - 1)} \sum_{i=1}^{N_C} \sum_{j=1}^{i-1} \left| [\mathbf{A}_M]_{ij} - [\mathbf{A}_L]_{ij} \right|,$$

where $N_C$ denote the number of clusters, and $\mathbf{A}_M$ and $\mathbf{A}_L$ denote the weighted adjacency matrix of $G_M$ and $G_L$ respectively. Since the edge-weight is normalized into [0,1], the minimum and maximum of $D(G_M, G_L)$ are 0 and 1. Moreover, to evaluate the goodness of $D(G_M, G_L)$, we randomly created the location semantic graph $G_R$ and the average is obtained from 30 times of running $D(G_M, G_R)$. From the experiments, we obtained $D(G_M, G_L) = 0.084$ and $\overline{D(G_M, G_R)} = 0.281$, which shows much closer results to the modeled location semantic graph.

### 5.2.2 *Evaluation on $\theta$-secure cloaking area*

**Attack models and measures:** The adversary is assumed to have location semantic information at best, e.g. the location semantic graph of the model used for generating the trajectory. Consequently, the adversary takes a better position than our method because our method uses the learned location semantic graph which slightly deviates from the actual location semantic graph.

First of all, we checked how much location semantic information the adversary would gain from a cloaking area, which is quantified in EMD of a cloaking area. Since the adversary and our method have a different location semantic graph, EMD computed from the same cloaking area would be different. This implies that the $\theta$-cloaking area returned by our algorithm may not guarantee $\theta$ degree of protection from the adversary's view. To be clear, we denote $\theta_t$ as a threshold for computing a cloaking area in our algorithm while $\theta_a$ is the EMD from the adversary's view.

To evaluate resistance against a location similarity attack launched by the adversary, we measure the number of $\theta_a$-*insecure cloaking areas*. A $\theta_a$-insecure cloaking area is a cloaking area which has EMD higher than $\theta_a$ from the adversary's point of view. It implies that the adversary would gain $\theta_a$ degree or more of location semantics information from the corresponding cloaking area.

The cost of using a cloaking area is measured by the size of the cloaking area. Since a mobile user uses the cloaking area for the anonymization, he/she needs to spend more network traffic and computing costs proportional to the size of the cloaking area.

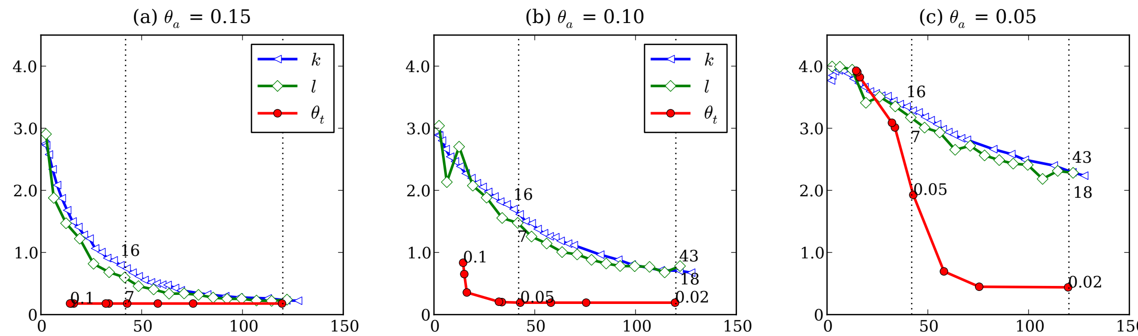



**Figure 10: The number of $\theta_a$-insecure cloaking areas by varying $k$, $l$, $\theta_t$. X-axis: Average size of cloaking areas; Y-axis: Number of $\theta_a$-insecure cloaking areas ($\times 10^3$).**

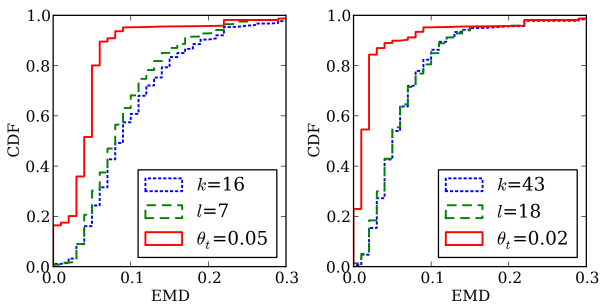



**Figure 11: EMD of all cloaking areas by varying $k$, $\ell$, and $\theta_t$**

**Safety of a $\theta$-secure cloaking area:** Figure 7 shows the cumulative distribution function (CDF) of all 4000 cloaking areas' EMD for $\theta_t = 0.05$. When EMD is computed from our algorithm (method's view), 90% of the cloaking areas' EMD is below 0.05, which implies that our algorithm successfully returns $\theta_t$-secure cloaking area in most cases. Some cloaking area's EMD is over 0.05 because our algorithm failed to find $\theta_t$-secure cloaking area under the given number of extending steps. When it comes to the adversary's view, it shows slightly higher EMD value than the method's view. Although computed EMD is different, EMD of most cloaking areas is below 0.06 from the adversary's view, which is close to the model's view.

Figure 8 shows the number of $\theta_a$-insecure cloaking areas while changing $\theta_t$. Overall, as $\theta_t$ increases the number of insecure cloaking areas also increases, because a lower privacy level is enforced for higher $\theta_t$. In addition, an actual $\theta_a$ degree of anonymity is obtained when $\theta_t < \theta_a$. For example, in 0.08-insecure case ($\theta_a = 0.08$), there are about 3000 insecure cloaking areas when $\theta_t = 0.08$, but fewer than 450 insecure cloaking areas when $\theta_t \leq 0.06$. Since this difference between $\theta_t$ and $\theta_a$ is a relative difference between the model's view and the adversary's view, we believe this does not indicate a weakness in our method. To achieve $\theta_a$ degree of anonymity, a smaller $\theta_t$ can be used, e.g. using $\theta_t$=0.06 to achieve $\theta_a$=0.08 degree of anonymity in this case.

**Cost of a $\theta$-secure cloaking area:** Figure 9 shows the average size of the cloaking area versus $\theta_t$. As $\theta_t$ decreases the average size of a cloaking area increases because a small $\theta_t$ implies more strict anonymization degree and causes a larger cloaking area. In addition, the curve follows a negative exponential function, which can be interpreted as follows. As $\theta_t$ approaches 1 (no location privacy guarantee), the average size would approach zero (the size of an exact position). On the contrary, as $\theta_t$ approaches zero (the maximum location privacy guarantee), the average size would approach infinity (the size of the maximum area).

**Comparison with $k$-anonymity and $\ell$-diversity:** The safety and the costs of our method are compared with baseline methods, location $k$-anonymity and $\ell$-diversity. 4000 requests are anonymized with our method as well as with baseline methods while changing the parameter of each method's privacy requirement ($k$, $l$, and $\theta_t$).

Figure 10 plots the number of $\theta_a$-insecure cloaking areas with the average size of cloaking areas for each parameter setting. In most cases, our method shows a lower number of $\theta_a$-insecure the cloaking areas when the average size of cloaking area is similar to the others. For example, in Figure 10-(c), when $k = 43$, $l = 18$, and $\theta_t = 0.02$, the number of insecure cloaking areas of $\theta_t$ is much lower than $k$ and $\ell$. This signifies that our method provides better safety when the costs for enforcing the location privacy are the same.

Moreover, our method also shows better performances in costs when the provided safety is the same. To be specific, when the number of $\theta_a$-insecure cloaking areas are the same, our method has a lower average size of cloaking areas than the others. Such superiority of our method becomes clearer as a safety requirement gets strict. In other words, as the criteria for insecure cloaking areas ($\theta_a$) decreases, the gap between our method ($\theta_t$) and baseline methods ($k$ and $\ell$) remarkably increases as shown from Figure 10-(a) to Figure 10-(c). Thus, a mobile user would get more benefits if location privacy preservation is the primary need.

In order to investigate more details, Figure 11 shows CDF of EMD of all cloaking areas when the costs, the average size of cloaking areas, are similar, e.g. ($k$=16, $\ell$=7, $\theta_t$=0.05) and ($k$=43, $\ell$=18, $\theta_t$=0.02). In both cases, $\theta_t$ is always more concentrated to lower EMD than $k$ and $l$, which implies that our method guarantees better protection of the location semantics against any level of a location similarity attack.

Figure 12 illustrates the two-dimensional distribution of each cloaking area's size and its EMD, when the costs are similar. Overall, the distribution of the cloaking area from our method is concentrated in the bottom but the distributions from baseline methods are dispersed upward, which indicates that our method provides better location semantic protection at the same costs. Moreover, the brightest cell in Figure 12-(b3) is located more left than in Figure 12-(b1) and (b2). As a result, a large portion of the cloaking area in our method has a smaller size. However, since the average size of the cloaking area in these three figures is similar, some cloaking areas of our method would have a bigger size. This suggests that a mobile user using our method would enjoy low costs in most cases, but in some cases would pay relatively high costs.

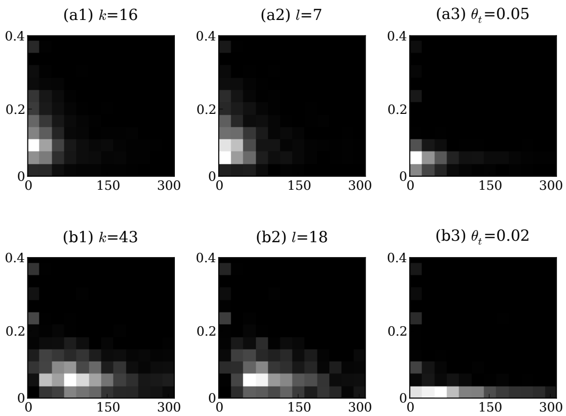



**Figure 12: Two-dimensional distribution of each cloaking area's size and its EMD. The brighter the cell is the more frequent the occurrence is. X-axis: Size of a cloaking area; Y-axis: EMD.**

# 6. DISCUSSION

Most previous work guarantees the anonymity and unlinkability based on $k$-anonymity and $\ell$-diversity. Since our work focuses on protecting the location semantics using $\theta$-secure cloaking area, the anonymity and unlinkability could be unprotected in some cases. For resolving this issue, all three parameters, $\theta$-secure cloaking area, $k$-anonymity, and $\ell$-diversity, could be used together, similarly *PrivacyGrid* [1] uses $k$-anonymity and $\ell$-diversity together. In addition, since our method does not have *reciprocity* [14], $\theta$-secure cloaking area may reveal the user location information caused by outliers; i.e. in peripheral areas where there are few semantically related locations, the cloaking area can become quite large and reveal that the user is in that peripheral area. To satisfy the reciprocity property, *HilbertCloak* [14] algorithm could be used.

Furthermore, our method relies on an EMD computation both in an offline step (clustering semantic locations) and an online step (computing a $\theta$-secure cloaking area). Thus, it is vital for our method to be able to efficiently compute EMD, especially in the online stage. The time complexity for computing EMD can be formalized using a minimum cost network problem, and it can be solved in $O(n^3 \log n)$ [22], where $n$ is either the number of locations or clusters in our method. Because the number of clusters would be quite small, we believe this is not the serious load for the online stage. For the offline stage, approximation algorithms can be adopted, which empirically lead to $O(n)$ or $O(n^2 \log n)$ [17, 21] with error bounds. Note that each approximation algorithm requires a specific setting for the ground distance; i.e. $L_1$ distances [17] or thresholded distances [21]. Since the ground distance in our method is on the non-euclidean space, more investigations should be done to adopt approximation algorithms.

# 7. RELATED WORK

**Anonymization for publishing relational database:** In order to protect published relational database data such as medical data, k-anonymity [23] was developed. $k$-anonymity guarantees the adversary cannot distinguish an individual record from at least '$k$-1' other tuples. However, since '$k$-1' other tuples may be the same sensitive values, $\ell$-diversity [19] was proposed which enforces tuples in the same group have at least '$\ell$-1' diverse sensitive values. Ninghui et al. proposed $t$-closeness [15] to resolve the semantic breaches of $k$-anonymity and $\ell$-diversity. $t$-closeness guarantees tuples in the same group are statistically similar to the entire data using EMD.

**Anonymization for location based services:** Gruteser et al. first proposed location privacy technique based on the $k$-anonymity concept and trusted-anonymization server [10]. The cloaking area, which is extended until '$k$-1' other users are included, is computed through a trusted anonymization server and used for LBS requests instead of exact coordinates. A series of work has improved the computation of a cloaking area under $k$-anonymity. *Clique-Cloak* [6] and *Casper* [20] proposed personalized location anonymization. CliqueCloak locates a clique in a graph to compute the cloaking area and Casper uses a quadtree-based pyramid data structure for fast computation of the cloaking area. *Probabilistic Cloaking* [3] proposed imprecise LBS requests which yield probabilistic results. The *HilbertCloak* [14] algorithm utilizes Hilbert space-filling curve and its cloaking area is independent of mobile user distribution. To reduce the size of the cloaking area, historical locations of mobile nodes are used for computing the cloaking area instead of the current mobile node's location [26]. Feeling-based location privacy [27] sets '$k$' using the location where a mobile user feels safe enough to disclose her location.

Using the $k$-anonymity based technique, the cloaking area may include only one meaningful location (e.g. a specific hospital or school) and disclose strong relationships to such a location. Thus, *PrivacyGrid* [1] proposed location $\ell$-diversity, which extends the cloaking area until '$\ell$-1' different locations are included. *PrivacyGrid* used both location $k$-anonymity and $\ell$-diversity so that the anonymization into different persons and locations can be done together. Similarly, XSTAR [24] attempted to achieve the optimal balance between high query-processing efficiency and robust inference-attack resilience while considering $k$-anonymity and $\ell$-diversity together. Several works [28, 25, 5] have identified semantic breach issues, but impractical assumptions were made to resolve such issues. *Location diversity* [28] assumed that location semantics are pre-labeled and $\ell$-diversity is able to protect the semantic breaches, and *PROBE* [5] also assumed the pre-labeled location semantics and requires many profile parameters for each user. *p-sensitivity* [25] assumed that a LBS request is classified into a sensitive or insensitive request.

In architectural perspectives, previous location privacy protection schemes mostly follow the trusted-server-based model in which an anonymization server delegates all LBS requests for mobile users. However, *SpaceTwist* [29] proposes an client-based model which uses a fake location instead of using 'exact location' for computing a cloaking area. [7] also proposes the client-based model based on Private Information Retrieval (PIR) with cryptographic techniques. The peer-to-peer model [4, 8, 13] attempts to remove trusted anonymization server. Based on a decentralized cooperative peer-to-peer model, location information among nearby peers is shared and used for computing the cloaking area.

**Location data mining:** A recommending system for travelers is proposed in [31, 30]. Trajectory data is analyzed and interesting locations are mined based on visited frequencies on each location. In [16], a similarity between users is mined based on the sequence property of people's movement behaviors, which also enables to identify correlations among locations. To the best of our knowledge, our research is the first to discover location semantics using staying duration and utilize it for computing semantically heterogeneous cloaking areas.

## 8. CONCLUSION

This paper proposes novel location privacy protecting techniques. Our proposed methods protect the location semantics from the LBS applications by performing the cloaking with *semantically heterogeneous* locations. Experimental results validate our proposed methods.

## 9. REFERENCES


[1] B. Bamba, L. Liu, P. Pesti, and T. Wang. Supporting anonymous location queries in mobile environments with privacygrid. In *Proceeding of the 17th International Conference on World Wide Web (WWW)*, 2008.

[2] T. Brinkhoff. A framework for generating network-based moving objects. *GeoInformatica*, 6(2):153–180, 2002.

[3] R. Cheng, Y. Zhang, E. Bertino, and S. Prabhakar. Preserving user location privacy in mobile data management infrastructures. In *Privacy Enhancing Technologies (PET)*, 2006.

[4] C.-Y. Chow, M. F. Mokbel, and X. Liu. A peer-to-peer spatial cloaking algorithm for anonymous location-based service. In *Proceedings of the ACM International Symposium on Advances in Geographic Information Systems (GIS)*, 2006.

[5] M. Damiani, E. Bertino, and C. Silvestri. The PROBE Framework for the Personalized Cloaking of Private Locations. *Transactions on Data Privacy*, 3(2):123–148, 2010.

[6] B. Gedik. Location Privacy in Mobile Systems: A Personalized Anonymization Model. In *IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2005.

[7] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan. Private queries in location based services: anonymizers are not necessary. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2008.

[8] G. Ghinita, P. Kalnis, and S. Skiadopoulos. PRIVE: anonymous location-based queries in distributed mobile systems. In *Proceedings of the International Conference on World Wide Web (WWW)*, 2007.

[9] M. C. González, C. a. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–82, June 2008.

[10] M. Gruteser and D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the International Conference on Mobile Systems, Applications and Services (MobiSys)*, 2003.

[11] M. M. Haklay and P. Weber. Streetmap: User-generated street maps. *IEEE Pervasive Computing*, 7:12–18, 2008.

[12] J. Han and M. Kamber. *Data mining: concepts and techniques*. Morgan Kaufmann, 2006.

[13] H. Hu and J. Xu. Non-Exposure Location Anonymity. In *IEEE International Conference on Data Engineering (ICDE)*, 2009.

[14] P. Kalnis, G. Ghinita, K. Mouratidis, and D. Papadias. Preventing Location-Based Identity Inference in Anonymous Spatial Queries. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 19(12):1719–1733, 2007.

[15] N. Li, T. Li, and S. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *IEEE International Conference on Data Engineering (ICDE)*, 2007.

[16] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. Mining user similarity based on location history. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS)*, 2008.

[17] H. Ling and K. Okada. An efficient Earth Mover's Distance algorithm for robust histogram comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 29(5):840–853, 2007.

[18] F. Liu, K. a. Hua, and Y. Cai. Query l-diversity in Location-Based Services. In *International Conference on Mobile Data Management: Systems, Services and Middleware (MDM)*, 2009.

[19] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy Beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):1–52, 2007.

[20] M. Mokbel, C. Chow, and W. Aref. The New Casper: Query Processing for Location Services without compromising privacy. In *Proceedings of the International Conference on Very large data bases (VLDB)*, 2006.

[21] O. Pele and M. Werman. Fast and robust earth mover's distances. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[22] Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision (IJCV)*, 40(2):99–121, 2000.

[23] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, 10(5):557–570, 2002.

[24] T. Wang and L. Liu. Privacy-aware mobile services over road networks. *Proceedings of the VLDB Endowment*, 2(1):1042–1053, 2009.

[25] Z. Xiao, J. Xu, and X. Meng. p-Sensitivity: A Semantic Privacy-Protection Model for Location-based Services. In *International Conference on Mobile Data Management Workshops (MDMW)*, 2008.

[26] T. Xu and Y. Cai. Exploring Historical Location Data for Anonymity Preservation in Location-Based Services. In *IEEE International Conference on Computer Communications (INFOCOM)*, 2008.

[27] T. Xu and Y. Cai. Feeling-based location privacy protection for location-based services. *Proceedings of the ACM conference on Computer and communications security (CCS)*, 2009.

[28] M. Xue, P. Kalnis, and H. Pung. Location Diversity: Enhanced Privacy Protection in Location Based Services. *Location and Context Awareness (LoCA)*, pages 70–87, 2009.

[29] M. Yiu, C. Jensen, X. Huang, and H. Lu. Spacetwist: Managing the trade-offs among location privacy, query performance, and query accuracy in mobile services. In *IEEE International Conference on Data Engineering (ICDE)*, 2008.

[30] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative location and activity recommendations with GPS history data. In *Proceedings of the International Conference on World Wide Web (WWW)*, 2010.

[31] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the International Conference on World Wide Web (WWW)*, 2009.