Semantic-Aware Location Privacy Preservation on Road Networks

Yanhui Li[1(✉)], Ye Yuan[1], Guoren Wang[1], Lei Chen[2], and Jiajia Li[3]

[1] Northeastern University, Shenyang, China
yliboneu@gmail.com
[2] Hong Kong University of Science and Technology, Hong Kong, China
[3] Shenyang Aerospace University, Shenyang, China

Abstract. In this paper, we address the topic of location privacy preservation of mobile users on road networks. Most existing techniques of privacy preservation rely on structure-based *spatial cloaking*, but pay little attention to location semantic information. Yet, location semantic information may disclose sensitive information about mobile users. Thus, we propose *CloSed*, a semantic-awareness privacy preservation model to protect users' privacy from violation. We design *cloaked sets* that should cover different semantic regions of road networks as well as satisfy quality of service (QoS). As the problem of calculating the optimal cloaked set is NP-hard, we design a greedy algorithm that balances QoS and privacy requirements. Extensive experiments evaluations demonstrate the efficiency and effectiveness of our proposed algorithm in providing privacy guarantees on large real-world datasets.

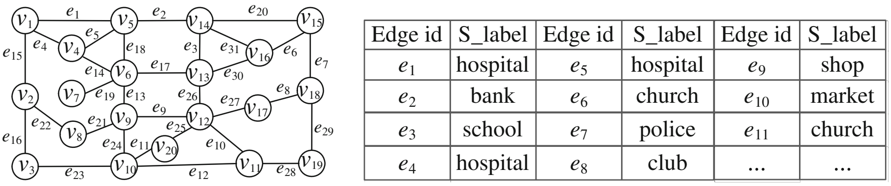
1 Introduction

Advances in positioning technologies along with the tremendous popularity of mobile devices have resulted in the widespread adoption of location-based services (LBS) on road networks. Examples of these applications include navigation services, identification of points of interest (POIs), and receiving traffic alerts or notifications. While enjoying the convenience of LBS, however, users also face significant risks of privacy leakage [23]. Adversaries can exploit user location information for such nefarious purposes as stalking, spamming, and inferring political/religious affiliations or alternative lifestyles.

The state-of-the-art for protecting the positions of LBS users over road networks is based on the model of *segment l-diversity* [3,26]. In this model, the actual user position is replaced by a set of *segments*, i.e., edges in a road network, and the number of segments indicates the degree of diversity. Though this solution can satisfy most privacy-preservation requirements, it cannot resist the types of *semantic homogeneity attacks* illustrated by the following example.

Example 1. Consider a scenario in Fig. 1. A patient, named Bob, asks for services through his GPS-enabled mobile phone from road e_1. To prevent Bob's location from leakage, the approach based on segment l-diversity cloaks Bob's walking

S.B. Navathe et al. (Eds.): DASFAA 2016, Part II, LNCS 9643, pp. 314–331, 2016.
DOI: 10.1007/978-3-319-32049-6_20



(a) Simplified Road Network

Edge id	S_label	Edge id	S_label	Edge id	S_label
e_1	hospital	e_5	hospital	e_9	shop
e_2	bank	e_6	church	e_{10}	market
e_3	school	e_7	police	e_{11}	church
e_4	hospital	e_8	club	...	...

(b) Mapping Table

Fig. 1. Semantic road network

road with other nearby roads [3,26]. In our example, we assume that $l = 3$, and then the *cloaked set* may be $\{e_1, e_4, e_5\}$. Unfortunately, it is easy for an adversary to infer that Bob is in the hospital, since all roads e_1, e_4, and e_5 have the homogeneous semantics, namely hospital. Hence, even though Bob's location is seemingly obfuscated, it can be inferred by a semantic homogeneity attack.

Although some techniques have been proposed to resist semantic homogeneity attacks [6,16,28] over road networks, they have different limitations. The solutions proposed in [16] have a deterministic property for its cloaked areas, so it is subject to *reverse engineering attacks*, e.g., a *replay privacy attack* [26]. The offline approaches of [6,28] cannot support the privacy requirement update due to their cloaked sets being generated a priori for a particular privacy requirement. Changes in mobile users' privacy requirements are frequent, thus seriously threatening the applicability of these approaches. The work [28] also presents an online cloaking algorithm to protect sensitive semantic information. However, the cloaking cost is expensive due to considering velocity-based linkage attacks which is out of our scope.

To solve the problem highlighted in Example 1 and overcome the drawbacks of methods above, we propose *CloSed*, a semantic-aware privacy-preserving model, and design a new solution to protect the location privacy of mobile users on road networks against semantic homogeneity attacks. We illustrate the model using Example 1 as a running example throughout the paper. In our approach, instead of cloaking Bob's current road e_1 with other nearby roads $\{e_4, e_5\}$ of the hospital, CloSed generates a cloaked road set consisting of Bob's current *semantic road* and other nearby semantic roads, e.g., the cloaked set is $\{e_1, e_2, e_3\}$. In this example, the semantic of Bob's current road may be hospital, bank or school. Therefore, the adversary can no longer infer the exact semantic of Bob's location.

Our primary goal is to protect location privacy while guaranteeing the quality of location-based services for snap-shot queries. Our strategy focuses on semantic diversity, which guarantees that it would be difficult to associate a specific user with a specific semantic with a high possibility. It also regards QoS as a critical measure for designing privacy preservation solutions, and supports personalized privacy requirements.

To achieve this goal, in *CloSed*, each mobile user can designate his location privacy requirement as *l-semantic diversity*. That is, rather than l-segment diversity, we focus on the cloaked road set possessing at least l different semantic types. Thus, mobile users can use location-based services without the need to reveal their private location and location semantic information. To implement l-semantic diversity over road networks, our solution, named *EIRank*, consists of two steps: pre-processing and online cloaking. The pre-processing phase reduces the cloaked space by roughly grouping roads into different clusters, called *buckets*, while the online cloaking phase generates the desirable cloaked set in each bucket. In the pre-processing phase, to guarantee the efficient generation of buckets, structure and semantic information should be integrated. The major challenge lies in how to combine them together seamlessly. We propose the concepts of *edge interaction (EI) network* and *virtual nodes* to embed structure and semantic information together.

In designing our solution to the problem of privacy preservation for mobile users on road networks, we thus make a number of contributions, as follows:

- CloSed's semantic-aware model extends existing solutions by offering protection against semantic homogeneity attacks.
- CloSed's approximation algorithm EIRank naturally balances privacy requirements with QoS.
- EIRank integrates structure and semantic information seamlessly by transforming road networks into EI networks and leveraging the idea of virtual nodes.
- EIRank's evaluation over large real-world datasets demonstrate its efficiency at cloaking the optimal road set and guaranteeing that exact location and semantic information cannot be leaked.

The remainder of the paper is organized as follows. We introduce our road network and privacy preservation model in Sect. 2. In Sects. 3 and 4, we describe the technique and algorithm for location anonymization. In Sect. 5, we report extensive experimental results. We briefly review the related work in Sect. 6. Finally, Sect. 7 concludes the paper.

2 Problem Definition

We begin this section by presenting the road network model. We then formally define our privacy preservation model, and the goals of the associated techniques. Finally, we present our algorithm framework.

2.1 Road Network

Definition 1 (Semantic Road Network). *A road network is modeled as an undirected graph $G = (V, E, \xi)$ with a node set V and an edge set E, such that (i) a node $v \in V$ represents a road intersection or a location (e.g., hospital); (ii) an edge $e = (u, v) \in E$, also called a **segment**, connects two nodes u and v; and (iii) L represents a semantic function, i.e., for each edge $e \in E$, $\xi(e)$ is the sensitive semantic label of segment e.*

Example 2 (*Semantic Road Network*). Fig. 1(a) shows an example of a semantic road network, in which each edge is associated with a semantic ID. Figure 1(b) gives semantic labels corresponding to the IDs. Nodes v_1, v_4 and v_5 in Fig. 1(a) are different buildings within the same hospital. Edges e_1, e_4 and e_5 connecting these three nodes would then have the same sensitive semantic label "hospital". Thus, the area represented by the triangle (v_1, v_4 and v_5) would indicate the hospital.

2.2 Privacy Preservation Model

To resist against semantic homogeneity attacks as given in the introduction, we propose the following privacy preservation model.

Definition 2 *(Cloaked Set l-semantic diversity (***CloSed***)). A user's published cloaked segment set S_c=$\{e_1, e_2, ..., e_i, ...\}$ is said to have l-semantic diversity, if (i) S_c contains at least l different types of semantic labels, i.e., $|\xi(S_c)| = |\bigcup_{\forall e \epsilon S_c} \xi(e)| \geq l$, and (ii) the possibility of distinguishing a user's semantic label among other semantic labels in S_c does not exceed $\frac{1}{l}$.*

Returning to Example 1, to achieve *CloSed*, Bob's published cloaked segment set can be $S_c = \{e_1, e_2, e_3\}$ or $S_c = \{e_1, e_2, e_6\}$. The cloaked segment set $S_c = \{e_1, e_2, e_3\}$ indicates that Bob may be in a hospital, a bank or a school. The cloaked segment set $S_c = \{e_1, e_2, e_6\}$ indicates that Bob may be in a hospital, a bank or a church.

Architecture. Similar to existing works [2,4,10,13,20,26], we adopt the classical centralized privacy-preserving architecture. In this architecture, the location anonymizer is a trusted entity that lies between mobile users and service providers (SP), and performs location anonymization and result filtering operations. More specifically, the location anonymizer first removes identity labels (e.g., id) and transforms the original query with an accurate location to another query with a cloaked set, according to users' privacy requirements. Next, SP computes and forwards the produced candidate results to the location anonymizer. At last, the location anonymizer extracts the exact answers from the candidate results by adequately filtering false hit information.

Based on the processing framework, in the CloSed model, a mobile user should specify his/her privacy profile (l, σ_t), where l indicates l-semantic diversity and σ_t is the maximum temporal tolerance to guarantee QoS. To preserve user privacy, we identify an important property that is sufficient for a cloaking technique.

Definition 3 *(*Segment Oblivious*). For a query user u in segment e, given u's profile (l,σ_t), his/her published cloaked segment set S_c satisfies the segment oblivious property iff (i) S_c contains at least l semantic labels; (ii) $e \in S_c$; and (iii) a query initiating in any segment of the cloaked set S_c will return the same cloaked set S_c as the cloaked set for the given l.*

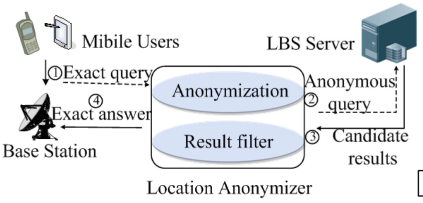


Fig. 2. Privacy-preserving architecture

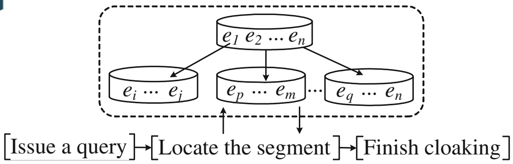


Fig. 3. Algorithmic framework

From Definition 3, it can be shown that any solution to Definition 2 satisfies the following theorem.

Theorem 1. *A cloaking technique for a road network G_r can achieve l-semantic diversity, if every cloaked set S_c published in G_r satisfies the* segment oblivious *property.*

Proof. According to Definitions 2 and 3, it is obviously to reach Theorem 1.

In addition to the preservation of cloaked set l-semantic diversity, the other objectives of our cloaking technique are that: (1) The cloaked set should not reveal the exact segment of any user; and (2) The cloaking technique should not compromise the QoS.

2.3 Algorithmic Framework of Anonymization

To achieve privacy preservation, in the location anonymizer (Fig. 2), the technique employed needs to blur the exact active segment of each mobile user to a cloaked set that satisfies user's privacy profile. A segment is marked as active segment if it is associated with at least one query.

To meet the requirement of privacy (i.e., Theorem 1) and achieve high QoS, our anonymization algorithm consists of two stages: an offline pre-processing phase and an online cloaking phase, as shown in Fig. 3. In the offline pre-processing phase, we allocate all segments of a road network to different buckets, so that we can perform anonymization in one bucket rather than search the entire road network in the cloaking process. In the online cloaking phase, we locate the buckets of active segments and anonymize segments based on user privacy profiles.

3 Segment Allocation

This section presents the offline pre-processing phase as introduced in the algorithmic framework. Specifically, the segments of a road network are allocated to different buckets according to users' privacy requirements. To achieve most user privacy requirements, we make the following observation.

Observation 1: The location semantic privacy requirements L of user privacy profiles follow a Gaussian distribution $L \sim N(\mu, \sigma^2)$, i.e., most user privacy requirements fall in the middle range, and fewer have higher privacy requirements. The parameter μ is the mean of the distribution, and the parameter σ is its standard deviation.

It follows that we can leverage the 3σ rule, also known as the 68-95-99.7 empirical rule, which states that about 99.7 % of values drawn from a Gaussian distribution are within three standard deviations from the mean. We accordingly set the semantic number of a bucket to $\mu + 3\sigma$ to satisfy all user location anonymization in one bucket. Definition 4 states the goal of segment allocation.

Definition 4 *(*Segment Allocation*). The segments of a road network $G = (V, E, \xi)$ are allocated to p buckets, $G_1, G_2, ...G_p$, where $p > 1$ and $G_i = (V_i, E_i, \xi_i)$, such that $V = \bigcup_{1 \leq i \leq p} V_i$, $E = \bigcup_{1 \leq i \leq p} E_i$, $\xi(E) = \bigcup_{1 \leq i \leq p} \xi_i(E_i)$, and the following conditions are satisfied.*
(i) The segments of all buckets are disjoint, i.e., $\forall 1 \leq i, j \leq p$, $E_i \cap E_j = \phi$.
(ii) The semantic number of a bucket must exceed the threshold $\mu + 3\sigma$, i.e., $|\xi(E_i)| = |\bigcup_{\forall e \epsilon E_i} \xi(e)| \geq \mu + 3\sigma$.

In addition to protecting the location privacy of mobile users, the cloaking algorithm should not compromise QoS, which mainly depends on communication cost. We use the number of candidate results to measure communication cost, which is formulated in Definition 5. Without loss of generality, we focus our attention on k-nearest neighbors (kNN) queries.

Definition 5 *(*LBS Server Processing*). [26] For a query q with associated anonymous segment set S_c, the candidate results of q consists of two parts: (1) the POIs on the segments of S_c, and (2) the results as q is issued on the boundary nodes of the boundary set S_{bn}, where the boundary set is a set of nodes whose some connected edges are not included in S_c. Formally, $CR(q, S_c) = (\bigcup_{s \in S_c} O(q, s)) \bigcup (\bigcup_{v \in S_{bn}} O(q, v))$*

Based on this query processing model, it can be seen that the communication cost $CR(q, S_c)$ is significantly influenced by parameters $|S_c|$ and $|S_{bn}|$. However, reducing $|S_c|$ and $|S_{bn}|$ imposes conflicting demands on $CR(q, S_c)$. This is explained by the fact that segments that are near each other tend to possess similar semantic labels.

For a user privacy profile, our objective is to find the optimal cloaked set which is minimized in terms of communication cost, while satisfying l-sematic diversity. In summary, our problem is equivalent to the following optimization problem:

Minimize $CR(q, S_c)$, subject to $|\xi(S_c)| = |\bigcup_{\forall e \epsilon S_c} \xi(e)| \geq l$.

According to the paper [27], the problem of computing an optimal cloaked set is NP-hard.

Solution. Based on above analysis, we propose a greedy solution called *EIRank*. Intuitively, cloaking adjacent segments with different semantic labels provides a compact structure and semantic preference simultaneously. In other words, we prefer cloaking the segments exhibiting *structure similarity* and *semantic label dissimilarity.* To measure the similarity of linkage structures and the dissimilarity of semantic labels, we introduce two scoring functions: $S(n_1, n_2)$ and $\mathit{Diff}(e_p.\varphi, e_q.\varphi)$, respectively.

In many applications, objects are considered similar if they are related to similar objects. Based on this intuition, we adopt a general similarity metric called SimRank to measure the similarity of linkage structures. SimRank is calculated by Eq. 1.

$$S(n_1, n_2) = \begin{cases} 1 & n_1 = n_2 \\ \frac{C}{|I_{n_1}||I_{n_2}|} \sum\limits_{j \in I_{n_2}} \sum\limits_{i \in I_{n_1}} S(i, j) & n_1 \neq n_2 \end{cases} \tag{1}$$

where C refers to as a decay factor, is a constant between 0 and 1, and I_n represents the set of neighbors of n. Note that Eq. 1 is defined to be 0 when $I_{n_1} = \emptyset$ or $I_{n_2} = \emptyset$.

To evaluate the dissimilarity of semantic labels of segments, we use the normalized edit distance. In this case, the dissimilarity of semantic labels $\mathit{Diff}(e_p.\varphi, e_q.\varphi)$ is measured by the edit distance between the semantic labels with regard to the length of the semantic label. The edit distance, $\mathit{Edit}(e_p.\varphi, e_q.\varphi)$, between two semantic labels, $e_p.\varphi$ and $e_q.\varphi$, is defined as the minimum number of basic operations required to transform one semantic label into the other. In this paper, the basic operations are defined as insertion, deletion and substitution of symbols, which is formalized as follows.

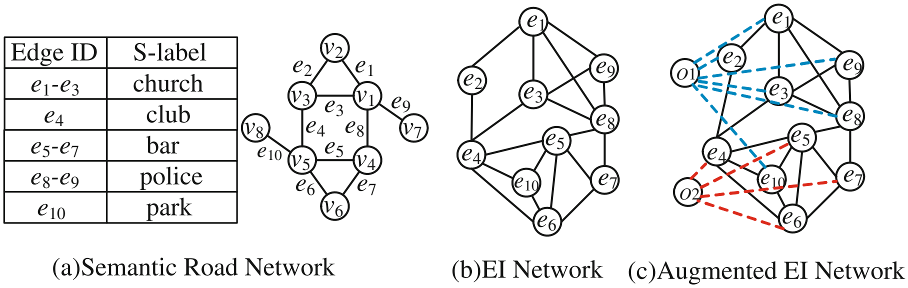
Let $T_s(b|a)$ represents the substitution of symbol a by symbol b ($a \neq b$), $T_i(a)$ represents the insertion of symbol a, and $T_d(a)$ represents the deletion of symbol a. Then,

$$\mathit{Diff}(e_p.\varphi, e_q.\varphi) = \frac{\mathit{Edit}(e_p.\varphi, e_q.\varphi)}{\mathit{Max}(|e_p.\varphi|, |e_q.\varphi|)} \tag{2}$$

where $e_p.\varphi$ denotes the label function of e_p, and $\mathit{Max}(|e_p.\varphi|, |e_q.\varphi|)$ represents a function that computes the larger length of the two labels $e_p.\varphi$ and $e_q.\varphi$.

To combine linkage structure and segment semantic information for segment allocation, we propose a solution, called *EIRank*, for simultaneously representing link-based similarity and semantic-based dissimilarity. Our solution consists of four steps: *EI network construction*, *label clustering*, *Augmented EI Network Construction* and *segment allocation.* Next, we will discuss each step in details.

EI Network Construction. For simplicity, we assume that the semantic label of each edge is unique. To integrate linkage structure and segment semantic, the semantic road network is transformed into an edge interaction (EI) network. An EI network node, called *e-node*, represents an edge in the original semantic road network, and two e-nodes are adjacent if their corresponding edges share a common node in the original semantic road network. The labels of e-nodes in the EI network are given by the semantic labels of the corresponding edges in the road network. For example, edges e_1 and e_2 share a common node v_2 in the semantic road network (Fig. 4(a)), and thus e-nodes e_1 and e_2 are linked together in the EI network (Fig. 4(b)). Since the segment id itself represents the semantic label of the segment, we do not mark the labels of the e-nodes anymore in the EI network.



Edge ID	S-label
e_1-e_3	church
e_4	club
e_5-e_7	bar
e_8-e_9	police
e_{10}	park

(a)Semantic Road Network (b)EI Network (c)Augmented EI Network

Fig. 4. Example of EIRank strategy

Label Clustering. The problem of computing the dissimilarity of two segment labels is equivalently converted into the one of computing the dissimilarity of two e-node labels in the EI network. We use the method mentioned above to achieve this goal. Take the labels of the two e-nodes e_1 and e_4 in Fig. 4(b) as an example. By performing four basic operation $T_s(l|h)$, $T_s(b|r)$, $T_d(c)$ and $T_d(h)$, the label of e-node e_1 is transformed to the label of e-node e_4. Therefore, the dissimilarity of the two e-node labels is $\mathit{Diff}(e_1.\varphi, e_4.\varphi) = \frac{2}{3}$.

Based on the dissimilarity of the labels of e-nodes in the EI network, we perform a generalized k-medians clustering [19] for the labels of the e-nodes in the EI network. The result of label clustering for Fig. 4(b) is $\{church, police, park\}$ and $\{bar, club\}$.

Augmented EI Network Construction. In this step, we create a virtual node for each cluster and connect the e-nodes whose labels are in the same cluster to the virtual node. This new generated network is called *augmented EI network*. The original e-nodes in a label cluster have higher structure similarities by adding the virtual nodes. For example, Fig. 4(c) shows the updated EI network corresponding to Fig. 4(b). Two virtual e-nodes o_1 and o_2 are added to represent the clusters $\{church, police, park\}$ and $\{bar, club\}$, respectively. Then, the e-nodes in the set $\{e_1, e_2, e_3, e_8, e_9, e_{10}\}$ are connected to the virtual node o_1. In the same manner, virtual node o_2 is connected to the e-nodes in the set $\{e_4, e_5, e_6, e_7\}$.

Segment Allocation. As stated above, the segments of a cloaked set needs to have the structure similarity and semantic label dissimilarity. Based on the above steps, the dissimilarity of the original e-node labels has been transformed to the similarity of the linkage structure. This is consistent with the similarity of the

Algorithm 1. Baseline Algorithm

```
Input: Semantic road network $G = (V, E, \xi)$, Bucket scale $N_l$
Output: Buckets $G_1, G_2, ..., G_p$
Transform the $G$ into EI network;
Execute the label clustering for e-nodes;
Compute $S(e_p, e_q)$ for all e-node pairs;
Allocate($e_p, e_q$, $G_S$);
Merge buckets $G_i$ where $|\xi(G_i)| < N_l$;
return non-empty buckets $G_1$,$G_2$,...,$G_p$;
```

linkage structure. Next, we use the function $S(e_p, e_q)$ to measure the similarity for every pair of non-virtual e-nodes.

To compute SimRank efficiently, we adopt the method in [7]. In this case, the similarity of e-nodes is measured by Eq. 3, which states that the similarity of two e-nodes is the expectation of the total time which is the time taken by two random walkers starting from two different nodes to finally meet.

$$S(e_p, e_q) = E(C^{\tau(e_p,e_q)}) \tag{3}$$

Once the similarity has been computed for all e-node pairs, we use the single-linkage hierarchical clustering [24] to perform the segment allocation. The function Allocate (e_p, e_q, G_S) is used to describe this process.

The complete description of our EIRank strategy is given in Algorithm 1.

4 Online Cloaking Phase

In the previous section, we have described the pre-processing phase of our approach. Once partitioned buckets have been obtained, the remaining work is to generate a cloaked set according to a user's online request. Before detailing our cloaking algorithm, we present several index structures, namely *Ordered Locating Index* (*OLI*), semantic-aware order preserving list (*SOPlist*), and cloaked *l*-diverse segment set (*Cloaked l-maplist*), used in the online cloaking.

4.1 Index Structure

Ordered Locating Index. In order to quickly locate the position of a segment in a segment allocation, we design a novel index structure called *OLI* based on the hash table for organizing the segments in order. We keep a record of each entry in the form of (*Seg*,*Sid*,*Cid*,*Pointer*) where Seg is the segment identifier, Sid is the bucket identifier of the segment Seg, Cid is the position identifier of segment Seg in bucket Sid, and Pointer is a pointer to the next entry. We use Eq. 4 to compute the sequence of segment $Seq(e_{i,j})$ in the ordered linked list to obtain the Sid and Cid of $e_{i,j}$. The first three entries of this equation are used to compute the number of segments before segment $e_{i,j}$. Note that $e_{i,j}$ connects the nodes i and j.

$$\begin{aligned} Seq(e_{i,j}) = & \sum_{k=1}^{i-1} degree(k) - |S_{overlap}|_{S_{overlap}=\{e_{lt}\},t<i,l<i} \\ & + |S_{prior}|_{S_{prior}=\{e_{ip}\},i<p<j} + 1 \end{aligned} \tag{4}$$

Set the segment $e_{6,9}$ in Fig. 1 as an example. Using Eq. 4, we compute its segment sequence $Seq(e_{6,9})$ = degree(v_1) + degree(v_2) + degree(v_3) + degree(v_4) + degree(v_5) - $|\{e_{1,2}, e_{1,4}, e_{1,5}, e_{2,3}, e_{4,5}\}|$) + $|\{e_{6,7}\}|$ +1 = $3+3+2+3+4-5+1+1 = 12$. Then, searching for the 12th record in OLI which is shown in Fig. 5, we get Sid = 1 and Cid = 2. We conclude that the segment $e_{6,9}$ is in bucket 1, at position 2.

SOPlist and Cloaked l-maplist. To facilitate the execution of the cloaking algorithms, we also propose two other data structures. *SOPlist* is a 2-semantic diversity index whose objective is to speed up the computation of the cloaked set. Each record of SOPlist is represented as ((seman1, n_1), (seman2, n_2), Pointer), where (seman1, n_1) ((seman2,n_2)) denotes n_1 (n_2) adjacent segments of semantic label seman1 (seman2), while Pointer is a pointer to the next record.

The role of *Cloaked l-maplist* is to record the cloaked sets that have been formed for distinct semantic requirements so far. This is achieved by re-using the mapping between segments and cloaked sets. A basic cell of Cloaked l-maplist is represented as(l_i, *npointer*, *spointer*) and l_i_set, where l_i indicates l_i-semantic diversity, *npointer* and *spointer* are pointers to the next basic cell and l_i_set, respectively, and l_i_set records the last position of each cloaked set with regard to semantic requirement l_i. l_i_set is dynamically maintained to keep track of the current maximum position of cloaked sets of semantic requirement l_i in a bucket.

Example 3 (SOPlist and Cloaked l-maplist). Suppose the content of a bucket is $\{e_{23}, e_{13}, e_{22}, e_{21}, e_{17}, e_4, e_1, e_5, e_{18}, e_{14}, e_{19}\}$. Then, Fig. 5 shows the SOPlist and Cloaked l-maplist corresponding to the bucket.

4.2 The Cloaking Algorithm

We introduce our online cloaking algorithm, which is summarized in Algorithm 2. It mainly uses of the segment oblivious property which is stated in Definition 3.

The algorithm first initiates an empty cloaked set and computes the sequence of specified segments to locate the position of the segment in the segment allocation (lines 1–2). The algorithm then finds the maximum value l_{max} in Cloaked l-maplist and compares the segment location *Cid* in the bucket with l_{max} (line 3). If the value l_{max} is larger than *Cid*, the algorithm simply searches for the Cloaked l-maplist to find the range of the cloaked set (lines 4–5). In this case, it means that the cloaked set has been computed. Otherwise, it is necessary to execute the operations of lines 6–12. Finally, the algorithm searches for the segments in the bucket range from $[x_1, x_2]$, and returns the corresponding cloaked set.

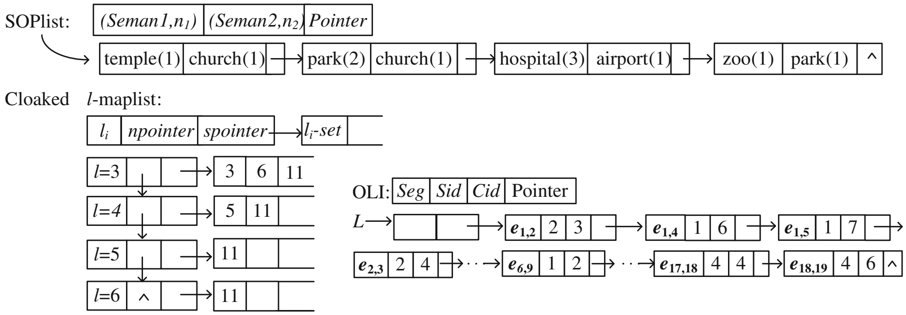


Fig. 5. Index structure

Example 4 (Online Cloaking). Continuing with Example 3, we assume that l_3_set={3}, $l_{max} = 3$, and two users u_1 and u_2 with the same privacy profile (3, 1) located in segment e_{13} and e_{17}, respectively. Since e_{13}.Cid=2< $l_{max} = 3$, we can compute $x_1 = 0, x_2 = 3$ according to Cloaked l-maplist and return S_c ={e_{23}, e_{13}, e_{22}}. Since e_{17}.Cid=5> $l_{max} = 3$, we cannot compute the interval $[x_1, x_2]$ directly. So we continue to cloak from $l_{max} + 1 = 4$ in the SOPlist, and obtain $l_{max} = 6$. When we checks the item zoo(1), we stops traversing in the SOPlist. Then,we can conclude that the residual semantic number exceeds 3. So we can safely set $l_{max} = 6$. As the condition $e_{17}.Cid < l_{max}$ is satisfied, we set $x_1 = 3$, $x_2 = 6$, and obtain S_c={e_{21}, e_{17}, e_4}.

5 Experimental Evaluation

In this section, we evaluate the performance of our proposed location anonymization algorithms through extensive experiments. Our methods are implemented on a machine with CPU Inter(R) Core(TM)i7-2600, 8.00 GB memory, 3.40 GHz frequency, 500 GB hard disk. All programs are coded in C++.

5.1 Experimental Setup

(1) Datasets. We use two real road network datasets[1]: California and Oldenburg road networks. These datasets contain POIs of various categories, e.g., church, hospital, airport, which we used as query objects in our experiment. Table 1 gives the parameters of the two real road networks.

(2) Query Generator. For each real dataset, we randomly pick 2000 query points from the positions of trajectories. To simulate different traffic condition, these trajectories are derived from real trajectories and synthetic trajectories which are generated by a traffic simulator[2]. The parameters of queries are listed in Table 2. In each experiment, we run 2000 queries and report the average result.

[1] http://www.cs.utah.edu/~lifeifei/SpatialDataset.htm.
[2] http://www.fh-oow.de/institute/iapg/personen/brinkhoff.

Algorithm 2. Online Cloaking

```
Input: Location$(x,y)\epsilon e_i$, Privacy profile$(l,\sigma_t)$, OLI OSI, Soplist $SL$, Cloaked
       $l$-maplist $CL$
Output: Cloaked set $S_c$
Initialize $S_c = \Phi$ ;
Compute Seq$(e_i)$ to acquire $Sid_0$,$Cid_0$ in OSI ;
Compute maximum value $l_{max}$ of $l_i$_set where $l_i$=$l$ ;
if $l_{max} \geq Cid_0$ then
    compute interval$(x_1, x_2)$ in Cloaked $l$-maplist $CL$ ;
else
    while $l_{max} < Cid_0$ do
        $l_{old-max} = l_{max}$;
        update $l_{max}$=cloak($l_{max}$+1,$l$, $SL$);
        if $residualsemantic(l_{max}, SL) < l$ then
            update $l_{max}$=end position of Soplist $SL$ ;
        insert $x_2$ into $CL$ ;
    $x_1 = l_{old-max}$, $x_2 = l_{max}$ ;
$S_c$=$\bigcup\{e_k\}_{x_1<e_k.Cid\leq x_2, Seq(e_k).sid=Sid_0}$;
return $S_c$ ;
```

Table 1. Real dataset parameters

Name of dataset	Vertex count	Edge count	Semantic types count	POIs count
OLdenburg (OL) road network	6,105	7,035	10	600
California (CA) road network	21048	21693	62	104,771

(3) Algorithms. We evaluate the following algorithms. (a) EIRank: The algorithm is our proposed solution for protecting location privacy on road networks. (b) SA: This is an algorithm proposed in [17]. To compare with our approach, we modify this solution. That is, we don't consider identity protection (k-anonymity), and are only interested in protecting location and location semantic information. More specially, the algorithm first achieves a *Voronoi-partition graph* from the road network. Then, it determines the initial vertex's Voronic-partition according to the query user's location. Next, it gradually merges neighboring vertex's Voronic-partitions until the semantic requirement is satisfied.

Table 2. Parameters setting

Parameters	Default values	Range
l: semantic diversity	**5**	[2,10]
k: kNN query	**5**	[2,10]
t: semantic type count	**62**	[62,100]

(4) Metrics. In our experiments, we evaluate the following metrics. (a) Cloaking Size: this metric measures the size of a cloaked set. It is defined as the count of the segment that contains in a cloaked set. (b) Relative Semantic level: this metric measures the achieved semantic diversity l' for the cloaking algorithm normalized by the user specified sematic diversity level l, i.e., $\frac{l'}{l}$. (c) Cloaking Time: the cloaking time is used to measure the runtime of the cloaking algorithm.

Besides, we also use the following two metrics to measure QoS. (c) Query Time (PT): this metric is measured by the execution time of processing a query at the server side. (d) Communication Cost (CC): we use the size of the candidate results set to measure the communication cost.

5.2 Experimental Results

In the first three experiments, we examine the efficiency of our cloaking algorithms. In the last two experiments, we examine PT and CC.

Cloaking Size. Figure 6(a) shows the effect of varying semantic diversity on the cloaking size. From the results, we can observe that with the increase of semantic diversity, the cloaking sizes all increase. In addition, the cloaking size of SA is always larger than that of EIRank. The main reason is that the cloaking strategies of the two algorithms are different. EIRank performs segment-based perturbation, which stops just after obtaining user specified semantic requirement. In contrast, SA performs vertex Voronoi-based perturbation. Based on this difference, a cloaked set of SA contains more segments than that of EIRank.

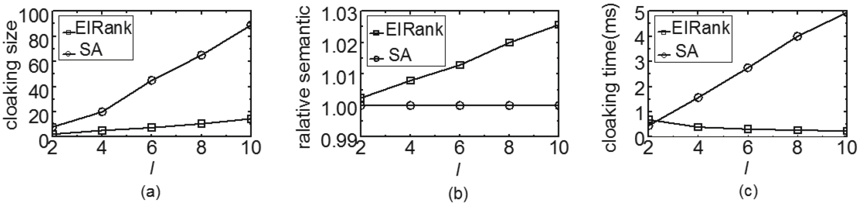


Fig. 6. The efficiency of the cloaking algorithms on the California road network

Relative Semantic Level. Figure 6(b) shows the relative semantic level with regard to semantic diversity. It can be seen that as the semantic diversity increases, the relative semantic level of SA remains unchanged and that of *EIRank* increases. This is because the semantic number of a cloaked set exactly equals to the user-defined semantic diversity for SA algorithm. To resist reverse engineering attacks, the lastest cloaked set of each bucket contains more than l semantics for EIRank.

Cloaking Time. Figure 6(c) shows the impact of varying semantic diversity on the cloaking time for the two algorithms. From the figure, we observe that with the increase of semantic diversity, the time cost of EIRank drops significantly and the time cost of SA increases dramatically. It also can be seen that the cloaking time cost of EIRank is always less than that of SA.

A large semantic diversity l results in a relative large cloaked set. For the other segments other than query segment, the cloaked sets are generated by searching the *Cloaked l-maplist* directly. As we do not need to reconstruct the cloaked sets, which greatly decreases the cloaking time. In contrast, each cloaked set of SA is generated completely dependently. With the increase of semantic diversity, SA needs to search more vertex's Voronoi-partition to achieve the cloaked set, which increases the cloaking time. As the cloaked set of SA is larger than that of EIRank, EIRank runs faster than SA.

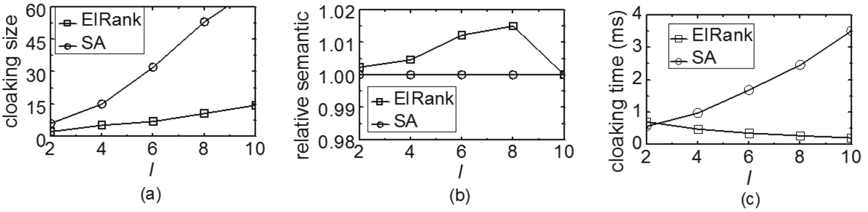


Fig. 7. The efficiency of the cloaking algorithms on the Oldenburg road network

Figure 7 shows the performance of the cloaking algorithms on the Oldenburg road network. We observe that the trendy is the same as that of California road network. Based on this fact, in the following experiments, we just show the performance of the algorithms on the California road network.

Query Processing Cost. Figure 8(a) illustrates the query time of the two algorithms with different values of semantic diversity. From the results, it is clear that the query time all increases as the semantic diversity increases. Furthermore, the query processing of SA run quite slowly in comparison to EIRank as expected. The results above are reasonable, the query time mainly depends on *cloaking size*. For the same semantic diversity, cloaking size of EIRank is smaller than that of SA. The cloaked set becomes large for a big semantic diversity, and hence the query time increases.

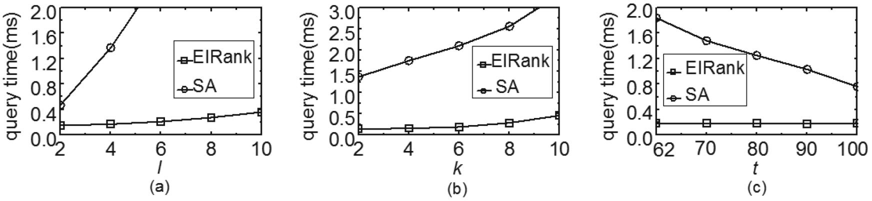


Fig. 8. Query time vs parameters l, k and t.

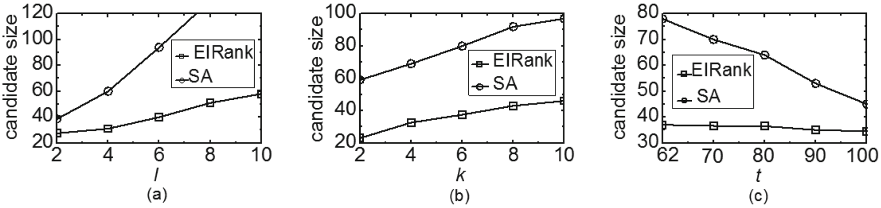


Fig. 9. Communication cost vs parameters l, k and t

Figure 8(b) shows the effect of varying k on the query time. It can be seen that with the increase of k, the query costs of two algorithms increase. We also observe that the algorithm EIRank outperforms SA in most cases. The reasons are as follows. On the one hand, based on our query processing model, a larger k needs to search more segments to acquire the k-nearest neighbors for boundary nodes. On the other hand, the cloaked set size of EIRank is far smaller than that of SA.

Figure 8(c) shows the effect of semantic type count t on query time. As observed, with the semantic type count increases, the query time of two algorithms degrades. We also notice that the parameter t has stronger influence on SA algorithms than on EIRank algorithms.

These phenomena are explained by the following facts: (1) With the increase of semantic types, a part of semantic types are replaced by smaller granularity semantic types. As cloaking of SA algorithm is based on vertex's Voronoi-partition, the size of cloaked sets is smaller than before. and (2) The parameter t has little impact on the cloaked set size of EIRank, and hence it almost have no effect on query time of EIRank.

Communication Cost. Figure 9 shows the impact of different parameters on communication cost. As mentioned above, we measure communication cost indirectly in terms of the size of the candidate results sets, since each result set must be transmitted from server to location annoymizer. From these graphs, we can see that the trend is the same as for query processing cost, and can be explained in a similar way.

6 Related Work

Our work relates to two main streams of research, concerning location privacy and location semantics, respectively.

Location privacy. Location anonymization has attracted much interest as a solution to protect user location privacy in LBS. It mainly makes use of location obfuscation techniques to hide an user's exact location. Examples include space transformation [2,9,25], fake location [14,29], mix-zones [21], and spatial cloaking [1,5,8,11–13,20]. Among various anonymization techniques, spatial cloaking is the prominent. It enlarges an user's exact location to a cloaked region until some privacy conditions are satisfied, such as k-anonymity [8]. Unfortunately, most existing cloaking techniques are no longer applicable in road networks, because the area granularity of measurement tends to fail.

Recently, there exists several research on location privacy over road networks [3,15,16,18,26]. The most famous technique is based on the model of segment l-diversity [3,26]. As mentioned above, this solution cannot prevent the location semantic information leakage.

Location Semantics. In generally, the sensitive information is disclosed using two kinds of published information: query semantics [22,27] and location semantics. In the first case, it means that the query contents issued from a cloaked set are at least l different types. Our paper concentrates on protecting the sensitive information using location semantics over road networks.

Location l-diversity is first introduced in [1], However, it doesnt distinguish the place type. Lee et al. [16] proposes mining the place semantics using Earth Mover's Distance to avoid location semantic leakages, but it does not consider the road networks environment. Yigitoglu et al. [28] extends the *semantic location cloaking model* [6] to protect semantic location in urban settings. Due to the cloaked sets being generated a priori for a particular privacy requirement, this approach cannot support the privacy requirement updating. As the limitations mentioned above, we don't make comparison with them. Li et al. [17] solves the location semantic leakages in road networks based on the vertex Voronoi-partition. Unfortunately, this solution is subject to *reverse engineering attacks*. Our solution overcomes these drawbacks.

7 Conclusion

In this paper, we propose a semantic-aware privacy preservation model named *CloSed* to preserve user privacy on road networks. In our model, the cloaked set provides semantic protection without compromising QoS. To achieve this goal, we design an advanced algorithm to balance the privacy requirement and QoS. Extensive experiment evaluations show the efficiency and effectiveness of our proposed algorithms on large-scale real datasets.

Acknowledgments. This research is supported by the National Natural Science Foundation of China under Grant Nos. 61572119, 61173029, 61332006, 61332014, 61328202, 61502317 and U1401256.

References

1. Bamba, B., Liu, L., Pesti, P., Wang, T.: Supporting anonymous location queries in mobile environments with privacygrid. In: WWW, pp. 237–246. ACM (2008)
2. Chor, B., Kushilevitz, E., Goldreich, O., Sudan, M.: Private information retrieval. IEEE Symp. Found. Comput. Sci. **45**(6), 41–50 (1998)
3. Chow, C., Mokbel, M.F., Bao, J., Liu, X.: Query-aware location anonymization for road networks. GeoInformatica **15**(3), 571–607 (2011)

4. Chow, C., Mokbel, M., Aref, W.: Casper*: query processing for location services without compromising privacy. Trans. Database Syst. (TODS) **34**(4), 24 (2009)
5. Chow, C., Mokbel, M., Liu, X.: A peer-to-peer spatial cloaking algorithm for anonymous location-based services. In: GIS, pp. 171–178 (2006)
6. Damiani, M., Silvestri, C., Bertino, E.: Fine-grained cloaking of sensitive positions in location-sharing applications. Pervasive Comput. **10**(4), 64–72 (2011)
7. Fogaras, D., Rácz, B.: A scalable randomized method to compute link-based similarity rank on the web graph. In: Lindner, W., Fischer, F., Türker, C., Tzitzikas, Y., Vakali, A.I. (eds.) EDBT 2004. LNCS, vol. 3268, pp. 557–567. Springer, Heidelberg (2004)
8. Gedik, B., Liu, L.: Location privacy in mobile systems: a personalized anonymization model. In: Distributed Computing Systems, pp. 620–629. IEEE (2005)
9. Ghinita, G., Kalnis, P., Khoshgozaran, A., Shahabi, C., Tan, K.: Private queries in location based services: anonymizers are not necessary. In: SIGMOD, pp. 121–132. ACM (2008)
10. Ghinita, G., Zhao, K., Papadias, D., Kalnis, P.: A reciprocal framework for spatial k-anonymity. Inf. Syst. **35**(3), 299–314 (2010)
11. Gruteser, M., Grunwald, D.: Anonymous usage of location-based services through spatial and temporal cloaking. In: Proceedings of the 1st International Conference on Mobile Systems, Applications and Services, pp. 31–42. ACM (2003)
12. Hu, H., Xu, J.: Non-exposure location anonymity. In: ICDE, pp. 1120–1131. IEEE (2009)
13. Kalnis, P., Ghinita, G., Mouratidis, K., Papadias, D.: Preventing location-based identity inference in anonymous spatial queries. TKDE **19**(12), 1719–1733 (2007)
14. Kido, H., Yanagisawa, Y., Satoh, T.: An anonymous communication technique using dummies for location-based services. In: ICPS, pp. 88–97 (2005)
15. Ku, W., Zimmermann, R., Peng, W., Shroff, S.: Privacy protected query processing on spatial networks. In: Data Engineering Workshop, pp. 215–220 (2007)
16. Lee, B., Oh, J., Yu, H., Kim, J.: Protecting location privacy using location semantics. In: SIGKDD (2011)
17. Li, M., Qin, Z., Wang, C.: Sensitive semantics-aware personality cloaking on road-network environment. Int. J. Secur. **8**(1), 133–146 (2014)
18. Li, P., Peng, W., Wang, T., Ku, W., Xu, J., Hamilton, J., et al.: A cloaking algorithm based on spatial networks for location privacy. In: Sensor Networks, Ubiquitous and Trustworthy Computing, pp. 90–97 (2008)
19. Martłnez-Hinarejos, C.D., Juan, A., Casacuberta, F.: Generalized k-medians clustering for strings. In: Perales, F.J., Campilho, A.J.C., de la Blanca, N.P., Sanfeliu, A. (eds.) IbPRIA 2003. LNCS, vol. 2652, pp. 502–509. Springer, Heidelberg (2003)
20. Mokbel, M., Chow, C., Aref, W.: The new casper: query processing for location services without compromising privacy. In: VLDB, pp. 763–774. VLDB Endowment (2006)
21. Palanisamy, B., Liu, L.: Mobimix: protecting location privacy with mix-zones over road networks. In: ICDE, pp. 494–505. IEEE (2011)
22. Pan, X., Wu, L., Piao, C., Xu, X.: P^3RN:personalized privacy protection using query semantics over road networks. In: Li, F., Li, G., Hwang, S., Yao, B., Zhang, Z. (eds.) WAIM 2014. LNCS, vol. 8485, pp. 323–335. Springer, Heidelberg (2014)
23. Pedreschi, D., Bonchi, F., Turini, F., Verykios, V.S., Atzori, M., Malin, B., Moelans, B., Saygin, Y.: Privacy protection: regulations and technologies, opportunities and threats. In: Giannotti, F., Pedreschi, D. (eds.) Mobility, Data Mining and Privacy, pp. 101–119. Springer, Heidelberg (2008)

24. Sibson, R.: Slink: an optimally efficient algorithm for the single-link cluster method. Comput. J. **16**(1), 30–34 (1973)
25. Stavros, P., Spiridon, B., Dimitri, P.: Nearest neighbor search with strong location privacy. PVLDB **3**, 619–629 (2010)
26. Wang, T., Liu, L.: Privacy-aware mobile services over road networks. PVLDB **2**(1), 1042–1053 (2009)
27. Xiao, Z., Xu, J., Meng, X.: p-Sensitivity: a semantic privacy-protection model for location-based services. In: MDMW 2008, pp. 47–54. IEEE (2008)
28. Yigitoglu, E., Damiani, M., Abul, O., Silvestri, C.: Privacy-preserving sharing of sensitive semantic locations under road-network constraints. In: MDM, pp. 186–195. IEEE (2012)
29. Yiu, M., Jensen, C., Huang, X., Lu, H.: Spacetwist: managing the trade-offs among location privacy, query performance, and query accuracy in mobile services. In: ICDE, pp. 366–375. IEEE (2008)