ATLANTIS PRESS
*Advances in Engineering Research (AER), volume 61*
International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2017)

# Generating Perturbations with Hilbert Curves and Differential Privacy for Location Privacy

Na Wang [a], Haiyang Yu [b]
School of Beijing University of Technology, Beijing 100124, China
[a]M13051182363@163.com, [b]billyukiwi@163.com

**Keywords:** Location-based Services, k-anonymity, Differential privacy, Privacy protection.

**Abstract.** Location privacy protection method of Location-based Services (LBS) system mostly depends on the trusted third party anonymity server. When the attacker has sufficient background knowledge, it is proved that Location privacy can't be adequately protected by k-anonymity based on location obfuscation enforced using cloaking regions. But perturbation-based mechanisms with differential privacy have been proven effective to defend attackers with any background knowledge. In this paper, k-anonymity and differential privacy are used to generate a perturbation, according to that the incremental nearest neighbor query is enforced, so as to achieve the purpose of LBS privacy protection. Experiments are presented to demonstrate how perturbation-based mechanisms provide a well-balanced tradeoff between privacy and service accuracy.

## 1. Introduction

With the development of wireless communication technology and mobile positioning technology, more and more mobile devices have GPS precise positioning function, which makes the location services become more and more popular. LBS refer to information and entertainment services [1] for mobile users based on the geographical location and other information of mobile devices. Typical applications include map-based applications, POIs, coupons or discounts, GPS navigation and location-aware society Network and so on [1]. Users in the easy access to a variety of LBS, will inevitably left in the network a large number of digital traces and service attributes, attached to the digital trail or service attributes on the context to expose the user's personal habits, interests, interpersonal, physical condition and other personal information. Therefore, exposing these personal information to untrustworthy third parties (such as LBS providers) is bound to cause serious privacy concerns. If the adversary collect such information and do fusion analysis, the user's privacy will be out of the question.

Differential privacy is a new privacy definition proposed by Dwork in 2006 for the privacy disclosure of statistical databases [2]. It described that the calculation results for the data set are not sensitive to the change of an element. Therefore, an element of the privacy disclosure risk caused by its accession to the data set is controlled in the acceptable range. The attacker can't observe the results of the calculation to obtain accurate individual information. Differential privacy can solve two shortcomings of the traditional privacy protection model. First of all, the differential privacy model assumes that an attacker can obtain all the elements except the target element, the sum of which can be understood as the maximum knowledge of the attacker can master the background. Secondly, it has a solid mathematical foundation; the privacy protection is strictly defined and provides a quantitative assessment method, so that different parameters of the data set under the protection provided by the level of privacy can be compared. Therefore, the theory of differential privacy has gradually become a hot topic in the field of privacy protection.

Therefore, how to prevent the attacker from using all the available data and combined with the available background knowledge to fully speculate on the privacy of users is to deal with the new situation of LBS privacy protection issues. Many generalization techniques with k-anonymity have been shown to not adequately protect LBS privacy [3], but a perturbation technique with differential privacy guarantees has been shown to be effective against attackers with arbitrary background knowledge [4,5]. Differential privacy, regardless of any possible background knowledge of the

attacker, is currently the most effective technique for general protection against prior knowledge of attackers. So this paper is used to prevent the attacker from inferring external links by using user activity, and reduce the query error rate. Query with a perturbation cannot assume any direct link to a single user risk and avoid the high computational and communication overhead of region queries. Disadvantageously, a query with perturbation results in an inaccurate result set. However, if the perturbation location is fairly close to the true location, the query result may be close enough to the real result set. There must be an inherent tradeoff between the accuracy of the result set involved and the choice of perturbation. This paper focus on the generation of perturbation and analysis later trade-offs.

## 2. Related Work

### 2.1 System Architecture.

The LBS privacy protection architecture proposed in this paper is composed of mobile terminal, trusted anonymity server and LBS server [6], its system architecture is illustrated in Fig.1 Anonymity server is responsible for collecting the exact location information of users and responding to the location update; converting the precise location information of the user to a cloaking region through an anonymous algorithm; returning the candidate result set from the LBS server and selecting the correct query result to the corresponding mobile user.

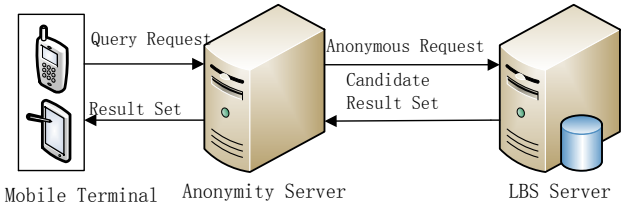



Fig. 1 Overall system architecture of LBS

The central server architecture has the advantages of global information of the user, good privacy protection effect, the smaller communication overhead between the mobile terminal and the anonymity server. But the disadvantage is that the anonymity server needs to be responsible for a large number of users frequent location updates, anonymous processing a query result refinement. So it is the bottleneck of the system to deal with; anonymity server to master all the mobile users and location-related knowledge, when attacked will lead to extremely serious privacy disclosure; the deployment of a large number of users on the trusted anonymity server is very difficult.

The framework of data protection based on differential privacy is an interactive framework, as shown in Fig.2. The data manager designs the corresponding differential privacy function K according to the data application requirements. When the user sends out the query to the data server, result will be returned to user processed by the function K. Database is in anonymity server.

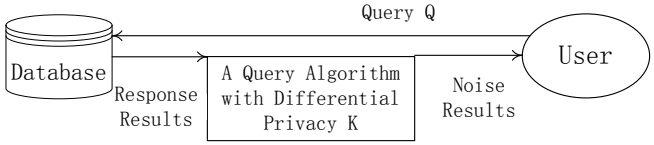



Fig. 2 An interactive framework for differential privacy

### 2.2 Related Definitions.

**Definition 1** LBS query is represented by a four-tuple$< U, t, loc, U_{poi} >$, where U is user ID, t is current time, loc is the location where user commits query at time t, and $U_{poi}$ is service attribute.

LBS privacy consists of query privacy and location privacy. The query privacy is related to the disclosure of $U_{poi}$, and the location privacy is related to disclosure and abuse of loc.

**Definition 2** The Hilbert curve [7] is transformed into a one-dimensional by user's two-dimensional [8, 9]. The moving user is indexed by B+-tree, and the cloaking region is constructed by selecting k users with serial Hilbert value.

Hilbert transform ensure that near two points in two-dimensional is also closer one-dimensional. In Fig.3, Hilbert curve in two-dimensional using $8 \times 8$ partition. In Fig.4, Hilbert curve in two-

dimensional with $4 \times 4$ division. If $U_1$ sends query with anonymous request k = 3, the Hilbert cloak algorithm divides 10 users into 3 buckets. So that each bucket contains at least 3 users, the return result constituting a cloaking region is $\{U_1, U_2, U_3\}$. If user $U_1$ sends an anonymous request with k = 4, the result is $\{U_1, U_2, U_3, U_4\}$.

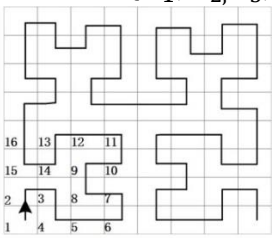
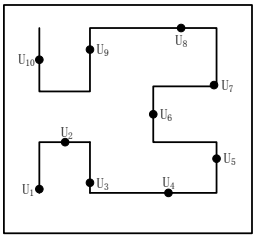



Fig. 3 Hilbert Curve (8×8)



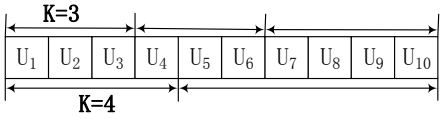




Fig. 4 Hilbert Curve (4×4), k=3 and k=4

**Definition 3** Reciprocity requires algorithms to satisfy: generated cloaking region contains at least k users; for the same k, each user in the cloaking region sends query, the algorithm can return the same cloaking region.

Hilbert cloak is to meet reciprocity, so my privacy protection to select the algorithm to disguise.

**Definition 4** (Differential privacy [10]) A function K gives ε-differential privacy, if for data sets $D_1$ and $D_2$ differing on at most one element, and if the function K can output the result $D \subseteq$ Range (K) arbitrarily on the datasets $D_1$ and $D_2$,

$$\Pr[K(D_1) = D] \leq e^{\varepsilon} \times \Pr[K(D_2) = D]$$

The probability is taken is over the coin tosses of K.

The probability $\Pr[\cdot]$ is controlled function K, which also indicates the risk of privacy disclosure. Parameter ε denotes the degree of privacy protection. Attack model defined by differential privacy is: attacker knows all the sensitive attributes except one element. The sensitive attribute information of this element can still be protected. The ε in the framework of differential privacy determines their similarity to some extent. The smaller ε, the higher similarity, the higher degree of privacy protection.

**Definition 5** (Δf) For $f: D_1 \rightarrow R^d$, the sensitivity of f is $\Delta f = \max\|f(D_1) - f(D_2)\|_p$, for set $D_1$, $D_2$ differing in at most one element [11].

R represents the mapped real numbers space, d represents the query dimension of function f, and p represents the distance $L_p$ used by metric Δf, usually using $L_1$(the sum of absolute values).

**Definition 6** (Laplace) given data set D with function $f: D \rightarrow R^d$ and sensitivity $\Delta f$, the function $K(D) = f(D) +$ Yprovides ε-differential privacy, where $Y \sim Lap(\Delta f/\varepsilon)$ is random noise, and it is a Laplace distribution with the parameter $\Delta f/\varepsilon$ [12].

The definition indicates that the value of the Laplace noise parameter λ depends on Δf and ε. The amount of noise is proportional to Δf and inversely proportional to ε. The greater Δf of function K, the greater noise required.

## 3. Selecting perturbation approach

Anonymity using k-anonymity is mainly divided into two steps: (1) When the user submits a query$< U, t, loc, Upoi >$, anonymity server generates a cloaking region containing at least k users using Hilbert cloak algorithm to meet the reciprocal according to the user loc. (2) The z-score normalization is then performed on the $(x, y)$ coordinates of the k users as a data set. Then the normalized datasets are averaged, and then Laplace noise is added. The average value of $x_p$ and $y_p$after adding noise is used as the perturbation and incremental nearest neighbor query is carried out for LBS server [13].

### 3.1 Hilbert Curves for Cloaking Algorithm.

**Algorithm1** Partitioning Cloaking Region using Hilbert Curves

Input: Query user U, anonymous degree k

Output: Location information of at least k users in cloaking region

**Step1.** Fill space with Hilbert curve, number each cell in direction of the Hilbert curve for hID.

**Step2.** Use B+-tree and hID to index users. Traversing B+-tree, each sorted by Hilbert value node has a rank value (ranks start from 0), and returns the $\mathrm{rank_u}$ of the cell corresponding to U.

**Step3.** According to the $\mathrm{rank_u}$ calculated k-barrel start position and end position:

$\mathrm{start} = \mathrm{rank_u} - (\mathrm{rank_u} \bmod \mathrm{k})$

$\mathrm{end} = \mathrm{start} + \mathrm{k} - 1$

If end> the rank value of the last node:

$\mathrm{start} = \mathrm{rank_u} - (\mathrm{rank_u} \bmod \mathrm{k}) - \mathrm{k}$

$\mathrm{end} =$ the rank value of the last node

**Step4.** Traversing B+-tree returns users location information contained in interval $[\mathrm{start}, \mathrm{end}]$.

Boundary point issue in algorithm: For location on boundary, the principle of upper bound lumped together with bottom and left bound lumped together with right is adopted. For example, if user's location is on border numbered 1 and 2, user is considered to be in cell 1. The user's location is on the boundaries numbered 4 and 5, and the user is considered to be in a cell numbered 5.

Time complexity of constructing cloaking region is $O(\log N + K)$, and time complexity of searching, inserting and deleting is $O(\log N)$. Hilbert cloak algorithm is suitable for a lot of mobile users to update their location frequently, and have different anonymity degree k requirements.

### 3.2 Generating Perturbation.

**Algorithm2** Generating a perturbation

Input: Algorithm1 returns user's location information in cloaking region, privacy parameter ε

Output: The coordinates x and y of perturbation

**Step1.** The users coordinate location information x, y in the cloaking region returned by the Hilbert cloak algorithm is processed:

$X = \{x_1, x_2, \cdots x_k\}$

$Y = \{y_1, y_2, \cdots y_k\}$

**Step2.** Z-score normalization ($z = \frac{x-\mu}{\sigma}$) is performed on the data sets X and Y, respectively.

**Step3.** The handled set X and Y are averaged to obtain values $\bar{x}$, $\bar{y}$. According to definition 5, we obtain Δf of the set X and Y, respectively, and add the Laplace noise to get $x_l$, $y_l$.

**Step4.** The Laplace noise is added according to the definition 6 and the probability of $x_l$ and $y_l$ is similar to the k-1 data of any one of the set X and Y respectively.

**Step5.** Generating $(x_p, y_p)$ according to the inverse of the normalization of step2 using $x_l$and $y_l$.

In the above algorithm, it is assumed that there are exactly k users in the cloaking region for processing. Δf is the key parameter to determine the amount of noise added, and it is a proportional relationship. So the standardization of sets will be smaller Δf. The amount of noise is inversely proportional to ε. In order to obtain the result of adding the noise with high precision, we should reduce the privacy protection level; the larger values of ε will be0.3, 0.5, 0.75, 1 and 1.25.

## 4. Evaluation

### 4.1 Evaluating Perturbation.

Nearness. Distance between perturbation and true location.

Resemblance. Let $O = \{o_1, o_2, \cdots, o_k\}$be the objects retrieved by a KNN-query relative to the true location of user U, and the perturbed location will be $O' = \{o_1', o_2', \cdots, o_k'\}$.The resemblance is the fraction of common objects between O and $O'$given as $\frac{|O \cap O'|}{|O|}$.

Displacement. The displacement is measured as average difference in object distance (from the true location) across the set of mismatched objects, given as

$\begin{cases} \frac{\sum_{i=1}^{k} dist(o_i', u) - \sum_{i=1}^{k} dist(o_1, u)}{|O| - |O \cap O'|}, O \neq O' \\ 0, O = O' \end{cases}$, dist(·) show distance between query user and other users.

### 4.2 Privacy Evaluation.

The experimental environment is Intel (R) Core (TM) i5-4590, 3.30GHZ, 8GRAM, Windows 7 Professional PC. The algorithm is implemented in Java language. The road network was modeled

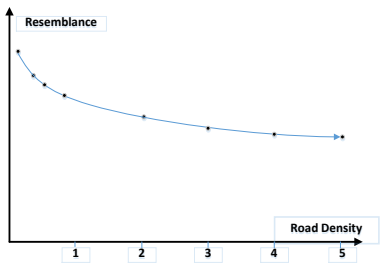
using the Thomas Brink off Road Network Data Generator [14]. The data set used in experiment is the real traffic network of Oldenburg. The area is 26915m * 23572m, the number of vertices is 6105, and the number of edges is 7035.

Experiments randomly selected 5 users in the Oldenburg road network dataset as query users, and repeated the algorithm 30 times for each ε to observe the value of Nearness. It is shown the percentage of generating perturbation falling within 500 m and 1000 m from the actual user location in Table 1. It can be seen from Table 1, the greater ε, the lower effect of privacy protection. We can choose a better value of 0.5 or 0.75.

Table 1. The percentage of the Nearness falling within a certain range with different ε

| ε | ≤ 1000m | ≤ 500m |
|---|---|---|
| 0.3 | 85.33 | 41.33 |
| 0.5 | 92.67 | 52 |
| 0.75 | 94.67 | 62 |
| 1.0 | 97.33 | 69.33 |
| 1.25 | 98 | 72.67 |

When 0.75 is selected, the privacy protection effect is better. Then the values of Resemblance and Displacement are observed for different k values and different road densities. As shown in Fig.5 (a), the change curve of Resemblance under different road densities under k = 5 is selected. The greater road density, the smaller Resemblance, because the greater density, the smaller intercrossing between k users generated by the Hilbert cloak algorithm and k users of the incremental nearest neighbor query. Fig.5 (b) shows Resemblance's curve with the same density and different k. The value of Resemblance becomes larger as the value of k increases.



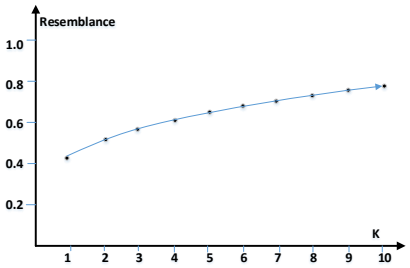
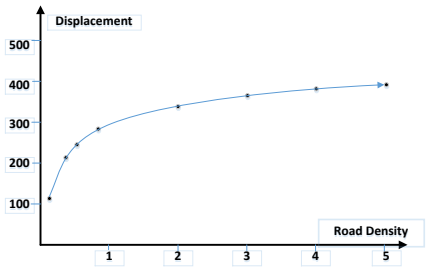




Fig. 5 (a) k=5, different road density Fig. 5 (b) Road density is 2, different K

Fig. 6 (a) shows the Displacement's curve under different road densities at k = 5 selected. The greater road density, the larger Displacement. Fig.6 (b) shows the Displacement's curve with the same density of different values of k value. The value of Displacement becomes larger as the value of k increases, but the float will not be too large.



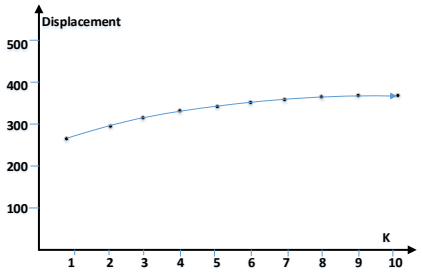




Fig. 6 (a) k=5, different road density Fig. 6 (b) Road density is 2, different k

From the above experimental results, the model can avoid the failure of incremental nearest neighbor query, and according to the user's different privacy requirements, combined with the density of the network to select a suitable value of ε and k, we can have corresponding treatment. This model is not suitable for road density is too small.

## 5. Conclusion and Future Work

The location obfuscation technique is inadequate and cannot resist an attacker with sufficient background knowledge, but sending a query from a perturbation has not a risk of being linked to a

single user. In this paper, the k-anonymity is combined with differential privacy to generate a perturbation, and then the incremental nearest neighbor query is performed according to the perturbation, which can resist the attacks with the maximum background knowledge. But it is better protection of privacy to mobile node around user's location is evenly distributed. When Hilbert cloak algorithm to generate perturbations is used, the incremental nearest neighbor query can be ensured. How to apply this scheme to a continuous query is a future problem to be solved.

## References

[1]. Lee B, Oh J, Yu H, et al. Protecting location privacy using location semantics[C].Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, 2011, p.1289-1297.
[2]. Dwork C. Differential privacy: A survey of results[C].International Conference on Theory and Applications of Models of Computation. Springer Berlin Heidelberg, 2008, p.1-19.
[3]. Shokri R, Troncoso C, Diaz C, et al. unravelling an old cloak: k-anonymity for location privacy[C].Proceedings of the 9th annual ACM workshop on Privacy in the electronic society. New York, 2010, p. 115-118.
[4]. Dewri R. Local differential perturbations: Location privacy under approximate knowledge attackers [J]. IEEE Transactions on Mobile Computing. Vol. 12(2013) No. 12, p. 2360-2372.
[5]. Andrés M E, Bordenabe N E, Chatzikokolakis K, et al. Geo-indistinguishability: Differential privacy for location-based systems[C].Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security. New York, 2013, p.901-914.
[6]. Zhang XJ, Gui XL, WU ZD. Privacy preservation for location-based services: A survey [J]. Ruan Jian Xue Bao/ Journal of Software. Vol. 26(2015) No. 9, p.2373-2395.
[7]. Rakib S S, Zarai Y. Method of providing space filling patterns: US, US8185568 [P]. 2012.
[8]. Kalnis P, Ghinita G, Mouratidis K, et al. Preventing location-based identity inference in anonymous spatial queries[j].IEEE Transactions on Knowledge and Data Engineering. Vol. 19(2007) No. 12, p.1719-1733.
[9]. Takabi H, Joshi J B D, Karimi H A. A collaborative k-anonymity approach for location privacy in location-based services[C]. International Conference on Collaborative Computing: Networking, Applications and Worksharing. COLLABORATECOM, 2009, p1-9.
[10]. Dwork C. A firm foundation for private data analysis [J]. Communications of the ACM. Vol. 54(2011) No. 1, p.86-95.
[11]. Zhang XJ, Meng XF. Differential Privacy in Data Publication and Analysis [J].CHINESE JOURNAL OF COMPUTERS. Vol. 37(2014)No. 4,p.927-949
[12]. Xiong P, Zhu TQ, Wang XF.A Survey on Differential Privacy and Applications [J].CHINESE JOURNAL OF COMPUTERS. Vol. 37(2014)No. 1,p.101-122
[13]. Hu Demin, Zheng Xia. Space Twist-based k-anonymity Incremental Nearest Neighbour Query Algorithm for Location Privacy Protection. Application Research of Computers. Vol. 33(2016)No. 8,p.2402-2405
[14]. Brinkhoff T. A framework for generating network-based moving objects [J]. GeoInformatica. Vol. 6(2002) No. 2, p.153-180.