Protection of Location Privacy using Dummies for Location-based Services

Hidetoshi Kido[†] Yutaka Yanagisawa[††] Tetsuji Satoh[†,††]
†*Graduate School of Information Science and Technology, Osaka University*
††*NTT Communication Science Laboratories, NTT Corporation*
h-kido@ist.osaka-u.ac.jp yutaka@cslab.kecl.ntt.co.jp satoh.tetsuji@lab.ntt.co.jp

Abstract

Recently, highly accurate positioning devices enable us to provide various types of location-based services. On the other hand, because position data obtained by such devices include deeply personal information, protection of location privacy is one of the most significant issues of location-based services. Therefore, we propose a technique to anonymize position data. In our proposed technique, the psrsonal user of a location-based service generates several false position data (dummies) sent to the service provider with the true position data of the user. Because the service provider cannot distinguish the true position data, the user's location privacy is protected. We conducted performance study experiments on our proposed technique using practical trajectory data. As a result of the experiments, we observed that our proposed technique protects the location privacy of users.

1. Introduction

Recently, based on sensing technology developments, we can easily obtain position data using highly accurate positioning devices such as GPS [3]. Such position data is used in various types of location-based services (LBS) [6]. For example, LBSs provide the nearest restaurant information to users, including its location, menu, hours of operation, and so on.

In LBSs, as shown in Figure 1, users can get a service from service providers in return for true position data. After sending data, users cannot delete or modify it. In other words, they cannot prevent service providers from analyzing motion patterns using the stored true position data [1]. To avoid this problem, it is necessary to develop a system to prevent service providers from learning the user's true position data.

We propose a new anonymous communication technique to protect the location privacy of people using LBSs. In our proposed technique, a user sends true position data with several false position data (dummies) to a service provider, who creates a reply message for each received position data. The user only extracts the necessary information from the reply message. In this manner, service providers cannot distinguish true position data from a set of position data if all dummies have temporal consistency. To achieve temporal consistency, we also propose dummy generation algorithms to control dummy generating positions.

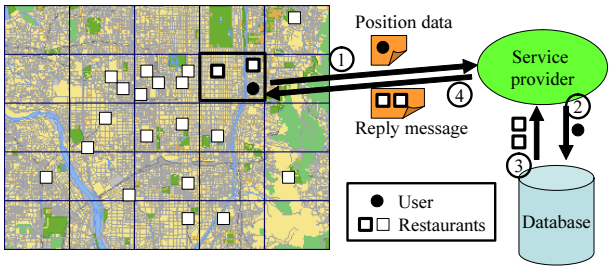

Figure 1. Example of an LBS.

Moreover, we explain our proposed technique by first defining three evaluation functions based on *Anonymity Set* [5] that evaluates anonymity about positions. To assess our technique, we implemented a simulation system. As a result of experiments with actual trajectory data, we conclude that our technique protects the location privacy of people using LBSs.

The rest of our paper is organized as follows. Section 2 describes location privacy and defines evaluation functions based on *Anonymity Set*. Our proposed techniques are presented in Section 4. After that, we describe some performance studies of our proposed technique and offer some conclusions.

2. Location privacy and Anonymity Set

A. Beresford and F. Stajano defined location privacy as "the ability to prevent other parties from learning one's current or past location" [1]. They also argued that a system that can obtain position data invades location privacy.

In this section, we describe location-based services that need to protect location privacy. After that we define some evaluation functions based on *Anonymity Set*.

2.1. Location privacy for LBS

We discuss privacy protection for LBSs shown in Figure 1. In LBSs, the sent message of the user is comprised of

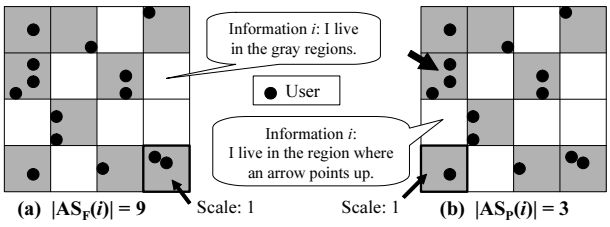

Figure 2. Examples of $AS(i)$.

at least a user ID and true position data. In this paper, we assume that a user ID cannot be connected to the user's privacy information because of pseudonyms. However, even if the user ID is hidden, privacy may be invaded by position data. Here we show an example.

An LBS gives a user information about the times when buses will arrive at the nearest stop in a current vicinity. For example, a person goes to a clinic every week and every time uses this service at his house and the clinic. If such position data are accumulated and analyzed, a staff or a patient of the clinic may learn the person's address.

This example shows that based on position data, a person's location privacy can be invaded. To protect location privacy, it is necessary to anonymize the position data. Before explaining how to anonymize position data, we define location anonymity, which is the degree of anonymity about position data. We consider the following two requirements to enhance location anonymity in an LBS:

- **Ubiquity**
  Ubiquity means that subjects exist in an entire area. When all users live in the same region, service providers can specify users. On the other hand, when users live in various regions, service providers have difficulty specifying users. Thus, ubiquity enhances location anonymity of users in the entire area.

- **Congestion**
  Congestion means that a large number of subjects exists in a region, an idea originated from *k-anonymous* proposed by M. Gruteser and D. Grunwald [4]. Users send position data to service providers in a region. When a large number of users live in the region, service providers have difficulty specifying a user. Thus, congestion enhances location anonymity of users in the region.

Ubiquity guarantees location anonymity of every user. Congestion guarantees location anonymity of local users. Thus, we consider ubiquity more significant than congestion.

2.2. Extended Anonymity Set for LBS

Anonymity Set, a measure that evaluates anonymity, was originated by D. Chaum [2]. A. Pfitzmann and M. Kohntopp define an *Anonymity Set* as "the set of all possible subjects." [5]

Table 1. Location anonymity for Figure 3.

		(a)	(b)	(c)
Ubiquity	**F**	○	×	○
Congestion	**P**	○	○	×

We extend it to a location anonymization technique for LBSs. The extended definition of an *Anonymity Set* is "the set of all subjects determined by information about position." Here, we define it symbolically. First, we define the following symbols.

a	:	a subject
A	:	a set of subjects, $A = \{a_1, a_2, \ldots, a_n\}$
i	:	information about A
I	:	a set of information
$\|A\|$	:	cardinality of A
$\hat{A}$	:	power set of A (2^A)

Each i is represented as a *sentence* that shows information that limits a set belonging to A. For example, assume that A is a set of people. When it provides i to each element included in A who live in Japan, i restricts a set of all people living on earth to a set of all people residing in Japan.

Based on the symbols, we propose a function $AS(i)(i \in I)$ and its cardinality of as follows:

$$\begin{aligned} AS(i) &= 2^A = \hat{A} \quad (AS : I \to \hat{A}) \\ |AS(i)| &= |\hat{A}| \end{aligned}$$

Next, we define the following two functions to evaluate location anonymity.

- $AS_F(i)$: $AS_F(i)$ is a function that returns α_F, which is a set of regions limited by i. $|AS_F(i)|$ denotes the number of α_F, shows the total scale of α_F, or the number of α_F if the regions are of the same scale. $AS_F(i)$ is defined as follows. r_j shows the region.

$$\begin{aligned} A_F &= \{r_1, r_2, \ldots, r_m\} \subset A \quad (\forall r_j \in A_F) \\ |A_F| &= |\{r_1, r_2, \ldots, r_m\}| = m \\ AS_F(i) &= \alpha_F (\in \hat{A}_F) \quad (AS_F : I \to \hat{A}_F) \\ |AS_F(i)| &= |\alpha_F| \end{aligned}$$

  As shown in Figure 2 (a), when assuming that the scale of a region equals 1, $|AS_F(i)|$ equals 9 if i is provided with "I live in the gray regions."

- $AS_P(i)$: $AS_P(i)$ is a function that returns α_P, which is a set of persons limited by i. $|AS_P(i)|$ denotes the number of α_P and shows the number of α_P. $AS_P(i)$ is defined as follows. p_j shows a person.

$$\begin{aligned} A_P &= \{p_1, p_2, \ldots, p_m\} \subset A \quad (\forall p_j \in A_P) \\ |A_P| &= |\{p_1, p_2, \ldots, p_m\}| = m \\ AS_P(i) &= \alpha_P (\in \hat{A}_P) \quad (AS_P : I \to \hat{A}_P) \\ |AS_P(i)| &= |\alpha_P| \end{aligned}$$

  As shown in Figure 2 (b), $|AS_P(i)|$ equals 3 if i is provided with "I live in the region where an arrow points up."

2.3. Quantification of location anonymity

We use function $AS(i)$ to evaluate the location anonymity of LBSs. All areas that provide the service are divided into regions, as shown in Figure 1. The precision

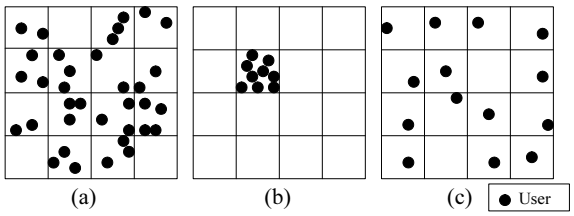

Figure 3. Example of position data.

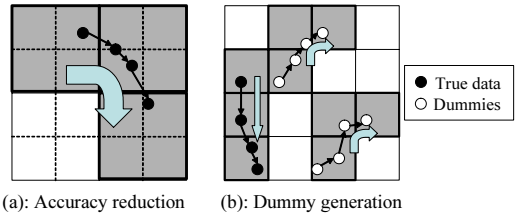

Figure 4. Two anonymous communication techniques for LBS.

of the position data is the same scale as the regions. We must define more two symbols: **F** and **P**. **F**, derived from $|AS_F(i)|$, is represented as a scale of all regions where people live. **P**, derived from $|AS_P(i)|$, is represented as the number of people in a specific region. Now, we describe the relationship between **F** and **P** and two elements: ubiquity and congestion. As explanation, Figure 3 shows some distribution examples of position data. Table 1 shows the degree of location anonymity for the examples.

- Ubiquity—**F**
  F corresponds to ubiquity. In other words, an increase of **F** enhances location anonymity. As shown in Figure 3 and Table 3, when there are many regions where people live, LBSs have ubiquity, and user location anonymity is high.

- Congestion—**P**
  P corresponds to congestion. In other words, an increase of **P** enhances location anonymity. An exception is the regions at $\mathbf{P} = 0$, which are not considered because no people live in that region. As shown in Figure 3 and Table 3, the region where many people live has congestion, and user location anonymity in the region is high.

3. Anonymous communication technique

M. Gruteser and D. Grunwald proposed anonymous usage of a location-based service [4], in which a user sends position data whose precision has beed reduced to the service providers shown in Figure 4 (a). In Figure 4, note that the precision of the position data is shown by the gray regions. The service provider can only learn vague details of the position of users in the usage. Thus, usage enhances location anonymity. However, such usage has a problem: observers can easily comprehend user moves when tracing

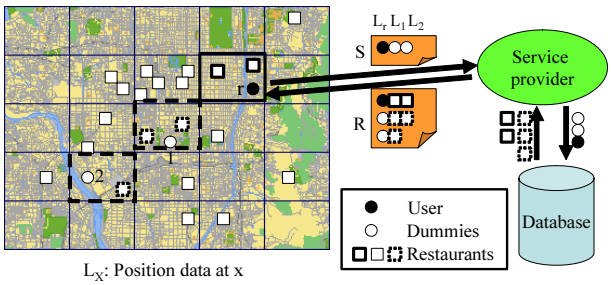

Figure 5. Example of anonymous LBS using our technique.

data for several minutes because the position data chain creates a rough trajectory, as in Figure 4 (a). In this section, we describe the basic idea of our proposed anonymous communication technique that deals with the problem of 3.1. Then in 3.2, we propose dummy generation algorithms.

3.1. Basic idea

To address the traceable problem, we propose a new anonymous communication technique for LBSs in which the user sends position data including noise to the service provider. The noise consists of a set of false position data called 'dummies.' Here, we describe how to use anonymous LBSs with our technique as shown in Figure 5. The service's procedure from beginning to end follows:

1. An LBS user obtains his own position data *r* from a device such as GPS.
2. Dummies are generated at positions *1* and *2*.
3. The user creates a service requiring message **S** that includes position data at *r*, *1*, and *2* and sends **S** to the service provider.
4. The service provider creates a service answered message **R** that responds to receiving all position data and sends **R** to the user.
5. The user receives **R** and only picks up necessary data from **R**.

The user knows the true position data, but the service provider does not. So the service provider cannot distinguish true position data from a set of received position data. In this way, anonymous service is complete.

Figure 4 (b) shows an example of our technique: The user generates two dummies not identical to (a) that can move in different directions from the true position data. Consequently, to comprehend user moves is more difficult. Actually, because the other users simultaneously send position data, the user is more secure.

3.2. Dummy generation algorithm

We describe a dummy generation algorithm from which observers cannot discern true position data and dummies. Generally speaking, the among which each subject can move in a fixed time is limited. If dummies are generated

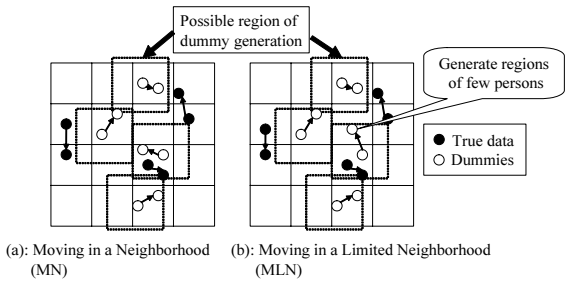

Figure 6. Illustration of two dummy generation algorithms.

randomly, we can easily find differences between true position data and dummies when using LBSs that need position data continuously, such as a road navigation service. If an observer finds true position data, location anonymity is reduced. To avoid this, the dummy must not behave completely different from the true position data. We present the following two dummy generation algorithms to prevent the service provider from finding the true position data.

- Moving in a Neighborhood (MN)
  (Figure 6 (a), Table 2)
  In this algorithm, the next position of the dummy is decided in a neighborhood of the current position of the dummy. The communication device of the user memorizes the previous position of each dummy. Then the device generates dummies around the memory.
- Moving in a Limited Neighborhood (MLN)
  (Figure 6 (b), Table 3)
  In this algorithm, the next position of the dummy is also decided in the neighborhood of the current position of the dummy. However, the next position is limited by the density of the region. This algorithm is adaptable in cases where the communication device of the user can get other user's position data. First, the device of the user gets the other user's position data. Next, the device generates dummies around the memory that are the same as the MN algorithm. Then, if there are many users in the generated region, the device generates the dummy again. The process is repeated several times.

We define $\mathrm{Shift}(\mathbf{P})$ as a measure that evaluates the two algorithms. $\mathrm{Shift}(\mathbf{P})$ expresses a shift of **P** in each region between times `t` and `t+1`. If the number of persons changes greatly in a region, there is a high possibility that the dummy moves strangely compared with the true position data, creating a risk that observers may find the true position data. To that end, it cannot enhance location anonymity. In other words, when $\mathrm{Shift}(\mathbf{P})$ decreases, location anonymity is high.

4. Evaluation

To evaluate our technique shown in Section 3, we experimented with actual trajectory data. In this section, we show experimental outlines and results.

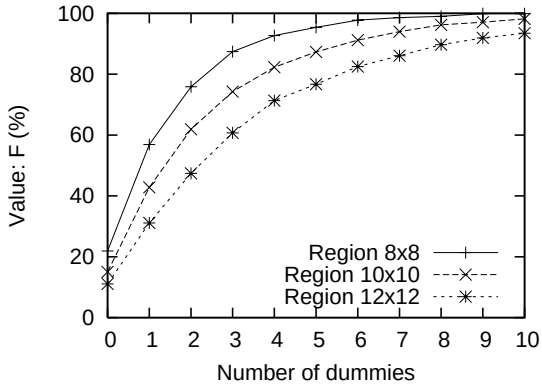

Figure 7. Comparison of location anonymity and number of dummies.

4.1. Settings

We did two experiments to evaluate the following:

- Comparison of location anonymity and number of dummies.
- Effectivity of the dummy generation algorithm.

We use **F** as a measure to evaluate the location anonymity of people and $\mathrm{Shift}(\mathbf{P})$ as a measure to compare the MN and MLN algorithms to the random generation of dummies.

We implemented a simulation system for the experiments that can deal with coordinates x and y and time t and display them. Moreover, the system has a module that generates dummies based on true position data. We can calculate the value of **P**, **F**, and $\mathrm{Shift}(\mathbf{P})$ using the system. For the experiments, we gave the system 39 rickshaw trajectories from Nara, Japan.

For simplification, we added the following assumption: All users generated the same number of dummies. That is, if a user generates two dummies, other users also generate two.

4.2. Results

Figure 7 shows the relationship between the number of dummies and ubiquity, **F**. In this figure, as location anonymity, a setting in which one dummy is generated in 8×8 regions is higher than another setting in which a dummy is not generated in 12×12 regions. In other words, the dummy generation technique enhances location anonymity more effectively than the accuracy reduction technique. Moreover, as expected, the more dummies, the larger the value of **F**. As shown in Figure 7, if a user achieves 80% of **F**, we conclude that the user needs three dummies in the 8×8 regions, four dummies in the 10×10 regions, and six dummies in the 12×12 regions.

Figure 8 shows the value of $\mathrm{Shift}(\mathbf{P})$ for MN, MLN, and the random algorithm. We set the number of regions

Table 2. Moving in a Neighborhood (MN) algorithm.

```
// Input:  positions of dummies at t-1
// Output:  positions of dummies at t
// random(x,y):  generate a random number between x and y

struct dummy {
       double x;         // x coordinate
       double y;         // y coordinate
       double t;         // time
};
void RangeLimit (double m, int n) {
      struct dummy prev[100], next[100];

      (Assignment prev[] to the Input);
      for (i=1;i<n;i++) {
            next[i]->x = random( (prev[i]->x)-m, (prev[i]->x)+m);
            next[i]->y = random( (prev[i]->y)-m, (prev[i]->y)+m);
            next[i]->t = (prev[i]->t)++;
      }
      (Output the contents of the next[]);
}
```

Table 3. Moving in a Limited Neighborhood (MLN) algorithm.

```
// Input:  positions of dummies at t-1
// Output:  positions of dummies at t
// random(x,y):  generate a random number between x and y
// position(x,y):  return the amount of position data where (x,y,t-1) belongs

struct dummy { (defined in Table 2) };
void RangeNumberLimit (int aveP, double m, int n) {
      struct dummy prev[100], next[100];
      int k = 0;

      (Assignment prev[] to the Input);
      for (i=1;i<n;i++) {
            next[i]->x = random( (prev[i]->x)-m, (prev[i]->x)+m);
            next[i]->y = random( (prev[i]->y)-m, (prev[i]->y)+m);
            next[i]->t = (prev[i]->t)++;
            if (position(next[i]->x, next[y]->y) > aveP) {
                  if (k<=3) { k++; continue; } else { k=0; }
            }
      }
      (Output the contents of the next[]);
}
```

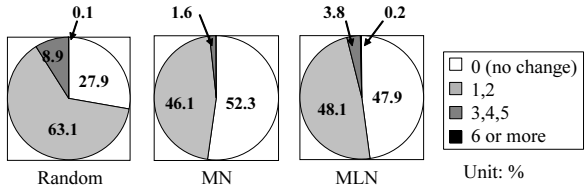

Figure 8. Relationship between dummy generation algorithms and $\mathrm{Shift}(\mathbf{P})$.

at 10×10 and the number of dummies at three. The results reveal that both algorithms have less $\mathrm{Shift}(\mathbf{P})$ than the random algorithm. Thus, we conclude that the two proposed algorithms are more effective than the random.

5. Conclusions

In this paper, we proposed a new anonymous communication technique for location-based services to protect location privacy using dummies. In the technique, a client system generates several false position data, which the system sends with the true information of the user to the service provider.

References

[1] A. R. Beresford and F. Stajano. Location privacy in pervasive computing. *IEEE Pervasive Computing*, 2(1):46–55, 2003.

[2] D. Chaum. The dining cryptographers problem: Unconditional sender and recipient untraceability. *Journal of Cryptology*, 1:65–75, 1988.

[3] I. Getting. The global positioning system. In *IEEE Spectrum*, volume 30, pages 36–47, December 1993.

[4] M. Gruteser and D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the First International Conference on Mobile Systems, Applications, and Services*, pages 31–42, 2003.

[5] A. Pfitzmann and M. Kohntopp. Anonymity, unobservability, and pseudeonymity: a proposal for terminology. In *International workshop on Designing privacy enhancing technologies*, pages 1–9. Springer-Verlag New York, Inc., 2001.

[6] O. Wolfson, P. Sistla, B. Xu, J. Zhou, and S. Chamberlain. DOMINO: Databases fOr MovINg Objects tracking. In *SIGMOD'99 Conference Proceedings*, pages 547–549, 1999.