Location Anonymization With Considering Errors and Existence Probability

doi: 10.1109/TSMC.2016.2564928

Location Anonymization with Considering Errors and Existence Probability

Yuichi Sei and Akihiko Ohsuga, *Member, IEEE*

***Abstract*—Mobile devices that can sense their location using GPS or Wi-Fi have become extremely popular. However, many users hesitate to provide their accurate location information to unreliable third parties if it means that their identities or sensitive attribute values will be disclosed by doing so. Many approaches for anonymization, such as k-anonymity, have been proposed to tackle this issue. Existing studies for k-anonymity usually anonymize each user's location so that the anonymized area contains k or more users. Existing studies, however, do not consider location errors and the probability that each user actually exists at the anonymized area. As a result, a specific user might be identified by untrusted third parties. We propose novel privacy and utility metrics that can treat the location and an efficient algorithm to anonymize the information associated with users' locations. This is the first work that anonymizes location while considering location errors and the probability that each user is actually present at the anonymized area. By means of simulations, we have proven that our proposed method can reduce the risk of the user's attributes being identified while maintaining the utility of the anonymized data.**

***Index Terms*—ubiquitous computing, privacy, location information, anonymization.**

I. Introduction

Many research studies have analyzed mobile users' consumer behavior by connecting the users' location information and their attributes such as gender and age. As a result, organizations can create good marketing programs and yield optimized advertisement delivery.

In this paper, we consider two types of organizations. The first type includes organizations that collect users' attributes and location data directly. We call these organizations "data holders." We assume that the data holder can be trusted and wants to anonymize and publish users' information to other organizations. The other type includes organizations that do not collect user attributes and location data by themselves but want to analyze this type of data. We call these organizations "data analyzers." Data analyzers may not be trusted, but they receive users' data from the trusted organizations if the data is anonymized. The anonymized data need not contain explicit identifiers such as name and address.

In this paper, our goal is not only to protect location data but also to protect other sensitive attribute values such as diseases associated with location data. We provide an example of such data in Table I. Our goal is to protect sensitive attribute values, even if adversaries are aware of the fact that a particular user is included in the anonymized database and they know the exact location data of the user.

Even if the anonymized data does not contain any explicit identifiers, the data analyzers have the possibility of identifying a specific user's attribute values. For example, let us assume that a data analyzer knows that Alice is present in a location (x_1, y_1) at time t and receives information that an anonymous user has an attribute value such as "disease is cancer" located at (x_1, y_1) at time t. In this case, the data analyzer can very easily identify that Alice has cancer.

To tackle this problem, several research studies on the anonymization of location data have been proposed. The k-anonymity [1] is one of the most popular privacy metrics. Many studies on the anonymization of location data have considered k-anonymity (e.g., [2]–[4]). They claim that they can create an anonymized area that contains k or more users.

However, most of these studies do not consider location errors. Even if we use GPS localization, the error range can be over 80 meters in urban canyons [5]. Moreover, there are many other methods for estimating the user's location since the GPS function consumes a lot of energy. If we use these other methods, the error range can include hundreds or thousands of meters [6], [7].

If we use existing methods of k-anonymity for anonymizing location data, the resulting anonymized areas may contain less than k users. Moreover, there is the possibility that the anonymized areas contain no users or only one user in the worst case. A few studies, such as NWA [8], [9], have considered location errors, but have not considered the existence probability of users and the attacker's knowledge. As a result, if we assume that attackers can have knowledge about a user's location, NWA does not satisfy k-anonymity [10], [11]. Our previous paper, written in Japanese [12], considered location errors, but the anonymization cost was large, and the evaluation was not sufficient.

Moreover, there is another issue related to location error. The existing studies on the k-anonymity of location data use the average size of the anonymized areas as a utility metric. However, this is not sufficient. Suppose that a data analyzer received information that Alice, Bob, and Charlie were located at an anonymized area L_1 at time t_1. There is a possibility that no one existed at L_1. In this case, the data analyzer will perform an incorrect analysis. Therefore, we should use another utility metric that can consider location error. In this paper, we also propose a novel utility metric that considers the probability of users being present at each anonymized area.

Our contributions are as follows:

Y. Sei and A. Ohsuga are with the University of Electro-Communications, Tokyo 182-8585, Japan (e-mail: seiuny@uec.ac.jp; ohsuga@uec.ac.jp).

This work was supported by JSPS KAKENHI Grant Numbers 24300005, 26330081, 26870201, and also supported by the Telecommunications Advancement Foundation.

Manuscript received XXX, 2005; revised XXX, 2015.

TABLE I
AN EXAMPLE OF RAW INFORMATION THAT A DATA HOLDER KNOWS.

Name	Location	Time	Attr1-Salary	Attr2-Disease
Alice	(x_1, y_1)	t_1	3,200	Cancer
Bob	(x_2, y_2)	t_1	750	Sty
Charlie	(x_3, y_3)	t_1	470	Cold
Dave	(x_4, y_4)	t_1	20	HIV
Eric	(x_5, y_5)	t_1	530	Cut
Flora	(x_6, y_6)	t_1	500	Fever

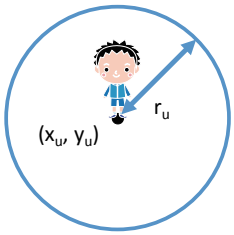


Fig. 1. Raw location data of user u.

- We clarify the problems of current studies for k-anonymity of location data.
- We propose a novel privacy metric and a utility metric that can treat the location error.
- We propose an efficient anonymization algorithm for the proposed metrics.

The rest of the paper is organized as follows. Section II presents the models of applications and attacks. Privacy and utility, as used in this paper, are defined in Section III. Section IV presents the design of our algorithm. The results of our simulations are presented in Section V. Section VI discusses several design issues in our method. Section VII discusses the related methods. Finally, Section VIII concludes the paper.

II. BACKGROUND AND PROBLEM DEFINITIONS

We will first describe the scenarios considered in this paper, the representation of the location error, and the problem we will tackle.

A. Background

A data holder has a database that contains user data consisting of each user's attribute and corresponding location data. Location data may be trajectory data. To simplify our discussion, we anonymize a snapshot of trajectory data. Anonymizing trajectory data will be reserved for future work, but we can, nevertheless, use the results of this paper to anonymize trajectory data.

We assume that the data holder wants to publish the database for collaboration with data analyzers. Owing to privacy concerns, the data holder wants to anonymize the database in order to ensure that any adversaries cannot identify each individual's attribute values. We assume that the data holder is an honest entity, but the data analyzers are semi-honest entities, which means they might try to identify the sensitive values of arbitrary users.

Operating systems such as iOS, Android, and Windows Phone OS express location by latitude, longitude, and accuracy [13]–[15]. Accuracy means a radius of a circle centered at the location's latitude and longitude, and the user might exist in the circle. Following these operating systems, we assume that the data holders have each user's central coordinates and the radius of the circle, where the user may exist at each time. Let (x_u, y_u) denote the central coordinates of the circle, where user u may exist, and let r_u denote the radius of the circle (Figure 1). We call this circle an "existence probability circle."

TABLE II
AN EXAMPLE OF ANONYMIZED INFORMATION GENERATED BY EXISTING STUDIES OF k-ANONYMITY FOR LOCATION DATA.

Location	Time	Attr1-Salary	Attr2-Disease
L_1	t_1	3,200	Cancer
L_1	t_1	750	Sty
L_1	t_1	470	Cold
L_2	t_1	20	HIV
L_2	t_1	530	Cut
L_2	t_1	500	Fever

B. Problem

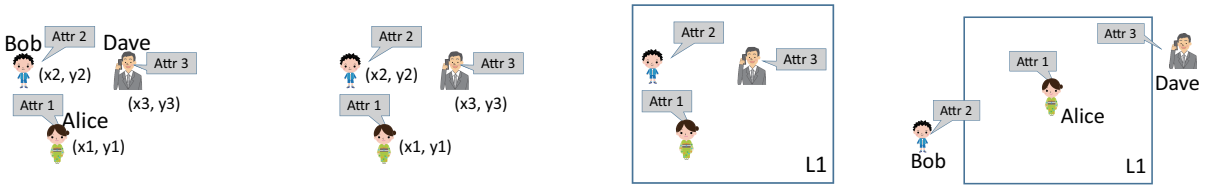
Suppose that a data holder has the information contained in Figure 2(a). If the data holder provides this information directly to a data analyzer, the data analyzer can specify each user's location and attribute values. Even if the data holder removes the user names, the problem will not be solved completely (Figure 2(b)). The reason behind this fact is described here. Suppose the data analyzer knows that Alice existed at (x_1, y_1). In this case, the location data related to the provided information can be used in identify users even if the provided information does not have any explicit identifiers that specify Alice.

Many studies described in Section VII generate "anonymized areas" that contain k or more users. Figure 2(c) represents the resulting anonymized area named L_1 where we set $k = 3$. The data holder provides the information that k users exist in anonymized area L_1 to the data analyzer. The data analyzer cannot know exactly which of the k users is Alice.

However, in reality, the data holder's location data has errors. For example, there is the possibility, in reality, that only Alice exists (Figure 2(d)).

Suppose that the data analyzer knows, from another data source, that only Alice exists in the area L_1. If the data analyzer concludes that, in the information provided by the data holder, some user attribute values have a relatively high probability of being owned by users existing in area L_1, the data analyzer can determine with high probability that Alice owns the attribute values.

For example, Table I shows the raw information that the data holder knows. Table II gives an example of anonymized information generated by existing studies where $k = 3$. Suppose that Alice existed in an exclusive shop, where only rich people tend to go, and suppose that the anonymized area L_1 is part of that shop. If the raw information, which the data holder believes, is correct, Alice, Bob, and Charlie must be in



(a) Sensed data (b) Inadequate anonymization (c) k-anonymization based on sensed data by existing studies (d) True information

Fig. 2. An example of anonymization.

that shop at time t_1. In this case, even if the data analyzer, who knows that Alice existed in the shop at time t_1, receives the anonymized information, the data analyzer cannot know what Alice's salary is and cannot know whether Cancer, Sty, or Cold is her disease.

Let us consider that the raw information, which the data holder believes, is not true. Suppose that only Alice existed in the exclusive shop at time t_1 in reality, and the data analyzer knows the fact. When the data analyzer receives the anonymized information shown in Table II from the data holder, the data analyzer can consider that the user, whose salary is 3,200, has a relatively high probability of existing in the exclusive shop. As a result, the data analyzer can consider that Alice has a relatively high probability of existing in the shop. Therefore, the data analyzer can conclude that the first record of Table II represents Alice, and Alice's salary is 3,200 and her disease is Cancer.

If we consider k-anonymized information, the data analyzer must not identify each user's record with a probability greater than $1/k$. However, if the data holder's information contains some errors, and the data analyzer knows more correct information about the specific user, the data analyzer can identify the user's record with a probability greater than $1/k$.

In this paper, we propose a novel privacy metric (k, w)-anonymity, which requires that k or more users exist in an anonymized location with probability greater than or equal to w.

Moreover, we should propose a novel utility metric while considering the location error. If the anonymized information indicates that users exist in the anonymized area, the probability that the information is true should be high.

Therefore, our goal is to ensure that k or more users exist in each anonymized area with a probability greater than or equal to w. We want to then try to minimize the size of each anonymized area and, at the same time, increase the probability that the users exist in the anonymized area in reality. We will define the privacy and utility metrics more specifically in Section III.

III. Metrics

We propose a privacy metric and a utility metric while considering the location error.

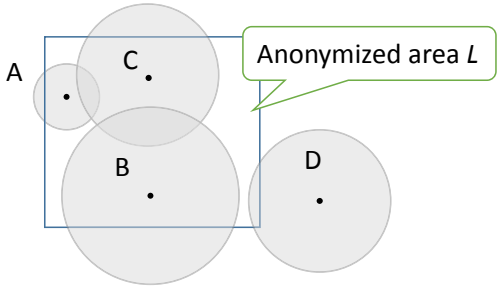


Fig. 3. $(1, 0.99)$-anonymity, $(2, 0.91)$-anonymity, $(3, 0.55)$-anonymity, and $(4, 0.027)$-anonymity when the existence probability of A, B, C, and D within L is 0.9, 0.75, 0.8, and 0.05, respectively.

A. Privacy Metric

We propose (k, w)-anonymity as follows:

DEFINITION III.1 ((k, w)-anonymity)**.** *The database satisfies (k, w)-anonymity if and only if more than or equal to k users exist in each anonymized area in the database with a probability greater than or equal to w.*

For example, in Figure 3, there are four users and an anonymized area. The circles in Figure 3 represent the existence probability circles of these users. Suppose that the probability that users A, B, C, and D exist in the anonymized area is 90%, 75%, 80%, and 5%, respectively. In this case, the probability that all four users exist in the anonymized area is $0.9 \times 0.75 \times 0.8 \times 0.05 = 0.027$. Therefore, the anonymization satisfies $(4, 0.027)$-anonymity. In a similar way, the anonymization also satisfies $(2, 0.91)$-anonymity and $(3, 0.55)$-anonymity.

Although existing studies that do not consider the errors of location data indicate that the anonymization satisfies 3-anonymity, the probability that more than or equal to three users exist in the anonymized area is only 0.55.

B. Utility Metric

Let U denote the set of users the data holder has, and let $L(u)$ represent the anonymized area where user u existed at the specific time. Let $p_{u,_L}$ denote the probability that user u existed in area L at the specific time, and let g_u represent the existence probability circle of user u. That is, the central point of g_u is (x_u, y_u) and the radius is r_u.

TABLE III
LIST OF NOTATIONS

Notation	Description
N	Number of users in the database the data holder has
(x_u,y_u)	Observed central point where the data holder believes user u exists
g_u	Circle where user u may exist (existence probability circle of u)
r_u	Radius of g_u
$L(u)$	Anonymized area where user u may exist
$\|L\|$	Size of area L
U_L	Set of users who may exist in anonymized area L
α	Degree of importance of existence probability

Let $L \cap g_u$ represent the overlapped area of L and g_u. The value of $p_{u,_L}$ is calculated by

$$p_{u,_L} = \frac{|L \cap g_u|}{|g_u|},$$

where $|S|$ represents the size of area S.

We define the utility with the following equation:

$$Utility = \sum_{u \in U} \frac{(p_{u,L(u)})^\alpha}{|L(u)|}, \tag{1}$$

where the parameter α is used for adjusting the weight of the size of the anonymized areas and the possibility of users existing in the corresponding anonymized area. The parameter α has a value greater than or equal to 0, and we set α to a larger value if we think that the existence probability is more important.

IV. ANONYMIZATION ALGORITHM

We describe the anonymization algorithm by considering the proposed privacy and utility metrics. Main notations are shown in Table III.

First, we analyze the characteristics of the location data to be considered in anonymization, and then we describe the outline of the anonymization algorithm. We will then, explain the details of the algorithm.

A. Overview

We introduce two characteristics of location data. First, users are unevenly distributed in the area. In some locations, users are densely located, but, in other locations, users may be sparsely located. If the borderline of the anonymized areas is set in the area where many users exist, the existence probability that many users exist near the borderline decreases. Second, the accuracy of each person is different. We can increase the utility by giving priority to users measured with high accuracy.

Based on this analysis, the proposed method consists of three phases: the area division phase, the area expansion phase, and the area reduction phase.

In the area division phase, we use Mondrian algorithm [16], which is widely used in existing studies, as a baseline method. Our proposed method first generates a full-anonymized area where all users exist, and then recursively divides the anonymized area until the resulted anonymized areas do not satisfy (k, w)-anonymity.

Then, the proposed method tries to increase the utility by expanding the anonymized area.

Our proposed method repeats the area division phase and the area expansion phase until the resulted anonymized areas do not satisfy (k, w)-anonymity. Our proposed method then executes the area reduction phase once.

We describe the details of these three phases in the following subsections.

B. Area Division Phase

We divide the full-anonymized area iteratively. Each iteration chooses the longer dimension to be divided and divides the chosen dimension so that each divided area contains half of the users. For example in Figure 4, we try to divide Area L by the line A_0 and get Areas L_1 and L_2. In this figure, the black points represent the observed points of users, and the circles represent the existence probability circles of the users. We divide each anonymized area only when the resulted areas satisfy (k, w)-anonymity. If we cannot divide an anonymized area at the longer dimension, we try to divide the anonymized area at the other dimension.

Let U_L denote the set of users who exist in area L with probability greater than 0. The probability $P(L, k)$ that k or more users exist in area L is calculated by

$$P(L,k) = 1 - \sum_{\{S|S \in \mathfrak{P}(U_L) \wedge |S| < k\}} \left[\prod_{u \in S} p_{u,_L} \cdot \prod_{u \notin S \wedge u \in U_L} (1 - p_{u,_L}) \right], \tag{2}$$

where $\mathfrak{P}(U_L)$ represents a power set of U_L.

Here, the amount of calculation becomes very heavy when the size of U_L is large. For example, the number of possible S that satisfies $\{S \in \mathfrak{P}(U_L) \wedge |S| < k\}$ is about 3 billion when $U_L = 50$ and $k = 10$. The algorithm that can be calculated faster is described in IV-E.

Let L_1 and L_2 denote the two areas generated by dividing one area. When the condition $P(L_1, k) < w$ or $P(L_2, k) < w$ holds, the division is canceled.

C. Area Expansion Phase

We get two anonymized areas as a result of the area division phase. The process described below is executed for each area. For example in Figure 4, we try to expand Areas L_1 and L_2 separately. We explain the detailed process for L_1 here.

We call the boundary line between L_1 and L_2, which obtained in the previous area division phase a "boundary side." Let A_0 denote the position of the boundary side in Figure 4. The objective of this phase is to expand the boundary side to the position that can obtain the maximum utility. As the anonymized area L_1 only expands, it will satisfy (k, w)-anonymity if it satisfies (k, w)-anonymity before executing this phase.

Initially, we consider the area that contains L_1, and the existence probability circles of all the users are included in

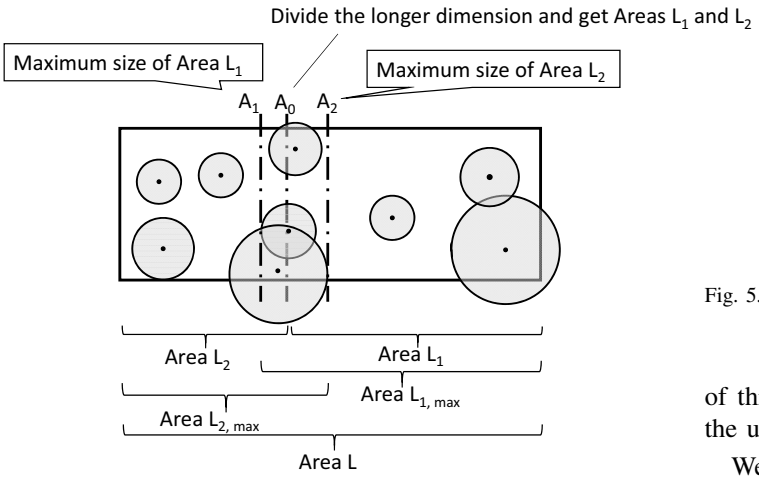


Fig. 4. Area division and area expansion.

area L_1. Let $L_{1,max}$ denote the expanded area and A_1 denote the position of the expanded boundary side. We then search for the optimal position of the boundary side that maximizes the utility. The search space is the range from A_0 to A_1.

The probability that each user actually exists in area L_1 increases by expanding the boundary side. However, the rate of increase decreases by expanding the boundary side more and more. Therefore, the utility function, which has a parameter that represents the position of the boundary side in the range from A_0 and A_1, is a unimodal function. We can obtain the maximum value of the unimodal function by using golden section search [17], [18].

Then, we define the positions A_{n1} and A_{n2} as

$$A_{n1} = \frac{\phi \cdot A_0 + A_1}{\phi + 1}, \qquad A_{n2} = \frac{\phi \cdot A_1 + A_0}{\phi + 1}, \tag{3}$$

where ϕ is the golden ratio, and $\phi = (1 + \sqrt{5})/2$.

Assume that the utility with position A_{n1} is smaller than that with position A_{n2}. In this case, the boundary side of L_1 is updated from A_0 to A_{n1}. If the utility with position A_{n2} is smaller than that with position A_{n1}, the boundary side of L_{max} is updated from A_1 to A_{n2}.

This process is repeated based on the updated L_1 and $L_{1,max}$. Finally, we get the optimal position of the boundary side that maximizes the utility.

The overall process is conducted also for L_2. The optimal position of the boundary side of L_2 is determined between A_0 to A_2 in the same way.

D. Area Reduction Phase

The area division phase and the area expansion phase are repeated until the resulting anonymized areas do not satisfy (k, w)-anonymity, and the area reduction phase is executed once at the end. The process of the area reduction phase is executed for every anonymized area. Let L denote one of the anonymized areas (Figure 5).

Let L_{min} denote the minimum area that intersects the existence probability circles of all users in L. The objective

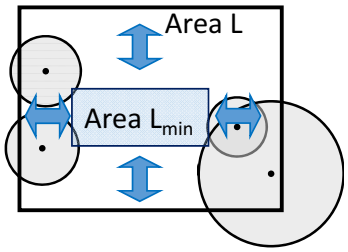


Fig. 5. Area reduction.

of this phase is to obtain the optimized area that maximizes the utility within the range of L to L_{min}.

We can narrow the anonymized area L by moving one of its four sides. We calculate two positions for golden section search for each side and calculate $P(L, k)$ and the utility for each position of each side. We then adopt the position that satisfies (k, w)-anonymity and maximizes the utility. This process is repeated to get the optimal anonymized area that maximizes the utility.

E. Semi-optimal Algorithm

The amount of calculation of (2) might become too heavy to obtain the result in a realistic time period. In this subsection, we describe the semi-optimal algorithm that reduces the amount of calculation greatly.

First, we can know immediately that (k, w)-anonymity is not satisfied when $|U_L| < k$.

Then, we count the number of probability existence circles included completely in anonymized area L, and let m denote the number. In this case, we can say that area L satisfies (k, w)-anonymization when the probability that $k-m$ or more users, except for the m users, existing area L is greater than or equal to w.

Furthermore, we introduce a semi-optimal algorithm that can be calculated faster. We discretize the value range of probability that each user actually exists in area L into d levels, for example, from 0 to 0.1, from 0.1 to 0.2, and so on. As an example, assume that $d = 10$ and the probability that a user actually exists in area L is 0.223. We can consider that the probability is not 0.223 but 0.2. Although we cannot maximize the utility, as we underestimate the existence probability, we can ensure the resulting anonymized area satisfies (k, w)-anonymity.

Let c_j represent the number of users for whom $p_{u,L}$ is greater than or equal to j/d and less than $(j + 1)/d$. That is, c_j is calculated by

$$c_j = \left| \left\{ u | u \in U \ \wedge \ \frac{j}{d} \leq p_{u,L} < \frac{j+1}{d} \right\} \right|. \tag{4}$$

The probability $P(L, k)$ that k or more users exist in area

L has the condition:

$$P(L,k) \geq 1 - \sum_{q=0}^{k-m-1} Q(q), \text{ where } Q(q) = \sum_{i_1=\underline{\Lambda}(1)}^{\bar{\Lambda}(1)} \cdot \sum_{i_2=\underline{\Lambda}(2)}^{\bar{\Lambda}(2)} \cdots \sum_{i_{d-1}=\underline{\Lambda}(d-1)}^{\bar{\Lambda}(d-1)} \cdot \prod_{j=1}^{d-1} \left[{}_{c_j}C_{i_j} \cdot (\frac{j}{d})^{i_j} \cdot (1-\frac{j}{d})^{c_j - i_j} \right], \quad (5)$$

$$\bar{\Lambda}(s) = \min(c_s, q - \sum_{j=1}^{s-1} i_j),$$

$$\underline{\Lambda}(s) = \max(0, q - \sum_{j=1}^{s-1} i_j - \sum_{j=s+1}^{d-1} c_j).$$

Here, we have the following theorem:

Theorem IV.1. *When $|U_L| \geq k$, the following equation always holds for $\bar{\Lambda}(s)$ and $\underline{\Lambda}(s)$ in (5);*

$$\bar{\Lambda}(s) \geq \underline{\Lambda}(s).$$

The proof of Theorem IV.1 is described in the Appendix A.

The part of $\prod_{j=1}^{d-1}$ in (5) represents the probability that i_j users out of c_j users exist in area L. The part of $\sum_{i_1} \sum_{i_2} \ldots \sum_{i_{d-1}}$ represents the all combinations of i_j $(j = 1, \ldots, d-1)$ that satisfy the summation of i_j $(j = 1, \ldots, d-1)$ is equal to q. The value q represents the number of users who exist in area L, and we change the value of q from 0 to $k - m - 1$.

We give an example for calculating $P(L,k)$. Assume that there are 10 users and the probability that each of the 10 users exists in L is greater than 0, and assume that $k = 7$, $c_3 = 5$, $c_5 = 1$, and $c_7 = 3$. Assume that the existence probability circle of one user is completely included in L. In this case, we can calculate (5) as follows:

$$P(L,7) \geq 1 - \sum_{q=0}^{7-1-1} \sum_{i_3=\max(0,q-(1+3))}^{\min(5,q)} \sum_{i_5=\max(0,q-i_3-3)}^{\min(1,q-i_3)} \sum_{i_7=\max(0,q-(i_3+i_5))}^{\min(3,q-(i_3+i_5))} \left[{}_5C_{i_3} (\frac{3}{10})^{i_3} (1-\frac{3}{10})^{5-i_3} \cdot {}_1C_{i_5} (\frac{5}{10})^{i_5} (1-\frac{5}{10})^{1-i_5} \cdot {}_3C_{i_7} (\frac{7}{10})^{i_7} (1-\frac{7}{10})^{3-i_7} \right].$$

As a result, we get $P(L,7) \geq 0.15$.

We can determine whether or not the target area satisfies (k,w)-anonymity based on (5) faster than the calculation based on (2). Furthermore, we introduce two theorems in order to terminate the calculation of (5) in the process of summing $Q(q)$.

Theorem IV.2. *When the value of 1-(sum of $Q(q)$ for $q = 0, \ldots, z$) where $z < k - m - 1$ becomes less than w in the process of summing $Q(q)$, we can terminate the calculation of the sum of $Q(q)$ and conclude that the target area does not satisfy (k,w)-anonymity.*

Theorem IV.3. *When the value of $Q(z)$, where $z < k-m-1$, is less than $Q(z-1)$ and 1-(sum of $Q(q)$ for $q = 0, \ldots, z$ plus $Q(z) \times (k - m - 1 - z)$) is greater than or equal to w, we can terminate the calculation of the sum of $Q(q)$ and conclude that the target area satisfies (k,w)-anonymity.*

The proofs of Theorems IV.2 and IV.3 are described in the Appendix B and C.

The overall algorithm for (k,w)-anonymity is shown in Algorithm 1.

Algorithm 1 Anonymization for (k,w)-anonymity

Input: Privacy parameters k and w, Target area L, Set of users U
Output: Set of anonymized areas $\mathfrak{L}$

```
𝔏 ⇐ division&expansion(k, w, L, U)
for L ∈ 𝔏 do
    L' ⇐ areaReduction(k, w, L, U)
    𝔏 ⇐ (𝔏 \ {L}) ∪ {L'}
end for
return 𝔏
```

Algorithm 2 division&expansion

Input: Privacy parameters k and w, Target area L, Set of users U
Output: Set of anonymized areas $\mathfrak{L}$

```
𝔏 ⇐ ∅
dim ⇐ chooseDimension(L)
A_0 ⇐ getMedianPoint(L, dim, U)
L_1 ⇐ one of the two areas generated by deviding L at A_0
L_2 ⇐ the other area
if both L_1 and L_2 satisfy (k,w)-anonymity then
    L_1 ⇐ areaExpansion(L_1, U)
    L_2 ⇐ areaExpansion(L_2, U)
    𝔏 ⇐ 𝔏∪ division&expansion(k, w, L_1, U)
    𝔏 ⇐ 𝔏∪ division&expansion(k, w, L_2, U)
else
    𝔏 ⇐ {L}
end if
return 𝔏
```

V. Evaluation

No existing studies have anonymized the location data while considering location error and users' existence probability. We compared our proposed method with Mondrian [16]. Although Mondrian is not the state-of-the-art method, a lot of studies still use it as a baseline method. Therefore, the characteristics of the proposed method can be easily understood by comparing with Mondrian.

A. Data Set

We conducted an experiment using the data set generated from the Siafu tool [19], which is used in many other studies such as [20], [21]. The Siafu tool is an open-source simulator, where the person behavior is modeled based on the typical behavior of a living person. We prepared the location map that has an 8.4km × 8.4km area (Figure 6 shows part of the

Fig. 6. Snapshot of Siafu simulator.

map). The number of users was set from 1,000 to 10,000, and they were programmed to go to their companies, restaurants, and/or parks on foot or by car depending on the distance from their houses to their destinations. We used the location data obtained by the simulator with an interval of 5 minutes for 4 hours, which means there were 48 data sets.

As the location data obtained by the Siafu tool has no errors, we added some errors to each location data randomly. The amount of the errors, i.e., the radius of each existence probability circle, was determined randomly with a range of R_{min} meters to R_{max} meters.

All experiments were conducted on an Intel Xeon CPU E5-2687W v2 @ 3.40GHz 3.40GHz workstation with 128 GB of RAM.

B. Evaluation Measurements

We evaluated the utility based on (1). Moreover, we introduced a privacy measurement; k-Persons Ratio (KPR). Let A_L denote the number of anonymized areas. Let $pos(u)$ denote the *true* position of user $u \in U$. Note that the true value of $pos(u)$ is unknown to the data holder because the locations of the data holder's database contain some errors. Let $\mathcal{A}$ denote the set of anonymized areas, and let $U(\mathfrak{a})$ represent the true set of users who exist in $\mathfrak{a} \in \mathcal{A}$; that is, $U(\mathfrak{a}) = \{u | u \in U \wedge pos(u) \text{ is included in } \mathfrak{a}\}$. We define A'_L as $|\{\mathfrak{a} | \mathfrak{a} \in \mathcal{A} \wedge |U(\mathfrak{a})| \geq k\}|$.

The KPR is defined as

$$\text{KPR} = A'_L / A_L. \tag{6}$$

C. Results

In this evaluation, the default values for parameters were $d = 10$, $k = 10$, $w = 0.9$, $\alpha = 1$, $N = 5000$, $R_{min} = 5m$, and $R_{max} = 500m$.

The results of 4-hours data at 5-minute intervals are shown in Figure 7. The *average KPR* represents the average of KPR of each anonymized area, and the *minimum KPR* represents the minimum value of KPR of all the anonymized areas.

The utility values of the proposed method are smaller than that of Mondrian. However, the average KPR and the minimum KPR of the proposed method are greater than those of Mondrian. In particular, the minimum KPR of Mondrian is almost 0.

The averages KPR of our proposed method are greater than the value of w. This is because we have used the semi-optimal algorithm based on (5), which underestimates the existing probability of each user, and the utility function defined by (1) considers the existence probability of users.

We then conducted experiments with varying parameters of k, α, N, and R_{max}.

Figure 8 shows the average results for varying the value of k. The larger k becomes, the smaller the utility becomes. The reason is that the size of an anonymized area narrows according to k. On the other hand, the larger k becomes, the larger the minimum KPR becomes. As the minimum KPR measures the minimum value, it tends to increase when the number of anonymized areas is small.

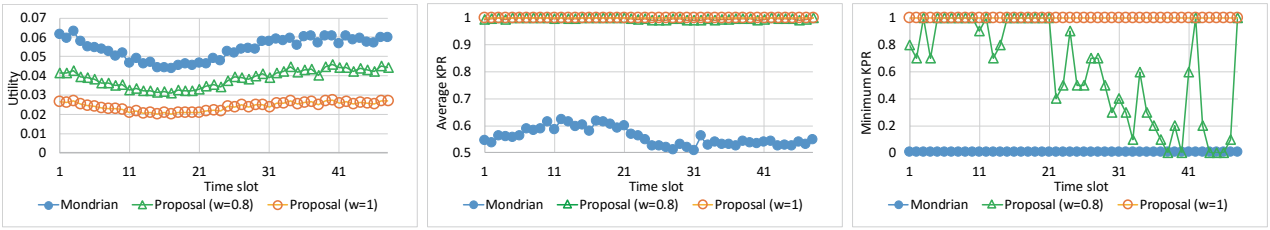
Figure 9 shows the average results for varying the value of α. The larger α becomes, the more the value of (1) tends to be reduced, and the utility values of both methods decrease according to α. Although the average KPR of each method does not change significantly according to α, the minimum KPR of our proposed method has improved significantly. As the size of each anonymized area tends to be large when the value of α is large, the probability that many users actually exist in the anonymized area becomes high.

Next, we conducted simulations for varying the value of N, and the average results are shown in Figure 10. The larger N becomes, the larger the utility values of both methods become. On the other hand, the larger N becomes, the smaller the average and minimum KPR and the minimum KPR become. As the population density is high when N is large, the impact of the location error is relatively large.

The average results for varying the value of R_{max} are shown in Figure 11. We know from the figure that the values of the utility, the average KPR, and the minimum KPR decrease when the location error is large.

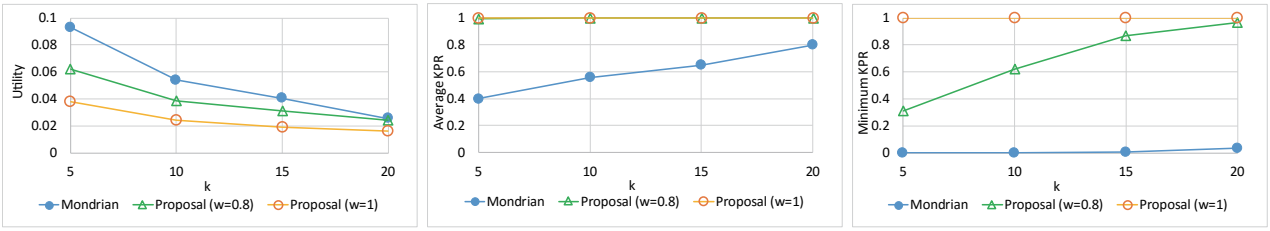
Finally, the average results of the anonymization time are shown in Figure 12. The anonymization time of the proposed method increases exponentially with k. This is because the time complexity of (5) increases exponentially with k. However, many existing studies assume that k ranges from about 3 to 20. We think that our proposed method can anonymize an area in a realistic period if we set k within this range. The anonymization time increases linearly with N.

On the other hand, the anonymization time of Mondrian is less than one second. Mondrian is better in terms of the anonymization time. However, Mondrian and all existing studies do not consider the location error, although the utility and privacy measures do consider the location error. Therefore, we think that it is difficult for existing studies to increase the utility and privacy values even if we give existing studies much more time for anonymization. Of course, if we give the existing studies the mechanisms that consider the location error like this study, the performance should be improved. Existing studies can use the proposed method to obtain such a mechanism.



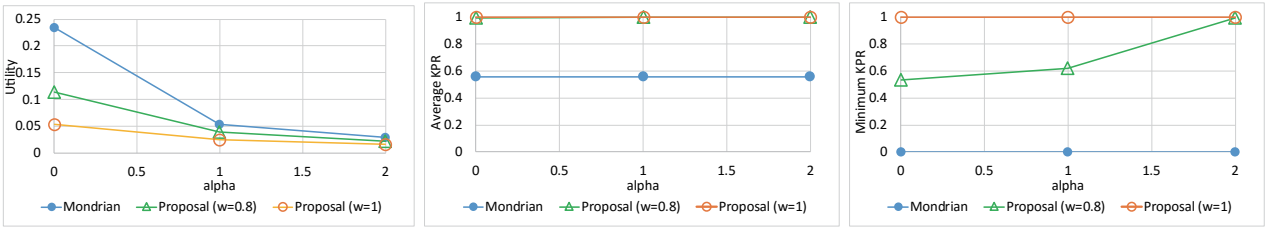
(a) Utility (b) Average KPR (c) Minimum KPR

Fig. 7. Utility and KPR every 5 minutes.



(a) Utility (b) Average KPR (c) Minimum KPR

Fig. 8. Average results with varying k.



(a) Utility (b) Average KPR (c) Minimum KPR

Fig. 9. Average results with varying α.

VI. DISCUSSION

A. Performance Analysis

Equation 5 has the biggest impact on the time complexity. The part of $\prod_{j=1}^{d-1}$ repeats multiplication $O(d)$ times. The part of $\sum_{q=0}^{k-m-1}$ repeats summation $O(k)$ times and the part of $\sum_{i_1}\cdots\sum_{i_{d-1}}$ repeats summation $O(k^d)$ times.

The number of repetitions of calculating the entire (5) for anonymizing one area is proportional to the number of users. Therefore, the time complexity of the proposed method is $O(dk^dN)$.

B. Validity of Our Scenarios

- A malicious data analyzer might know the actual location of a user.

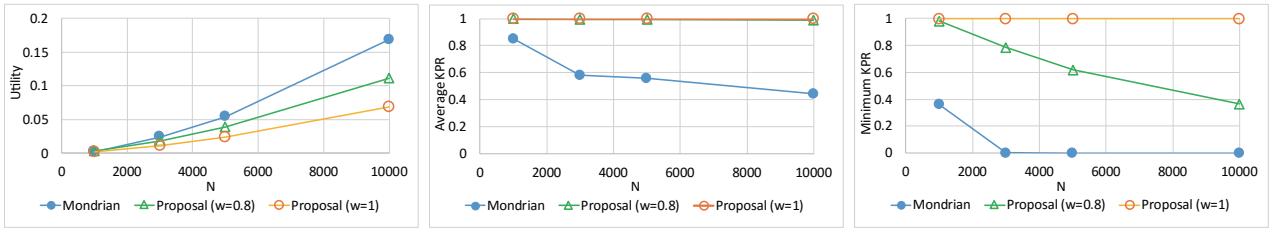
We give some examples of reasons why a malicious data analyzer knows that Alice existed in area L_1 at a certain time.

1) The malicious data analyzer saw Alice physically in area L_1 at that time.
2) The malicious data analyzer received the location information of Alice being in area L_1 at that time from another data holder.

In the future, many organizations will share personal data among them. Therefore, the risk that malicious data analyzers combine several personal data and then the privacy information is leaked will increase.

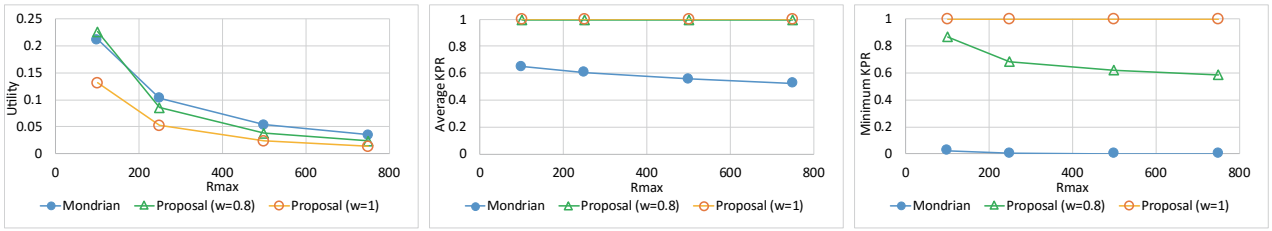
- Location has some error, and the amount of error varies largely.

We can get the precise location data of persons when they pass through ticket gates or buy items in shops. However, the GPS localization error ranges from a few meters to over 80 meters [5], and the error ranges over 500 m [22] with Wi-Fi based localization.



(a) Utility (b) Average KPR (c) Minimum KPR

Fig. 10. Average results with varying N.



(a) Utility (b) Average KPR (c) Minimum KPR

Fig. 11. Average results with varying R_{max}.

As mobile phones are battery-powered, mobile phone users usually obtain the location at a very low frequency. Therefore, the data holder should estimate the location of users by location estimation methods, such as [6], [7], if the data holder and the data analyzer want to analyze the location data in detail. In this case, the amount of error would be increased.

The accuracy of location data will be improved in the future, but the variation of the amount of the error will still exist.

C. Parameter Setting

We set k from 5 to 20, which covers the range of k in existing studies [23]–[25]. Determining the appropriate value of k is our future work, but out of the scope of this paper.

Determining the appropriate values of α and w is also our future work.

Moreover, we can modify the proposed privacy metric and the proposed method so that we can set several values of w. For example, by introducing k' $(k' < k)$, the modified method would be able to ensure that k' or more users exist in an anonymized area with probability 1, and k or more users exist in the area with probability w. Although we should extend our proposed method, we can realize this by changing only the check process that determines whether or not the area satisfies the required privacy condition.

D. Application to Other Privacy Metrics

Although k-anonymity can protect individual identities, there are times when it still cannot protect the sensitive attributes of these individuals. The l-diversity [26] ensures that the probability of identifying an individual's sensitive attribute is less than or equal to $1/l$ [1].

For example, we can define (l, w)-diversity as follows:

DEFINITION VI.1 ((l, w)-diversity)**.** *The database satisfies (l, w)-diversity if and only if more than or equal to l users exist in each anonymized area, and the relative frequency of each of the sensitive attribute values does not exceed $1/l$ for each anonymized area, with probability greater than or equal to w.*

Differential privacy [27], [28] is another privacy metric, and it makes user data anonymous by adding noise to a dataset so that an attacker cannot determine whether or not a particular point of user data is included.

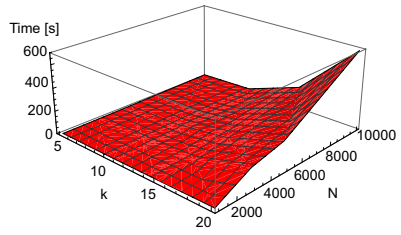
Although existing studies for differential privacy also do not consider the location error, we can extend it as follows:

DEFINITION VI.2 (differential privacy considering the error)**.** *Let D represent all users' actual location, which is unknown, even to the data holder. Let D' be database differing on, at most, one record. A randomized mechanism $\mathcal{A}$ satisfies ϵ-differential privacy considering the error if and only if for all sets Y of outputs, the following equation holds:*

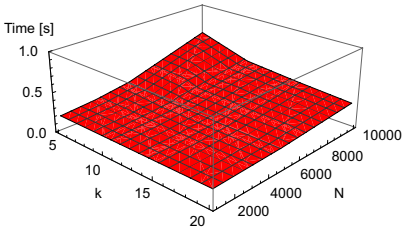
$$P(\mathcal{A}(D) \in Y) \leq e^{\epsilon} P(\mathcal{A}(D') \in Y) \quad \text{for all} \quad D, D'.$$

This definition is similar to the original one. The difference is that D is known to the data holder, that is, the data holder has D, in the definition of the original differential privacy, wheras D is unknown in the definition of the defferential

[1]There are several definitions of l-diversity.



(a) Proposal



(b) Mondrian

Fig. 12. Calculation time.

privacy considering the error. The definition of differential privacy considering the error has nothing to do with the database that the data holder has.

Proposing an algorithm to realize differential privacy considering the error is our future work.

VII. Related Work

Many studies on the anonymization of location data have used k-anonymity [29] as a privacy metric. When the original position of Alice is (x, y), an anonymized area that contains (x, y) is generated. Each anonymized area should contain k or more users.

Minimizing the size of the anonymized areas is a common objective of existing studies on the anonymization of location data (e.g., [2]–[4]). Due to the fact that finding an optimal k-anonymity is NP-hard [30], many existing algorithms for k-anonymity search for better anonymization through heuristic approaches. Mondrian [16], which is an efficient top-down greedy approach, is widely used in many other studies as a base method [8], [25].

Research studies on the anonymization of trajectory data such as [31] also have been proposed. Each trajectory data of a user consists of several location data represented by $\{(x_1, y_1, t_1), (x_2, y_2, t_2), \ldots\}$ where x_i and y_i represent longitude and latitude, respectively, and t_i represents the observed time. These studies aim at preventing data analyzers, who know Alice existed in a specific location at time t, from identifying where Alice existed at other times. Many studies have conducted k-anonymization for each specific time; hence, the results of our study, in this paper, can be used by these existing studies to anonymize trajectory data that contain the location error.

Abul et al. considered the location error and proposed a method called NWA for k-anonymity of the location data [8], [9]. NWA ensures that k or more users are located within the circle with the radius δ, which represents the uncertainty threshold. However, [10], [11] proved that NWA does not offer k-anonymity if location data has errors and the data analyzer knows the true location of the user.

Many algorithms [32], [33] for protecting location data and attributes of users other than previously mentioned research studies have been proposed, but they have not considered the location error.

The k-anonymity does not protect a location's privacy in many cases, as shown in [34], and we should consider the inference attack addressed in [35] to protect location privacy. These studies are important for protecting location data. However, it should be noted that the main goal of our proposed method is to protect sensitive attributes such as a disease and a salary associated with the location data, even if the attacker knows the true location of a user. To the best of our knowledge, this is the first work that protects sensitive attributes associated with location data by considering location errors.

Because several organizations have started collecting location data associated with sensitive attributes [36], we believe that algorithms for protecting the location data and the sensitive attributes associated with the location data have become important. In our future work, we plan to combine our study with other studies on protecting location privacy, such as [34], [35], to protect location data and sensitive attribute values while at the same time considering location errors.

In regard to utility metrics for location anonymization, a lot of existing studies, such as [24], [37], use the utility metric that increases as the sizes of anonymized areas decrease. This utility metric does not consider the location error.

VIII. Conclusion

We showed that data analyzers might identify a specific user from a k-anonymized location database with a probability greater than $1/k$, if the location has an error. Hence, it is unclear how the existing studies for k-anonymity protect against untrusted third parties. To tackle this problem, we proposed a novel privacy metric (k, w)-anonymity, which ensures that more than or equal to k users exist in each anonymized area with probability greater than or equal to w. Moreover, we proposed a novel utility metric, which can consider not only the anonymized area size but also the probability that each user actually exists in the area.

We proposed an anonymization algorithm based on the proposed metrics. We have proved that our algorithm can realize (k, w)-anonymization in a realistic time period, although the utility became less than that of existing studies.

Future work will include the evaluation of other real datasets. We also plan to apply our approach to other state-of-the-art methods. The main contributions of this paper are algorithms of area division, area expansion, and area reduction. As these algorithms can be performed after using other existing anonymization methods, we think that it is not difficult to combine our proposed method with them.

REFERENCES

[1] L. Sweeney, "k-anonymity: a model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[2] C.-Y. Chow, M. F. Mokbel, and T. He, "Tinycasper: a privacy-preserving aggregate location monitoring system in wireless sensor networks," in *Proc. ACM SIGMOD*, 2008, pp. 1307–1310.

[3] G. Ghinita, P. Kalnis, and S. Skiadopoulos, "PRIVE: anonymous location-based queries in distributed mobile systems," in *Proc. WWW*. ACM, may 2007, pp. 371–380.

[4] B. Bamba, L. Liu, P. Pesti, and T. Wang, "Supporting anonymous location queries in mobile environments with privacygrid," in *Proc. WWW*. ACM, apr 2008, pp. 237–246.

[5] N. M. Drawil, H. M. Amar, and O. A. Basir, "GPS Localization Accuracy Classification: A Context-Based Approach," *IEEE Trans. Intelligent Transportation Systems*, vol. 14, no. 1, pp. 262–273, 2013.

[6] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang, "Map-matching for low-sampling-rate GPS trajectories," in *Proc. 17th ACM SIGSPATIAL GIS*, 2009, pp. 352–361.

[7] K. Zheng, Y. Zheng, X. Xie, and X. Zhou, "Reducing Uncertainty of Low-Sampling-Rate Trajectories," in *Proc. IEEE ICDE*, 2012, pp. 1144–1155.

[8] O. Abul, F. Bonchi, and M. Nanni, "Never Walk Alone : Uncertainty for Anonymity in Moving Objects Databases," in *Proc. IEEE ICDE*, 2008, pp. 376–385.

[9] ——, "Anonymization of moving objects databases by clustering and perturbation," *Information Systems*, vol. 35, no. 8, pp. 884–910, 2010.

[10] R. Trujillo-Rasua and J. Domingo-Ferrer, "On the privacy offered by (k, δ)-anonymity," *Information Systems*, vol. 38, no. 4, pp. 491–494, 2013.

[11] G. Poulis, S. Skiadopoulos, G. Loukides, and A. GkoulalasDivanis, "Apriori-based algorithms for km-anonymizing trajectory data," *Trans. Data Privacy*, vol. 7, no. 2, pp. 165–194, 2014.

[12] Y. Sei and A. Ohsuga, "Location Anonymization on the Basis of Accuracy," *IEICE trans. Information and Systems (Japanese Edition)*, vol. J97-D, no. 5, pp. 964–974, 2014.

[13] Microsoft Inc., "Windows Phone Dev Center, http://dev.windowsphone.com/en-us/develop."

[14] Apple Inc., "iOS Developer Library, http://developer.apple.com/library/ios."

[15] Google Inc., "Android Developers, http://developer.android.com/."

[16] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," in *Proc. IEEE ICDE*, 2006, pp. 25–25.

[17] J. Kiefer, "Sequential minimax search for a maximum," *Proceedings of the American Mathematical Society*, vol. 4, no. 3, pp. 502–506, mar 1953.

[18] L. Nazareth and P. Tseng, "Gilding the Lily: A Variant of the Nelder-Mead Algorithm Based on Golden-Section Search," *Computational Optimization and Applications*, vol. 22, no. 1, pp. 133–144, 2002.

[19] M. Martin and P. Nurmi, "A Generic Large Scale Simulator for Ubiquitous Computing," in *Proc. MobiQuitous*, 2006, pp. 1–3.

[20] D. Cassou, J. Bruneau, C. Consel, and E. Balland, "Toward a Tool-Based Development Methodology for Pervasive Computing Applications," *IEEE Transactions on Software Engineering*, vol. 38, no. 6, pp. 1445–1463, nov 2012.

[21] D. C. Nazario, I. V. B. Tromel, M. A. R. Dantas, and J. L. Todesco, "Toward assessing Quality of Context parameters in a ubiquitous assisted environment," in *Proc. EEE Symposium on Computers and Communications (ISCC)*, jun 2014, pp. 1–6.

[22] K. Lin, A. Kansal, D. Lymberopoulos, and F. Zhao, "Energy-accuracy trade-off for continuous mobile device location," in *Proc. MobiSys*, 2010, pp. 285–298.

[23] L. Yao, G. Wu, J. Wang, F. Xia, C. Lin, and G. Wang, "A Clustering K-Anonymity Scheme for Location Privacy Preservation," *IEICE Trans. Information and Systems*, vol. E95-D, no. 1, pp. 134–142, 2012.

[24] T. Takahashi and S. Miyakawa, "CMOA: continuous moving object anonymization," in *Proc. 16th International Database Engineering & Applications Sysmposium (IDEAS)*. ACM, 2012, pp. 81–90.

[25] H. Hu, J. Xu, S. T. On, J. Du, and J. K.-Y. Ng, "Privacy-aware location data publishing," *ACM Trans. Database Systems*, vol. 35, no. 3, pp. 1–42, 2010.

[26] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy Beyond k-Anonymity," in *Proc. IEEE ICDE*, 2006, pp. 24:1–24:12.

[27] C. Dwork, "Differential Privacy," in *Automata, Languages and Programming*, 2006, vol. 4052, pp. 1–12.

[28] J. Domingo-Ferrer, "On the Connection between t-Closeness and Differential Privacy for Data Releases," in *Proc. International Conference on Security and Cryptography (SECRYPT)*, 2013, pp. 478–481.

[29] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proc. ACM SIGMOD*, 2005, pp. 49–60.

[30] A. Meyerson and R. Williams, "On the complexity of optimal K-anonymity," in *Proc. ACM PODS*, 2004, pp. 223–228.

[31] M. Terrovitis and N. Mamoulis, "Privacy Preservation in the Publication of Trajectories," in *Proc. IEEE MDM*, 2008, pp. 65–72.

[32] A. Parate, M.-C. Chiu, D. Ganesan, and B. M. Marlin, "Leveraging graphical models to improve accuracy and reduce privacy risks of mobile sensing," in *Proc. AMC Mobisys*, 2013, pp. 83–96.

[33] G. Maganis, E. Shi, H. Chen, and D. Song, "Opaak: Using Mobile Phones to Limit Anonymous Identities Online," in *Proc. AMC Mobisys*, 2012, pp. 295–308.

[34] R. Shokri, C. Troncoso, C. Diaz, J. Freudiger, and J.-P. Hubaux, "Unraveling an old cloak: k-anonymity for location privacy," in *Proc. ACM workshop on Privacy in the electronic society*, 2010, pp. 115–118.

[35] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying Location Privacy," in *Proc. IEEE Security & Privacy*, 2011, pp. 247–262.

[36] J. Iio, K. Yoshida, A. Koike, H. Shimizu, Y. Shirai, K. Kuwayama, K. Kuriyama, H. Konami, and S. Takayama, "Location-based Personal Information Log Reveals the Relation between Behavioral Characteristics and Consumption Tendencies," *IPSJ Journal*, vol. 52, no. 7, pp. 2256 – 2267, 2011.

[37] A. Gkoulalas-Divanis, P. Kalnis, and V. S. Verykios, "Providing K-Anonymity in location based services," *ACM SIGKDD Explor. Newsl.*, vol. 12, no. 1, pp. 3–10, 2010.

APPENDIX A
PROOF OF THEOREM V.1

Proof. From $|U_L| \geq k$, we get

$$m + \sum_{j=1}^{d-1} c_j >= k. \tag{7}$$

We will prove that Theorem IV.1 holds by mathematical induction when (7) holds.

1) We show that it is true for $s = 1$.

When $s = 1$, the values of $\bar{\Lambda}(s)$ and $\underline{\Lambda}(s)$ are expressed by

$$\begin{aligned} \bar{\Lambda}(1) &= \min(c_1, q), \\ \underline{\Lambda}(1) &= \max(0, q - \textstyle\sum_{j=2}^{d-1} c_j). \end{aligned} \tag{8}$$

Put

$$H_1 = q - \sum_{j=2}^{d-1} c_j. \tag{9}$$

Because $q \geq 0$ and $c_j \geq 0$ for all j ($j = 1, \ldots, d-1$), we can prove that Theorem IV.1 holds when $s = 1$ by proving that

$$H_1 \leq c_1 \tag{10}$$

always holds.

The maximum value of q is $q = k - m - 1$ because of (5). Therefore, we get from (9),

$$H_1 \leq k - m - 1 - \sum_{j=2}^{d-1} c_j. \tag{11}$$

Because

$$\sum_{j=2}^{d-1} c_j = \sum_{j=1}^{d-1} c_j - c_1, \tag{12}$$

we get

$$H_1 \leq k - m - 1 - \sum_{j=1}^{d-1} c_j + c_1. \tag{13}$$

From (7) and (13), we get

$$H_1 \leq c_1 - 1. \tag{14}$$

Therefore, (10) always holds. Hence, Theorem IV.1 holds when $s = 1$.

2) We show that it is true for $s = s' + 1$ when we assume it is true for $s = s'$.

When $s = s'$, the values of $\bar{\Lambda}(s)$ and $\underline{\Lambda}(s)$ are expressed by

$$\begin{aligned} \bar{\Lambda}(s') &= \min(c_{s'}, q - \textstyle\sum_{j=1}^{s'-1} i_j), \\ \underline{\Lambda}(s') &= \max(0, q - \textstyle\sum_{j=1}^{s'-1} i_j - \sum_{j=s'+1}^{d-1} c_j). \end{aligned} \tag{15}$$

Put

$$H_{s'} = q - \textstyle\sum_{j=1}^{s'-1} i_j - \sum_{j=s'+1}^{d-1} c_j. \tag{16}$$

Because we assume that the theorem IV.1 holds when $s = s'$, the following equation

$$H_{s'} \leq c_{s'} \tag{17}$$

always holds.

When $s = s' + 1$ the values of $\bar{\Lambda}(s)$ and $\underline{\Lambda}(s)$ are expressed by

$$\begin{aligned} \bar{\Lambda}(s'+1) &= \min(c_{s'+1}, q - \textstyle\sum_{j=1}^{s'} i_j), \\ \underline{\Lambda}(s'+1) &= \max(0, q - \textstyle\sum_{j=1}^{s'} i_j - \sum_{j=s'+2}^{d-1} c_j). \end{aligned} \tag{18}$$

Put

$$H_{s'+1} = q - \textstyle\sum_{j=1}^{s'} i_j - \sum_{j=s'+2}^{d-1} c_j. \tag{19}$$

Because $q \geq 0$ and $c_j \geq 0$ for all j ($j = 1, \ldots, d-1$), we can prove that Theorem IV.1 holds when $s = s' + 1$ by proving that

$$H_{s'+1} \leq c_{s'+1} \tag{20}$$

always holds.

From (16) and (19), we get

$$H_{s'+1} = H_{s'} - i_{s'} + c_{s'+1}. \tag{21}$$

From (17), we get

$$H_{s'+1} \leq c_{s'} - i_{s'} + c_{s'+1}. \tag{22}$$

Because of (5), we get $i_{s'} \leq c_{s'}$. Therefore

$$H_{s'+1} \leq c_{s'+1}. \tag{23}$$

Hence, (20) always holds. Therefore, Theorem IV.1 holds when $s = s' + 1$.

Based on mathematical induction, Theorem IV.1 holds. □

Appendix B
Proof of Theorem V.2

Proof. Because the value of $Q(q)$ for $q > 0$ has a positive value, the sum of $Q(q)$ for $q = 0, \ldots, z+1$ is always greater than the sum of $Q(q)$ for $q = 0, \ldots, z$.

Therefore, the value of 1-(sum of $Q(q)$ for $q = 0, \ldots, k - m - 1$) is always smaller than the value of 1-(sum of $Q(q)$ for $q = 0, \ldots, z$) where $z < k - m - 1$. □

Appendix C
Proof of Theorem V.3

Proof. The function $Q(q)$ is a unimodal function with respect to q. Therefore, if the value of $Q(z+1)$, where $z < k-m-1$ is less than $Q(z)$, the value of $Q(z+i)$ for all $i = 1, \ldots, k - m - 1 - z$ is less than the value of $Q(z)$.

Hence, the sum of $Q(q)$ for $q = 0, \ldots, k - m - 1$ is less than the sum of $Q(q)$ for $q = 0, \ldots, z$ plus $Q(z) \times (k - m - 1 - z)$. □

Yuichi Sei received his Ph.D. degree in Information Science and Technology, from the University of Tokyo, Japan, in 2009. From 2009 to 2012 he was working at Mitsubishi Research Institute. Since 2013, he has been an assistant professor at the University of Electro-Communications. His research interests include pervasive computing, privacy-preserving data mining, and software engineering.

Akihiko Ohsuga received a B.S. degree in mathematics from Sophia University in 1981 and a Ph.D. degree in computer science from Waseda University in 1995. He is a professor in the Graduate School of Informatics and Engineering, the University of Electro-Communications (UEC). He is also a visiting professor in National Institute of Informatics (NII). His research interests include agent technologies, web intelligence, and software engineering. He is a member of IEEE Computer Society (IEEE CS), Information Processing Society of Japan (IPSJ), Institute of Electronics, Information and Communication Engineers (IEICE), Japanese Society for Artificial Intelligence (JSAI), Japan Society for Software Science and Technology (JSSST), and Institute of Electrical Engineers of Japan (IEEJ). He was a chair of IEEE CS Japan Chapter. He was a member of the board of directors of JSAI and JSSST. He received the 1986 Paper Award from IPSJ.