PWS Cup 2019: Location Data Anonymization Competition

Takao Murakami[1,a)] Hiromi Arai[2] Makoto Iguchi[3] Hidenobu Oguri[4] Hiroaki Kikuchi[5] Atsushi Kuromasa[6] Hiroshi Nakagawa[2] Yuichi Nakamura[7] Kenshiro Nishiyama[8] Ryo Nojima[9] Takuma Hatano[10] Koki Hamada[11] Yuji Yamaoka[4] Takayasu Yamaguchi[12] Akira Yamada[13] Chiemi Watanabe[14]

Abstract: The amended act on the protection of personal information, which has been enforced since May 2017, states that personal data can be provided to a third party without users' consent if the data are anonymized as "anonymously processed information." However, anonymization methods are not clear, and hence we annually hold PWS Cup to clarify secure and appropriate anonymization methods. This year, we focus on location data and hold location data anonymization competition. This paper describes its contents.

Keywords: location privacy, anonymization, ID-disclosure, trace inference

1. Introduction

The amended act on the protection of personal information, which has been enforced since May 30 2017, states that personal data can be provided to a third party without users' consent if the data are anonymized appropriately (e.g., pseudonymization data processing such as adding noise, generalizing, and deleting). The method for anonymizing personal data securely and appropriately, however, is not clear. Various anonymization methods and standards have been proposed for a long time [1]. A guideline covering "What kind of anonymous processing is appropriate" is published by the Personal Information Protection Commission [2]. Yet, the information is not sufficient to clarify secure and appropriate data anonymization methods.

In the Opinion 05/2014 on Anonymisation Techniques [3] issued by the EU's Article 29 Working Party, the following three risks are mentioned as criteria for determining if data is anonymized or not.

Singling out: A risk of records being identified (singled out).

Linkability: A risk of multiple records (in the same database or in two different databases) concerning the same data subject being linked.

Inference: A risk of data subject's attributes being deduced with high probability.

The opinion states that while it is desirable to be secure against these risks, each data anonymization technology has its own merits and demerits. The opinion concludes that the appropriate data anonymization technique depends on the situation.

In an attempt to clarify secure and appropriate data anonymization methods, we have been holding the "PWS Cup" contest since 2015. In the contest, we compete for the utility and privacy level of anonymized data. In the 2015 contest, we used pseudo micro-data (National consumption survey) as the anonymization target. We used purchase history data as the anonymization target in the contest from 2016 to 2018.

1.1 PWS Cup 2019

In the 2019 contest, we use location data as the anonymization target[*1]. In this section, we describe the

[1] AIST
[2] RIKEN
[3] Kii Corporation
[4] Fujitsu Laboratories Ltd.
[5] Meiji University
[6] Fujitsu Cloud Technologies
[7] Waseda University
[8] BizReach, Inc.
[9] NICT
[10] NS Solutions Corporation
[11] NTT Secure Platform Laboratories
[12] NTT DOCOMO, Inc.
[13] KDDI Research, Inc.
[14] Tsukuba University of Technology
[a)] takao-murakami@aist.go.jp

[*1] Note that the standards for processing anonymously processed information stated in the amended act on the protection of personal information differ from the privacy standards for data anonymization adopted in this contest. In the former case, the standards presented in a guideline [2] require that data should be anonymized so that data subject is not identifiable even by the data provider who has the original data (i.e., maximum knowledge attacker model) while assuming that the attacker has only ordinary skills. In the latter case, we assume that the data receiver is a potential attacker (i.e., partial knowl-

characteristics of our contest.

1.1.1 Location data contest

As far as we know, this contest is the first location data anonymization contest in the world. The motivation of holding such location data anonymization contest is to accommodate the expectation for the utilization of the location data.

Recently, location-based services (LBS) such as POI (Point of Interest) search of nearby restaurants and route search are widely used. As a result, a large amount of "traces/trajectories" (movement history that is a chronological list of location data) is accumulating in such LBS providers. These traces are sometimes called as location big data and are expected to be leveraged in various applications such as analyzing popular spots [4], auto-tagging POI categories (like restaurants and hotels) [5], and analyzing the movement of foreign tourists [6].

On the other hand, there is a possibility that the traces are used to identify the patient's home and hospital where they are attending. The traces could also be used to infer the relationships that people do not want to disclose. Some studies show that original user ID can be identified with high probability even if the traces are pseudonymized [7], and this fact indicates that the pseudonymization alone is not sufficient. Therefore, anonymizing traces become essential before providing them to a third party, but it is not easy to find a method to anonymize the traces securely and appropriately. An anonymization mechanism based on "differential privacy," for example, is attracting attention recently. The mechanism, however, may not be suitable for time-series data like long traces. A study shows that a parameter called privacy budget ϵ becomes large as the length of trace becomes long, which could yield to low privacy protection [8].

An anonymization method for traces that realizes high utility and privacy is still not known. The aim of our location data anonymization contest is to contribute in clarifying such the anonymization method.

1.1.2 Partial knowledge attacker model

In order to evaluate the security and privacy levels, it is necessary to consider the background knowledge of attackers. There are two types of background knowledge models. The first model is a "maximum knowledge attacker model" in which an attacker knows the original data before it is being anonymized. The second model is a "partial knowledge attacker model," in which an attacker does not have any knowledge of the original data before it is being anonymized but has partial knowledge of other data. The former model assumes a sender of anonymized data as an attacker, and the latter model assumes a receiver of anonymized data as an attacker.

Location data is known to have a high singularity. A study indicates that about 95% of traces, composed of four location data, are unique (i.e., a trace is unique out of 1.5 million people) [9]. This implies that we need to delete almost all location data from original traces in order to prevent user ID disclosure from the anonymized traces. This implication, however, assumes the maximum knowledge attacker model. An attacker already knows the original traces in the maximum knowledge model, so there is no need to for the attack to attempt user ID disclosure attack (i.e., the privacy of the original traces is completely broken from the first place). In other words, the maximum knowledge attacker model allows us to evaluate the privacy in a situation where the attacker has full background knowledge, which is the worst case, but the model is somewhat unrealistic.

In the contest, therefore, we evaluate the privacy under the partial knowledge attacker model. Attackers do not have any knowledge of original traces but have partial knowledge of other traces. We call these traces as "reference traces" in this paper. When attacking anonymized traces, they use the reference traces. Location data people publish on SNS is an example of the reference trace that is available to attackers. The amount of information published in this manner is limited in general, so we must assume that the reference traces available in reality are limited.

We set the length of the reference traces in this contest accordingly. (Refer Section 3.1 for more details).

1.1.3 Evaluating ID disclosure and trace inference

In this contest, we evaluate "ID disclosure" and "trace inference" as risks. ID disclosure is an attack that attempts to infer the original user IDs from anonymized traces, and trace inference is an attack that attempts to restore the original traces completely (e.g., if the number of users is n and each user has t location data, the attacker is to restore nt location data). The former attack is also called re-identification attack, and the latter attack is also known as tracking attack [10].

The amended act on the protection of personal information considers ID disclosure as a risk for anonymously processed information. For traces processed by noise addition and generalization, ID disclosure attack only infers user IDs and does not link users with original location data. On the contrary, trace inference attack tries to restore the original traces completely and thus links users with original location data. It is, therefore, essential to consider trace inference also as a risk. Anonymized traces are usually pseudonymized, which correspond to shuffling traces by users. This implies that ID disclosure attempt is required to some extent before attempting trace inference attack, but this does not mean that anonymized traces that are resistant to ID disclosure are also resistant to trace inference. For example, k-anonymization ensures the ratio of successful ID disclosure to be below $1/k$, but it opens a room for attackers to perform attribute inference without performing ID disclosure [11]. Likewise, attackers could perform trace inference without performing ID disclosure (refer Section 3 of [12]). Thus, the correlation between "protection level against ID disclosure" and "protection level against trace inference" is not still evident.

edge attacker model) and assume that they possess professional/researcher level skills (refer Section 1.1.2 for more details). It is a future work to clarify the relationship between the data anonymization techniques that can resist attacks in this contest and the data anonymization techniques that can resist attacks performed by ordinary-skilled attackers with the maximum knowledge.

Based on the above discussion, we evaluate anonymized traces in two axis: ID disclosure resistant and trace inference resistant. Specifically, we distribute two sets of data, one for "ID disclosure challenge" and another for "trace inference challenge." For each data set, we assess teams who anonymize original trace set and make it tolerant against ID disclosure or trace inference attack. We also let teams perform ID disclosure and trace inference attacks to all anonymized trace sets and assess teams who succeed to break the tolerance of anonymized trace sets to ID disclosure and trace inference. From the assessment results, we try to clarify the correlation between "protection level against ID disclosure" and "protection level against trace inference" for each anonymized trace set. Refer Section 2.3 for the details.

For the utility measure, we set a "requirement value" in the contest. If the utility score of an anonymized trace set is below the requirement value, we regard the set as "invalid." When evaluating anonymized trace sets in the context of ID disclosure and trace inference, we only assess anonymized trace sets that are "valid," or the ones that have their utility score above the requirement value.

We hope that the knowledge obtained in this contest will serve as a reference for future discussion of the legal system.

2. Contest Overview

2.1 Notation

Throughout this paper, we denote the set of natural numbers and non-negative real numbers as $\mathbb{N}$ and $\mathbb{R}_{\geq 0}$, respectively. We also use the notation $[a] = \{1, \cdots, a\}$ for $a \in \mathbb{N}$.

Let $z \in \mathbb{N}$ denotes the number of teams participating in this contest. The set of teams is represented as $\mathcal{P} = \{P_1, \cdots, P_z\}$. Before starting the contest, each team will receive two sets of location data. The first location data set is for ID disclosure challenge, and the second location data set is for trace inference challenge. All location data sets will have the same number of users, but users covered in each data set are different (refer Section 3.1 for more details). Given the number of users $n \in \mathbb{N}$, we denote a set of users, covered in the j-th ($j \in \{1, 2\}$) location data set distributed to team P_i ($i \in [z]$), as $\mathcal{U}^{(i,j)} = \{u_1^{(i,j)}, \cdots, u_n^{(i,j)}\}$. The subscript of the user $u_k^{(i,j)}$ is called "user ID". The user ID k will thus be a natural number between 1 and n (i.e., $k \in [n]$).

The location information is discretized by dividing the target area into smaller regions. In the contest, the Tokyo metropolitan area is divided equally into 32×32 regions. We denote the number and set of the discretized regions as $m \in \mathbb{N}$ and $\mathcal{X} = \{x_1, \cdots, x_m\}$, respectively. The subscript of the location data x_k is called "Region ID." The region ID k will thus be a natural number between 1 and m (i.e., $k \in [m]$).

The time is also discretized by a set of equally spaced time interval (30 minutes interval in the contest). Each time interval is represented by a natural number.

Each location data set is composed of an "original trace set" and "reference trace set." The former trace set is the one to be anonymized by a team. The latter trace set is the reference for other teams when they are to attempt the ID disclosure and trace inference against the anonymized data. The reference trace set includes traces from time 1 to t, while the original trace set includes traces from time $t+1$ to $2t$.

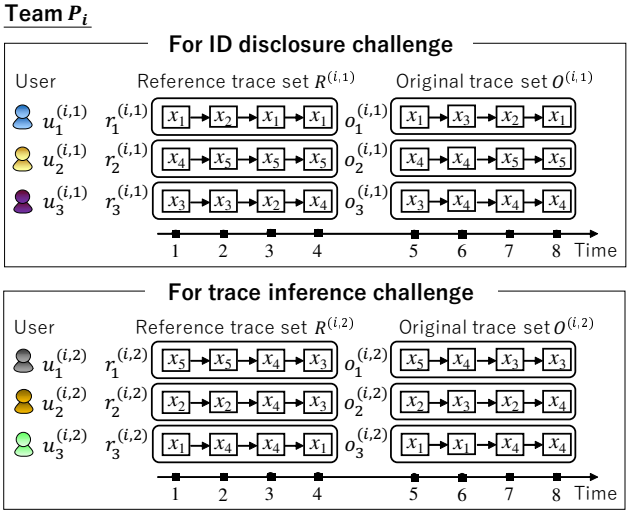

Fig. 1 An example of location data set ($i \in [z]$, $n = 3$, $m = 5$, $t = 4$)

For $k \in [n]$, we denote the reference and original trace of user $u_k^{(i,j)}$ as $r_k^{(i,j)}$ and $o_k^{(i,j)}$, respectively ($r_k^{(i,j)}, o_k^{(i,j)} \in \mathcal{X}^t$). $R^{(i,j)}$ and $O^{(i,j)}$ are the reference and original trace sets included in the j-th ($j \in \{1, 2\}$) location data set that are distributed to team P_i ($i \in [z]$). Both trace sets are generated by aligning corresponding user traces in order of User ID. Thus, they are both composed of nt location data and are set as $R^{(i,j)} = \{r_1^{(i,j)}, \cdots, r_n^{(i,j)}\}$ and $O^{(i,j)} = \{o_1^{(i,j)}, \cdots, o_n^{(i,j)}\}$.

Figure 1 illustrates an example of location data sets distributed to team P_i. In this example, the number of users is $n = 3$, the number of regions is $m = 5$, and the time is $t = 4$. You can see that $o_1^{(i,1)} = (x_1, x_3, x_2, x_1)$ and $O^{(i,1)} = \{o_1^{(i,j)}, o_2^{(i,j)}, o_3^{(i,j)}\}$.

2.2 Contest Flow

Figure 2 illustrates the contest flow. Team $P_1, \cdots, P_z$ and a judge Q participate the contest. The contest has the following four phases: **data anonymizing phase**, **utility assessment phase**, **ID disclosure & trace inference phase**, and **privacy assessment phase**. We will describe each phase briefly.

2.2.1 Data anonymizing phase

In this phase, the data is anonymized by processing location data and by shuffling traces (pseudonymization).

First, the judge Q distributes the j-th ($j \in \{1, 2\}$) original trace set $O^{(i,j)}$ to each team P_i ($i \in [z]$). Note that the reference trace set $R^{(i,j)}$ is not distributed at this phase; they will be distributed in the ID disclosure & trace inference phase.

Next, team P_i processes location data in the original trace set $O^{(i,j)}$ and creates the anonymized trace set $A^{(i,j)}$ that is a set of anonymized location data listed in order of user ID. Team P_i then submits the anonymized trace set $A^{(i,j)}$ to the judge Q. Team P_i can process both $O^{(i,1)}$ and $O^{(i,2)}$

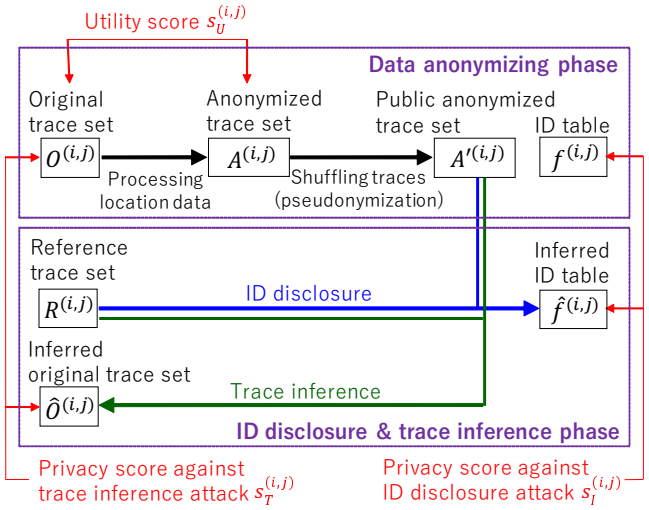

Fig. 2 Contest flow ($i \in [z]$, $j \in \{1, 2\}$)

original trace sets or can process just one of them. The team will not be disqualified even if it does not process any of them.

After receiving the anonymized trace set $A^{(i,j)}$, the judge Q shuffles n traces in the trace set randomly. The judge then assigns pseudo ID sequentially, from $n+1$ to $2n$, to the shuffled traces. As a result, the judge creates a public anonymized trace set $A'^{(i,j)}$ containing anonymized location data listed in order of pseudo ID and an ID table $f^{(i,j)}$ showing the correspondence between original user IDs and pseudo IDs. The public anonymized trace set $A'^{(i,j)}$ and the reference trace set $R^{(i,j)}$ will be disclosed to all teams during the ID disclosure & trace inference phase. The ID table $f^{(i,j)}$, however, is kept secret only to the judge Q.

We will describe more on the data anonymizing rule (location data processing and trace shuffling) in Section 3.2. More information on the anonymized trace set $A^{(i,j)}$, the public anonymized trace set $A'^{(i,j)}$, and the ID table $f^{(i,j)}$ will be also covered in Section 3.2.

2.2.2 Utility assessment phase

In this phase, the judge Q calculates the utility score $s_U^{(i,j)} \in [0, 1]$ of the anonymized trace set $A^{(i,j)}$ using the corresponding original trace set $O^{(i,j)}$. High $s_U^{(i,j)}$ score means the utility of the anonymized trace set is high, and low $s_U^{(i,j)}$ score means the utility of the anonymized trace set is low. Refer Section 3.4 for more details on the utility score.

The judge Q then compare the utility score $s_U^{(i,j)}$ with a requirement value $s_{req} \in [0, 1]$. If the condition $s_U^{(i,j)} \geq s_{req}$ is met, the judge regards $A^{(i,j)}$ as "valid." If $s_U^{(i,j)} < s_{req}$, then the judge regards $A^{(i,j)}$ as "invalid." The judge will announce the requirement value s_{req} beforehand.

Note that the utility score $s_U^{(i,j)}$ can be calculated from the original trace set $O^{(i,j)}$ and the anonymized trace set $A^{(i,j)}$. Team P_i can therefore check the validity of their anonymized trace set $A^{(i,j)}$ before submitting the set by checking if the utility score $s_U^{(i,j)}$ is more than or equal to the requirement value s_{req}. The judge Q will provide a program for calculating the utility score beforehand.

2.2.3 ID disclosure and trace inference phase

First, the judge Q provides all "valid" public anonymized trace sets and reference trace sets of all teams.

Next, each team executes ID disclosure and trace inference attacks against other teams' public anonymized trace sets using the reference trace sets as hints. A team can execute both ID disclosure and trace inference attacks to each public anonymized trace set, or they can execute only one of these attacks. The team will not be disqualified even if they do not execute any attack.

Let's take an example of team P_h ($h \neq i$) attacking the public anonymized trace set $A'^{(i,j)}$ of P_i. Team P_h executes the following actions:

ID disclosure attack Team P_h predicts user IDs that correspond to pseudo IDs in the public anonymized trace set $A'^{(i,j)}$ using the reference trace set $R^{(i,j)}$ as hints. The team creates an inferred ID table $\hat{f}^{(i,j)}$ and submits it to the judge Q.

Trace inference attack Team P_h predicts nt location data in the original trace set $O^{(i,j)}$, the trace set that correspond to the public anonymized trace set $A'^{(i,j)}$, using the reference trace set $R^{(i,j)}$ as hints. The team creates an inferred original trace set $\hat{O}^{(i,j)}$ and submits it to the judge Q.

Team P_h can submit only one $\hat{f}^{(i,j)}$ and $\hat{O}^{(i,j)}$ for P_i. This means that each team can perform at most one ID disclosure attack and one trace inference attack on another team.

We will explain more details on the inferred ID table $\hat{f}^{(i,j)}$ and the inferred original trace set $\hat{O}^{(i,j)}$ in Section 3.3.

2.2.4 Privacy assessment phase

In this phase, the judge Q scores the privacy level of the public anonymized trace set $A'^{(i,j)}$ against ID disclosure and trace inference attacks.

The privacy score against ID disclosure, $s_I^{(i,j)} \in [0, 1]$, is calculated by comparing the ID table $f^{(i,j)}$ and the inferred ID table $\hat{f}^{(i,j)}$. The privacy score against trace inference, $s_T^{(i,j)} \in [0, 1]$, is calculated by comparing the original trace set $O^{(i,j)}$ and the inferred original trace set $\hat{O}^{(i,j)}$. For both $s_I^{(i,j)}$ and $s_T^{(i,j)}$, the privacy level is higher if the scores are larger. We will explain the privacy score against ID disclosure and trace inference in more depth in Section 3.5.

A public anonymized trace set $A'^{(i,j)}$ gets at most $z-1$ ID disclosure attacks and at most $z-1$ trace inference attacks from other teams. Note that a public anonymized trace set for ID disclosure challenge $A'^{(i,1)}$ will get both ID disclosure and trace inference attacks. This holds true for a public anonymized trace set for trace inference challenge $A'^{(i,2)}$. The trace set also gets ID disclosure and trace inference attacks from sample programs [12]. The judge calculates the privacy scores against ID disclosure and trace inference for all attacks and find the minimum privacy score for ID disclosure attack ($s_{I,min}^{(i,j)} \in [0, 1]$) and trace inference attack ($s_{T,min}^{(i,j)} \in [0, 1]$). These scores become the final privacy scores of the public anonymized trace set $A'^{(i,j)}$. In other words, the final privacy scores are the ones marked when $A'^{(i,j)}$ gets the strongest attack.

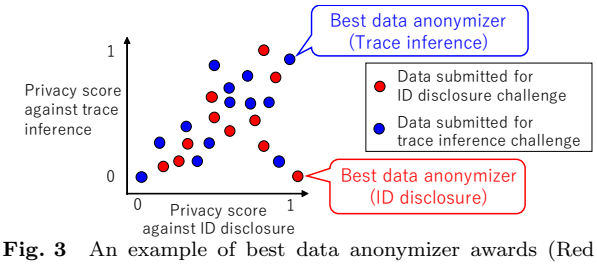

Fig. 3 An example of best data anonymizer awards (Red and blue dots represent public anonymized trace sets submitted for ID disclosure and trace inference challenge, respectively).

2.3 Awards

This section explains prizes given in the contest and how these prizes are designed.

2.3.1 Prizes

The following awards will be given in the contest. The award-winning teams will be selected after adding corresponding scores (privacy score against ID disclosure and/or trace inference attacks) from the preliminary and final rounds at a ratio of 1:9.

Overall winner, 2nd place, and 3rd place: These prizes will be given to teams that protect their data well against ID disclosure and trace inference attacks.

More specifically, the prizes will be given to the top-3 teams whose public anonymized trace sets mark the highest $s^{(i,1)}_{I,min} + s^{(i,2)}_{T,min}$, where $s^{(i,1)}_{I,min}$ is $A'^{(i,1)}$'s privacy score against ID disclosure attack and $s^{(i,2)}_{T,min}$ is $A'^{(i,2)}$'s privacy score against trace inference attack. Note that invalid or unsubmitted public anonymized trace set will get the privacy score of 0.

Best data anonymizer (ID disclosure): The prize will be given to a team whose public anonymized trace set for ID disclosure challenge ($A'^{(i,1)}$) marks the highest privacy score against ID disclosure attack ($s^{(i,1)}_{I,min}$).

Best data anonymizer (trace inference): The prize will be given to a team whose public anonymized trace set for trace inference challenge ($A'^{(i,2)}$) marks the highest privacy score against trace inference attack ($s^{(i,2)}_{T,min}$).

Figure 3 illustrates an example of how the best data anonymizer prizes are given for ID disclosure and trace inference.

Best risk assessor (ID disclosure): The prize will be given to a team who contributes the most in lowering the privacy score against ID disclosure attack. More details are described in the contest rule [13].

Best risk assessor (trace inference): The prize will be given to a team who contributes the most in lowering the privacy score against trace inference attack. More details are described in the contest rule [13].

Best presenter: Each team will make a presentation in CSS2019 venue to explain the algorithms used for anonymizing their data and for performing ID disclosure and trace inference attacks. The best presenter prize will be given to a team who makes the best presentation. More details are described in the contest rule [13].

2.3.2 Prize design

The purpose of awarding an overall winner, 2nd place, 3rd place, and best data anonymizers (one for ID disclosure and another for trace inference) is to collect good anonymized data sets that are resistant to ID disclosure and trace inference attacks. In the contest, we specify that the 1st location data set distributed to each team is for ID disclosure challenge and the 2nd location data set is for trace inference challenge. Hence, the anonymized trace sets submitted by each team are clear on their anonymizing intentions (i.e., if the trace set is meant to be resistant to ID disclosure attack or to trace inference attack). When selecting the best anonymizers, therefore, we only use the corresponding anonymized data sets.

In contrast, we use all public anonymized trace sets when selecting the best risk assessors, regardless of if the trace set is anonymized against ID disclosure or trace inference. Each team basically performs both ID disclosure and trace inference attacks to all public anonymized trace sets, and the team who contributes the most in lowering the privacy score against these attacks will get the best risk assessor prizes. This approach is chosen to explore if there is a correlation between the privacy scores against ID disclosure and the privacy score against trace inference (as shown in Fig 3).

In order not to put too much burden on each team, we allow a team to submit only one anonymized trace set (i.e., for ID disclosure challenge or trace inference challenge) or no anonymized trace set. Likewise, we allow a team to perform only one attack (i.e., ID disclosure attack or trace inference attack) or no attack. This rule allows each team to tailor their strategy to suit their own goals, such as focusing on one specific award (e.g., best risk assessor).

3. Contest details

3.1 Data set

An artificial data (PWSCup2019 artificial data) used in the contest is newly-generated on the basis of the SNS-based people flow data [14] that is an open data available to the public. The SNS-based people flow data is a public data set that contains artificially generated traces covering Tokyo metropolitan area for six days (7/1, 7/7, 10/7, 10/13, 12/16, and 12/22 in 2013). The traces are generated on the basis of real trace data, but they do not contain any information that is easily collated with real users.

As shown in Fig. 4, we divide Tokyo metropolitan area (latitude from 35.65 to 35.75 and longitude from 139.68 to 139.8) equally into $32 \times 32 = 1024$ regions and assign region ID sequentially from lower-left to upper-right. The number of regions (location data) is $m = 1024$, and the size of each region is approximately 347m (height) $\times$ 341m (width) with an assumption that one degree longitude and latitude (in Tokyo area) correspond to 111km and 91km, respectively.

Next, we extract traces of all 10181 users who have more than 10 location data in Tokyo metropolitan area and use them as learning data to formulate a Markov model-based trace generator. Using the trace generator, we create a reference trace set $R^{(i,j)} = (r^{(i,j)}_1, \cdots, r^{(i,j)}_{2000})$ and an original trace set $O^{(i,j)} = (o^{(i,j)}_1, \cdots, o^{(i,j)}_{2000})$ for a user set $\mathcal{U}^{(i,j)} = \{u^{(i,j)}_1, \cdots, u^{(i,j)}_{2000}\}$ with $n = 2000$ users. These data sets are created independently per team $i \in [z]$ and

Table 1 PWSCup2019 artificial data

Number of users	$n = 2000$
Target area	Latitude: 35.65 to 35.75 Longitude: 139.68 to 139.8
Number of location data	$m = 1024$ (divided into 32×32 regions)
Length of trace	$t = 40$ (from 8 : 00 to 17 : 59 for 2 days with time interval of 30 minutes)

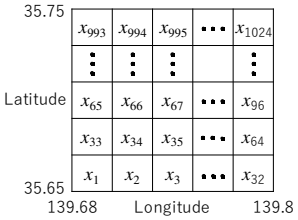

Fig. 4 Location data in the contest

per data set number $j \in \{1, 2\}$. The trace generator refers a feature of a user $u_k^{(i,j)}$ $(1 \leq k \leq 2000)$ when generating his reference trace $r_k^{(i,j)}$ and original trace $o_k^{(i,j)}$. $r_k^{(i,j)}$ and $o_k^{(i,j)}$ have high correlation, so teams can perform ID disclosure and trace inference attacks on public anonymized trace sets while referring to the corresponding reference trace sets.

The time is discretized by dividing the time from 8:00 to 17:59 in 30-minute interval. In the preliminary round, traces from the 1st and 2nd days are used for a reference trace, and traces from the 3rd and 4th days are used for an original trace. The length of each trace is thus $t = 40$, with time 1 representing 8:00 of the 1st day, time 40 representing 17:30 of the 2nd day, time 41 representing 8:00 of the 3rd day, and time 80 representing 17:30 of the 4th day. Table 1 illustrates the overview of PWSCup2019 artificial data. Note that the trace length in the table is for the preliminary round; we might change the length (i.e., the days covered in the traces) in the final round.

We will not disclose the detailed specification of our trace generator. The generator creates PWSCup2019 artificial data in the following manners:

(1) **Preserve population distribution:** The artificial data is generated in a way that the population distribution (i.e., the probability distribution on $m = 1024$ regions) in each hour from 6 o'clock to 17 o'clock is close to the distribution shown in the original SNS-based people flow data.

(2) **Preserve transition matrix:** The artificial data is generated in a way that the transition matrix of the Markov model (1024×1024 matrix) is closed to that of the original SNS-based transition flow data.

(3) **Model user's home:** The artificial data is generated in a way that most users stay in regions where their home reside in the morning (about 95% at 6 to 7 o'clock, and about 30% at 8 o'clock). A home region is assigned randomly to each user while preserving the population distribution. Note that if we include regions at 6 to 7 o'clock in the reference/original traces, the adversary could know the home regions for almost all of the users. Since this setting is too favorable to the adversary and also unrealistic, we include regions from 8 o'clock in the reference/original traces in our contest.

Increasing the number of users n will reduce the variation in utility and privacy scores among different data sets and thus will increase the fairness among teams. Making n too large, however, will make the computational complexity of data anonymizing, ID disclosure, and trace inference too difficult to handle and make the burden of teams too heavy. Considering the balance between fairness and the burden, we decide to set the user number to $n = 2000$. Refer our documentation [15] for more details on PWSCup2019 artificial data.

3.2 Data anonymization

Data is anonymized by processing location data and by shuffling traces (pseudonymization).

Processing location data: Team P_i processes nt location data in their original trace set $O^{(i,j)}$. In the contest, the allowed processing methods are "do nothing," "add noise," "generalize," and "delete."

(1) **Do nothing:** Use the original location data as is with no processing. For example, do $x_1 \to x_1$.

(2) **Add noise:** Exchange the original location data with another location data (i.e., another element in the set $\mathcal{X}$). For example, do $x_1 \to x_3$.

(3) **Generalize:** Exchange the original location data with a set of two or more location data. The set does not have to include the original location data. For example, do $x_1 \to \{x_1, x_3\}$ or $x_1 \to \{x_2, x_3, x_5\}$. The former operation is a generalization including the original location data while the latter operation is a generalization that does not include the original data information.

(4) **Delete:** Exchange the original location data with an empty set $\emptyset$. For example, do $x_1 \to \emptyset$.

If we denote a set of possible location data after anonymization as $\mathcal{Y}$, the set can be represented as a power set of $\mathcal{X}$ $(\mathcal{Y} = 2^{\mathcal{X}})$. This means that we accept all possible operation on each location data in our contest.

Figure 5 illustrates an example of data anonymization. The original trace set $O^{(i,1)}$ in the figure is the same as that in Fig. 1. Each trace set is depicted as a table with a user ID/pseudo ID, time, and region ID. The generalization and deletion are shown as a list of region ID (space separated) and an asterisk (*), respectively. For example, "2 4 5" means $\{x_2, x_4, x_5\}$.

Shuffling traces (pseudonymization): The judge Q shuffles the anonymized trace set $A^{(i,j)}$ to pseudonymize the traces. More specifically, the judge replaces user IDs $(1, 2, \cdots, n)$ randomly and assign pseudo IDs sequentially from $n + 1$ to $2n$. For example, pseudo ID $2001, 2002, 2003$ correspond to user ID $2, 3, 1$ in Fig. 5.

Anonymized trace set $A^{(i,j)}$ and public anonymized trace set $A'^{(i,j)}$: $A^{(i,j)}$ and $A'^{(i,j)}$ are defined as:

- $A^{(i,j)} = (a_1^{(i,j)}, \cdots, a_n^{(i,j)})$, and
- $A'^{(i,j)} = (a'^{(i,j)}_1, \cdots, a'^{(i,j)}_n)$.

where $a_k^{(i,j)} \in \mathcal{Y}^t$ is an anonymized trace of user ID k $(k \in [n])$ and $a'^{(i,j)}_k \in \mathcal{Y}^t$ is an anonymized trace of pseudo ID k $(k \in \{n+1, \ldots, 2n\})$.

In Fig. 5, for example, we can see that:

- $a_1^{(i,1)} = (x_2, x_3, \{x_2, x_4, x_5\}, \emptyset)$,
- $A^{(i,1)} = (a_1^{(i,1)}, a_2^{(i,1)}, a_3^{(i,1)})$,

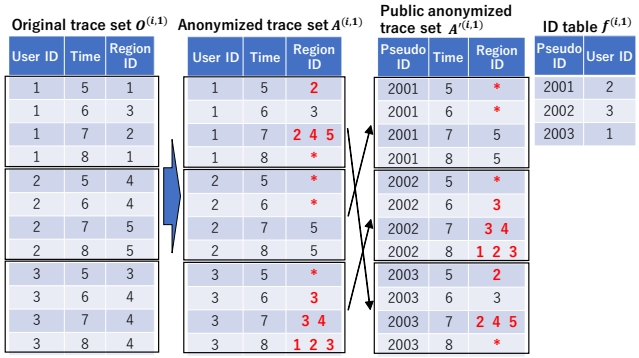

Fig. 5 An example of data anonymizing

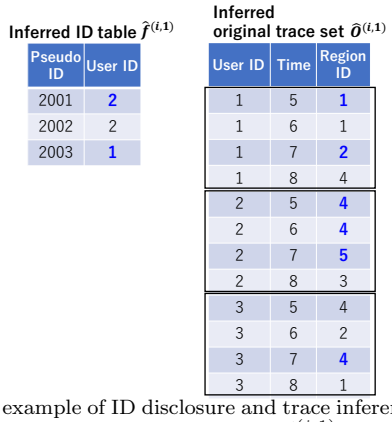

Fig. 6 An example of ID disclosure and trace inference results on the public anonymized trace set $A'^{(i,1)}$ in Fig. 5. User and region ID in blue match exactly with the original data.

- $a'^{(i,1)}_{2001} = (\emptyset, \emptyset, x_5, x_5)$, and
- $A'^{(i,1)} = (a'^{(i,1)}_{2001}, a'^{(i,1)}_{2002}, a'^{(i,1)}_{2003})$

ID table $f^{(i,j)}$: An ID table $f^{(i,j)}$ is a pair of pseudo ID and user ID. For example, $f^{(i,j)} = \{(2001, 2), (2003, 3), (2003, 1)\}$ in Fig 5.

3.3 ID disclosure and trace inference

In an ID disclosure and trace inference attempt, each team derives an inferred ID table $\hat{f}^{(i,j)}$ and an inferred original trace set $\hat{O}^{(i,j)}$ from a public anonymized trace set $A'^{(i,j)}$ using a reference trace set $R^{(i,j)}$ as a hint. Figure 6 shows an example of ID disclosure and trace inference attempt.

Inferred ID table $\hat{f}^{(i,j)}$: $\hat{f}^{(i,j)}$ is a table summarizing inferred user ID for each pseudo ID. Just as the ID table $f^{(i,j)}$, $\hat{f}^{(i,j)}$ is a set of pairs whose components are pseudo ID and user ID. Note that each user ID appears only once in the ID table $f^{(i,j)}$, but the same user ID can appear multiple times in the inferred ID table $\hat{f}^{(i,j)}$. In Fig. 6, for example, user ID "2" appears twice in $\hat{f}^{(i,j)} = \{(2001, 2), (2002, 2), (2003, 1)\}$.

Inferred original trace set $\hat{O}^{(i,j)}$: $\hat{O}^{(i,j)}$ is an inferred trace set that aligns inferred traces sequentially by user ID. In other words, $\hat{O}^{(i,j)} = (\hat{o}^{(i,j)}_1, \cdots, \hat{o}^{(i,j)}_n)$, where $\hat{o}^{(i,j)}_k \in \mathcal{X}^t$ is an inferred trace of user ID k ($k \in [n]$). In Fig. 6, for example, $\hat{o}^{(i,1)}_1 = (x_1, x_1, x_2, x_4)$ and $\hat{O}^{(i,1)} = (\hat{o}^{(i,1)}_1, \cdots, \hat{o}^{(i,1)}_3)$.

3.4 Utility score

A utility score $s^{(i,j)}_U \in [0, 1]$ is calculated from an original trace set $O^{(i,j)}$ and anonymized trace set $A^{(i,j)}$. Anonymized trace sets can be used in various purposes. To accommodate a variety of purposes, we adopt a general utility score in the contest.

Some examples of anonymized trace set applications include analyzing popular spots [4], auto-tagging POI categories [5], and analyzing the movement of foreign tourists [6]. Another example is an LBS provider model in which users do not trust the LBS provider; a user first anonymizes their location data and then sends the information to the LBS provider to receive a service like getting a POI search result (e.g., find restaurants near me) [8].

For all examples, the utility degrades as the location data is anonymized more. The utility will be lost entirely when the information is anonymized beyond a certain level. Taking an example of a POI search, a user will not be able to retrieve the POI around the original location if too much noise is added to the location data (e.g., shifting the location more than 5km) or if the location data is deleted.

We define the utility score based on these assumptions. First, we define a function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ that takes two location data $x_k, x_l \in \mathcal{X}$ as an input and their Euclidean distance $d(x_k, x_l) \in \mathbb{R}_{\geq 0}$ as an output. The location data x_k and x_l in our contest are regions as shown in Fig. 4, so we define $d(x_k, x_l)$ as an Euclidean distance between a center point of region x_k and region x_l. Since the size of each region is 347m (height) × 341m (width), $d(x_1, x_1) = 0$m, $d(x_1, x_2) = 341$m, and $d(x_1, x_{33}) = 347$m.

Next, we calculate the Euclidean distances between nt location data in the original trace set $O^{(i,j)}$ and corresponding nt location data in the anonymized trace set $A^{(i,j)}$. If the location data is deleted, the distance is interpreted as $r \in \mathbb{R}_{\geq 0}$. We denote the average of the Euclidean distance between l-th location data ($l \in [t]$) in the original and anonymized traces for user ID k ($k \in [n]$) as $c_{k,l} \in \mathbb{R}_{\geq 0}$. In Fig. 5, for example, we can see that: $c_{1,1} = d(x_1, x_2)$, $c_{1,2} = d(x_3, x_3) = 0$, $c_{1,3} = \frac{d(x_2,x_2)+d(x_2,x_4)+d(x_2,x_5)}{3}$, and $c_{1,4} = r$.

Finally, we use a piecewise linear function g shown at the left of Fig. 7 to transform each $c_{k,l}$ into a score value from 0 to 1 (the score for the deleted location data is treated as 0) and calculate the utility score $s^{(i,j)}_U$ by taking the average of nt scores.

In other words, we define function $g : \mathbb{R}_{\geq 0} \to [0, 1]$ as a function that takes $c \in \mathbb{R}_{\geq 0}$ as an input and calculate the following score $g(c)$ as an output:

$$g(c) = \begin{cases} 1 - \frac{c}{r} & (\text{if } x < r) \\ 0 & (\text{if } x \geq r) \end{cases} \tag{1}$$

Then, use this function as follow to calculate the utility score $s^{(i,j)}_U$:

$$s^{(i,j)}_U = \frac{1}{nt} \sum_{k=1}^{n} \sum_{l=1}^{t} g(c_{k,l}) \tag{2}$$

The piecewise function g is designed so that the score becomes 0 (i.e., the utility is completely lost) if location data is anonymized in a way that their Euclidean distances are

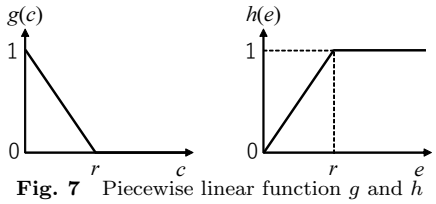

Fig. 7 Piecewise linear function g and h

larger than or equal to r (or the information is deleted). We set r as 2km in the contest; we did an experiment and confirmed that the utility score $s_U^{(i,j)}$ is highly correlated (the correlation coefficient of 0.9 or more) with the precision of a POI search application mentioned earlier. The details of this experiment are omitted in this paper.

3.5 Privacy score

Privacy score against ID disclosure $s_I^{(i,j)}$: $s_I^{(i,j)} \in [0,1]$ is calculated by comparing an inferred ID table $\hat{f}^{(i,j)}$ with the corresponding ID table $f^{(i,j)}$.

In the contest, we calculate $s_I^{(i,j)}$ by subtracting the rate of successful ID disclosure from 1. In other words, we calculate the privacy score $s_I^{(i,j)}$ as:

$$s_I^{(i,j)} = 1 - \alpha^{(i,j)} \tag{3}$$

where $\alpha^{(i,j)} \in [0,1]$ is the rate of successful ID disclosure that is derived by comparing $\hat{f}^{(i,j)}$ with $f^{(i,j)}$. In examples shown in Fig. 5 and Fig. 6, $\alpha^{(i,j)} = \frac{2}{3}$ and $s_I^{(i,j)} = 1 - \frac{2}{3} = \frac{1}{3}$.

Privacy score against trace inference $s_T^{(i,j)}$: $s_T^{(i,j)} \in [0,1]$ is calculated by comparing an inferred original trace set $\hat{O}^{(i,j)}$ with the corresponding original trace set $O^{(i,j)}$.

In the contest, we calculate the Euclidean distances between nt location data in an inferred original trace set $\hat{O}^{(i,j)}$ and the corresponding nt location data in the original trace set $O^{(i,j)}$ and use these distances to derive the privacy score against trace inference attack. Let $e_{k,l} \in \mathbb{R}_{\geq 0}$ be an Euclidean distance between l-th location data ($l \in [t]$) in the original and inferred original traces for user ID $k \in [n]$. In Fig. 5 and Fig. 6, for example, $e_{1,1} = d(x_1, x_1)$, $e_{1,2} = d(x_3, x_1)$, $e_{1,3} = d(x_2, x_2)$, and $e_{1,4} = d(x_1, x_4)$.

We can say that the privacy level is higher if the Euclidean distance between an original and inferred location data is bigger. We can also say the trace inference is completely failed if the distance goes over beyond a certain value (e.g., over 5km). Based on these assumptions, we use a piecewise linear function h shown at the right of Fig. 7 to transform a distance $e_{k,l}$ into a score value from 0 to 1.

There can be some sensitive location data, such as hospitals, that need to be handled carefully in traces. To accommodate such a need, we adjust the privacy score against trace inference $s_T^{(i,j)}$ by taking a weighted average of nt scores in a way that regions that have hospital POI (hospital regions) weight 10 times more than other regions. Figure 8 shows regions including POI of the "hospital" category extracted from a POI data in the SNS-based people flow data [14]. We set these 37 regions marked by blue dots as "hospital" regions.

Now we formulate how to calculate the privacy score against trace inference $s_T^{(i,j)}$. We define a function h : $\mathbb{R}_{\geq 0} \to [0,1]$ as the function that takes $e \in \mathbb{R}_{\geq 0}$ as an input and calculate the following score $h(e)$ as an output:

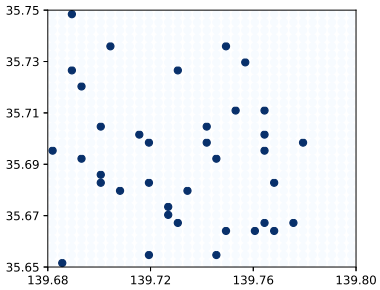

Fig. 8 Region with hospital category POI (blue dots, total 375 regions

$$h(e) = \begin{cases} \frac{e}{r} & (\text{if } x < r) \\ 1 & (\text{if } x \geq r) \end{cases} \tag{4}$$

The privacy score against trace inference $s_T^{(i,j)}$ is calculated with the following equation:

$$s_T^{(i,j)} = \frac{\sum_{k=1}^{n} \sum_{l=1}^{t} w_{k,l} h(e_{k,l})}{\sum_{k=1}^{n} \sum_{l=1}^{t} w_{k,l}} \tag{5}$$

where $w_{k,l} \in \{1, 10\}$ is a weight variable whose value is:

- 10 if the l-th location data ($l \in [t]$) in user ID k's trace ($k \in [n]$) in the original trace set $O^{(i,j)}$ is a hospital region, and
- 1 if otherwise.

4. Conclusion

This paper explores the PWS Cup 2019 contest focusing on the location data anonymization.

Acknowledgments This research was funded by JSPS Research Grants (Kakenhi No. 18H04099 and No. 19H04113).

References

[1] C. C. Aggarwal, P. S. Yu, Privacy-Preserving Data Mining, Springer, 2008. Springer

[2] Personal Information Protection Commission, Guidelines for the Act on the Protection of Personal Information (Anonymously Processed Information): `https://www.ppc.go.jp/files/pdf/guidelines04.pdf` (in Japanese).

[3] Article 29 Data Protection Working Party, "Opinion 05/2014 on Anonymisation Techniques," WP 216, 2014.

[4] Y. Zheng *et al.*, "Mining interesting locations and travel sequences from GPS trajectories," Proc. WWW'09, pp.791–800, 2009.

[5] M. Ye *et al.*, "On the Semantic Annotation of Places in Location-Based Social Networks," Proc. KDD'11, pp.520–528, 2011.

[6] FY 2016 Report of the Project for Promoting Revisiting Foreign Tourists in Hokkaido (Hokkaido LOVERS Expansion Project): `https://www.visit-hokkaido.jp/company/material/detail/44` (in Japanese).

[7] S. Gambs *et al.*, "De-anonymization attack on geolocated data," Journal of Computer and System Sciences, vol.80, no.8, pp.1597-1614, 2014.

[8] M. E. Andrés *et al.*, "Geo-Indistinguishability: Differential Privacy for Location-based Systems," Proc. CCS'13, pp.901–914, 2013.

[9] Y.-A. Montjoye *et al.*, "Unique in the Crowd: The privacy bounds of human mobility," Scientific Reports, vol.3,

no.1376, pp.1–5, 2013.
[10] R. Shokri *et al.*, "Quantifying location privacy," Proc. IEEE S&P'11, pp.247–262, 2011.
[11] A. Machanavajjhala *et al.*, "L-diversity: privacy beyond k-anonymity," Proc. ICDE'06, pp.24–35, 2006.
[12] PWS Cup 2019 Sample Programs: `https://www.iwsec.org/pws/2019/cup19-sample-e.pdf`
[13] PWS Cup 2019 Contest Rule: `https://www.iwsec.org/pws/2019/cup19-rule-e.pdf`
[14] Nightley and Center for Spatial Information Science (CSIS), SNS-based People Flow Data, [online] Available: `http://nightley.jp/archives/1954`.
[15] PWS Cup 2019 data set: `https://www.iwsec.org/pws/2019/cup19-dataset-e.pdf`