PWS Cup 2019 Contest Rule Ver1.2

PWS Cup Organizing Committee

September 3, 2019

1 Introduction

The notations used in this rule, the detailed information of the algorithms, and the data set are given in the following documents.

- PWS Cup 2019 rule paper submitted to CSS2019 [1].
- PWS Cup2019 artificial data [3] that is newly-generated on the basis of the SNS-based people flow data [2].

2 Contest rule

2.1 Overview

1. Each team is to submits designated data (anonymized trace sets in the data anonymizing phase, inferred ID tables and inferred original trace sets in the ID disclosure & trace inference phase) following the contest flow described in Section 2.2 of the PWS Cup 2019 rule paper [1].
2. Each team is to follow the data format and file name rules described in Section 2.4 and 2.5 in this paper when submitting their data.
3. Each team is not to perform the prohibited actions described in Section 2.6 in this paper.

2.2 Contest flow

Refer Section 2.2 of the PWS Cup 2019 rule paper [1].

2.3 Awards

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.

1. **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 $s_{I,min}^{(i,1)} + s_{T,min}^{(i,2)}$, where $s_{I,min}^{(i,1)}$ is $A'^{(i,1)}$'s privacy score against ID disclosure attack and $s_{T,min}^{(i,2)}$ 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.

2. **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 ($s_{I,min}^{(i,1)}$).

3. **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 ($s_{T,min}^{(i,2)}$).

4. **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. We will select this prize from all teams except the overall winner and the best data anonymizer (ID disclosure).

   More specifically, the prize will be given to the team who marks the lowest $s_I^* + \sum_{i=1}^{z} \sum_{j=1}^{2} s_I^{(i,j)}$, where s_I^* and $s_I^{(i,j)}$ are the privacy score against ID disclosure of the pseudonymized trace set presented by the committee (refer Section 2.3.6) and of the public anonymized trace sets submitted by other teams ($A'^{(i,j)}$), respectively. Note that public anonymized trace sets for trace inference challenge are also included in the summation of $s_I^{(i,j)}$. The privacy score is counted as 1 for all valid public anonymized trace sets that were not attacked.

   [Implication] The team who attempts many ID disclosure attacks against public anonymized trace sets (both for ID disclosure and trace inference challenges) will have high possibility of winning this prize.

5. **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. We will select this prize from all teams except the overall winner and the best data anonymizer (trace inference).

   More specifically, the prize will be given to the team who marks the lowest $s_T^* + \sum_{i=1}^{z} \sum_{j=1}^{2} s_T^{(i,j)}$, where s_T^* and $s_T^{(i,j)}$ are the privacy score against trace inference of the pseudonymized trace set presented by the committee (refer Section 2.3.6) and of the public anonymized trace sets submitted by other teams ($A'^{(i,j)}$), respectively. Note that public anonymized trace sets for ID disclosure challenge are also included in the summation of $s_T^{(i,j)}$. The privacy score is counted as 1 for all valid public anonymized trace sets that were not attacked.

   [Implication] The team who attempts many trace inference attacks against public anonymized trace sets (both for ID disclosure and trace inference challenges) will have high possibility of winning this prize.

6. **Pseudonymized trace set presented by the committee**

The pseudonymized trace set is generated by the PWS Cup 2019 committee. The set is generated from an original trace set having a user set that is different from the ones contained in the original trace sets distributed to participating teams. The pseudonymized trace set is generated with no location data processing; the committee just pseudonymizes the original trace set.

The committee will distribute this pseduonymized trace set and the corresponding reference trace set to all teams in the ID disclosure & trace inference phase. Each team is expected to attempt ID disclosure and trace inference attacks against the pseudonymized trace set. As already described in the previous sections, the best risk assessor (ID disclosure/trace inference) will be selected with the consideration of the privacy score of the pseudonymized trace sets against ID disclosure (s_I^*) and trace inference (s_T^*). Note that s_I^* and s_T^* will be counted as 1 if no ID disclosure or trace inference attack is made to the pseudonymized trace set.

7. **Best presenter**

   Each team will give an oral presentation in CSS2019 venue to explain the algorithms used for anonymizing their data and for performing ID disclosure and trace inference attacks. Each team is also expected to make a poster that summarizes the details of their algorithms.

   Multiple examiners selected by the judge Q will vote for the best presenter, and the team who receives the most votes will receive the best presenter prize.

   The judge Q will later collect oral and poster presentation materials. Consult the organizing committee if your team cannot make a presentation in the venue or cannot submit their presentation materials for some reasons.

2.4 Data format

- **User ID:** In this contest, we use reference trace set $R^{(i,j)}$ and original trace set $O^{(i,j)}$ covering a set of 2000 users. A user set is different for each team i ($1 \leq i \leq z$) and for data set j ($1 \leq j \leq 2$). A user ID, a natural number from 1 to 2000, is assign to each user in the original trace set $O^{(i,j)}$.
- **Pseudo ID:** A natural number from 2001 to 4000 assigned to a public anonymized trace set $A'^{(i,j)}$.
- **Region ID:** Tokyo metropolitan area (latitude from 35.65 to 35.75 and longitude from 139.68 to 139.8) is equally divided into $32 \times 32 = 1024$ regions. Region ID is assigned to each region sequentially from lower-left to upper-right; Region ID is thus a natural number from 1 to 1024.
- **Time:** 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. Traces from the 3rd and 4th days are used for an original trace. The length of each trace is thus $t = 40$ (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 2nd day). Note that we might change the length of the reference and original traces from 2-day long in the final round.
- **Reference trace set $R^{(i,j)}$ and original trace set $O^{(i,j)}$:** A csv file with user ID, time, and region ID. The first line has a header "user_id,time_id,reg_id." The second line and after cover a

list of user ID, time, and region ID in ascending order of (user ID, time).
For example, the original trace set in Fig. 1 is represented as follows:

```
user_id,time_id,reg_id
1,5,1
1,6,3
1,7,2
1,8,1
...
3,8,4
```

- **Anonymized trace set $A^{(i,j)}$:** A csv file with processed region ID listed in ascending order of (user ID, time). Note that user ID and time themselves are not stored in the file. The first line has a header "reg_id." The second line and after cover processed region ID in ascending order of (user ID, time).
  The possible data processing methods are adding noise, generalizing, and deletion. The generalized and deleted data are to be represented as a list of region IDs (space separated) and an asterisk (*), respectively.
  For example, the anonymized trace set in Fig. 1 is represented as follows:

```
reg_id
2
3
2 4 5
*
...
1 2 3
```

  We can see that a noise is added to user ID 1's trace at time 5 to shift the region as $x_1 \to x_2$. The same user's trace is generalized at time 7 to make the region vague as $x_2 \to \{x_2, x_4, x_5\}$. Finally, the same user's trace is deleted at time 8 to hide the region x_1.
- **Public anonymized trace set $A'^{(i,j)}$:** A csv file with pseudo ID, time, and processed region ID. The first line has a header "pse_id,time_id,reg_id." The second line and after cover a list of pseudo ID, time, and processed region ID in ascending order of (pseudo ID, time).
  For example, the public anonymized trace set in Fig. 1 is represented as follows:

```
pse_id,time_id,reg_id
2001,5,*
2001,6,*
2001,7,5
2001,8,5
...
```

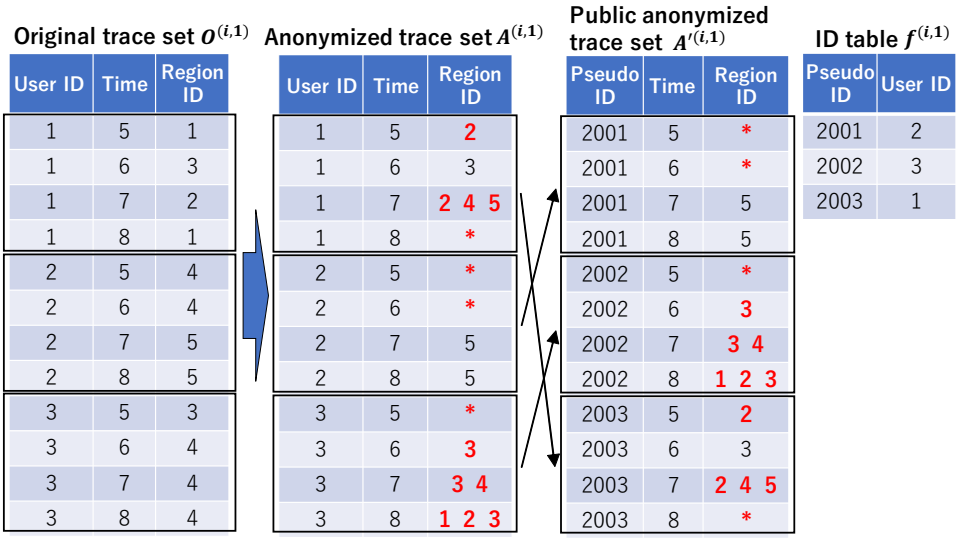

Figure1 An example of data anonymizing. The generalization and deletion are shown as a list of region ID (space separated) and an asterisk (*), respectively. For example, "2 4 5" means $\{x_2, x_4, x_5\}$

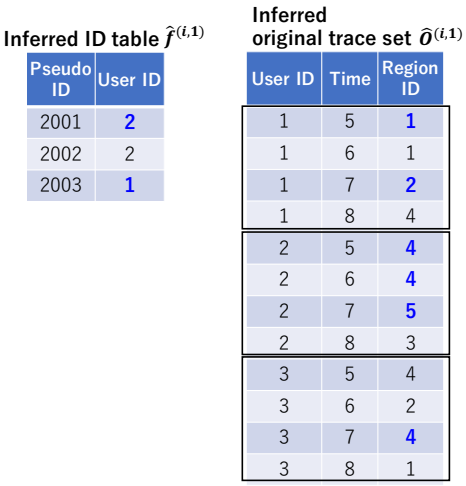

Figure2 An example of ID disclosure and trace inference results on the public anonymized trace set $A'^{(i,1)}$ in Fig. 1. User and region ID in blue match exactly with the original data.

```
2003,8,*
```

- **ID table $f^{(i,j)}$:** A csv file with pseudo ID and user ID. The first line has a header "pse_id,user_id." The second line and after cover a list of pseudo ID and user ID in ascending order of pseudo ID. For example, the ID table in Fig. 1 as represented as follows:

```
pse_id,user_id
2001,2
2002,3
```

```
2003,1
```

- **Inferred ID table $\hat{f}^{(i,j)}$:** A csv file with inferred user ID listed in ascending order of pseudo ID. Note that pseudo ID itself is not stored in the file. The first line has a header "user_id." The second line and after cover a list of inferred user ID in ascending order of pseudo ID.
  For example, the inferred ID table in Fig. 1 is represented as follows:

```
user_id
2
2
1
```

- **Inferred original trace set $\hat{O}^{(i,j)}$:** A csv file with inferred region ID listed in ascending order of (user ID, time). Note that user ID and time themselves are not stored in the file. The first line has a header "reg_id." The second line and after cover inferred region ID in ascending order of (user ID, time).
  For example, the inferred original trace set in Fig. 1 is represented as follows:

```
reg_id
1
1
2
4
...
1
```

2.5 File name

1. **For all files:**
   - A 3-digit number XXX represents a team number i ($1 \leq i \leq z$) of the team who anonymizes the data.
   - A 2-digit number YY represents a data set number j ($1 \leq j \leq 2$) which is either 01 or 02.
   - A 3-digit number WWW represents a team number i' ($i' \neq i$) who attacks the data (either ID disclosure attack or trace inference attack).
   - A flag ZZZ indicates if the data is anonymized for ID disclosure or trace inference challenge. The flag is to be added in front of an extension (.csv). If the data set number is $j = 1$, add the flag ZZZ=IDP (IDP is an abbreviation for "ID Protection"). If the data set number is $j = 2$, add the flag ZZZ=TRP (TRP is an abbreviation for "Trace Protection").
2. **File name of a reference trace set $R^{(i,j)}$: reftraces_teamXXX_dataYY_ZZZ.csv**
   For example, the file name should be:
   - reftraces_team012_data01_IDP.csv for the team number $i = 12$ and the data set number $j = 1$.

- reftraces_team012_data02_TRP.csv for the team number $i = 12$ and the data set number $j = 2$.

3. **File name of an original trace set $O^{(i,j)}$: orgtraces_teamXXX_dataYY_ZZZ.csv**
   For example, the file name should be:
   - orgtraces_team012_data01_IDP.csv for the team number $i = 12$ and the data set number $j = 1$.
   - orgtraces_team012_data02_TRP.csv for the team number $i = 12$ and the data set number $j = 2$.
4. **File name of an anonymized trace set $A^{(i,j)}$: anotraces_teamXXX_dataYY_ZZZ.csv**
   For example, the file name should be:
   - anotraces_team012_data01_IDP.csv for the team number $i = 12$ and the data set number $j = 1$.
   - anotraces_team012_data02_TRP.csv for the team number $i = 12$ and the data set number $j = 2$.
5. **File name of a public anonymized trace set $A'^{(i,j)}$: pubtraces_teamXXX_dataYY_ZZZ.csv**
   For example, the file name should be:
   - pubtraces_team012_data01_IDP.csv for the team number $i = 12$ and the data set number $j = 1$.
   - pubtraces_team012_data02_TRP.csv for the team number $i = 12$ and the data set number $j = 2$.
6. **File name of an ID table $f^{(i,j)}$: ptable_teamXXX_dataYY_ZZZ.csv**
   For example, the file name should be:
   - ptable_team012_data01_IDP.csv for the team number $i = 12$ and the data set number $j = 1$.
   - ptable_team012_data02_TRP.csv for the team number $i = 12$ and the data set number $j = 2$.

   Note that only the judge Q can view the ID table file.
7. **File name of an inferred ID table $\hat{f}^{(i,j)}$: etable_teamWWW-XXX_dataYY_ZZZ.csv**
   Concatenate the team number i' (the team who attacks the data) and the team number i (the team whose data is being attacked) with a hyphen. For example, the file name should be:
   - etable_team020-012_data01_IDP.csv if the team number $i' = 20$ performs ID disclosure attack against the team number $i = 12$'s public anonymized trace set (data set number $j = 1$).
   - etable_team020-012_data02_TRP.csv if the team number $i' = 20$ performs ID disclosure attack against the team number $i = 12$'s public anonymized trace set (data set number $j = 2$).
8. **File name of an inferred original trace set $\hat{O}^{(i,j)}$: etraces_teamWWW-XXX_dataYY_ZZZ.csv**
   Concatenate the team number i' (the team who attacks the data) and the team number i (the team whose data is being attacked) with a hyphen. For example, the file name should be:
   - etraces_team020-012_data01_IDP.csv if the team number $i' = 20$ performs trace inference attack against the team number $i = 12$'s public anonymized trace set (data set number $j = 1$).
   - etraces_team020-012_data02_TRP.csv if the team number $i' = 20$ performs trace inference attack against the team number $i = 12$'s public anonymized trace set (data set number $j = 2$).

2.6 Prohibited actions

2.6.1 Prohibited actions in data anonymizing phase

A team could be disqualified if they violate the following prohibitions.

1. A team submits more data than the specified limit number.
2. A team submits data that do not follow the specified format.
3. A team colludes with other teams. Note that sharing programs, modules and algorithms among teams are allows as long as the information disclosed only to the team is not inferred by other teams.

2.6.2 Prohibited actions in ID disclosure and trace inference phase

A team could be disqualified if they violate the following prohibitions.

1. A team submits data that do not follow the specified format.
2. A team colludes with other teams. Note that sharing programs, modules and algorithms among teams are allows as long as the information disclosed only to the team is not inferred by other teams.
3. A team submits more than two inferred ID tables for a public anonymized trace set (i.e., attempting more than two ID disclosure attacks on the same public anonymized trace set).
4. A team submits more than two inferred original trace sets for a public anonymized trace set (i.e., attempting more than two trace inference attacks on the same public anonymized trace set).

2.6.3 Prohibited actions for organizing committee members

1. A committee member colludes with teams (i.e., leaking information that is obtained using the privilege of the organizing committee).
2. A committee member hides any information while knowing that the information will be advantageous to teams.
3. A committee member leverages their privilege during the data submission period when he participates the contest as one of competitors (e.g., using the knowledge obtained using the organizing committee privileges when anonymizing data and when executing ID disclosure and trace inference attacks).

2.7 Open information

The judge makes the following information available to all teams.

2.7.1 Sample programs

Sample programs for anonymizing data, executing ID disclosure attack, and executing trace inference attack are available. Refer the documentation [4] for the details.

2.7.2 Utility assessment program

This program takes an original trace set $O^{(i,j)}$ and an anonymized trace set $A^{(i,j)}$ as an input. The program gives the utility score $s_U^{(i,j)}$ as an output.

2.7.3 Privacy score (against ID disclosure attack) assessment program

This program takes an ID table $f^{(i,j)}$ and an inferred ID table $\hat{f}^{(i,j)}$ as an input. The program gives the privacy score against ID disclosure attack $s_I^{(i,j)}$ as an output.

2.7.4 Privacy score (against trace inference attack) assessment program

This program takes an original trace set $O^{(i,j)}$ and an inferred original trace set $\hat{O}^{(i,j)}$ as an input. The program gives the privacy score against trace inference attack $s_T^{(i,j)}$ as an output.

2.7.5 Region assignment file

This file describes the information on each region (region ID, ID in a y-axis (latitude), ID in a x-axis (longitude), a latitude of the center of the region, a longitude of the center of the region, if the region is a hospital region (1: yes, 0: no)).

The content of the file is as follows (the first line is a header):

```
reg_id,y_id,x_id,y(center),x(center),hospital
1,1,1,34.6415625,135.441875,0
2,1,2,34.6415625,135.445625,1
3,1,3,34.6415625,135.449375,0
4,1,4,34.6415625,135.453125,0
...
1024,32,32,34.7384375,135.558125,0
```

2.7.6 Time assignment file

This file describes the information on each time (if the time belongs to a reference trace set (ref) or an original trace set (org), time (a natural number), day, hour, minute).

The content of the file is as follows (the first line is a header):

```
ref/org,time_id,day,hour,min
ref,1,1,8,0
ref,2,1,8,30
ref,3,1,9,0
```

```
ref,4,1,9,30
...
org,80,4,17,30
```

2.7.7 PWSCup2019 artificial data (Osaka data set)

This data is generated on the basis of the the SNS-based people flow data [2]. Just like the Tokyo data set used in the contest, Osaka metropolitan area (latitude from 34.64 to 34.74 and longitude from 135.44 to 135.56) in the SNS-based people flow data is divide equally into 32×32 regions and used as learning data to formulate a Markov model-based trace generator. The trace generator is then used to create reference trace sets $R^{(i,j)}$ and original trace sets $O^{(i,j)}$ for two teams (i.e., $1 \leq i \leq 2$, $1 \leq j \leq 2$). These trace sets are made open to all teams so that team members can understand how the reference and original trace sets look like.

We also conduct several experiments using this data set with sample programs. Refer the documentation [4] for more details.

2.7.8 PWSCup2019 artificial data (Tokyo data set)

Reference and original trace sets for two "imaginary" teams are made open to all teams during the data anonymizing phase so that a team can compare their trace sets with the sets distributed to other teams. User sets in these trace sets are different from the ones contained in the data sets distributed to participating teams.

In other words, we will create and publish reference trace sets $R^{(i,j)}$ and original trace sets $O^{(i,j)}$ for team $z+1$ and $z+2$ (i.e., $z+1 \leq i \leq z+2$, $1 \leq j \leq 2$) where z is the number of teams participating in the contest.

2.8 Preliminary round

The following rules apply to the preliminary round.

1. The requirement value for the utility score is set as follows: $s_{req} = 0.7$.
2. After the preliminary round, we will fully disclose original trace sets, anonymized trace sets, ID tables, and all inferred ID tables submitted by all teams.

2.9 Additional rules in the final round

The following rules might be added in the final round of the contest, after the preliminary round is finished. If we decide to add them, we will make an announcement to all teams before we start the final round.

1. The requirement value s_{req} of the utility score could be changed from the one of the preliminary round.

2. The length of reference and original trace sets could be changed from the ones of the preliminary round (i.e., 2-day long for each trace sets).

References

[1] Takao Murakami, Hiromi Arai, Makoto Iguchi, Hidenobu Oguri, Hiroaki Kikuchi, Atsushi Kuromasa, Hiroshi Nakagawa, Yuichi Nakamura, Kenshiro Nishiyama, Ryo Nojima, Takuma Hatano, Koki Hamada, Yuji Yamaoka, Takayasu Yamaguchi, Akira Yamada, Chiemi Watanabe, "PWS Cup 2019: Location Data Anonymization Competition," CSS2019.

[2] Nightley and Center for Spatial Information Science (CSIS), SNS-based People Flow Data, [online] Available: `http://nightley.jp/archives/1954`.

[3] PWS Cup 2019 Data Set: `https://www.iwsec.org/pws/2019/cup19-dataset-e.pdf`

[4] PWS Cup 2019 Sample Program: `https://www.iwsec.org/pws/2019/cup19-sample-e.pdf`