PWS CUP 2017 Rules – The Data Anonymization and Re-identification Competition Rules

PWS CUP Working Group

Ver. 1.3
October 17, 2017

Note: The technical details of algorithms and definitions are described in the following papers.

[1] H. Kikuchi, H. Oguri, H. Nakagawa, R. Nojima, T. Hatano, H. Hamada, T. Murakami, M. Monda, Y. Yamaoka, A. Yamada, C. Watanabe, PWSCUP2017: Evaluate re-identification risk on long time history data, Privacy Workshop 2017, IPSJ, 2017 (in Japanese).

[2] UCI Machine Learning Repository, Online Retail Data Set (`https://archive.ics.uci.edu/ml/datasets/Online+Retail`)

Rules are subject to change without prior notice.

Rules

1. In competition, data anonymizing players, data re-identifying players, and the judge are involved.
2. (Data Anonymizing Player) is given customer master table M, and transaction record table T. Each anonymizing player performs an anonymization algorithm A, e.g., replacing identities by pseudonyms, suppression of records, perturbation of date, or, swapping goods, to produce anonymized table $A(T)$ which will be submitted to the judge.
3. (Judge) generates the pseudonym table f which specifies a mapping from the set of customer identities and the set of pseudonyms based on the relationship between T and $A(T)$. After deleting specified records from $A(T)$, the judge publishes randomly permuted version of $A(T)$, as anonymized data S. Note that f is hidden to data re-identifying players.
4. (Data Re-identifying Players) estimate the hidden pseudonym table based on the received S and some part of T and submit the estimated pseudonym table $\hat{f}$.
5. (Data Set)
   (a) Transaction record table T and customer master table M are sampled from [2]. The public description is available in [1]. There are unique 500 customers in M.
   (b) T and M are available from official PWSCUP 2017 website (`https://pwscup.personal-data.biz/web/pws2017/index.php`).
   (c) Data set are public information. Researchers are allowed to distribute it and to publish paper

using data if [1] and [2] are cited as reference.

(d) Transaction record table T are partitioned into twelve monthly tables $T^1, \ldots, T^{(12)}$

(e) Partial knowledges $T_{\alpha_1}, \ldots, T_{\alpha_4}$ of the transaction record table are generated with sampling rate $\alpha_1 = 0.25, \alpha_2 = 0.5, \alpha_3 = 0.75, \alpha_4 = 1.0$ and will be provided to the re-identifying players. ~~Note that there are exact n customers for all subtables $T_{\alpha_1}, T_{\alpha_2}, T_{\alpha_3}$.~~ All stock codes are replaced with the first-two digits.

6. (PPC Rule Article #19) Japanese privacy commission, the Protection of Personal Information (PPC) publishes the enforcement rules in 2016 (`https://www.ppc.go.jp/files/pdf/PPC_rules.pdf`). In the competition, algorithms used by any data anonymizing player need to be satisfied the rules as follows:

   (a) (iii) (deleting those codes linking mutually plural information handled by business operator) Delete all customer identities $c_{\cdot,1}$ and assign pseudonyms $s_{\cdot,1}$ that does not belong to the set of original customer identities.

   (b) (i) (deleting descriptions which identify a specific individual), (iv) (deleting idiosyncratic descriptions), (v) (taking appropriate action). It is up to players. Arbitrary processing can be taken and players are required to present the action that they take and the interpretation of the rules at the final presentation. Note there is no additional point for talks. Be relax.

7. (Process Format)

   (a) Deleting records from transaction record table T as follows:

   `[17551,0,2010/12/15,14:12,22693,1.25,24]` → `[DEL,,,,,,]`

   Note that the total number of rows of $A(T)$ is identical to T. Record which begin with "`DEL`" is deleted even if the other columns has any values.

   (b) Let $A(T^{(\ell)})$ be the anonymized table for ℓ-th monthly transaction record $T^{(\ell)}$. The whole anonymized table $A(T)$ is the concatenation of all monthly anonymized table, i.e.,

   $$A(T) = \begin{pmatrix} A(T^{(1)}) \\ \vdots \\ A(T^{(12)}) \end{pmatrix}.$$

   (c) The order of anonymizing table $A(T)$ is identical to that of transaction record table T.

   (d) No new record can be added to $A(T)$.

8. (Pseudonym Assignment)

   (a) Anonymization is performed within the duration (monthly). No move records beyond the original month.

   (b) Assignment must be consistent, i.e., a customer has unique pseudonym in the month.

   (c) Monthly assignments of pseudonyms are independent, i.e., a customer may have multiple pseudonyms and no need to change every month.

   (d) Pseudonym named "`DEL`" is prohibited.

9. (Prohibited Actions in Anonymization)

   (a) Each team submit at most three anonymized data for a trial phase, and at most one anonymized data for the final phase. Note that arbitrary number of update is allowed and the latest three ones are used the "submission".

(b) Data format of $A(T)$ should be the same to the original T. The regular expression version is provided as "PWS CUP 2016 transaction data format" (in Japanese).

(c) Anonymized Table $A(T)$ are satisfied

$$|T| = |A(T)|$$
$$|T|/2 < |S|$$

where $|T|$, $|A(T)|$, $|S|$ are the number of rows (records) for T, $A(T)$, and S.

(d) The set of product ID $t_{\cdot,5}$ of anonymized table $A(T)$ subsets the set of product IDs of T. Note that arbitrary values can be specified for unit price $t_{\cdot,6}$ and quantities $t_{\cdot,7}$.

10. (Anonymized Data) The judge produces the anonymized data S from each of submitted anonymized tables $A(T)$ as follows.

(a) Remove rows that begin with $[DEL,,,,,,]$, and permute rows randomly. Note that $|T| = |A(T)| \geq |S|$.

(b) Identify pseudonym table F based on T and $A(T)$ as

$$F = \begin{pmatrix} c_{1,1} & f^{(1)}(c_{1,1}) & \cdots & f^{(12)}(c_{1,1}) \\ \vdots & & \ddots & \vdots \\ c_{n,1} & f^{(1)}(c_{n,1}) & \cdots & f^{(12)}(c_{n,1}) \end{pmatrix}$$

where $f^{(\ell)}(c_{i,1})$ is the pseudonym for i-th customer's identity $c_{i,1}$ in duration ℓ.

11. (Utility) The utility of anonymized data S is defined as

$$U(S) = \max_{i=1,\ldots,6} E_i(S)$$

where E_1, E_2, E_3 are item-based similarities defined in Section 3.6 in [1]. E_4 and E_5 are the means of difference of inter-records and unit price between T and $A(T)$. E_6 is the fraction of deleted records of $A(T)$.

12. (Re-identification) players submit the estimated pseudonym records based on the anonymized data S and the partial knowledge T_α as

$$\hat{F} = \begin{pmatrix} c_{1,1} & \hat{f}^{(1)}(c_{1,1}) & \cdots & \hat{f}^{(12)}(c_{1,1}) \\ \vdots & & \ddots & \vdots \\ c_{n,1} & \hat{f}^{(1)}(c_{n,1}) & \cdots & \hat{f}^{(12)}(c_{n,1}) \end{pmatrix}$$

13. (Re-identification of Pseudonym)

(a) Re-identification rate of Pseudonym is the fraction of correctly identified records for all 12 months out of $12n$ pairs in F. Namely,

$$\mathrm{reid}(F, \hat{F}) = \frac{|\{\forall \ell \in \{1, \ldots, 12\}\ f^{(\ell)}(c_{i,1}) = \hat{f}^{(\ell)}(c_{i,1})\}|}{n}$$

(b) Let $\mathrm{reid}(F, \hat{F_\alpha})$ be the re-identification rate for partial knowledge T_α. The aggregated re-identification rate for the whole knowledge is the sum of all re-identification rates, i.e.,

$$\mathrm{reid}(F, \hat{F}_*) = \sum_{\alpha \in \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}} \mathrm{reid}(F, \hat{F_\alpha})$$

Note that symbol “DEL” is treated as same as the other values.

14. (Privacy Metrics) Privacy metrics of anonymized data S is defined by

$$reid(S) = \max_{i=1,\ldots,} S_i(S)$$

where S_i is re-identification algorithm as follows:

S_1-datenum	identify records by date and quantity
S_2-itemprice	identify records by (two-digits) product and unit price
S_3-itemnum	identify records by (two-digits) product and quantity
S_4-itemdate	identify records by (two-digits) product and date
S_5-itemdate	identify records by (two-digits) product, unit price and quantity
S_6-itemdate	identify records by (two-digits) product, date and quantity
others	arbitrary algorithm used by other players

15. (Aggregation) The ranks in the trial phase and the final phase are aggregated with $1 : 9$ weight.
16. (WINNER) The WINNER is the player who submit $A(T)$ with the lowest aggregated rank.
17. (Judge) Any member of working group (judge) can be players under the following conditions are met:
   (a) No colluding with any players.
   (b) All knowledge given from working group (judge) must be transparent.
   (c) Any information accessed from judge must not be disclosed to any players.
18. (Prohibited Actions in Identification) The following actions are prohibited.
   (a) Collusion with any anonymizing players.
   (b) Invalid format of estimated psuedonym matrix $\hat{F}$. (The matrix is of the form n rows and 13 columns (customer ID plus 12 months), corded in CSV format. Uniqueness of pseudonym is not required. Valid pseudoym including DEL comes to element of the matrix.)
   (c) Submit the estimated matrix per team more than 11 times.
   (d) Identify anonymized data using any auxiliary knowledge other than given partial knowledge T_{α_i}. (E.g., using the Online Retail dataset, or using the full transaction data T (or T_{α_j} $(j \neq i)$) when identifing with given T_{α_i}.
19. (Rule of Final phase)
   (a) New transaction record table T and new customer master table M will be provided few days before the final phrase. Note T and M are replaced by new ones.
   (b) New partial knowledge $T_{\alpha_1}, \ldots, T_{\alpha_4}$ will be given at the final phase.
   (c) Each team is allowed to submit only one anonymized data ($A(T)$) for the final phrase.
   (d) Teams are ranked based on the sum of utility and privacy metrics, $U + E$.
   (e) The first, second and third aggregated ranked teams are awarded as anonymized awards. The team that re-identifies the first ranked anonymized data with the most accurate ratio is awarded as re-identifying award. All tied teams are awarded.
   (f) Finalists to “2F2: PWS Cup FINAL” are chosen based on the aggregated score and the recommendations from the PWSCUP committee.
   (g) Team that most carefully considered the Japanese regulations for de-identified data is chosen

from votes of audience in 2F2 and awarded as **regulation award**.

20. (Platform) No restriction to platform, operating system, and computer language. Remote participation is allowed.