The Data Anonymization and Re-identification Competition (DARC) Rules

DARC Working Group

Ver. 1.2
September 25, 2018

Notes:

- Rules might be subject to change during the competition if some serious issues are detected that requires it (e.g., unfair ways to "cheat" the system).
- The rules of the competition are strongly inspired from the rules of the PWSCUP 2017 (`https://pwscup.personal-data.biz/web/pws2017`).
- The details of the dataset are described in the following paper.

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

Rules

1. During the competition, the participants involved are the data anonymizing players, the data re-identifying players and the judge.
2. A Data Anonymizing Player (or Team) is given a transaction record table T. Each anonymizing player performs an anonymization algorithm A, which can for instance replace identities by pseudonyms, suppress records, perturb dates, swap goods, ... to produce an anonymized table $A(T)$ that will be submitted to the judge.
3. The Judge generates the pseudonym table F specifying 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 a randomly permuted version of $A(T)$, as the anonymized data S. Note that F is hidden from data re-identifying players.

4. A Data Re-identifying Player (or Team) estimate the hidden pseudonym table based on the received S and the information about the original table T and submit the estimated pseudonym table $\hat{F}$.

5. (Dataset)

   (a) The transaction records table T is sampled from [1] by keeping all the records that have at least 5 transactions and no more that 500 at the maximum.

   (b) There are n unique customers in T.

   (c) The transaction record table T is partitioned into twelve monthly transaction tables $T^1, \ldots, T^{(12)}$. All stock codes are replaced with the first-two digits. Note that there are exactly n customers for all monthly tables.

   (d) T can be downloaded from the website of the competition.

   (e) This dataset is considered as public information. Researchers are allowed to distribute them and to publish paper using the data provided that [1] is cited as reference.

6. (Anonymized Table Format)

   (a) Deleted records from transaction record table T are specified as follows:

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

   Remark that the total number of rows of $A(T)$ should be identical to T. A record that begins with "`DEL`" is automatically deleted without taking into account the values of the other columns.

   (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 records table T.

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

7. (Pseudonym Assignment)

   (a) Arbitrary pseudonym can be assigned to customers provided that they are consistent within a month. More precisely, the identifier of a customer may or may not have many pseudonyms in $A(T)$ but if

he is assigned multiple ones, each pseudonym has to be consistent during one particular month (i.e., his identifier cannot be replaced by two or more different pseudonyms within the same month).

(b) The use of a pseudonym named "`DEL`" is prohibited.

8. (Production of Pseudonym Table and Anonymized Data) From each of submitted anonymized tables $A(T)$ and from the raw table T, the judge produces the pseudonym table F, and the anonymized data S using the following procedure.

   (a) Compute the pseudonym table F from T and $A(T)$ as

   $$F = \begin{pmatrix} c_1 & f^{(1)}(c_{1,1}) & \cdots & f^{(12)}(c_{1,1}) \\ \vdots & & \ddots & \vdots \\ c_n & f^{(1)}(c_{n,1}) & \cdots & f^{(12)}(c_{n,12}) \end{pmatrix}$$

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

   (b) A record in T that begins with "`DEL`" is automatically deleted without taking into account the values of the other columns.

   (c) Let $\overline{A(T)}$ be the anonymized table $A(T)$ minus all deleted records. Anonymized dataset S is obtained by permuting randomly all rows in $\overline{A(T)}$.

   (d) Note that as a result $|S| \leq |T| = |A(T)|$, where $|T|$, $|A(T)|$, $|S|$ are the number of rows (i.e., records) for T, $A(T)$ and S.

9. (Prohibited Actions and Mandatory Constraints for Anonymization)

   (a) Each team submits at most three anonymized data during the anonymization phase. Note that an arbitrary number of updates is allowed but that the latest three ones will be considered as the "submitted ones".

   (b) The data format of $A(T)$ should be the same to the original T.

   (c) Anonymized Table $A(T)$ must satisfy the following constraints

   $$\begin{aligned} |T| &= |A(T)|, \\ |T|/2 &< |S|. \end{aligned}$$

   (d) The set of product IDs $t_{\cdot,5}$ of the anonymized table $A(T)$ have to be a subset of the set of product IDs of T (i.e., it is not allowed to generate new product IDs that are not part of the original database T). However, arbitrary values can be specified for unit price $t_{\cdot,6}$ and quantities $t_{\cdot,7}$.

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

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

    in which

(a) E_1, E_2 and E_3 are item-based similarities. More specifically, we based those metrics on the collaborative filtering approach.
Given a table (anonymized or not), we denote by I the list of items, and $U_{i,j}$ the set of users who bought both items i and j. We can also define scores $r_{x,i,k}$ based on the quantity of item i bought by user x:

i. For $k = 1$, the score $r_{x,i,1}$ is the number of times the item i was bought by user x modulo 12.

ii. For $k = 2$, the score $r_{x,i,2}$ equals the number of times the item i was bought by user x if this number is less than the given threshold $\alpha =?$, and 0 otherwise.

iii. For $k = 3$, the score $r_{x,i,3}$ equals the number of times the item i was bought by user x if this number is among the top $k =?$ highest numbers. Otherwise, the score is 0.

Let us fix a similarity index $1 \leq k \leq 3$. For ground truth T (respectively anonymized data S), we compute a matrix M_k (resp. M'_k), of size $m \times m$, where m is the cardinal of I, the set of items. Matrix coefficient $(M_k)_{i,j}$ (resp. $(M'_k)_{i,j}$) encodes the similarity distance between the item i and the item j, with scores computed from table T (resp. S), and is defined as follow :

$$(M_k)_{i,j} = \frac{\sum_{x \in U_{i,j}} r_{x,i,k} \times r_{x,j,k}}{\sqrt{\sum_{x \in U_i} r_{x,i,k}^2} \times \sqrt{\sum_{x \in U_j} r_{x,j,k}^2}}$$

Finally, item-based similarity E_k is defined as the similarity between M_k and M'_k as :

$$E_k(S) = sim((M_k), (M_k)') = \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} |(M_k)_{i,j} - (M'_k)_{i,j}|}{\sum_{i=1}^{m} \sum_{j=1}^{m} (M_k)_{i,j}}.$$

(b) E_4 and E_5 are the means of difference of inter-records and unit prices between T and $A(T)$.

(c) E_6 is the fraction of deleted records of $A(T)$.

(d) Note that a low value of $U(S)$ is representative of a strong utility while a high value corresponds to the opposite.

11. (Submission of Re-identification Data) Re-identification players submit the estimated pseudonym records based on the anonymized data S and the knowledge T as

$$\hat{F} = \begin{pmatrix} c_1 & \hat{f}^{(1)}(c_{1,1}) & \cdots & \hat{f}^{(12)}(c_{1,1}) \\ \vdots & & \ddots & \vdots \\ c_n & \hat{f}^{(1)}(c_{n,1}) & \cdots & \hat{f}^{(12)}(c_{n,12}) \end{pmatrix}$$

12. (Re-identification Rate)

(a) For specific F and $\hat{F}$, a re-identification rate is computed as the fraction of correctly identified records for all 12 months out of $12n$ pairs in F. Namely,

$$\text{reid}(F, \hat{F}) = \frac{\sum_{i=1}^{n} \sum_{l=1}^{12} |f^{(\ell)}(c_{i,\ell}) = \hat{f}^{(\ell)}(c_{i,\ell})|}{12n}$$

(b) The symbol "DEL" is treated as same as the other values.

(c) Note that a low value of $\text{reid}(F, \hat{F})$ is representative of a strong privacy level while a high value corresponds to the opposite.

13. (Privacy)

(a) Let $n_{attacks}$ be the number of attacks on an anonymized data S. The overall re-identification rate of the anonymized data S is defined as

$$reid(S) = \max_{i=1,\ldots,n_{attacks}} reid(F, \hat{F}_i)$$

in which $\hat{F}_i$ is re-identification algorithm as follows (the first 6 are used as baseline):

$\hat{F}_1$-datenum	identify records by date and quantity
$\hat{F}_2$-itemprice	identify records by (two-digits) product and unit price
$\hat{F}_3$-itemnum	identify records by (two-digits) product and quantity
$\hat{F}_4$-itemdate	identify records by (two-digits) product and date
$\hat{F}_5$-itemdate	identify records by (two-digits) product, unit price and quantity
$\hat{F}_6$-itemdate	identify records by (two-digits) product, date and quantity
$\hat{F}_i$, $i > 6$	arbitrary algorithm submitted by a re-identifying player

(b) Note that $n_{attacks}$ and therefore $reid(S)$ will vary during the re-identification phase of the competition.

14. (Score) The score of a team is simply equals to

$$score(S) = \frac{U(S) + reid(S)}{2}.$$

15. (Winner) The winner is the player who submit $A(T)$ with the lowest score obtained from the derived S.

16. (Judge) Any member of the competition committee (i.e., judge) can be players under the following conditions are met:

(a) No collusion with any players.

(b) All knowledge known by the judge that can impact the competition will be disclosed publicly to ensure the transparency of the competition.

(c) Any information accessed from judge must not disclose to any players.

17. (Prohibited Actions during Re-Identification) The following actions are prohibited.
    (a) Collusion with any anonymizing players.
    (b) Invalid format of estimated pseudonym matrix $\hat{F}$. The matrix is of the form n rows and 13 columns (customer ID plus estimated pseudonyms for 12 months), recorded in CSV format. The uniqueness of a pseudonym is not required and a valid pseudonym including DEL can be used in the matrix.
    (c) Submit the estimated matrix per team more than 10 times.
18. (Ranking)
    (a) Teams are ranked based on the sum of utility and privacy metrics, $U + E$.
    (b) The first, second and third aggregated ranked teams are awarded as anonymized awards.
    (c) The team that re-identifies the first ranked anonymized data with the most accurate ratio is awarded as re-identifying award.
    (d) All tied teams are awarded.
19. (Platform) No restriction to platform, operating system, and computer language.

A summary of different roles in the competition, in line with the rules notations, is provided in the following table:

role	inputs	outputs	max number of submissions per team
anonymizer player	T	A(T)	?
re-identifier player	S, T	$\hat{F}$	10
judge	T, A(T)	F, S	no restriction