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

DARC Working Group

Ver. 1.4
September 09, 2019

Notes:

- The participants to the DARC challenge should carefully read the following rules.
- 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 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. (Participants) 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 taking as input the table, 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. In practice for this competition, the judge will be the competition platform.
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) Each transaction is characterized by the identifier of the customer, the date and time of the transaction, the identifier of the item bought (which is a 5-digit code) as well as the unit price of this item and the quantity bought by the customer for this transaction.

   `[id_user,date,time,id_item,unit_price,quantity]`

   (c) There are $n = 4034$ unique customers in T for a total number of transactions $|T| = 307054$.

   (d) The transaction record table T is partitioned into thirteen monthly transaction tables $T^1, \ldots, T^{(13)}$ from 2010/12 to 2011/12.

   (e) T can be downloaded as the file "ground_truth.csv" from the website of the competition or directly from the following url under the repository "data" : `https://gitlab.aicrowd.com/drayer34/darc`

   (f) 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 by replacing the identifier of a customer by the string "`DEL`" as follows:

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

   Remark that the total number of rows of $A(T)$ should be identical to T. During the anonymization process, 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.,

$$A(T) = \begin{pmatrix} A(T^{(1)}) \\ \vdots \\ A(T^{(13)}) \end{pmatrix}.$$

(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 $A(T)$. This precludes the possibility of adding fake records. However, existing record can be altered providing that the constraints described hereafter are respected.

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. (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) For the date, the anonymization should preserve the same month but the day and the time can be arbitrarly set within this month. In addition, the date format of $A(T)$ should be the same as the original database 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 $t_{\cdot,4}$ 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 $t_{\cdot,5}$ and quantities $t_{\cdot,6}$.

9. (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

$$F = \begin{pmatrix} c_1 & f^{(1)}(c_{1,1}) & \cdots & f^{(13)}(c_{1,13}) \\ \vdots & & \ddots & \vdots \\ c_n & f^{(1)}(c_{n,1}) & \cdots & f^{(13)}(c_{n,13}) \end{pmatrix}$$

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 $A(T)/\{all\_deleted\_records\}$ (i.e., without the deleted records). Anonymized dataset S is obtained by permuting randomly all rows in $\overline{A(T)}$.

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

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 used in collaborative filtering approach. Given a table (anonymized or not), we denote by I the list of items and by $U_{i,j}$ is the set of users who have bought both items i and j. We also define the score $r_{x,i}$ as the quantity of item i bought by user x. For each file, the ground truth and the anonymized one, we compute a matrix M (respectively M'), of size $m \times m$ in which m is the cardinal of I, the set of items. $M_{i,j}$ represent the similarity between the item i and the item j, which can be instantiated as follows using the cosine similarity :

$$M_{i,j} = sim(i,j) = \frac{\sum_{x \in U_{i,j}} r_{x,i} \times r_{x,j}}{\sqrt{\sum_{x \in U_{i,j}} r_{x,i}^2} \times \sqrt{\sum_{x \in U_{i,j}} r_{x,j}^2}}$$

Finally, we can compute the distance between the two matrices M and M' as follows :

$$dist(M, M') = \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} |M_{i,j} - M'_{i,j}|}{\sum_{i=1}^{m} \sum_{j=1}^{m} M_{i,j}}$$

The metric E_1 is computed exactly as presented. The main difference with E_2 is that all the scores $r_{x,i}$ whose value is 6 or more in the original matrix M are set to zero in both M and M' (the rationale behind this metric is that it captures the similarity between items that are bought in small quantity). Finally for E_3 similarly to E_2 the scores $r_{x,i}$ of the objects that are not in the top-k ones (for $k = 180$) in the original dataset M are set to zero.

(b) E_4 and E_5 are the means of difference of inter-records dates 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

$$\hat{F} = \begin{pmatrix} c_1 & \hat{f}^{(1)}(c_{1,1}) & \cdots & \hat{f}^{(13)}(c_{1,13}) \\ \vdots & & \ddots & \vdots \\ c_n & \hat{f}^{(1)}(c_{n,1}) & \cdots & \hat{f}^{(13)}(c_{n,13}) \end{pmatrix}$$

12. (Re-identification Rate)

(a) For an original mapping F and the corresponding guess by the adversary $\hat{F}$, a re-identification rate is computed as the fraction of correctly identified records for all 13 months out of $13n$ pairs in F. Namely,

$$\mathrm{reid}(F, \hat{F}) = \frac{\sum_{i=1}^{n} \sum_{l=1}^{13} |f^{(\ell)}(c_{i,\ell}) = \hat{f}^{(\ell)}(c_{i,\ell})|}{13n}$$

(b) Only non "DEL" in F are used in the comparision. Thus making a $\hat{F}$ with only "DEL" values give a 0 score.

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

13. (Prohibited Actions during Re-Identification) The following actions are prohibited.

(a) Collusion with other anonymizing players.

(b) Invalid format of estimated pseudonym matrix $\hat{F}$. The matrix is of the form n rows and 14 columns (customer ID plus estimated pseudonyms for 13 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.

14. (Privacy)

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

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

in which $\hat{F}_i$ is the file generated by the 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 item product (rounded to the 2 first digits) and unit price
$\hat{F}_3$-itemnum	identify records by item product (rounded to the 2 first digits) and quantity
$\hat{F}_4$-itemdate	identify by item product (rounded to the 2 first digits) and date
$\hat{F}_5$-itemdate	identify by item product (rounded to the 2 first digits), unit price and quantity
$\hat{F}_6$-itemdate	by item product (rounded to the 2 first digits), date and quantity
$\hat{F}_i$, $i > 6$	arbitrary algorithm whose output is submitted by a re-identifying player

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

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

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

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

17. (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.

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 anonymization award.

    (c) The team that re-identifies the first ranked anonymized data with the most accurate ratio is awarded as re-identification 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
Anonymizing player	T	A(T)	Arbitrary but only the 3 last are considered
Re-identifying player	S, T	$\hat{F}$	10
Judge	T, A(T)	F, S	no restriction

Synthesis of the Symbols Used in the Rules

In progress

Data Structures	
Original table T	Full transaction record table. Partitioned in 13 monthly partitions $T^{(1)},\ldots,T^{(13)}$.
Anonymized table $A(T)$	Anonymized transaction table submitted to the judge by an anonymization team. It is the concatenation of the anonymization of the 13 monthly partitions.
Correct mapping F	Mapping between the customer IDs in T and the pseudonyms in $A(T)$. It is computed by the judge from T and $A(T)$ based on the order of the rows.
Post-processed anonymized table S	Table post-processed by the judge from $A(T)$ by removing the records tagged `DEL` and by permuting randomly the remaining records.
Estimating mapping $\hat{F}$	Mapping between the customer IDs in T and the pseudonyms in S proposed by a re-identification team.
Functions	
A	Anonymization algorithm.

Table 1: Symbols

Synthesis of the Constraints

In progress

Anonymization	
Number of rows in $A(T)$	The transaction table and the anonymized table must contain the same number of rows.
Number of deleted rows	The number of deleted rows in S must be strictly lower than the half of the number of rows in T.
Order of rows in $A(T)$	Same order as the rows in T.
Pseudonyms in $A(T)$	Each customer has at most one pseudonym per month and at least one pseudonym in the full anonymized table. A pseudonym is an arbitrary character string, excluding `DEL`.
Number of anonymized tables submitted	Any number of anonymized tables can be submitted to the system but only the last three versions submitted will be considered by the judge.
Dates in $A(T)$	The date format should be the same as in T. The month of an anonymized record must be the same as the month of the original record, only the day and time can differ.
Product IDs in $A(T)$	No new product ID can be generated.
Re-Identification	
Format of $\hat{F}$	$\hat{F}$ is a matrix of 14 columns that contains the same number of rows as T. The first column contains for each row the customer IDs. The thirteen other columns contain the thirteen estimated pseudonyms for a given customer ID (one per month).
Number of submissions of $\hat{F}$	The estimation of a mapping can be submitted 10 times at most by a re-identification team for a given table S.

Table 2: Constraints