

Projet Sécurité

INSA Centre Val de Loire
4A STI Promotion 2021

Auteurs :

Groupe : On sait pas

2019/2020

Tables générées

Pour tous les tables, les attributs considèrent comme des données quasi-identifiantes (QI) sont id_item et sont utilisées pour le K-anonymat. L'attribut considère comme une donnée sensible (S) est ID_user, utilisée pour la L-diversité.

Table 1 : on_sait_pas_1.csv et Table 2 : on_sait_pas_2.csv

- Diviser la ground_truth en des bds chaque bd contient les transactions d'un mois de la ground_truth
- Pour chaque bd on fait un algo du k-anonymat et l-diversité les k et l sont différents pour la table 1 et table 2
- Puis on fait une concaténation de tous les résultats des bds pour obtenir la bd qui va être traitée par le script python mentionner en bas.

Table 3 : on_sait_pas_3.csv.

- Traitement du ground_truth en entier par ARX en utilisant k-anonymat et l-diversité les k et l sont différents du table 1 et 2
- Puis traitement du résultat obtenu par le script python mentionner en bas

Script Python pour les tables :

Puisque les règles de la compétition demandent un format très précis soit respecté par nos données, et que la pseudonymisation des identifiants est obligatoire, il est nécessaire d'utiliser un script pour manipuler les données qui vont être rendues par ARX.

Le script réalisé se charge donc de :

- Appliquer une fonction d'hachage et du sel aux id_users dans l'objectif de les rendre non identifiables, le sel est différent pour chaque user et pour chaque mois donc chaque user a un hash différent chaque mois.
- Fixer tous les jours de la date de la table au jour qui a eu le nombre de prix le plus grand pour chaque mois.
- Mettre toutes les valeurs de l'attribut heure à 00:00,
- Regroupement des différents prix selon un intervalle puis élaboration de la médiane de ce dernier.

- Regroupement des différentes quantités selon un intervalle, puis élaboration de la médiane de ce dernier.
- Remplacer les lignes supprimer par ARX, en utilisant le format donner par les règles (id_user remplacé par DEL).

A la sortie de ce script, un fichier au format .csv utilisable par le script d'évaluation est obtenu.