DARC: Data Anonymization and Re-identification Challenge

Antoine Boutet, Mathieu Cunche, Sébastien Gambs,
Antoine Laurent and Benjamin Nguyen

RESSI 2020, 16th-18th Décembre 2020

L’anonymat selon le RGPD

Il y a lieu d'appliquer les principes relatifs à la protection des données à toute information concernant une personne physique identifiée ou identifiable. Les données à caractère personnel qui ont fait l'objet d'une pseudonymisation et qui pourraient être attribuées à une personne physique par le recours à des informations supplémentaires devraient être considérées comme des informations concernant une personne physique identifiable. ***Pour déterminer si une personne physique est identifiable, il convient de prendre en considération l'ensemble des moyens raisonnablement susceptibles d'être utilisés par le responsable du traitement*** *ou par toute autre personne pour identifier la personne physique directement ou indirectement, tels que le ciblage.* ***Pour établir si des moyens sont raisonnablement susceptibles d'être utilisés pour identifier une personne physique, il convient de prendre en considération l'ensemble des facteurs objectifs, tels que le coût de l'identification et le temps nécessaire à celle-ci, en tenant compte des technologies disponibles au moment du traitement et de l'évolution de celles-ci****. Il n'y a dès lors pas lieu d'appliquer les principes relatifs à la protection des données aux informations anonymes, à savoir les informations ne concernant pas une personne physique identifiée ou identifiable, ni aux données à caractère personnel rendues anonymes de telle manière que la personne concernée ne soit pas ou plus identifiable. Le présent règlement ne s'applique, par conséquent, pas au traitement de telles informations anonymes, y compris à des fins statistiques ou de recherche.*

→ Obligation de moyens

→ Obligation d’évaluation du coût et de l’efficacité de la ré-identification

DARC

- UQAM
  - Challenge programmé pour PETS 2020, Montréal
- Beta testeurs: 3 promotions INSA
  - OT Vie Privée et Ethique, 5ème année, Informatique, Lyon
  - OT Sécurité et Vie Privée, 5ème année, Télécommunications, Lyon
  - Sécurité et Technologies Informatiques, 4ème année, Centre Val de Loire

Objectif Pédagogique

- Compétences de projets
  - Communication, organisation, interaction, ...
- Compétences techniques
  - Développement, analyse, méthodologie, ...
- Compétence sécurité
  - Techniques d’anonymisation, ré-identification, ...
- Compétences éthiques
  - Protection des données, fair play, ...

Règles du challenge

- Les données
- Les métriques d’évaluation
- Calcul du score

[1] Daqing Chen, Sai Liang Sain, and Kun Guo, Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining, Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp. 197-208, 2012

Les données

- Transactions d'un site e-commerce UK [1]
  - 307 054 enregistrements concernant 4034 clients sur 13 mois

id_user	date	hours	Id_item	price	qty
17850	2010/12/01	08:26	85123A	2.55	6
17850	2010/12/01	08:26	71053	3.39	6
...					
12583	2010/12/01	08:45	22492	0.65	36

[1] Daqing Chen, Sai Liang Sain, and Kun Guo, Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining, Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp. 197-208, 2012.

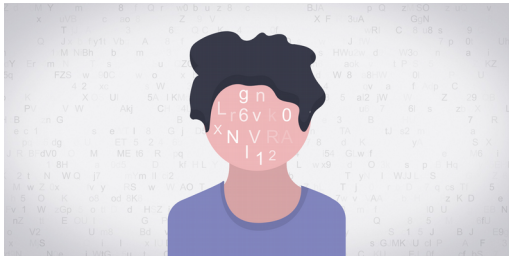
Évaluation utilité / ré-identification

- Fourniture de 6 métriques d’utilité de référence
  - 3 métriques basées sur les articles (métriques utilisées dans le filtrage collaboratif)
  - 2 métriques sur la moyenne des différences de dates et prix entre les enregistrements
  - 1 métrique sur la proportion d’enregistrement supprimés
- Fourniture d’une métrique de désanonymisation naïve

Calcul des scores

- Chaque groupe a pu soumettre jusqu'à 3 jeux de données anonymisés (S)
- Chaque groupe peut soumettre plusieurs attaques sur chaque jeu de données des autres groupes
- Le score de chaque soumission (S) est calculé par : $Score_D(S)$ = Utilité réelle * (1-$MAX_{groupes}$(score attaque))
- Le score de défense d'un groupe est le meilleur score de sa soumission (S) : $Score_G = MAX_S(Score_D(S))$
- Le score d'attaque d'un groupe est la somme de sa meilleure attaque sur chacun des autres groupes

Déroulement du Challenge

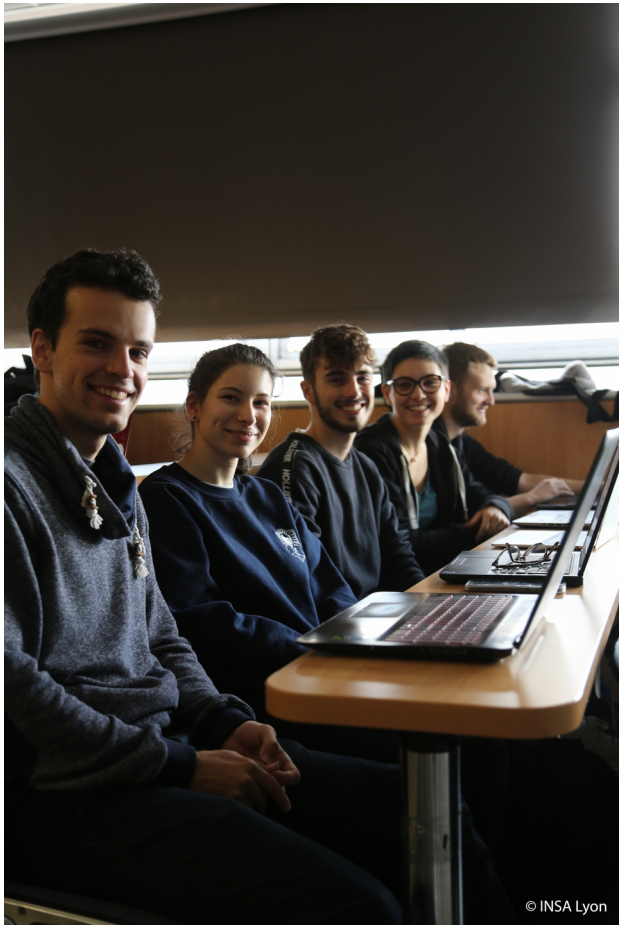


- Phase 1 : anonymisation des données

- Phase 2 : ré-identification

Résultats

1) INSA Centre Val de Loire
2) 5IF INSA-Lyon
3) 5TC INSA-Lyon

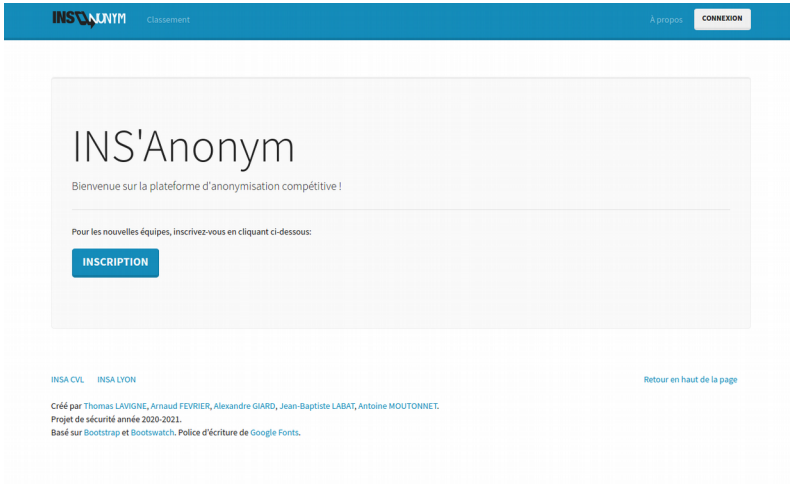


Retours d’Expérience

- Réduire les dépendances
  - Indisponibilité de la plate-forme crowdai.org

    → Tâches d’évaluation manuelles
- Plusieurs formations avec des contraintes d’EDT différents
  - Temps de préparation hétérogène
  - Rassemblement des étudiants pour la phase live

DARC nouvelle version



- Données de géolocalisation
- Développement d'une plate-forme de soumission

→ Phase 2 live le 18/12/2020

→ Organisation d'un challenge plus ouvert