GOUVERNEMENT
Liberté
Égalité
Fraternité

Pôle d'Expertise de la Régulation Numérique

Collecte de données volontairement contribuées et localement anonymisées

Joris Duguépéroux, Gaspard Defréville, Lucas Verney, Nicolas Deffieux - PEReN

https://www.peren.gouv.fr/

GOUVERNEMENT
Liberté
Égalité
Fraternité

Pôle d'Expertise de la Régulation Numérique

Le PEReN

- Service à compétence national, créé en août 2020

- Deux missions principales :
  - **Assistance aux administration** dans la régulation des plateformes numériques
  - **Projets exploratoires** (sans demandes spécifiques)
    - Pour des **expérimentations**
    - Pour de la **recherche académique**

Régulation des **plateformes** ! (pas des gens)

GOUVERNEMENT
Liberté
Égalité
Fraternité

Pôle d'Expertise de la Régulation Numérique

Contexte du projet

- Peu de données pour certaines plateformes

  => **Idée :** exploiter **le droit à la portabilité des données du RGPD**

- **Données personnelles** (risques pour les contributeurs, RGPD)

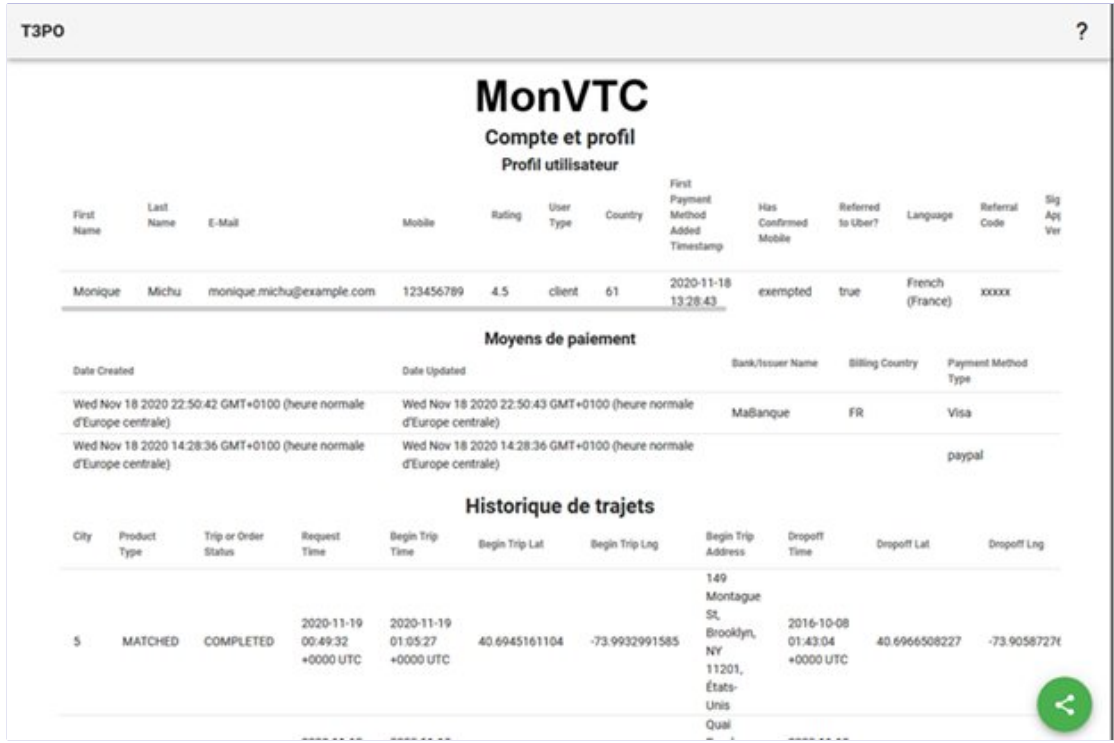
  => **Idée : anonymiser localement** (avant tout envoi)

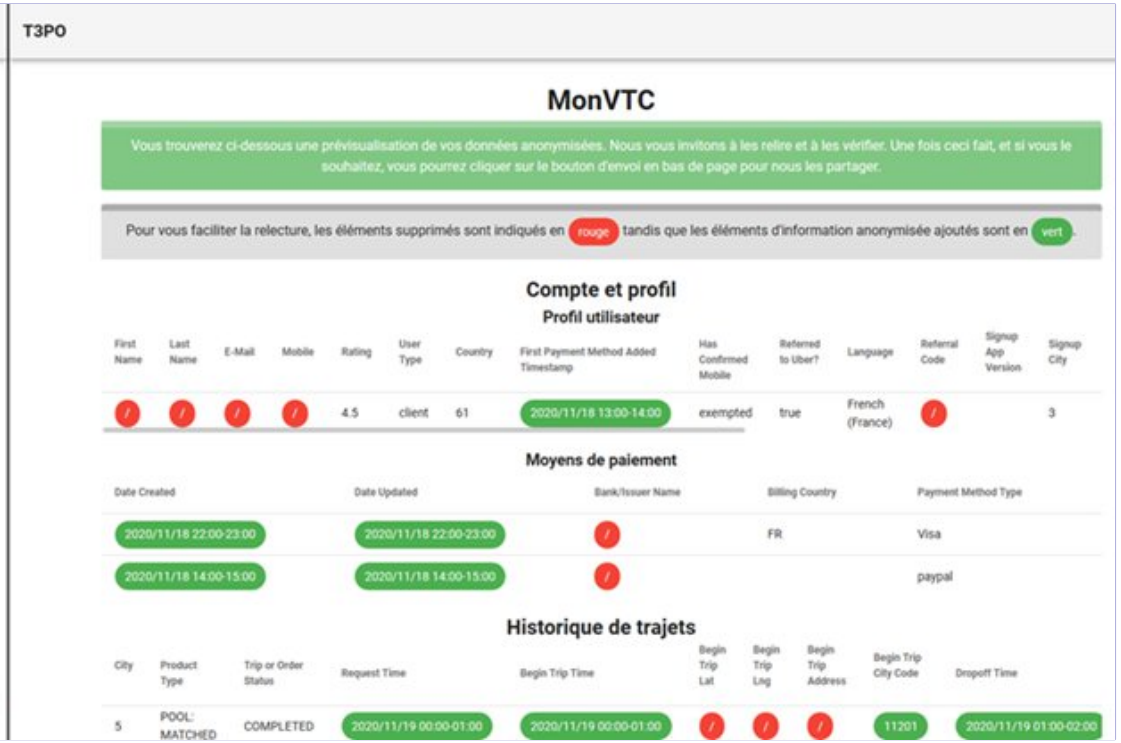
- Secteur d'expérimentation ? **VTC / livreurs à vélo**
  - **Indicateurs publiés pour comparaison** (mais insuffisants) + sécurité juridique de la portabilité (LOM)
  - Secteur organisé (associations représentatives et syndicats)
  - **Secteur demandeur**, avec **besoin de données** pour aborder le sujet

GOUVERNEMENT
Liberté
Égalité
Fraternité

Pôle d'Expertise de la Régulation Numérique

Illustration





Modèle de contribution et modèle d'attaque

- Modèle de contribution :
  - L'utilisateur suit la démarche pour **demander ses données**
  - Une fois ses données récupérées, il revient sur l'outil pour les **visualiser**
  - Il peut en partager un **extrait anonymisé** s'il le souhaite
  - **Interaction unique** avec l'utilisateur : au moment de l'envoi de ses données

- Modèle d'attaque :
  - Attaquant récupère le jeu de données (fuite) ou nos analyses
  - **La plateforme peut être attaquante** (cas limite de difficulté pour nous)

Challenge : anonymiser localement

- Anonymiser localement (au sein de l'enclave de l'appareil de l'utilisateur)
  - Visibilité sur l'intégralité des données de l'utilisateur
  - Aucune vision globale des données de tous les utilisateurs

- Construire ce processus d'anonymisation est un *challenge*

GOUVERNEMENT
Liberté
Égalité
Fraternité

Pôle d'Expertise de la Régulation Numérique

Un outil majeur : le déchiffrement par seuil

Idée générale : besoin d'un certain nombre de clés avant de pouvoir déchiffrer

- Travailleur : chiffre ses données, envoie les chiffrés
- Travailleur (avec des conditions à définir) : envoie une clé
- PEReN : attend de recevoir k clés pour pouvoir déchiffrer les envois
  - Le déchiffrement n'est possible que pour les messages chiffrés avec la bonne clé

Par commodité, on confondra déchiffrement par seuil et partage de secret par seuil

Du déchiffrement par seuil au k-anonymat : *STAR*

STAR: Distributed Secret Sharing for Private Threshold Aggregation Reporting (Davidson et al. 2022)

- Cas d'usage initial : télémétrie de navigateur (Brave), découverte de « heavy hitters »
  - Chaque internaute envoie ses données chiffrées
  - Clé déterminée par les données*
  - Si *k* données identiques, déchiffrement possible
  - Ajout de données « annexes » possibles (chiffrées avec la même clé)
  - Protocole rapide et fonctionnel

* directement par l'entropie des données si elle est suffisante, par un tiers indépendant en utilisant les données comme clé de génération d'aléatoire sinon.

STAR appliqué aux livreurs

- Travailleur : envoie ses valeurs chiffrées
  - Besoin d'un tiers indépendant pour fournir de l'aléatoire (ne manipulera aucune donnée)
- Lorsqu'on a au moins k valeurs identiques, on peut déchiffrer
- k-anonyme par construction ?

Ajouts envisagés (brique PEReN)

Problème : on ne connaît pas le nombre de répondants

- Choix de la granularité complexe…
- Vers une version récursive !
  - On chiffre les données à plusieurs granularités distinctes
  - On déchiffre à mesure des résultats reçus

Ajouts envisagés (brique PEReN)

Éviter l'auto-déchiffrement si un travailleur envoie plusieurs fois ses données ? Tout en autorisant les mises à jour ?

- Ajout d'un timestamp (e.g. mensuel)
- Avant chaque envoi, confronter les données à un hash de (mot de passe + données + timestamp)
- Note : auto-déchiffrement nuisible au travailleur (vie privée) mais aussi pour nous (qualité des données)

Problématiques *privacy*

- Est-ce du *k*-anonymat ?
  - **Non** : si tous les « voisins » et le parent sont déchiffrés, on peut avoir des infos
  - Solution ? **Déchiffrer tous les voisins en même temps** si tous les quorums sont réunis (problème : demande encore davantage de participants pour la même info)
  - Non publié et non peer-reviewé à ce jour

- Peut-on exploiter **les données en « annexe »** ?
  - Serait utile pour réduire l'impact de la granularité
  - Avec de la **local differential privacy** par exemple ?
  - Problème : **garanties extrêmement complexes** à quantifier…

- Berk, du *k*-anonymat, ça a été cassé mille fois…
  - On est preneur s'il y a mieux qui réponde à nos contraintes

GOUVERNEMENT
Liberté
Égalité
Fraternité

Pôle d'Expertise de la Régulation Numérique

Problématiques données

- **Priorisation** des valeurs à demander :
  - Par construction, **compromis finesse / quantité** des données obtenues
  - Comment analyser proprement des « heavy hitters » ?

- Maximiser la qualité des données :
  - Quelle **représentativité** pour les données ?
  - Ajout de **données « en annexe »** ? (et de quelle manière ?)

- Comment attirer la **confiance** ?
  - Travail de **vulgarisation**
  - Travail de **visualisation** (des données, de l'anonymisation)

GOUVERNEMENT
Liberté
Égalité
Fraternité

Pôle d'Expertise de la Régulation Numérique

Problématiques métiers

- Question liées au **contexte**
  - Distinguer livreur/compte du fait de la sous-location ?
  - Analyse sur les **profils** de travailleurs ? Sur les **trajets** ?
  - Questions **en dehors de la portabilité** ? (Source principale de revenus ? Utilisation de plusieurs plateformes ? Infos socio-démographiques ?)

- **Comparaison** avec les indicateurs publiés par la LOM

- Comment **attirer les travailleurs** ?
  - Passage par des **syndicats** ? → **augmente le risque** si l'anonymisation casse
  - **ARPE** ? → institution récente et **peu connue**
  - **Plateforme** ? → **défiance** de la part des travailleurs

Conclusion

- Un cas d'usage **compliqué à résoudre**
  - **Peu de contact** avec des utilisateurs qui ne se connaissent pas
  - **Attaquant fort**
  - Besoin en **données fiables**
- Une **solution satisfaisante** (on espère !)
  - Une base de **démonstration** pour la **visualisation**
  - Un protocole fonctionnel pour du ***k*-anonymat (récursif) *by design***
  - Implémentation à venir courant 2022-2023 (*open-source*)
- Mais **beaucoup de questions**
  - *Privacy* : ajout/substitution avec de la ***differential privacy*** ?
  - Données : **beaucoup d'incertitudes** sur le traitement de telles données
  - Métier : **difficulté de collecte**, de priorisation, de contact avec les acteurs