

# Collecte de données volontairement contribuées et localement anonymisées

---

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

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

## Le PEReN

---

- Service à compétence nationale, créé en août 2020
- Deux missions principales :
  - **Assistance aux administrations** 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)

## 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

# Illustration

T3PO
?
T3PO

## MonVTC

### Compte et profil

#### Profil utilisateur

First Name	Last Name	E-Mail	Mobile	Rating	User Type	Country	First Payment Method Added Timestamp	Has Confirmed Mobile	Referred to Uber?	Language	Referral Code	Sig App Ver
Monique	Michu	monique.michu@example.com	123456789	4.5	client	61	2020-11-18 13:28:43	exempted	true	French (France)	xxxxx	

#### Moyens de paiement

Date Created	Date Updated	Bank/Issuer Name	Billing Country	Payment Method Type
Wed Nov 18 2020 22:50:42 GMT+0100 (heure normale d'Europe centrale)	Wed Nov 18 2020 22:50:43 GMT+0100 (heure normale d'Europe centrale)	MaBanque	FR	Visa
Wed Nov 18 2020 14:28:36 GMT+0100 (heure normale d'Europe centrale)	Wed Nov 18 2020 14:28:36 GMT+0100 (heure normale d'Europe centrale)			paypal

#### Historique de trajets

City	Product Type	Trip or Order Status	Request Time	Begin Trip Time	Begin Trip Lat	Begin Trip Lng	Begin Trip Address	Dropoff Time	Dropoff Lat	Dropoff Lng	
	5	MATCHED	COMPLETED	2020-11-19 00:49:32 +0000 UTC	2020-11-19 01:05:27 +0000 UTC	40.6945161104	-73.9932991585	149 Montague St, Brooklyn, NY 11201, États-Unis	2016-10-08 01:43:04 +0000 UTC	40.6966508227	-73.9058727

## MonVTC

Vous trouverez ci-dessous une prévisualisation de vos données anonymisées. Nous vous invitons à les relire et à les vérifier. Une fois ceci fait, et si vous le souhaitez, vous pourrez cliquer sur le bouton d'envoi en bas de page pour nous les partager.

Pour vous faciliter la lecture, les éléments supprimés sont indiqués en rouge tandis que les éléments d'information anonymisée ajoutés sont en vert.

### Compte et profil

#### Profil utilisateur

First Name	Last Name	E-Mail	Mobile	Rating	User Type	Country	First Payment Method Added Timestamp	Has Confirmed Mobile	Referred to Uber?	Language	Referral Code	Signup App Version	Signup City
/	/	/	/	4.5	client	61	2020/11/18 13:00-14:00	exempted	true	French (France)	/		3

#### Moyens de paiement

Date Created	Date Updated	Bank/Issuer Name	Billing Country	Payment Method Type
2020/11/18 22:00-23:00	2020/11/18 22:00-23:00	/	FR	Visa
2020/11/18 14:00-15:00	2020/11/18 14:00-15:00	/		paypal

#### Historique de trajets

City	Product Type	Trip or Order Status	Request Time	Begin Trip Time	Begin Trip Lat	Begin Trip Lng	Begin Trip Address	Begin Trip City Code	Dropoff Time	
	5	POOL: MATCHED	COMPLETED	2020/11/19 00:00-01:00	2020/11/19 00:00-01:00	/	/	/	11201	2020/11/19 01:00-02:00

## 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*

## 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

## 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)

## 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 (récuratif) *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