Publication de données préservant la vie privée

Cas des données de mobilité

Ulrich Aïvodji

École de technologie supérieure

ulrich.aivodji@etsmtl.ca

Plan

- Publication de données préservant la vie privée
  - Notions clés
  - Menaces à la vie privée
  - Méthodes ``d'anonymisation''
  - Évaluation de la qualité des données publiées
  - Limites des méthodes d'anonymisation
  - Confidentialité différentielle
- Données de mobilité

Publication de données préservant la vie privée

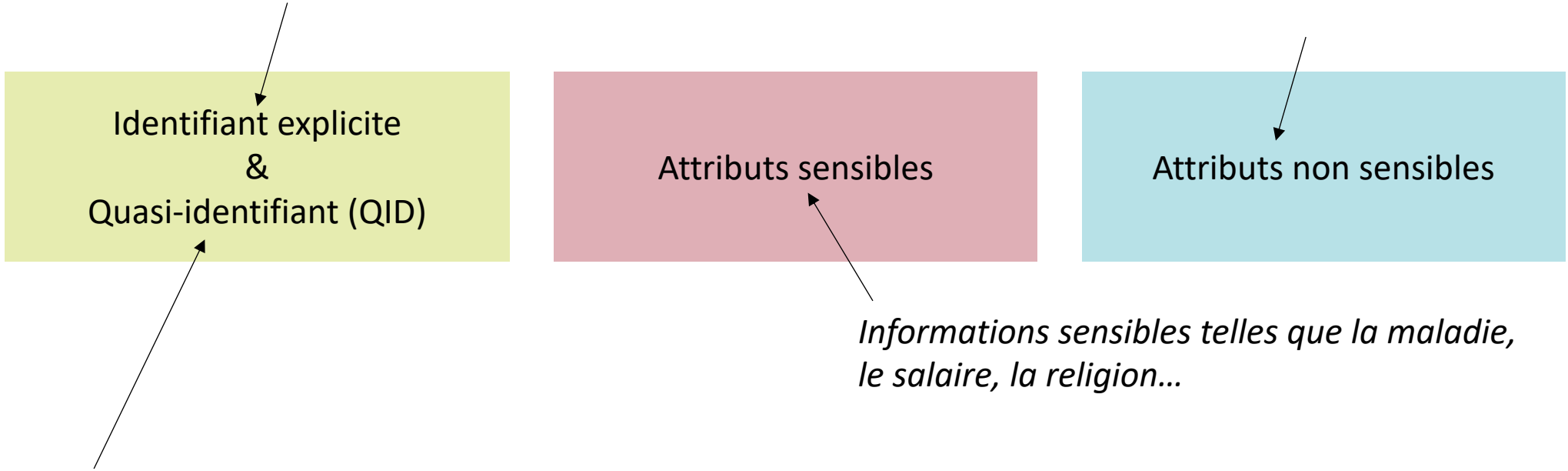
Adversaire

- **Toute entité** qui cherche à **récupérer** les **données personnelles** d'une personne concernée, **sans son consentement explicite**, pour établir un **profil** ou **déduire ses données privées**

Enregistrement

Attributs tels que le nom & prénom, le numéro de sécurité sociale et les enregistrements biométriques qui peuvent identifier directement un sujet

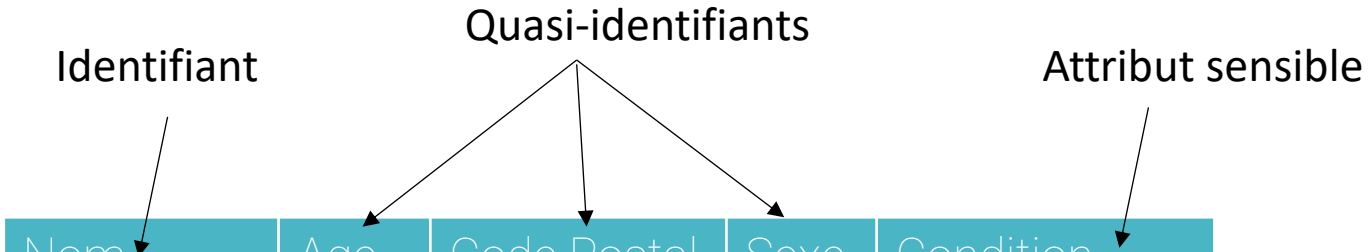
Autres que identifiants et attributs sensibles



Informations sensibles telles que la maladie, le salaire, la religion…

Attributs qui ne sont pas en eux-mêmes des identifiants uniques, mais qui combinés (par exemple, {code postal, date de naissance, sexe}), à d'autres quasi-identifiants pourraient potentiellement identifier un sujet.

Exemple



Nom	Age	Code Postal	Sexe	Condition
Yoshida	23	H2X1Y9	M	Asthme
Cohen	29	H2X3X2	F	Hypertension
Achebe	25	H2X3E2	F	Schizophrénie
Murphy	42	H2S3C7	M	Cancer
Bouchard	45	H2S2L8	F	Diabète
Smith	55	H2S2E7	M	Grippe

Menaces

Bob (la cible)

- **Couplage d'enregistrement**
  - l'enregistrement $x \in D'$ appartient à Bob. x a la propriété $P_x \Rightarrow$ Bob a la propriété P_x
- **Couplage d'attribut**
  - Bob appartient à un sous groupe $G \subset D'$ dont les membres ont les propriétés $P_1, \ldots, P_n$
- **Couplage de table**
  - Bob appartient au jeu de données D' dont les membres ont les propriétés $P_1, \ldots, P_n$
- **Attaque probabiliste**
  - L'adversaire améliore ses connaissances sur Bob après avoir obtenu D'

*Privacy-preserving data publishing: A survey of recent developments (Fung et al., 2010)

American Community Survey

B24082 | SEX BY CLASS OF WORKER AND MEDIAN EARNINGS IN THE PAST 12 MONTH

2020: ACS 5-Year Estimates Detailed Tables | Universe: Civilian employed population 16 years and over with earnings

Notes | Geos | Years | Topics (1) | Surveys | Codes | Hide | Transpose | Margin of Error | Restore | Excel | CSV | ZIP

	United States
Label	**Estimate**
Male:	46,020
Private for-profit wage and salary workers:	44,908
Employee of private company workers	43,976
Self-employed in own incorporated business workers	62,021
Private not-for-profit wage and salary workers	48,774
Local government workers	53,934
State government workers	53,255
Federal government workers	70,902
Self-employed in own not incorporated business workers and unpaid family workers	32,344
Female:	33,108
Private for-profit wage and salary workers:	30,999
Employee of private company workers	30,840
Self-employed in own incorporated business workers	38,076
Private not-for-profit wage and salary workers	40,256
Local government workers	42,387
State government workers	43,384
Federal government workers	60,096
Self-employed in own not incorporated business workers and unpaid family workers	19,937

Agrégation

- **Principe générale:** Combiner plusieurs enregistrements en un seul
- **Intuition:** Regrouper les enregistrements rend difficile l'inférence d'information spécifique à un individu

Agrégation

- **Attaques** exploitant informations auxiliaires

L'entreprise *Massive Dynamics* publie chaque année le salaire moyen de ces employés

Année	Salaire Moyen
2020	135,632
2021	136,758

Information auxiliaire

- 41 employés en 2020
- Carlos a été embauché entre les deux dates

Quel est le salaire de Carlos ?

$$\frac{\Sigma s_i}{41} = 135{,}632$$

$$\frac{s_{\text{Carlos}} + \Sigma s_i}{42} = 136{,}758$$

Le salaire de Carlos est 182,924

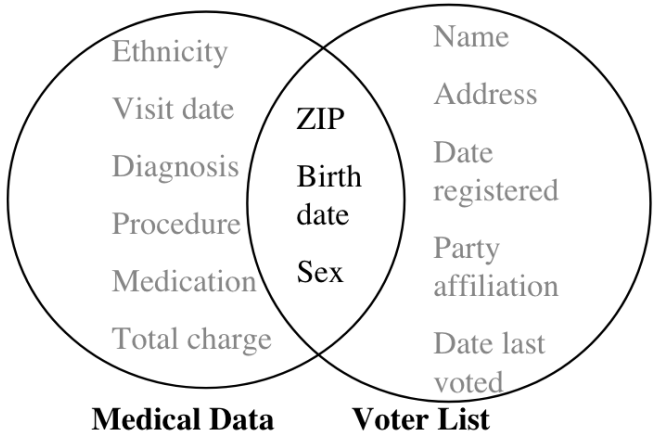
Nom	Age	Code Postal	Sexe	Condition
Yoshida	23	H2X1Y9	M	Asthme
Cohen	29	H2X3X2	F	Hypertension
Achebe	25	H2X3E2	F	Schizophrénie
Murphy	42	H2S3C7	M	Cancer
Bouchard	45	H2S2L8	F	Diabète
Smith	55	H2S2E7	M	Grippe

Nom	Age	Code Postal	Sexe	Condition
1	23	H2X1Y9	M	Asthme
2	29	H2X3X2	F	Hypertension
3	25	H2X3E2	F	Schizophrénie
4	42	H2S3C7	M	Cancer
5	45	H2S2L8	F	Diabète
6	55	H2S2E7	M	Grippe

Pseudonymisation

- **Principe générale:**
  Remplacer les identifiants par des pseudonymes

Unicité des QID



* Simple demographics often identify people uniquely (Sweeney, 2000)

In this document, I report on experiments I conducted using 1990 U.S. Census summary data to determine how many individuals within geographically situated populations had combinations of demographic values that occurred infrequently. It was found that combinations of few characteristics often combine in populations to uniquely or nearly uniquely identify some individuals. Clearly, data released containing such information about these individuals should not be considered anonymous. Yet, health and other person-specific data are publicly available in this form. Here are some surprising results using only three fields of information, even though typical data releases contain many more fields. It was found that 87% (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}. About half of the U.S. population (132 million of 248 million or 53%) are likely to be uniquely identified by only {place, gender, date of birth}, where place is basically the city, town, or municipality in which the person resides. And even at the county level, {county, gender, date of birth} are likely to uniquely identify 18% of the U.S. population. In general, few characteristics are needed to uniquely identify a person.

* Estimating the success of re-identifications in incomplete datasets using generative models (Rocher et al., 2019)

While rich medical, behavioral, and socio-demographic data are key to modern data-driven research, their collection and use raise legitimate privacy concerns. Anonymizing datasets through de-identification and sampling before sharing them has been the main tool used to address those concerns. We here propose a generative copula-based method that can accurately estimate the likelihood of a specific person to be correctly re-identified, even in a heavily incomplete dataset. On 210 populations, our method obtains AUC scores for predicting individual uniqueness ranging from 0.84 to 0.97, with low false-discovery rate. Using our model, we find that 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes. Our results suggest that even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization set forth by GDPR and seriously challenge the technical and legal adequacy of the de-identification release-and-forget model.

K-Anonymat

- Protège contre le **couplage d'enregistrement**

- **Principe générale: E**xige que chaque **classe d'équivalence** (ensemble d'éléments similaires par rapport à un QID) **contienne au moins k enregistrements. Généralisation et suppression.**
- **Intuition:** Chaque individu dans le jeu assaini est similaire à au moins k-1 autres sujets.

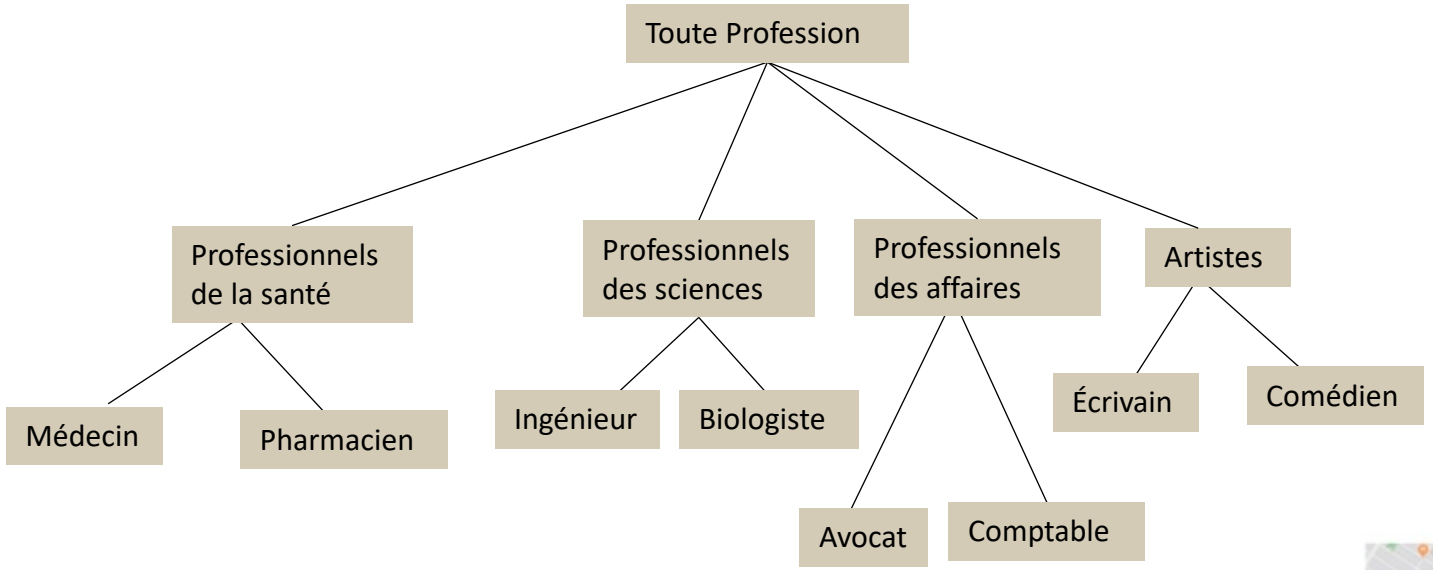
* Achieving k-anonymity privacy protection using generalization and suppression. (Sweeney, 2002)

K-Anonymat -- Exemple

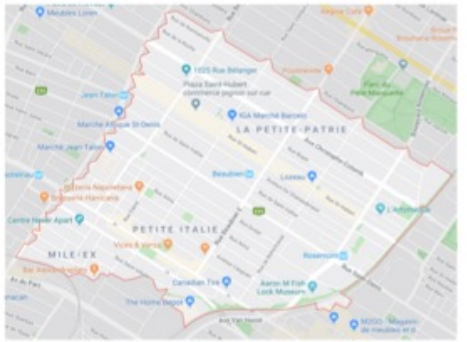
Nom	Age	Code Postal	Sexe	Condition
Yoshida	23	H2X1Y9	M	Grippe
Cohen	29	H2X3X2	F	Obésité
Achebe	25	H2X3E2	F	Grippe
Murphy	42	H2S3C7	M	Cancer
Bouchard	45	H2S2L8	F	Diabète
Smith	55	H2S2E7	M	Hypertension
Rachel	32	J3Y7G5	F	Cancer
James	38	J3Y3L1	M	Cancer
Carl	35	J3Y4W6	M	Cancer

Nom	Age	Code Postal	Sexe	Condition
*	< 30	H2X***	*	Grippe
*	< 30	H2X***	*	Obésité
*	< 30	H2X***	*	Grippe
*	> 40	H2S***	*	Cancer
*	> 40	H2S***	*	Diabète
*	> 40	H2S***	*	Hypertension
*	30-40	J3Y***	*	Cancer
*	30-40	J3Y***	*	Cancer
*	30-40	J3Y***	*	Cancer

K-anonymat -- Généralisation



H2X



H2S

K-Anonymat -- Attaque #1: Homogénéité

Nom	Age	Code Postal	Sexe	Condition
James	38	J3Y3L1	M	?

Nom	Age	Code Postal	Sexe	Condition
*	< 30	H2X***	*	Grippe
*	< 30	H2X***	*	Obésité
*	< 30	H2X***	*	Grippe
*	> 40	H2S***	*	Cancer
*	> 40	H2S***	*	Diabète
*	> 40	H2S***	*	Hypertension
*	30-40	J3Y***	*	Cancer
*	30-40	J3Y***	*	Cancer
*	30-40	J3Y***	*	Cancer

James a le Cancer

K-Anonymat -- Attaque #2: Information auxiliaire

Nom	Age	Code Postal	Sexe	Pays
Yoshida	23	H2X1Y9	M	Japon

Les Japonais ont parmi les taux d'obésité les plus bas au monde

Nom	Age	Code Postal	Sexe	Condition
*	< 30	H2X***	*	Grippe
*	< 30	H2X***	*	Obésité
*	< 30	H2X***	*	Grippe
*	> 40	H2S***	*	Cancer
*	> 40	H2S***	*	Diabète
*	> 40	H2S***	*	Hypertension
*	30-40	J3Y***	*	Cancer
*	30-40	J3Y***	*	Cancer
*	30-40	J3Y***	*	Cancer

Yoshida a la Grippe

L-Diversité

- Protège contre l'attaque #1 (*Homogénéité)*
  - *Les enregistrements contenue dans chaque groupe de taille k sont diversifié*
- Augmente la résistance à l'attaque #2 (Information auxiliaire)
  - *L'attaquant a besoin de l-1 propositions fausses* sur la cible
- Protège contre le **couplage d'enregistrement**
- Protège contre le **couplage d'attribut**

- **Principe générale: k-anonymat +** exige que **chaque classe d'équivalence** ait au moins **l valeurs bien représentées** pour chaque attribut sensible.

* l-diversity: Privacy beyond k-anonymity. (Machanavajjhala et al., 2007)

Métriques à usage général

- Dans beaucoup de cas, l'analyste ne sait pas comment le jeu de données publié sera utilisé
- **Mesures de similarité** entre le **jeu de données original** et le **jeu de données publié**
  - **Distorsion Minimale:** *Chaque généralisation ou suppression coûte une unité de distorsion*
  - **ILoss (Information Loss):** *Capture la perte d'information introduite par la généralisation en une valeur* v_g
    - $ILoss(v_g) = \frac{|v_g| - 1}{|D_A|}$
    - $|v_g|$: nombres de nœuds descendants de v_g
    - $|D_A|$: nombre de valeurs distinctes originales de l'attribut

Métriques à usage spécifique

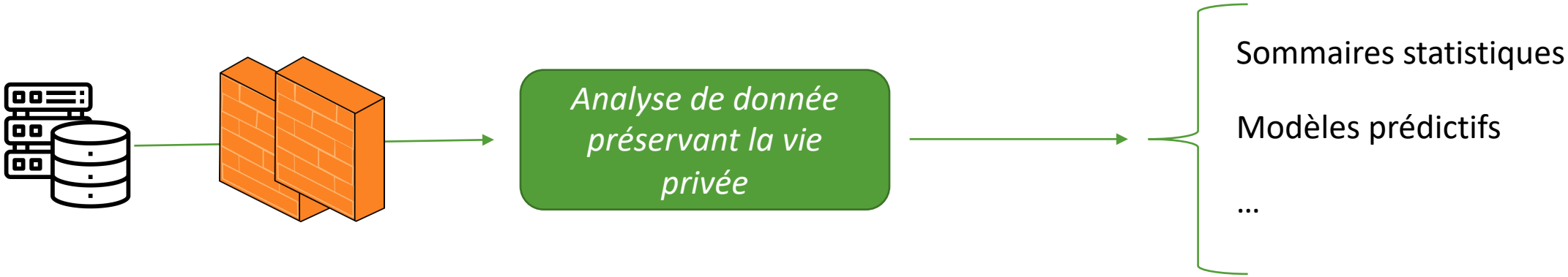
- Cas d’utilisation du jeu de données est connu à l’avance
  - Exploiter la connaissance de la tâche finale pour améliorer l’utilité
  - Exemple:
    - Coefficient de clustering
    - Taux d’erreur de classification …

Limites de l’anonymisation

- k-anonymat : couplage d’enregistrement
- l-diversité : couplage d’attribut
- Protection dépendante de la connaissance auxiliaire de l’adversaire
- Attaques par composition
  - publications indépendantes de jeux de données anonymisés
- Besoins
  - Protection contre le couplage de table et l’attaque probabiliste
  - Protection indépendante de la connaissance de l’adversaire
  - Quantification formelle du compromis vie-privée/utilité

- Composition Attacks and Auxiliary Information in Data Privacy. (Ganta et al., 2014)
- Attacks on Deidentification’s Defenses. (Cohen, 2020)

Changement de paradigme



Définition

- Jeux de données adjacents: ensembles de données qui ne diffèrent que par **l'ajout ou le retrait** d'un enregistrement.
- Un mécanisme aléatoire $\boldsymbol{M : X_n \rightarrow Y}$ fournit une **ε-confidentialité différentielle** si pour chaque paire d'ensembles de données adjacents D, D′ ∈ X_n et pour chaque sous-ensemble de sortie S ⊆ Y:

$$\frac{\mathrm{P\,r}[\mathrm{M(D)} \in \mathrm{S}]}{\mathrm{P\,r}[\mathrm{M(D')} \in \mathrm{S}]} \leq \exp(\epsilon)$$

Le résultat d'une fonction calculée sur un jeu de donnée soit (approximativement) la même qu'un individu particulier soit présent ou absent

- ε ≥ 0 est le paramètre de confidentialité

* Calibrating noise to sensitivity in private data analysis. (Dwork et al., 2006)
* The Algorithmic Foundations of Differential Privacy (Dwork and Roth, 2014)

Requêtes numériques

- Calculer une fonction à valeur numériques $\boldsymbol{f: N^{|X|} \rightarrow R^{k}}$
- La fonction $\boldsymbol{f}$ associe à une **base de données privée $\boldsymbol{D}$, $\boldsymbol{k}$ nombres réels**
- Comment obtenir un algorithme différentiellement confidentiel différentiel pour calculer $\boldsymbol{f(D)}$?
- On sait qu'il faut rajouter un bruit
  - Comment calculer la **quantité** de ce bruit?
  - **Où** insérer ce bruit?
  - Quelle **erreur** introduit l'ajout de ce bruit?
  - Peut-on la **quantifier**?

Sensibilité globale l_1

- Permet de **calculer la quantité de bruit**
- **Comment** un **enregistrement peut affecter** la valeur d'une fonction
  - La **quantité de bruit** nécessaire pour **camoufler** n'importe quelle **contribution individuelle**
- La sensibilité globale l_1 d'une fonction $\boldsymbol{f}: \boldsymbol{N}^{|X|} \rightarrow \boldsymbol{R}^{\boldsymbol{k}}$ est
  - $\Delta_1(f) = \max_{D,D':\left|\left|D-D'\right|\right|_1 \leq 1} ||f(D) - f(D')||_1$
- **Exemple**
  - $\Delta_1(Comptage) = 1$
    - L'ajout ou le retrait d'un individu change le comptage d'au plus 1

Mécanisme Laplace

$M_{Lap}(D, \boldsymbol{f}: \boldsymbol{N}^{|\boldsymbol{X}|} \rightarrow \boldsymbol{R}^{\boldsymbol{k}}, \epsilon)$

1. Calculer $\Delta = \Delta_1(f)$
2. Pour $k = 1, \dots, K$
   - $Y_k \sim Lap(0, \frac{\Delta}{\epsilon})$ indépendamment pour chaque k
3. Retourner $f(D) + Y$
   - $Y = (Y_1, \dots, Y_K) \in R^k$

Données de mobilité

Géolocalisation?

- La géolocalisation associe une **position** géographique **à un objet**, souvent personnel
  - La position, in fine, est donc associée **à un individu**
  - Si divulguée, peut mener à des bris de vie privée

Pourquoi il faut les protéger

- **Raison 1:** les données de mobilité ont un **fort potentiel d'inférence**

Unique in the Crowd: The privacy bounds of human mobility

Yves-Alexandre de Montjoye[1,2], César A. Hidalgo[1,3,4], Michel Verleysen[2] & Vincent D. Blondel[2,5]

[1]Massachusetts Institute of Technology, Media Lab, 20 Ames Street, Cambridge, MA 02139 USA, [2]Université catholique de Louvain, Institute for Information and Communication Technologies, Electronics and Applied Mathematics, Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium, [3]Harvard University, Center for International Development, 79 JFK Street, Cambridge, MA 02138, USA, [4]Instituto de Sistemas Complejos de Valparaíso, Paseo 21 de Mayo, Valparaíso, Chile, [5]Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. We coarsen the data spatially and temporally to find a formula for the uniqueness of human mobility traces given their resolution and the available outside information. This formula shows that the uniqueness of mobility traces decays approximately as the 1/10 power of their resolution. Hence, even coarse datasets provide little anonymity. These findings represent fundamental constraints to an individual's privacy and have important implications for the design of frameworks and institutions dedicated to protect the privacy of individuals.

Pourquoi il faut les protéger

- **Raison 2:** les données de mobilité sont **collectées en très grande quantité**

Projet Mobilitics
Inria, 2016

Exemple de jeu de données de mobilité

Identifiers	Spatiotemporal points *[with non-positioning information]*					Attributes		
Alice +39 320 191 7047	45.061679, 7.677888 2018/01/24 08:05	45.062518, 7.662191 2018/01/24 10:32	45.062288, 7.671960 2018/01/24 14:18	45.058935, 7.686642 2018/01/24 19:41	45.070908, 7.684926 2018/01/24 22:01	Female	Accountant	€ 36,000
Bob +39 339 205 3011	45.068962, 7.698691 2018/01/24 07:10	45.068962, 7.698691 2018/01/24 07:13	45.067780, 7.694743 2018/01/24 09:20	45.079630, 7.671697 2018/01/24 16:18	45.081024, 7.625563 2018/01/24 16:53	Male	Engineer	€ 74,000
Charlie +39 347 772 3345	45.068962, 7.698691 2018/01/24 07:10	45.068962, 7.698691 2018/01/24 07:13	45.033696, 7.675753 2018/01/24 11:53	45.040004, 7.676439 2018/01/24 11:54	45.081024, 7.625563 2018/01/24 16:53	Male	Lawyer	€ 74,000
Dave +39 328 055 4606	45.094756, 7.526836 2018/01/24 12:26	45.090878, 7.528896 2018/01/24 13:13	45.066634, 7.515850 2018/01/24 15:17	45.063240, 7.522717 2018/01/24 17:48	45.092090, 7.524776 2018/01/24 19:33	Female	Consultant	€ 103,000
Erin +39 348 223 1098	45.135216, 7.760983 2018/01/24 07:11 *[incoming SMS]*	45.008652, 7.532330 2018/01/25 03:44 *[outgoing call]*	45.109635, 7.640991 2018/01/24 20:32 *[location area update]*	45.109635, 7.640991 2018/01/24 20:58 *[outgoing SMS]*	45.105176, 7.641850 2018/01/24 21:48 *[outgoing call]*	Male	Plumber	€ 31,000
Frank +39 333 879 4903	45.064937, 7.641850 2018/01/24 23:29 *[outgoing call]*	45.008652, 7.532330 2018/01/25 03:44 *[incoming SMS]*	45.008652, 7.532330 2018/01/24 22:48 *[outgoing SMS]*	45.004768, 7.535076 2018/01/24 23:35 *[incoming SMS]*	45.003069, 7.532673 2018/01/25 00:42 *[incoming SMS]*	NA	NA	NA

Figure 1: Example of database of trajectory micro-data. Each record is composed of an identifier (left), a spatiotemporal trajectory (middle), and additional attributes (right). In this specific example, the person's name and phone address are the identifiers, and spatiotemporal points in the trajectory are GPS locations augmented with non-positioning information, within brackets, about their mobile communication activity. Attributes consist of gender, employment and revenue.

* Privacy in trajectory micro-data publishing: a survey. (Fiore *et* al., 2020)

Objectifs d’attaque

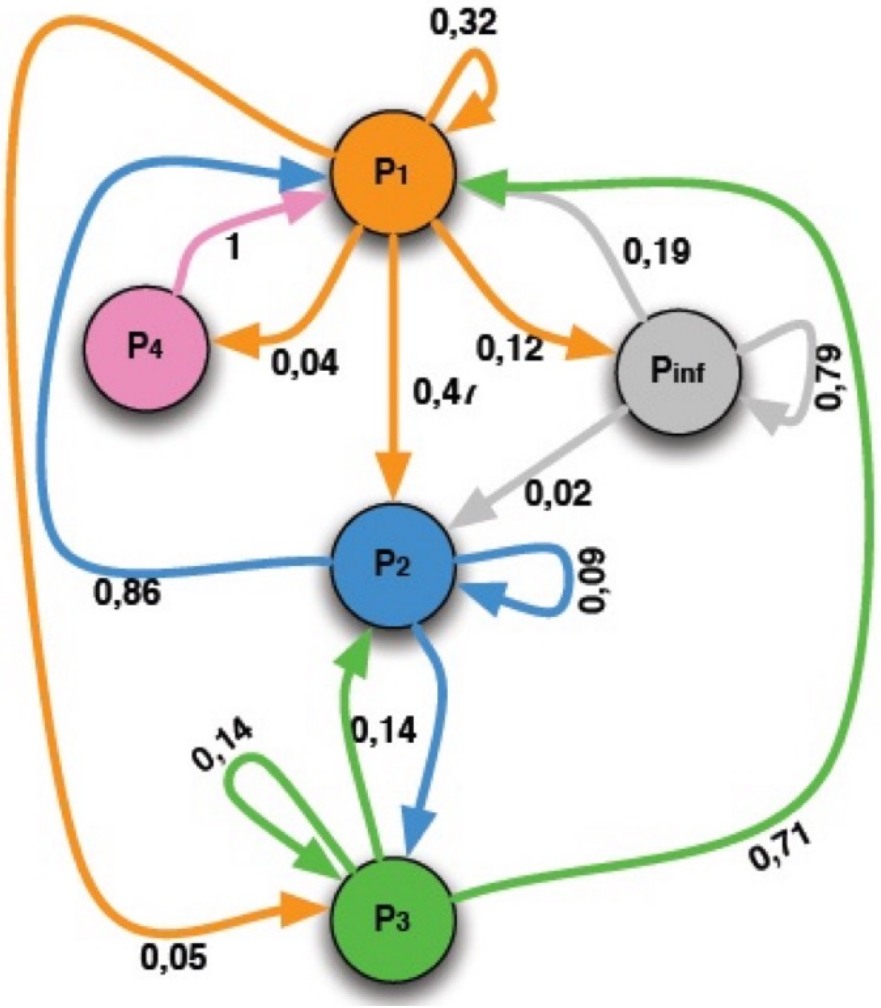
- **Identifier des points d’intérêts (POI)**
- **Prédire les déplacements**
- **Apprendre la sémantique** des localisations et des mouvements
- **Dé-anonymiser** des données géolocalisées
- **Chaîner** un individu dans **différentes bases de données**
- **Reconstruire un réseau social**
- **Prédire des attributs démographiques**

Points d’intérêts

- **POI**: un site ou un point digne d’intérêt pour un individu (domicile, lieu de travail, …) ou un groupe d’individus (boutique, boulangerie, stations de métro, …)
- Identifié par sa position (lat, lon), éventuellement avec une étiquette sémantique
- Obtenu de différentes façons:
  - Une BDD (connaissance externe): OpenStreetMap
  - Par questionnaire: où habitez/travaillez-vous?
  - Heuristiques à partir de données de localisation
    - **Heuristique Begin/End**
      - La mobilité quotidienne commence/s’arrête souvent au domicile
      - Premier/Dernier point de la journée -> Domicile
  - Attaques d’inférence (clustering, etc.)

Modéliser la mobilité d'un individu

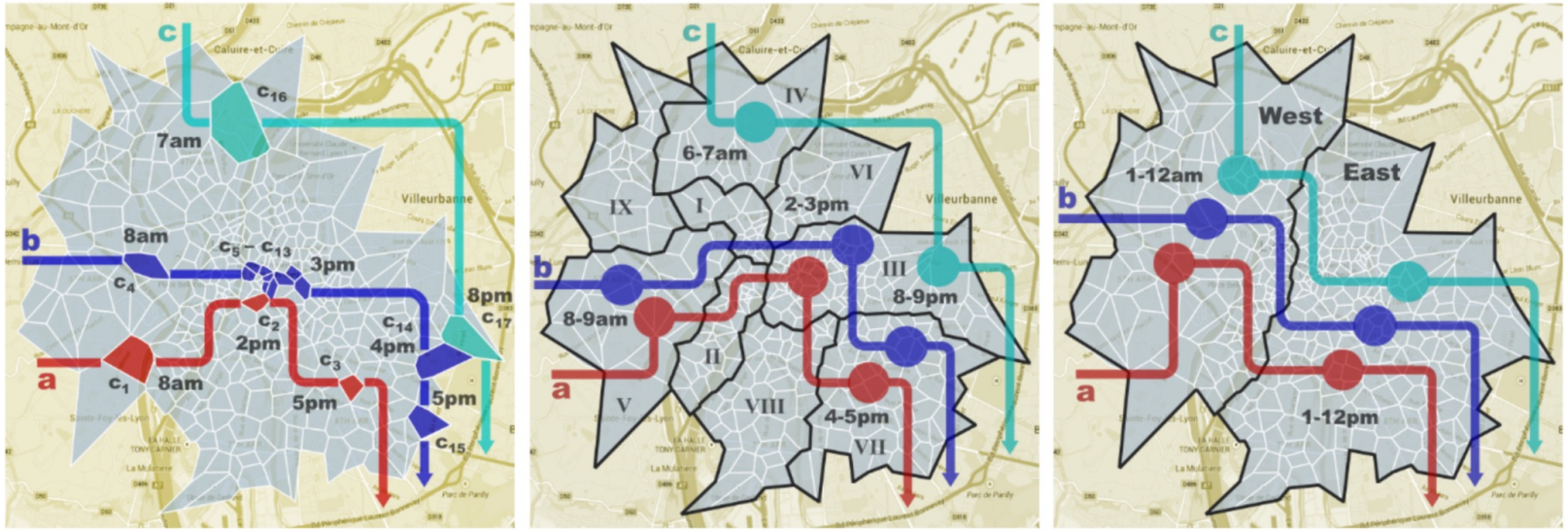
- Différentes formes possibles: **graphes pondérés**, **chaînes de Markov**, etc.
  - **États** = POIs
  - **Arrêtes** = transitions entre les POIs
  - **Poids** = probabilité de passer d'un POI à un autre
- Modèle de la **mobilité passée** qui peut être utilisé pour la prédiction de la **mobilité future**
  - **POI le plus probable** à partir de la position actuelle



Anonymisation de données de mobilité?

Agrégation spatio-temporelle

- Généralisation du couple (localisation, temps)

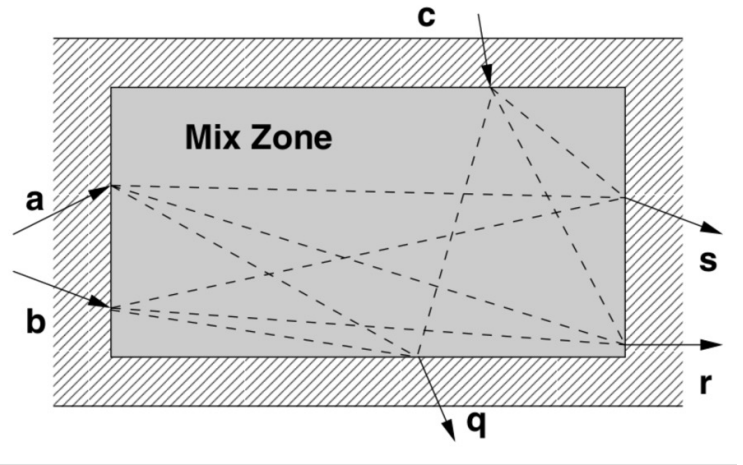


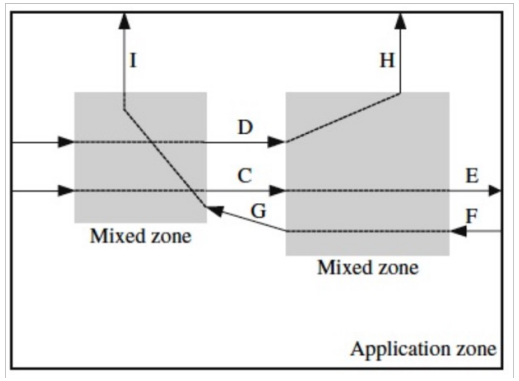
Autres transformations possibles

- **Échantillonner**
- **Échanger** les traces de deux individus
- **Enlever** des positions trop sensibles
- **Ajouter** de faux enregistrements ou de faux individus
- **Apprentissage machine**: apprendre un **modèle génératif** de la mobilité des individus de la BDD et **générer de nouvelles traces** qui ont de bonnes propriétés mais qui ne correspondent pas à des mobilités réelles

Mix-Zones

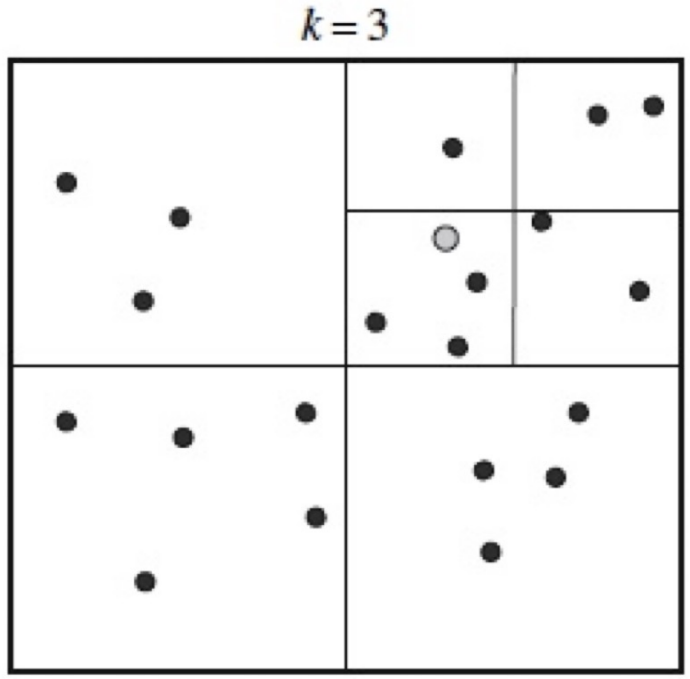
- Sur la base des Mix-nets, les zones de mixage
  - Non chaînabilité entre zones
  - Des zones ou aucune localisation n'est enregistrée
  - Changement de pseudo en sortie
- Besoin d'entropie donc problème si:
  - Majorité des traces ont le même parcours
  - Une seule trajectoire passe par zone
- Le choix des zones est primordial





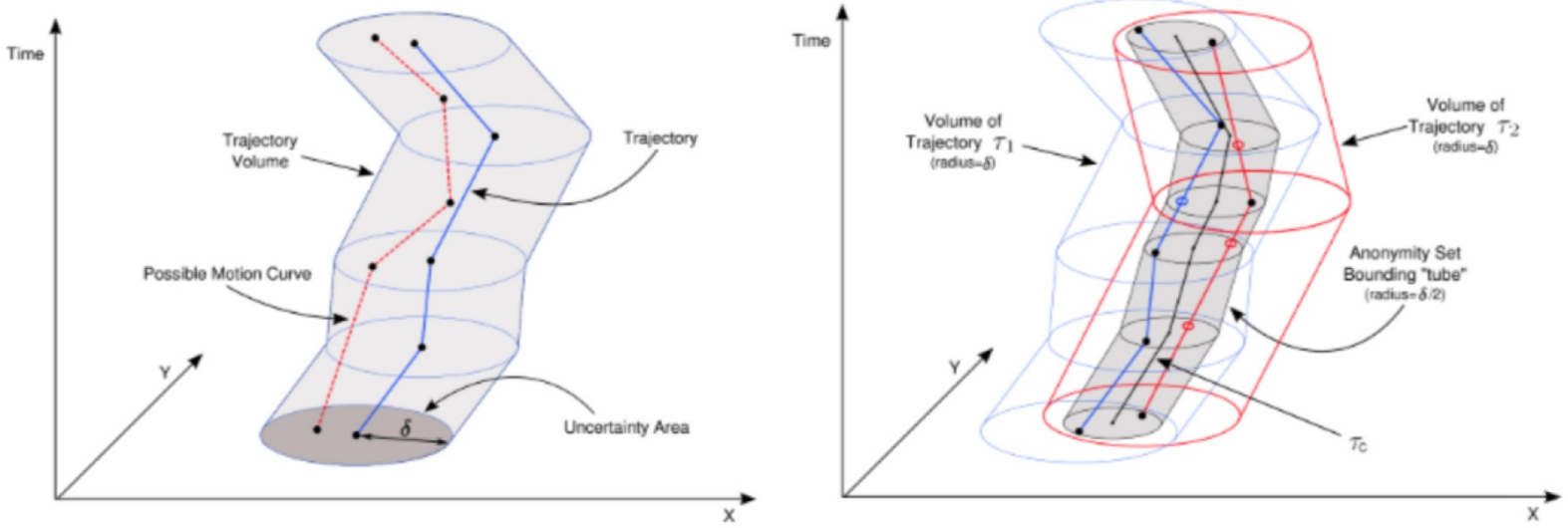
K-anonymat géographique

- **Couverture spatiale**: extension du **k-anonymat** aux <u>données spatio-temporelles</u>
- À chaque **unité de temps**, **chaque individu** est dans une **zone partagée** par au moins **k-1 autres individus**
- **<u>Exemple</u>**: **découpage récursive de l'espace**



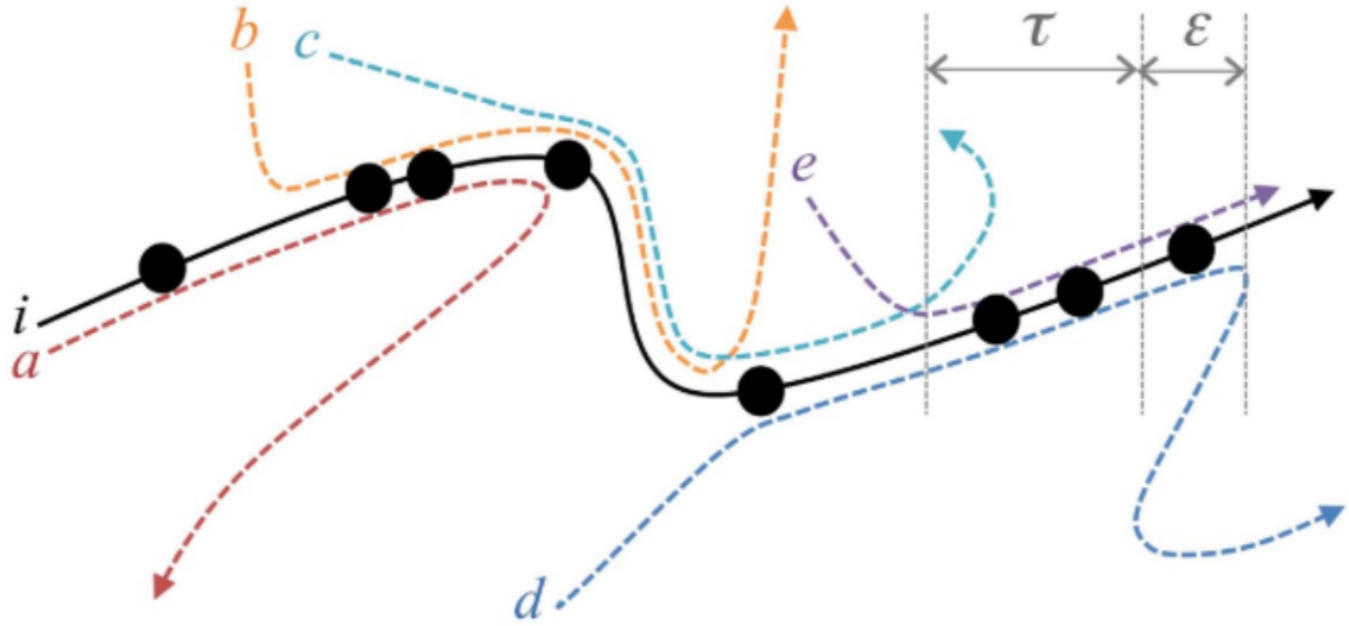
(k, δ)-anonymat

- Les **trajectoires** doivent se trouver **au plus à une distance δ** de **k-1 autres trajectoires**



$k^{\tau,\epsilon}$-anonymat

- Un adversaire qui **observe** la trajectoire pendant une **durée τ** (le passé) et possède une **capacité d'inférence ϵ** (le futur) est **incapable** de la **distinguer de k trajectoires**

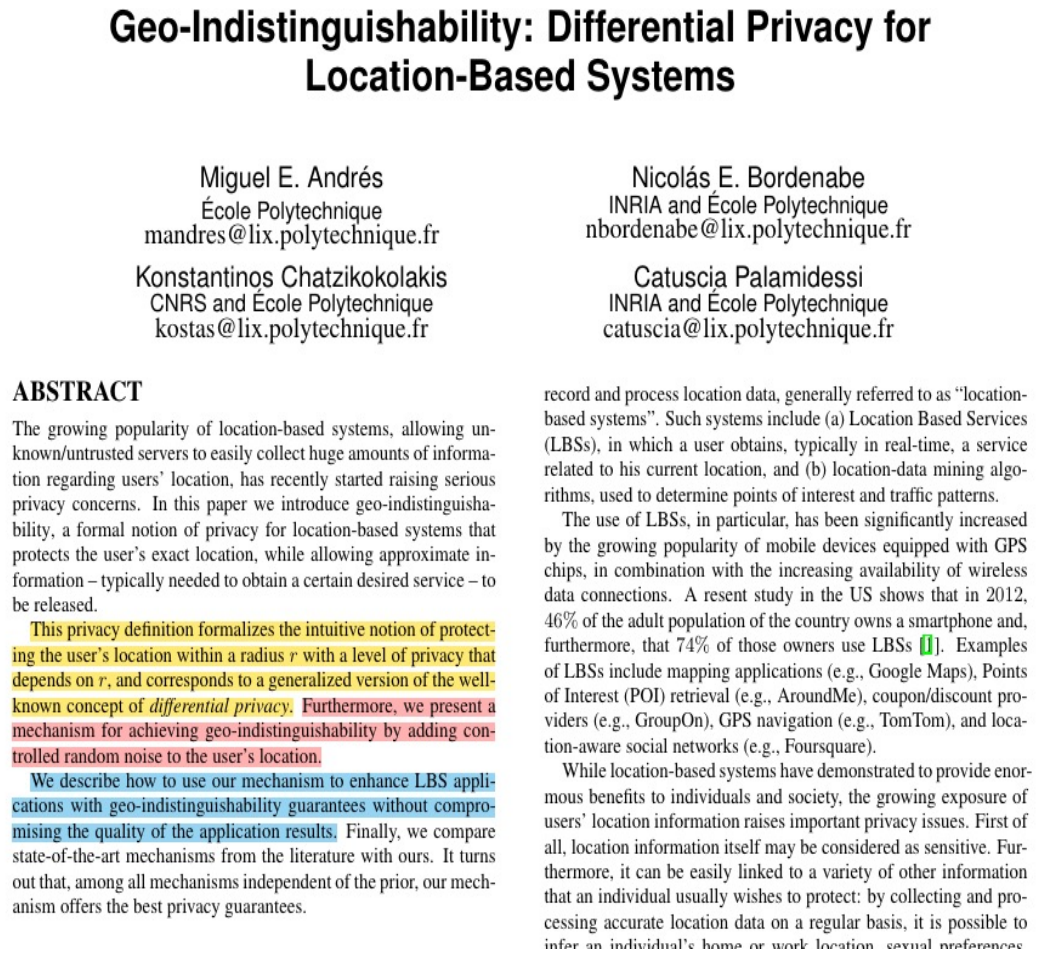


Limites du k-anonymat (et variantes)

- Garantie dépend fortement du facteur k
  - **Indistingabilité** = $\frac{1}{k}$
  - Quel k choisir pour ne pas trop diminuer l'utilité (souvent k=2 pour des trajectoires)?
- **Grande surface d'attaque**
  - Protection contre **l'unicité** des profils mais c'est tout!
    - Et si les k individus se rendent tous à l'hôpital, que puis-je inférer?
  - Restent possibles: attaques **d'appartenance**, de **co-localisation** et **divulgation des lieux visités**, etc
  - Attaques de composition: sensible à la sortie de nouveaux jeux de données (croisement, **chaînage**)

La confidentialité différentielle à la rescousse!

- Un utilisateur bénéficie d'une $\boldsymbol{l}$ **−confidentialité** si toute paire de localisations distante d'au plus **r** produit des observations ayant des distributions similaires, pour une notion de similarité dépendant de l
- **ϵ-géo-indistingabilité**
  - un mécanisme satisfait la ϵ-géo-indistingabilité si et seulement si pour tout rayon $r > 0$, l'utilisateur bénéficie d'une ϵr-confidentialité dans le rayon r
  - Efficace pour protéger des points uniques (POI)
  - Malheureusement pas pour une trajectoire
    - Trop de corrélation entre des position en mouvement
    - Idée: échantillonner?



Geo-Indistinguishability: Differential Privacy for Location-Based Systems

Miguel E. Andrés
École Polytechnique
mandres@lix.polytechnique.fr

Nicolás E. Bordenabe
INRIA and École Polytechnique
nbordenabe@lix.polytechnique.fr

Konstantinos Chatzikokolakis
CNRS and École Polytechnique
kostas@lix.polytechnique.fr

Catuscia Palamidessi
INRIA and École Polytechnique
catuscia@lix.polytechnique.fr

ABSTRACT

The growing popularity of location-based systems, allowing unknown/untrusted servers to easily collect huge amounts of information regarding users' location, has recently started raising serious privacy concerns. In this paper we introduce geo-indistinguishability, a formal notion of privacy for location-based systems that protects the user's exact location, while allowing approximate information – typically needed to obtain a certain desired service – to be released.

This privacy definition formalizes the intuitive notion of protecting the user's location within a radius r with a level of privacy that depends on r, and corresponds to a generalized version of the well-known concept of *differential privacy*. Furthermore, we present a mechanism for achieving geo-indistinguishability by adding controlled random noise to the user's location.

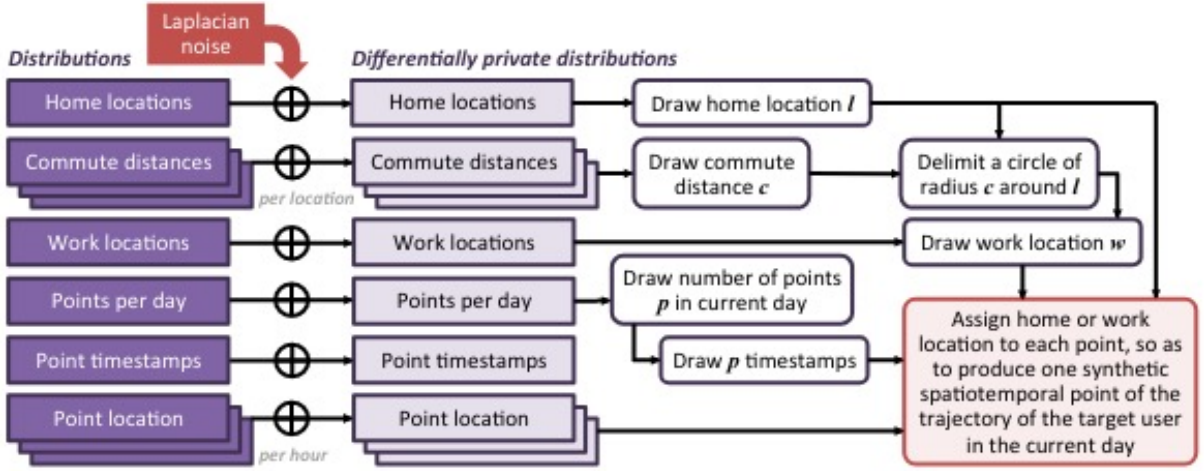
We describe how to use our mechanism to enhance LBS applications with geo-indistinguishability guarantees without compromising the quality of the application results. Finally, we compare state-of-the-art mechanisms from the literature with ours. It turns out that, among all mechanisms independent of the prior, our mechanism offers the best privacy guarantees.

record and process location data, generally referred to as "location-based systems". Such systems include (a) Location Based Services (LBSs), in which a user obtains, typically in real-time, a service related to his current location, and (b) location-data mining algorithms, used to determine points of interest and traffic patterns.

The use of LBSs, in particular, has been significantly increased by the growing popularity of mobile devices equipped with GPS chips, in combination with the increasing availability of wireless data connections. A resent study in the US shows that in 2012, 46% of the adult population of the country owns a smartphone and, furthermore, that 74% of those owners use LBSs [1]. Examples of LBSs include mapping applications (e.g., Google Maps), Points of Interest (POI) retrieval (e.g., AroundMe), coupon/discount providers (e.g., GroupOn), GPS navigation (e.g., TomTom), and location-aware social networks (e.g., Foursquare).

While location-based systems have demonstrated to provide enormous benefits to individuals and society, the growing exposure of users' location information raises important privacy issues. First of all, location information itself may be considered as sensitive. Furthermore, it can be easily linked to a variety of other information that an individual usually wishes to protect: by collecting and processing accurate location data on a regular basis, it is possible to infer an individual's home or work location, sexual preferences

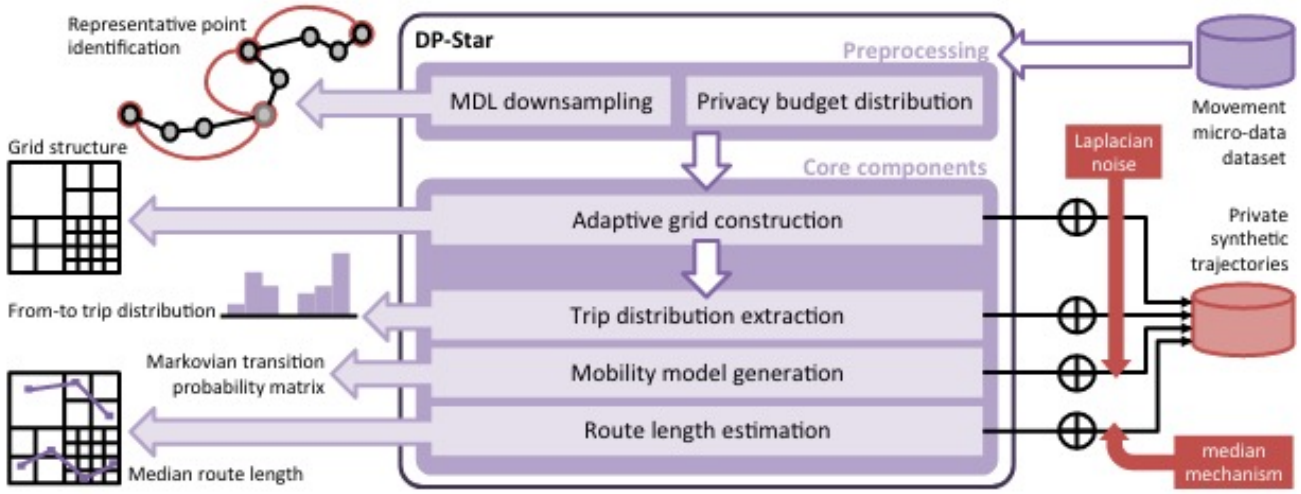
Trajets synthétique: DP-Where



1. Création de distributions pour décrire le jeu de données originale
   - Distribution spatiale des domiciles, lieux de travails, POIs
   - Nombre de localisation dans les trajectoires
2. Appliquer le mécanisme Laplace sur ces distributions
3. Génération de trajets synthétiques

* Dp-where: Differentially private modeling of human mobility. (Mir et al., 2013)

Trajets synthétique: DP-Star



1. Prétraitement des données + répartition du budget de confidentialité
2. Générations de différentes représentations
   - Espace: (e.g. KD-Tree)
   - Trajets Origine-Destination: Distributions de probabilité
   - Structure interne des trajets: Modèle de Markov
   - Longueur des routes: médiane des distances des trajectoires
3. Création des représentions différentiellement confidentielles
4. Génération de trajets synthétiques à partir des représentations bruitées

* Differentially private and utility preserving publication of trajectory data. (Gursoy et al., 2018)