

Introduction to Privacy and Surveillance

claude.castelluccia@inria.fr



So What is Privacy?

- A nice introductory video:
 - [*https://www.privacyinternational.org/node/568*](https://www.privacyinternational.org/node/568)

PRIVACY DEFINITIONS

What is Privacy?

- Abstract and subjective concept, hard to define
- Dependent on cultural issues
- A couple of popular definitions:
 - “The right to be let alone”
 - Focus on freedom from intrusion
 - “Informational self-determination”
 - Focus on control
- Privacy is a fundamental right!
 - Universal Declaration of Human Rights of UN (article 12), 1948
 - European directive 95 on the protection of data protection
 - Btw privacy > protection of data protection

The Universal Declaration of Human Rights

**"NO ONE SHALL BE SUBJECTED
TO ARBITRARY INTERFERENCE
WITH HIS PRIVACY, FAMILY,
HOME OR CORRESPONDENCE".**

Taxonomy of Privacy Threats [Solove]

Privacy **threats** we are trying to protect against (out of 16 identified by Solove)

- **Surveillance**: monitoring of electronic transactions
 - Preventive properties: anonymity, unobservability
- **Interrogation**: forcing people to disclose information
 - Preventive property: plausible deniability
- **Aggregation/Inference**: combining several sources of information
 - Preventive property: unlinkability
- **Identification**: connecting data to individuals.
 - Preventive properties: anonymity and unlinkability

What is Privacy for Computer Science?

- For CS, Privacy = Personal Data Protection
- How do we formalize privacy properties in computer systems?
 - Anonymity
 - Unlinkability
 - unobservability

Privacy properties from a technical point of view: **Anonymity**

- **Hiding link** between **identity and action**/piece of information.
 - Reader of a web page, person accessing a service
 - Sender of an email, writer of a text
 - Person to whom an entry in a database relates
- Pfitzmann-Hansen terminology:
 - “Anonymity is the state of being not identifiable within a set of subjects, the anonymity set”
 - “The anonymity set is the set of all possible subjects who might cause an action”
 - “Anonymity is the stronger, the larger the respective anonymity set is and the more evenly distributed the sending or receiving, respectively, of the subjects within that set is.”
 - *Probabilistic definition*

Source: *Anonymity, Unobservability, Pseudonymity, and Identity Management – A Proposal for Terminology*

Privacy properties from a technical point of view: **Unlinkability**

Hiding link between **two or more actions** / identities / pieces of information. Examples:

- Two anonymous letters written by the same person
- Two web page visits by the same user
- Entries in two databases related to the same person
- Two people related by a friendship link
- Same person spotted in two locations at different points in time
- Pfitzmann-Hansen terminology:
 - “Unlinkability of two or more items means that within a system, these items are no more and no less related than they are related concerning the a-priori knowledge”
 - Focus on the information leakage of a system

Privacy properties from a technical point of view: **Unobservability**

- **Hiding user activity (i.e. traffic analysis).** Examples:
 - Impossible to see whether someone is accessing a web page
 - Impossible to know whether an entry in a database corresponds to a real person
 - Impossible to distinguish whether someone or no one is in a given location
- Pfitzmann-Hansen terminology:
 - “Unobservability is the state of items of interest being indistinguishable from any item of interest at all”
 - “Sender unobservability then means that it is not noticeable whether any sender within the unobservability set sends.”

Anonymity < Unlinkability < Unobservability

PRIVACY METRICS: How to measure Privacy?

Can we “measure” privacy (Anonymity) ?

- Need to specify
 - Privacy properties we want to achieve
 - Adversary model: goals and capabilities
- Typically, adversaries are able to obtain probabilistic information.
 - Examples:
 - Probability of a person being the anonymous subject we want to identify (limited # of people in the world)
 - Probability of two information items being related to each other (e.g., two web page requests coming from the same user)
- Many proposals, open research field
 - Ex: information theoretic approach

A Primer on Info. Theory & Privacy

- There are around 7 billion humans on the planet:
 - the identity of a random, unknown person contains just under 33 bits of entropy ($2^{33} \sim 8$ billion).
 - When we learn a new fact about a person, that fact reduces the entropy of their identity by a certain amount.
- There is a formula to say how much:
 - $\Delta S = -\log_2 \Pr(X=x)$

Where ΔS is the reduction in entropy, measured in bits, and $\Pr(X=x)$ is simply the probability that the fact would be true of a random person.

A Primer on Info. Theory & Privacy

- For example:
 - Starsign: $\Delta S = -\log_2 \Pr(\text{STARSIGN}=\text{capricorn}) = -\log_2 (1/12) = 3.58$ bits of information
 - Birthday: $\Delta S = -\log_2 \Pr(\text{DOB}=2\text{nd of January}) = -\log_2 (1/365) = 8.51$ bits of information
- Note that if you combine several facts together, you might not learn anything new; for instance, telling me someone's starsign doesn't tell me anything new if I already knew their birthday.

A Primer on Info. Theory & Privacy

- Does fact/info about a person *identifies* that person?
 - If all I know about a person is their ZIP code, I don't know who they are.
 - If all I know is their date of birth, I don't know who they are.
 - If all I know is their gender, I don't know who they are.
 - But it turns out that if I know **these three things** about a person, I could probably deduce their identity!
- Entropy allows us to measure how close a fact comes to revealing somebody's identity uniquely.
 - entropy is the number of different possibilities there are for a random variable:
 - if there are two possibilities, there is 1 bit of entropy;
 - if there are four possibilities, there are 2 bits of entropy, etc. Adding one more bit of entropy doubles the number of possibilities.

A Primer on Info. Theory & Privacy

- Each starsign and birthday was assumed to be equally likely.
- The calculation can also be applied to facts which have non-uniform likelihoods.
- For instance, the likelihood that an unknown person's ZIP code is 90210 (Beverly Hills, California) is different to the likelihood that their ZIP code would be 40203 (part of Louisville, Kentucky).
 - As of 2007, there were 21,733 people living in the 90210 area, only 452 in 40203, and around 10 million in Moscow, 6,6 billion on the Planet.
 - Knowing my ZIP code is 90210: $\Delta S = -\log_2 (21,733/6,625,000,000) = 18.21$ bits
 - Knowing my ZIP code is 40203: $\Delta S = -\log_2 (452/6,625,000,000) = 23.81$ bits
 - Knowing that I live in Moscow: $\Delta S = -\log_2 (10524400/6,625,000,000) = 9.30$ bits

How much entropy is needed to identify someone?

- if we know someone's birthday, and we know their ZIP code is 40203, we have $8.51 + 23.81 = 32.32$ bits;
 - that's almost, but perhaps not quite, enough to know who they are
 - there might be a couple of people who share those characteristics.
- Add in their gender, that's 33.32 bits, and we can probably say exactly who the person is!

An Application To Web Browsers

- How would this paradigm apply to web browsers?
- In addition to the commonly discussed "identifying" characteristics of web browsers, like IP addresses and tracking cookies, there are more subtle differences between browsers that can be used to tell them apart (**fingerprints**).
- One significant example is the User-Agent string, which contains the name, operating system and precise version number of the browser, and which is sent every web server you visit.
- A typical User Agent string looks something like this:

```
Mozilla/5.0 (Windows; U; Windows NT 5.1; en-GB;  
rv:1.8.1.6) Gecko/20070725 Firefox/2.0.0.6
```

Application To Web Browsers (2)

- It turns out that that UA is quite useful for telling different people apart on the net.
- User Agent strings contain about 10.5 bits of identifying information,
 - if you pick a random person's browser, only one in 1,500 other Internet users will share their User Agent string.
- **Test your browser:**
<https://panopticlick.eff.org>
- So even if someone use TOR...the server can still get information about him!

Privacy metrics: challenges

- Modeling the background information available to the adversary
 - What kind of prior information / other sources of information does the attacker have access to?
- Modeling user behavior
 - Are users going to behave as we predict? What if they do not?
- Finding expressive metrics
 - How to interpret the result?
 - What is a “good” level of privacy?
- Metrics that generic enough for a variety of systems
 - Many proposals are ad-hoc

Legal Aspects

Legal Aspects

(slides from S. Tavernier- CNIL)



- I. RAPPEL DE QUELQUES NOTIONS CLES « INFORMATIQUE ET LIBERTES »
- II. LES PRINCIPES DE PROTECTION DES DONNEES ET LEUR APPLICATION A LA RECHERCHE
- III. FORMALITES ET CIL
- IV. QUELLES EVOLUTIONS POSSIBLES

A nice Introductory Video

<https://www.privacyinternational.org/node/570>

1

Quelques notions clés « Informatique et Libertés »

La protection des données personnelles en Europe et en France

- La directive européenne du 24 octobre 1995 : en voie de révision
- La loi informatique et libertés du 6 janvier 1978 modifiée en 2004:une éthique de l'informatique appliquée aux données personnelles
- La CNIL, une autorité administrative indépendante:
 - Collège de 17 membres
 - Budget: 17 millions d'euros
 - Effectifs: 180 personnes

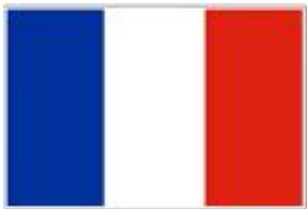
1. Donnée à caractère personnel: définitions

Directive européenne de 1995



« toute information concernant une personne identifiée ou identifiable ; est réputée identifiable une personne qui peut être identifiée, directement ou indirectement, notamment par référence à un numéro d'identification ou à un ou plusieurs éléments spécifiques, propres à son identité physique, physiologique, psychique, économique, culturelle ou sociale ; (...) pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens susceptibles d'être raisonnablement mis en oeuvre, soit par le responsable du traitement, soit par une autre personne, pour identifier ladite personne » (considérant 26 et article 2)

* Loi « Informatique et Libertés » :



« toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. Pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne » (article 2)

Donnée à caractère personnel: une notion large

toute information relative à une **personne physique** identifiée ou identifiable, **directement ou indirectement** par référence à un numéro d'identification ou un ou plusieurs éléments spécifiques ou par recoupement



La donnée à caractère personnel selon la CNIL

La CNIL: une interprétation large et une appréciation au cas par cas :

/ nature des données:

Données relatives à l'état civil: noms et prénoms

Données se rapportant indirectement à l'état civil :ex. initiales des noms et prénoms, date et lieu de naissance, n° de SS...

Données de localisation spatiotemporelle: commune de résidence, lieu de travail, données de geolocalisation, indications de dates (d'examens, d'hospitalisation, ...)

Données spécifiques, cas isolés (pathologies rares, nature d'emploi...)

Numéros n°s de tel, , n°s de CB, n°s aléatoires renvoyant à une liste de correspondance avec identités

Données biometriques, photos, voix...

Données techniques : adresse ip, adresse mac,, données de connexion ...
metadonnées, ...

...

/ l'importance relative de l'échantillon de population concernée;

/ le type de traitement effectué : ex. data mining, big data.

- **La loi ne s'applique pas aux données anonymisées!**

2

Les 5 règles d'or de la protection des données

Les règles de la protection des données

Finalité

Les DCP contenues dans un traitement ne sont recueillies et traitées que pour un usage déterminé et légitime, préalablement défini

Tout détournement de finalité est passible de sanctions pénales

Proportionnalité et pertinence

Seules les informations pertinentes et nécessaires au regard des objectifs poursuivis doivent être traitées

Durée limitée de conservation

Les informations ne peuvent être conservées de façon indéfinie dans les fichiers informatiques

Une durée de conservation précise doit être fixée en fonction de la finalité du fichier

Sécurité et confidentialité

Le responsable du traitement doit prendre les mesures nécessaires pour garantir la sécurité et la confidentialité des données

Les données peuvent néanmoins être communiquées à des « Tiers autorisés »

Respect des droits des personnes

Les personnes dont les données sont utilisées dans un traitement ont un droit d'information, d'accès, de rectification, de suppression et d'opposition/consentement sur leurs données

1. Finalité du traitement

- **PRINCIPE**

Les données « sont collectées pour des finalités déterminées, explicites et légitimes et ne sont pas traitées ultérieurement de manière incompatible avec ces finalités » (art 6)

- **Pierre angulaire de la réglementation I&L :**
 - fait le lien entre les données, les traitements et les missions de l'organisme qui les met en oeuvre
 - déclaration des traitements par finalité et non par logiciel/fichier utilisé
 - va permettre de déterminer les catégories de données susceptibles d'être traitées et leur durée de conservation

-

2. Pertinence et proportionnalité des données

• PRINCIPE

Les données doivent être « *adéquates, pertinentes et non excessives* au regard des finalités pour lesquelles elles sont collectées et de leurs traitements ultérieurs » (art 6 al.3)

→ seules peuvent être collectées les données strictement nécessaires

→ **protection particulière accordée à certaines catégories de données**

✓ Données dites « sensibles » :

- Origines raciales ou ethniques
- Opinions politiques, philosophiques ou religieuses, appartenances syndicales
- Données relatives à la santé ou à la vie sexuelle

✓ Données relatives aux infractions, condamnations et mesures de sûreté

✓ Numéro de sécurité sociale (NIR)

✓ Données comportant des appréciations sur les difficultés sociales des personnes

✓ Données biométriques



3. Conservation limitée des données

Principe 1 :

Les données « sont conservées sous une forme permettant l'identification des personnes concernées pendant une durée qui n'excède pas la durée nécessaire aux finalités pour lesquelles elles sont collectées et traitées »

Principe 2 :

Les DCP « ne peuvent être conservées au-delà de la cette durée qu'en vue d'être traitées à des fins historiques, statistiques ou scientifiques ; le choix des données ainsi conservées est opéré dans les conditions prévues à [l'article L. 212-4 du Code du patrimoine](#) »

- **En pratique, effacement ou anonymisation des données à l'issue de cette durée, ou archivage**

4. Obligation de sécurité

PRINCIPE (art. 34) :

« Le responsable du traitement est tenu de prendre toutes précautions utiles, au regard de la nature des données et des risques présentés par le traitement, pour préserver la sécurité des données et, notamment, empêcher qu'elles soient déformées, endommagées, ou que des tiers non autorisés y aient accès »

2 volets : respect de l'intégrité et de la confidentialité des données



5. Respect des droits des personnes

Droit à l'information

Droit d'opposition/consentement

**Droit d'accès, rectification,
suppression**

5. Droit à l'information

Il faut une information sur :

1. l'identité du responsable du traitement ;
2. la finalité du traitement ;
3. le caractère obligatoire ou facultatif des réponses ;
4. les conséquences d'un défaut de réponse ;
5. les destinataires des données ;
6. les modalités d'exercice de leurs droits ;



5. Droit d'opposition/consentement

- Principe :

Toute personne a le droit de s'opposer, pour des motifs légitimes, au traitement de ses données, sauf exceptions

→ cas des « enquêtes obligatoires », agréées par le CNIS et ayant reçu le visa des ministres compétents

→ pour les autres, nécessité d'informer les personnes du caractère facultatif des réponses

- Dans certains cas, il est nécessaire de recueillir un consentement explicite (case à cocher « j'accepte »)

→ en particulier pour traiter des données sensibles (sauf invocation de l'une des autres exceptions prévues à l'art. 8 de la loi « I&L »)



5. Droits d'accès et de rectification

- **Principe :**

Toute personne peut, directement auprès du responsable des traitements, avoir accès à l'ensemble des informations la concernant, en obtenir la copie et exiger qu'elles soient, selon les cas, rectifiées, complétées, mises à jour ou supprimées

3

Formalités préalables applicables et CIL

Déclarations, exonérations et dispenses

Les fichiers dispensés de déclaration par la CNIL: paie , fichiers de fournisseurs, information et communication externes, ...

Les traitements courants soumis à déclaration, sauf CIL: gestion RH, fichiers clients, **recherches hors données sensibles...**

Les traitements à risque soumis à autorisation ou avis de la CNIL

traitements de données sensibles (y compris à des fins de recherche), dispositifs biométriques, fichiers de police, téléservices, interconnexions sous certaines conditions



Le futur règlement européen: les orientations

- **Renforcer les droits des personnes** pour développer la confiance et contribuer à l'essor de l'économie numérique
- Assurer une plus grande harmonisation des règles de protection des données tout en renforçant la **responsabilité des entreprises**
- Affirmer la dimension mondiale de la protection des données
- Renforcer le rôle des autorités de protection des données (APD) et du groupe européen des APD, le G29

Data Anonymisation Techniques

Data Anonymization/De-Identification

- ❑ Anonymisation is NOT pseudo-anonymization!
- ❑ What is Pseudo-Anonymization?
 - ❑ *Personal information contains identifiers, such as **a name, date of birth, sex and address**. When personal information is pseudonymised, **the identifiers** are replaced by one pseudonym. Pseudonymisation is achieved, for instance, by encryption of the identifiers in personal data.*
- ❑ What is Anonymization?
 - ❑ *Data are anonymised if **all** identifying elements have been eliminated from a set of personal data (**all quasi-identifiers**). No element may be left in the information which could, by exercising reasonable effort, serve to re-identify the person(s) concerned. Where data have been successfully anonymised, they are no longer personal data.*

Source: **Handbook on European data protection law,**

http://www.echr.coe.int/Documents/Handbook_data_protection_ENG.pdf

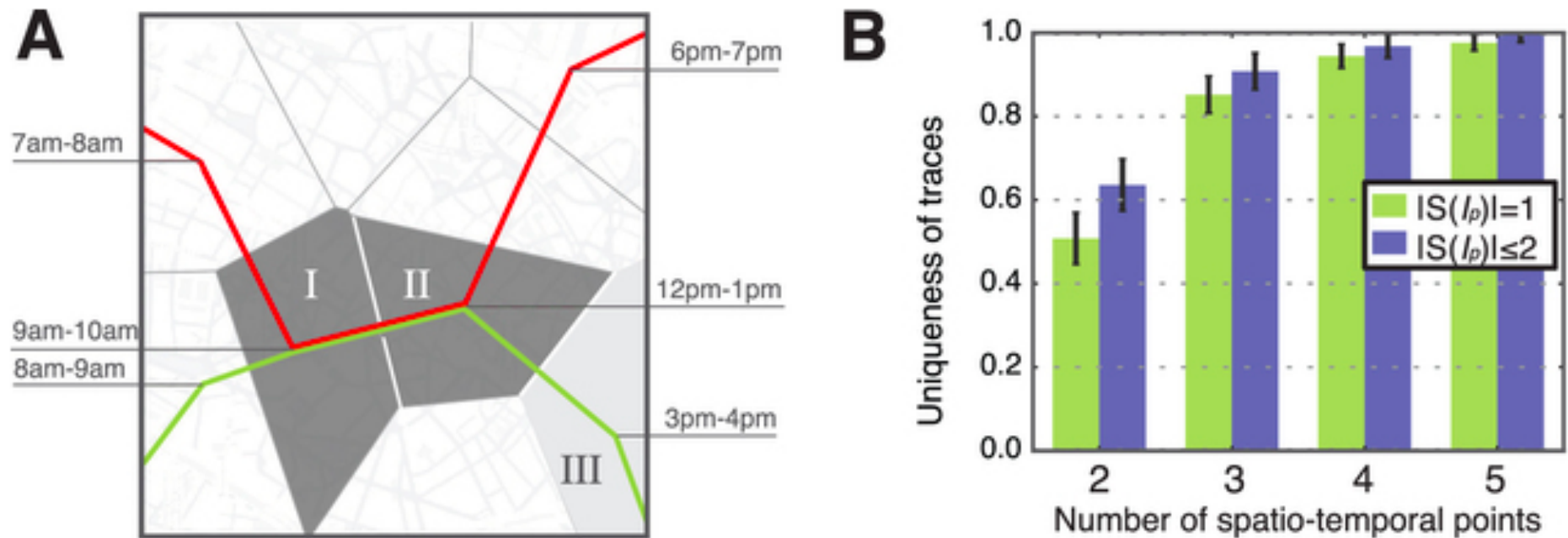
Pseudo-Ano. vs Anonymization (2)

- ❑ Why does Pseudo-Anonymization does not work?
 - ❑ It does not compose, i.e. several Pseudo-Anonymized data can be combined to de-anonymize...
 - ❑ External Information can also be exploited.
 - ❑ See Recent example with NY Taxi
 - ❑ **173 million** individual trips de-anonymized*
- ❑ Why is Data Anonymization Difficult?
 - ❑ Difficult (often impossible) to identify all quasi-identifiers!

Quasi-Identifiers

- ❑ Quasi-identifiers are difficult to identify exhaustively
- ❑ Many combination of attributes can be used to « single-out » a user
- ❑ We are all unique by different ways, we are full of Q.I.
 - ❑ See « Unicity me! * »
 - ❑ Mobility pattern, webhistory, .
 - ❑ Data (content) and meta-data
 - ❑ i.e. timing can betray you!
 - Google search timing pattern can tell when you were away!

Unique in the Crowd [Nature2013]



- Only 4 spatio-temporal points are necessary to uniquely identify a user with a probability $> 95\%$!

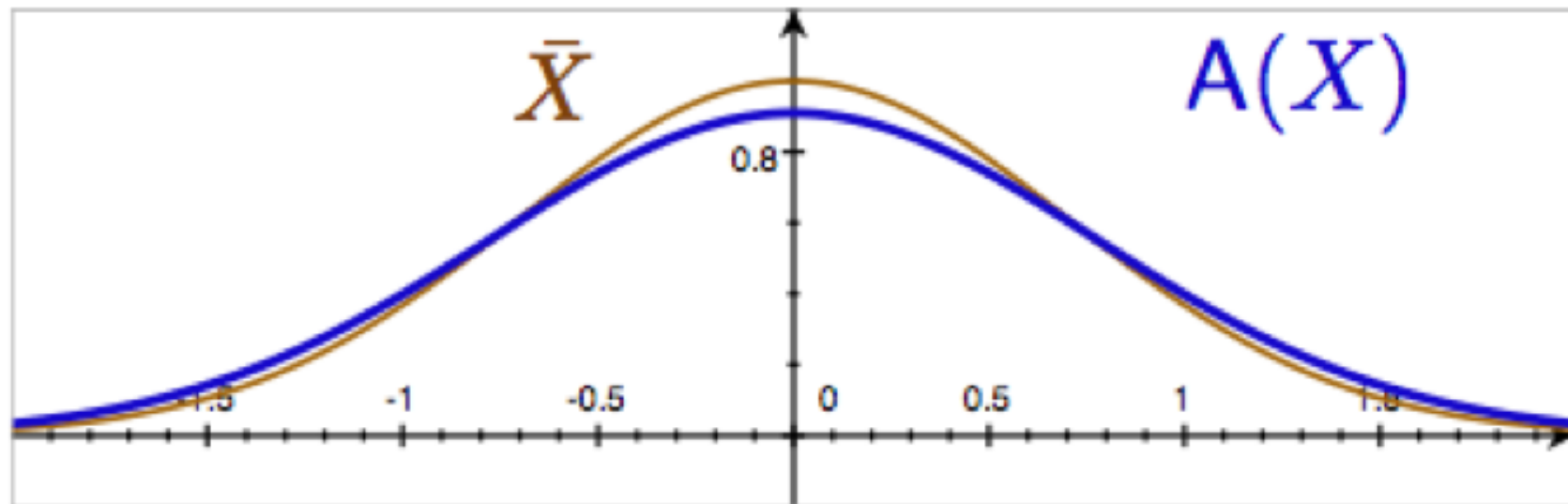
Some Data anonymization methods...

- ❑ Random perturbation
 - ❑ Input perturbation
 - ❑ Output perturbation
- ❑ Generalization
 - ❑ The data domain has a natural hierarchical structure.
- ❑ Suppression
- ❑ Permutation
 - ❑ Destroying the link between identifying and sensitive attributes that could lead to a privacy leakage.

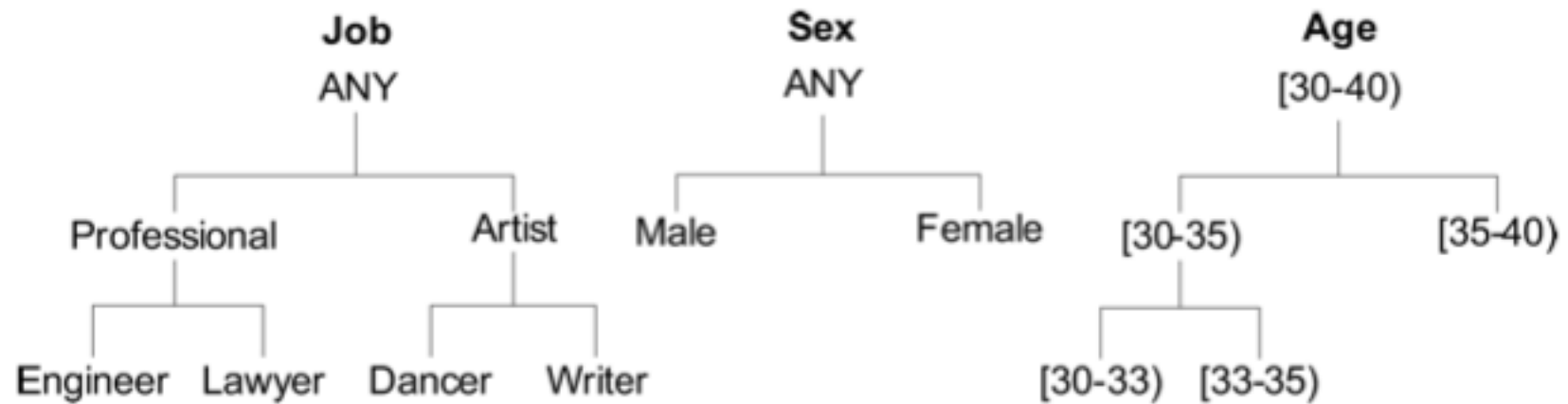
Randomization Methods

Randomization: add independent noise (such as Gaussian or uniform) to the values transmitted.

Goal: hide the specific values of attributes while preserving the joint distribution of the data.



Generalization Methods



Suppression Methods

| Age | Sex | Disease (sensitive) |
|-----|--------|------------------------|
| 30 | Male | Hepatitis |
| 30 | Male | Hepatitis |
| 30 | Male | HIV |
| 32 | Male | Hepatitis |
| 32 | Male | HIV |
| 32 | Male | HIV |
| 36 | Female | Flu |
| 38 | Female | Flu |
| 38 | Female | Heart |
| 38 | Female | Heart |

| Age | Sex | Disease (sensitive) |
|-----|--------|------------------------|
| 30 | Male | Hepatitis |
| 30 | Male | Hepatitis |
| 30 | Male | HIV |
| 32 | Male | Hepatitis |
| 32 | Male | HIV |
| 32 | Male | HIV |
| 36 | Female | Flu |
| 38 | Female | Flu |
| 38 | Female | Heart |
| 38 | Female | Heart |

K-anonymity

- **Privacy guarantee**: in each group of the sanitized dataset, each individual will be identical to a least $k - 1$ others.
- Reach by a combination of generalization and suppression.
- **Example of use**: sanitization of medical data.

| | Non-Sensitive | | | Sensitive |
|----|---------------|-----|-------------|-----------------|
| | Zip Code | Age | Nationality | Condition |
| 1 | 13053 | 28 | Russian | Heart Disease |
| 2 | 13068 | 29 | American | Heart Disease |
| 3 | 13068 | 21 | Japanese | Viral Infection |
| 4 | 13053 | 23 | American | Viral Infection |
| 5 | 14853 | 50 | Indian | Cancer |
| 6 | 14853 | 55 | Russian | Heart Disease |
| 7 | 14850 | 47 | American | Viral Infection |
| 8 | 14850 | 49 | American | Viral Infection |
| 9 | 13053 | 31 | American | Cancer |
| 10 | 13053 | 37 | Indian | Cancer |
| 11 | 13068 | 36 | Japanese | Cancer |
| 12 | 13068 | 35 | American | Cancer |

Figure 1. Inpatient Microdata

| | Non-Sensitive | | | Sensitive |
|----|---------------|------|-------------|-----------------|
| | Zip Code | Age | Nationality | Condition |
| 1 | 130** | < 30 | * | Heart Disease |
| 2 | 130** | < 30 | * | Heart Disease |
| 3 | 130** | < 30 | * | Viral Infection |
| 4 | 130** | < 30 | * | Viral Infection |
| 5 | 1485* | ≥ 40 | * | Cancer |
| 6 | 1485* | ≥ 40 | * | Heart Disease |
| 7 | 1485* | ≥ 40 | * | Viral Infection |
| 8 | 1485* | ≥ 40 | * | Viral Infection |
| 9 | 130** | 3+ | * | Cancer |
| 10 | 130** | 3+ | * | Cancer |
| 11 | 130** | 3+ | * | Cancer |
| 12 | 130** | 3+ | * | Cancer |

Figure 2. 4-anonymous Inpatient Microdata

But K-Ano. does not compose 😞!

- **Question**: suppose that Alice's employer knows that she is 28 years old, she lives in ZIP code 13012 and she visits both hospitals. What does he learn?

| | Non-Sensitive | | | Sensitive |
|----|---------------|-----|-------------|-----------------|
| | Zip code | Age | Nationality | Condition |
| 1 | 130** | <30 | * | AIDS |
| 2 | 130** | <30 | * | Heart Disease |
| 3 | 130** | <30 | * | Viral Infection |
| 4 | 130** | <30 | * | Viral Infection |
| 5 | 130** | ≥40 | * | Cancer |
| 6 | 130** | ≥40 | * | Heart Disease |
| 7 | 130** | ≥40 | * | Viral Infection |
| 8 | 130** | ≥40 | * | Viral Infection |
| 9 | 130** | 3* | * | Cancer |
| 10 | 130** | 3* | * | Cancer |
| 11 | 130** | 3* | * | Cancer |
| 12 | 130** | 3* | * | Cancer |


(a)

| | Non-Sensitive | | | Sensitive |
|----|---------------|-----|-------------|-----------------|
| | Zip code | Age | Nationality | Condition |
| 1 | 130** | <35 | * | AIDS |
| 2 | 130** | <35 | * | Tuberculosis |
| 3 | 130** | <35 | * | Flu |
| 4 | 130** | <35 | * | Tuberculosis |
| 5 | 130** | <35 | * | Cancer |
| 6 | 130** | <35 | * | Cancer |
| 7 | 130** | ≥35 | * | Cancer |
| 8 | 130** | ≥35 | * | Cancer |
| 9 | 130** | ≥35 | * | Cancer |
| 10 | 130** | ≥35 | * | Tuberculosis |
| 11 | 130** | ≥35 | * | Viral Infection |
| 12 | 130** | ≥35 | * | Viral Infection |

(b)


But K-ANO does not compose 😞!

- **Question**: suppose that Alice's employer knows that she is 28 years old, she lives in ZIP code 13012 and she visits both hospitals. What does he learn?



| | Non-Sensitive | | | Sensitive |
|----|---------------|-----|-------------|-----------------|
| | Zip code | Age | Nationality | Condition |
| 1 | 130** | <30 | * | AIDS |
| 2 | 130** | <30 | * | Heart Disease |
| 3 | 130** | <30 | * | Viral Infection |
| 4 | 130** | <30 | * | Viral Infection |
| 5 | 130** | ≥40 | * | Cancer |
| 6 | 130** | ≥40 | * | Heart Disease |
| 7 | 130** | ≥40 | * | Viral Infection |
| 8 | 130** | ≥40 | * | Viral Infection |
| 9 | 130** | 3* | * | Cancer |
| 10 | 130** | 3* | * | Cancer |
| 11 | 130** | 3* | * | Cancer |
| 12 | 130** | 3* | * | Cancer |

(a)



| | Non-Sensitive | | | Sensitive |
|----|---------------|-----|-------------|-----------------|
| | Zip code | Age | Nationality | Condition |
| 1 | 130** | <35 | * | AIDS |
| 2 | 130** | <35 | * | Tuberculosis |
| 3 | 130** | <35 | * | Flu |
| 4 | 130** | <35 | * | Tuberculosis |
| 5 | 130** | <35 | * | Cancer |
| 6 | 130** | <35 | * | Cancer |
| 7 | 130** | ≥35 | * | Cancer |
| 8 | 130** | ≥35 | * | Cancer |
| 9 | 130** | ≥35 | * | Cancer |
| 10 | 130** | ≥35 | * | Tuberculosis |
| 11 | 130** | ≥35 | * | Viral Infection |
| 12 | 130** | ≥35 | * | Viral Infection |

(b)

Other Attacks on k-Anonymity

- k-Anonymity does not provide privacy if
 - Sensitive values in an equivalence class lack diversity
 - The attacker has background knowledge

Homogeneity attack

| Bob | |
|----------------|------------|
| Zipcode | Age |
| 47678 | 27 |

A 3-anonymous patient table

| Zipcode | Age | Disease |
|---------|-----|---------------|
| 476** | 2* | Heart Disease |
| 476** | 2* | Heart Disease |
| 476** | 2* | Heart Disease |
| 4790* | ≥40 | Flu |
| 4790* | ≥40 | Heart Disease |
| 4790* | ≥40 | Cancer |
| 476** | 3* | Heart Disease |
| 476** | 3* | Cancer |
| 476** | 3* | Cancer |

Background knowledge attack

| Carl does not have heart disease | |
|----------------------------------|------------|
| Zipcode | Age |
| 47673 | 36 |

Some Other Anonymization Schemes

- ▶ *l*-diversity (MKGV¹ 07): maintain the diversity for each group with respect to the possible values of the sensible attributes.
- ▶ Can be instanced by a metric based on *entropy*.
- ▶ Prevent against attacks based on homogeneity and some other attacks.
- ▶ *t*-closeness (LLV² 07): the distribution of the attributes in each group must be close to that on the global population.
- ▶ *t* is a threshold that should not be exceeded and which represents the proximity between distributions.

I-Diversity: Preventing the Homogeneity attack

| | | |
|-------------|-------|----------|
| Caucas | 787XX | Flu |
| Caucas | 787XX | Shingles |
| Caucas | 787XX | Acne |
| Caucas | 787XX | Flu |
| Caucas | 787XX | Acne |
| Caucas | 787XX | Flu |
| Asian/AfrAm | 78XXX | Flu |
| Asian/AfrAm | 78XXX | Flu |
| Asian/AfrAm | 78XXX | Acne |
| Asian/AfrAm | 78XXX | Shingles |
| Asian/AfrAm | 78XXX | Acne |
| Asian/AfrAm | 78XXX | Flu |

Sensitive attributes must be “diverse” within each quasi-identifier equivalence class

Distinct I-Diversity

- ❑ Each equivalence class has at least I well-represented sensitive values
- ❑ Doesn't prevent probabilistic inference attacks

| | |
|-----|----------------|
| ... | Disease |
| | ... |
| | HIV |
| | HIV |
| | ... |
| | HIV |
| | pneumonia |
| | bronchitis |
| | ... |

10 records

8 records have HIV

2 records have other values

t-Closeness

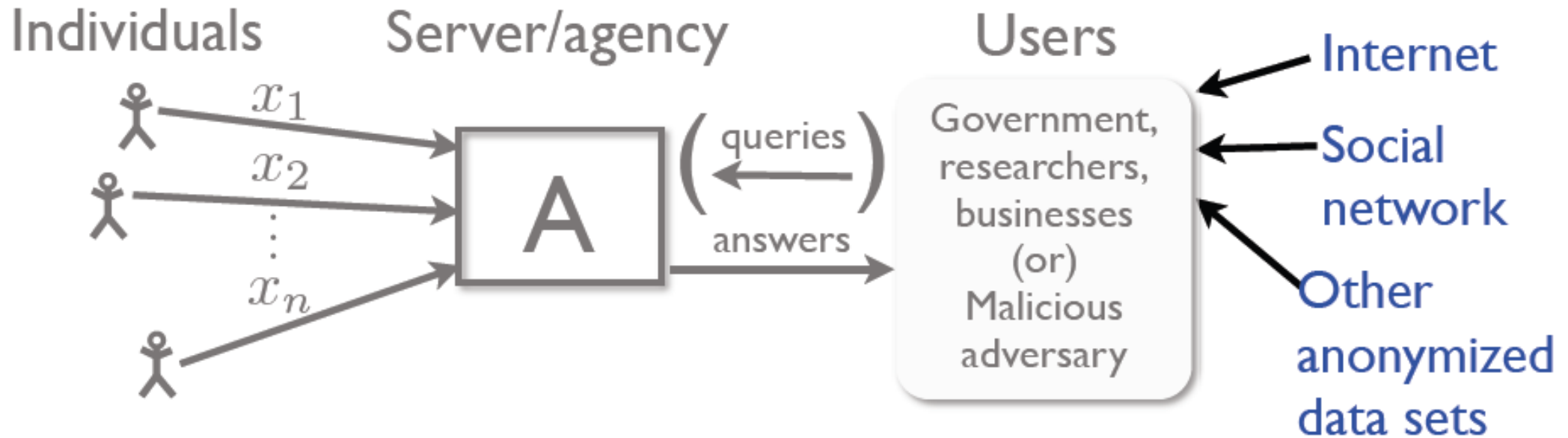
[Li et al. ICDE '07]

| | | |
|-------------|-------|----------|
| Caucas | 787XX | Flu |
| Caucas | 787XX | Shingles |
| Caucas | 787XX | Acne |
| Caucas | 787XX | Flu |
| Caucas | 787XX | Acne |
| Caucas | 787XX | Flu |
| Asian/AfrAm | 78XXX | Flu |
| Asian/AfrAm | 78XXX | Flu |
| Asian/AfrAm | 78XXX | Acne |
| Asian/AfrAm | 78XXX | Shingles |
| Asian/AfrAm | 78XXX | Acne |
| Asian/AfrAm | 78XXX | Flu |

Distribution of sensitive attributes within each quasi-identifier group should be “close” to their distribution in the entire original database

Why publish quasi-identifiers at all??

Why Data Anonymization is Hard: External Information



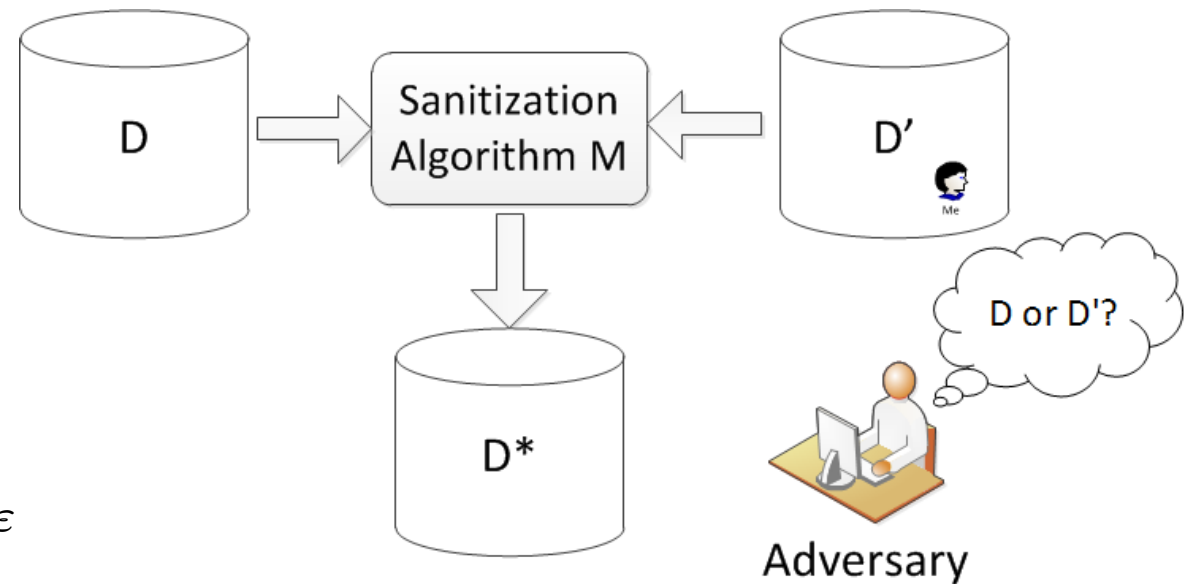
- Users have external information sources
 - Can't assume we know the sources
 - Can't ignore them!
- Anonymization schemes are regularly broken

Toward « Provable » Anonymization

- ❑ Stronger schemes are necessary
- ❑ Differential Privacy (DP)
 - ❑ Provides some strong and measurable guarantees
 - ❑ Secures even with external sources of data
 - ❑ Composes
- ❑ Intuition of DP:
 - ❑ Changes to my data not noticeable
 - ❑ Output is “independent” of my data

Privacy Model

- **Differential privacy**

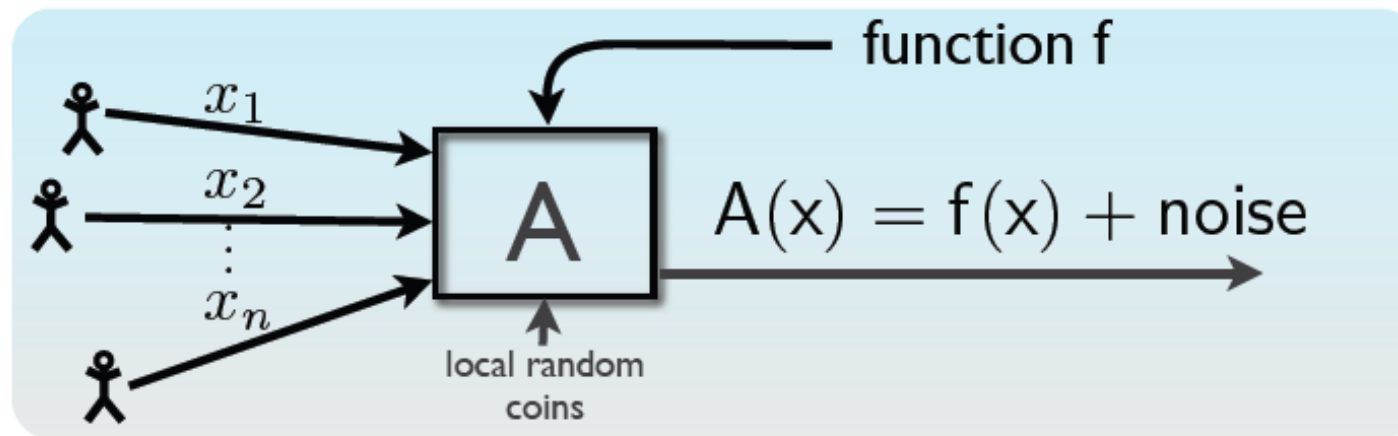


$$e^{-\epsilon} \leq \frac{\Pr(M(D) = D^*)}{\Pr(M(D') = D^*)} \leq e^{\epsilon}$$

- ***composes securely***: retain privacy guarantees in the presence of independent releases^[1]
- Secure even with arbitrary external knowledge!

[1] S.R. Ganta, S. Kasiviswanathan, A. Smith. *Composition Attacks and Auxiliary Information in Data Privacy*. KDD'08

Differential Privacy



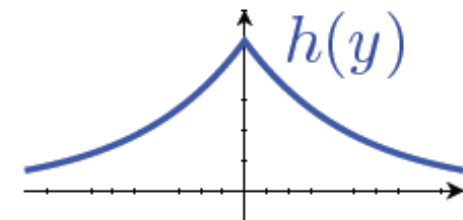
- **Global Sensitivity:** $GS_f = \max_{\text{neighbors } x, x'} \|f(x) - f(x')\|_1$

➤ Example: $GS_{\text{proportion}} = \frac{1}{n}$

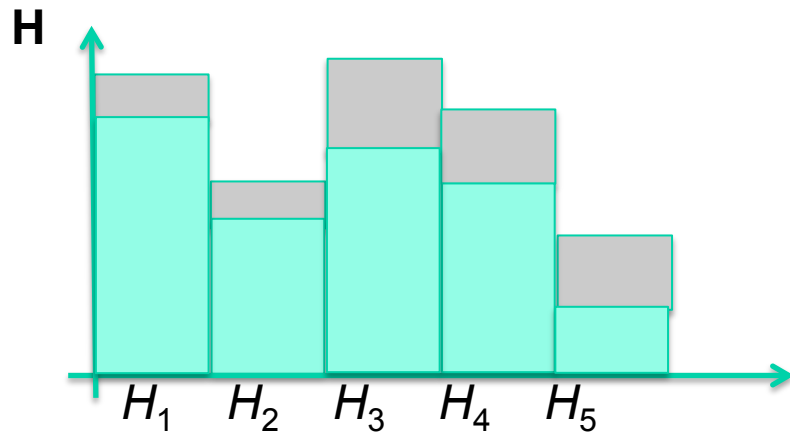
Theorem: If $A(x) = f(x) + \text{Lap}\left(\frac{GS_f}{\epsilon}\right)$, then A is ϵ -differentially private.

➤ Laplace distribution $\text{Lap}(\lambda)$ has density

$$h(y) \propto e^{-|y|/\lambda}$$



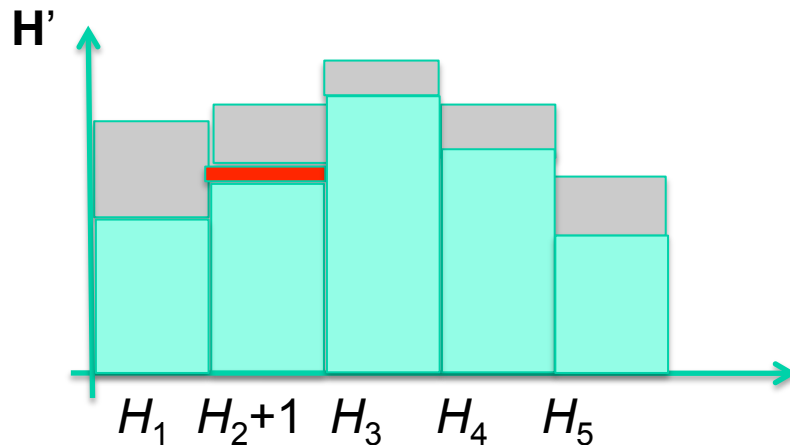
Histogram Release with Laplace Mechanism



Add random Laplace noise to each bin before publishing!

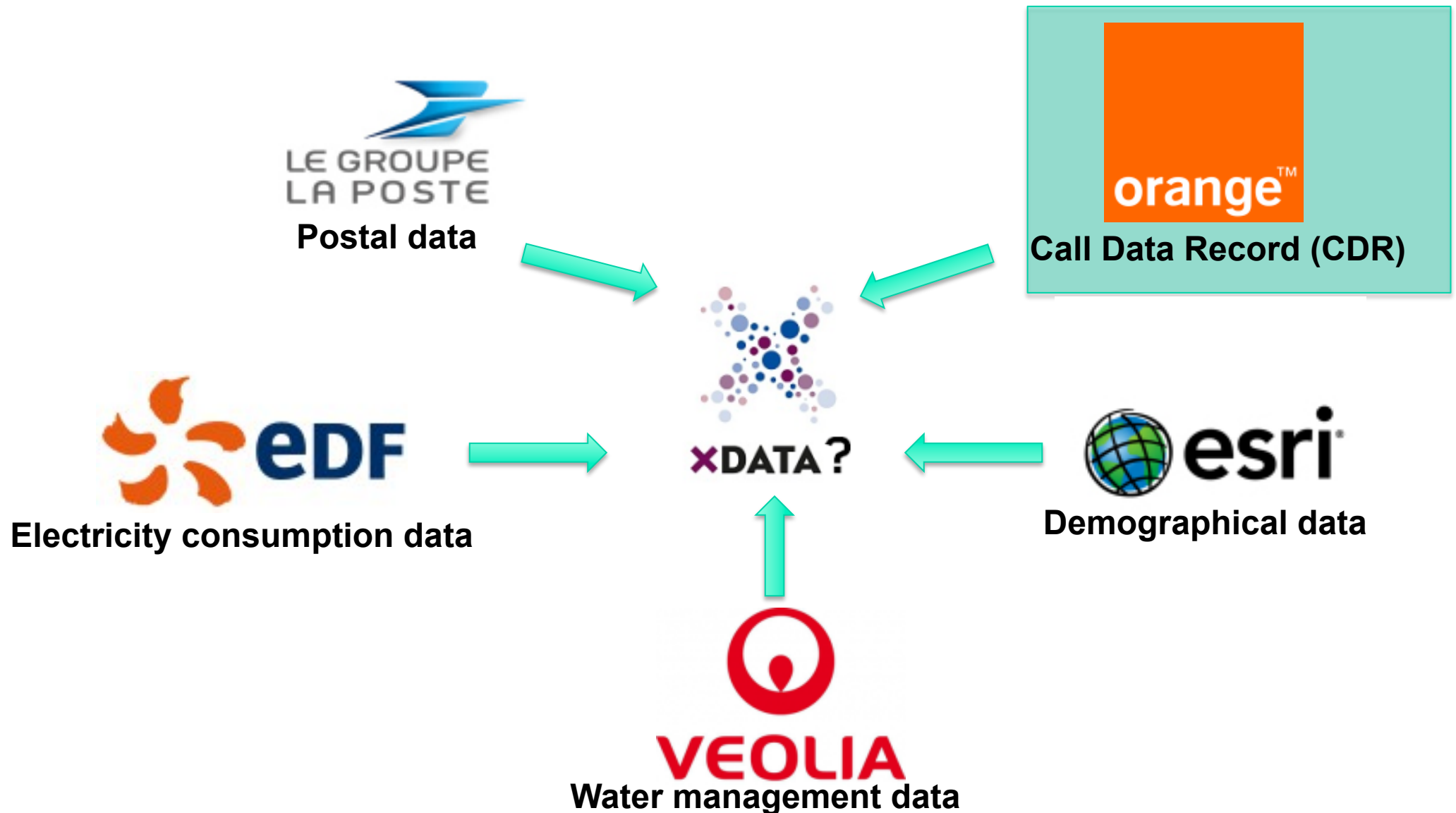


$$\frac{\prod_i \Pr(H_i + \text{Laplace}(\lambda) = H_i^*)}{\prod_i \Pr(H'_i + \text{Laplace}(\lambda) = H_i^*)} \leq \exp\left(\frac{\sum_i |H_i - H'_i|}{\lambda}\right) = e^{\frac{1}{\lambda}}$$



- **Global sensitivity:**
 $\Delta H = \sum |H_i - H'_i|$
- For histograms: $\Delta H = 1$
- If $\lambda = \Delta H / \epsilon$, we have ϵ -differential privacy

Example: Spatio-temporal density from CDR



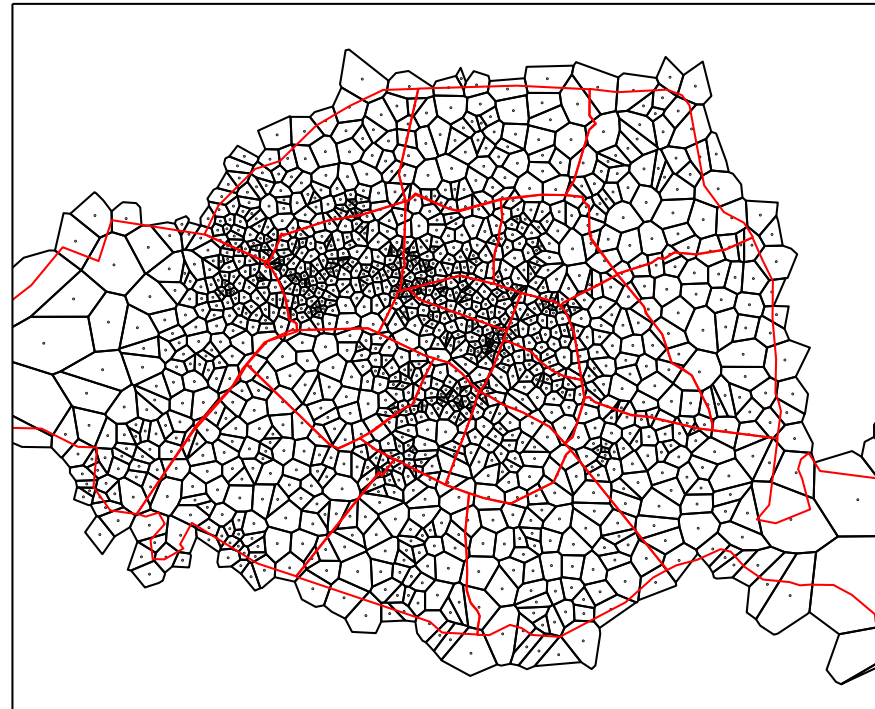
(Simplified) Call Data Record

| Rec # | Phone | Lat | Lon | Time | Event |
|-------|------------|--------|------|---------------------|---------------|
| 1 | 0644536701 | 46.345 | 2.32 | 13:34:12 01/09/2007 | Incoming SMS |
| 2 | 0634556702 | 47.123 | 1.65 | 14:31:02 02/09/2007 | Outgoing Call |
| ... | ... | | | | |

- 4 types of events:
 - Incoming SMS/Call
 - Outgoing SMS/Call
- Phone numbers are scrambled (No Personal Data in the dataset)

Paris CDR (provided by Orange™)

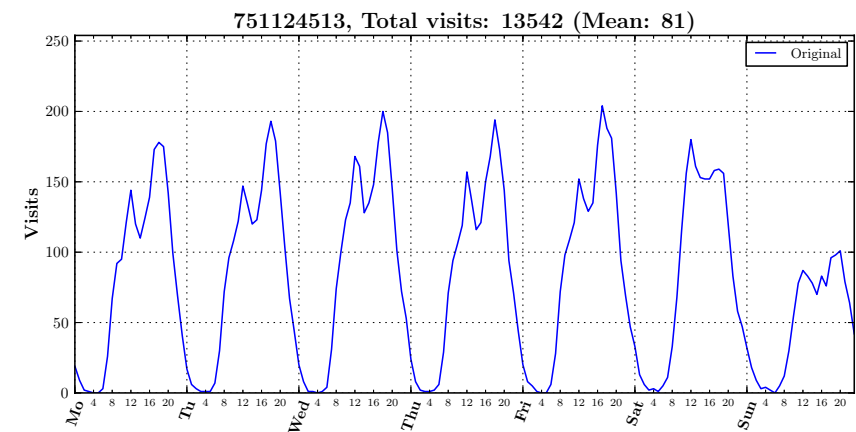
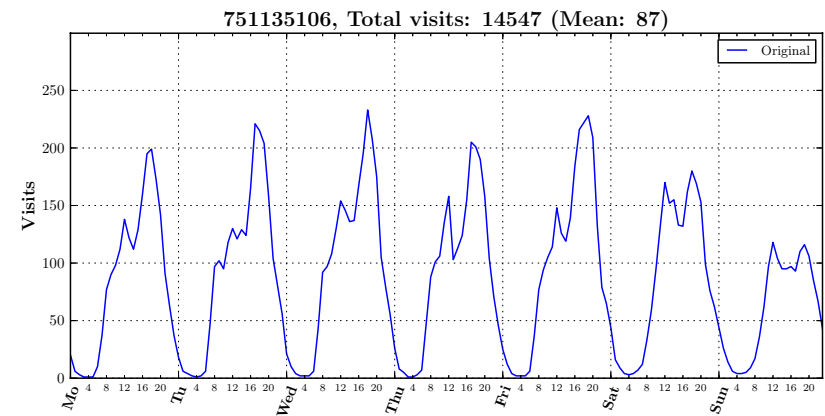
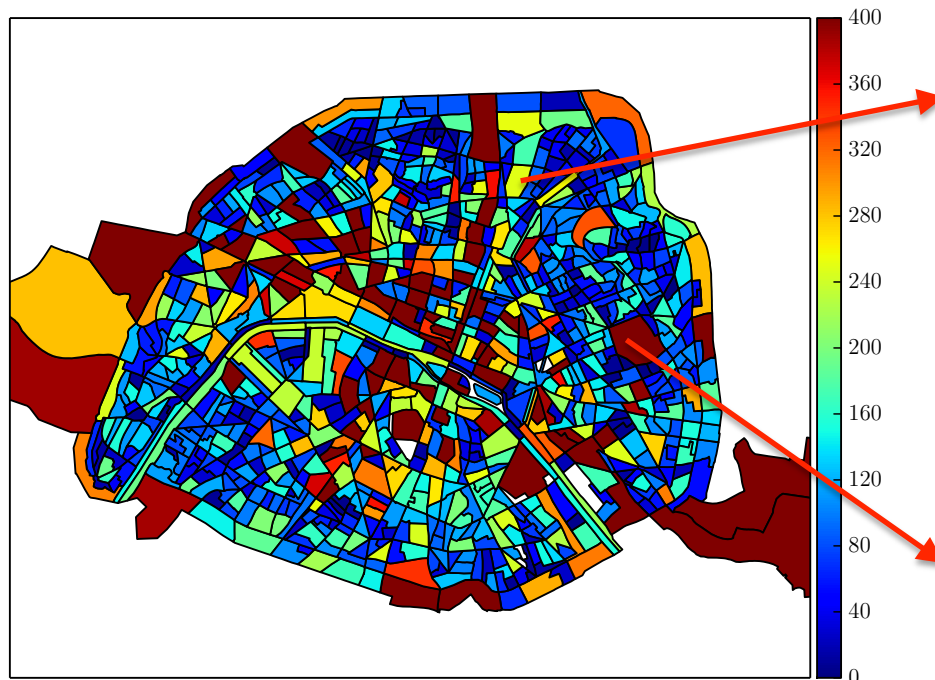
- 1,992,846 users
- 1303 towers
- 10/09/2007 – 17/09/2007
- Mean trace length: 13.55
(std.dev: 18)
- Max. trace length: 732



Goal: Release spatio-temporal density (and not CDR)

Number of individuals at a given hour at any IRIS cell in Paris

IRIS cells



Overview of our approach

1. Sample x (≈ 30) visits per user uniformly at random (to decrease sensitivity)
2. Create time-series: map tower cell counts to IRIS cell counts
3. Perturb these time-series to guarantee differential privacy

Overview of our approach

1. Sample x (≈ 30) visits per user uniformly at random
2. Create time-series: map tower cell counts to IRIS cell counts
3. Perturb these time-series to guarantee differential privacy

Perturbation of time series

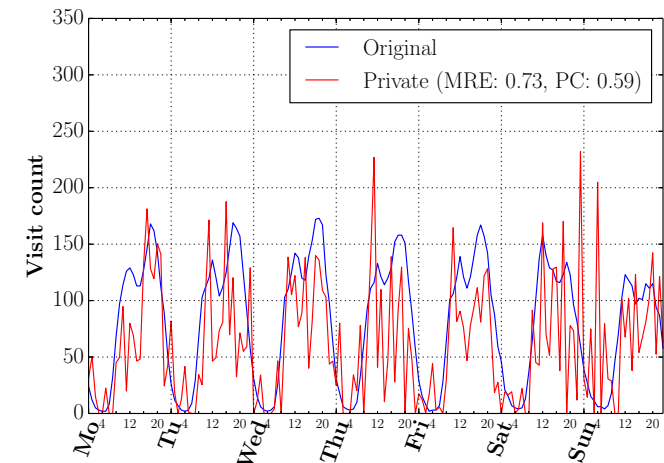
Naïve solution: add properly calibrated Laplace noise to each count of the IRIS cell (one count per hour over 1 week)

Problem: *Counts are much smaller than the noise!*

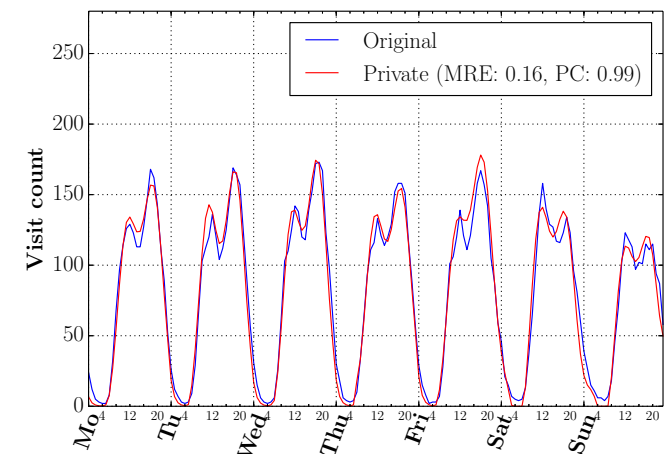
Our approach:

1. cluster nearby less populated cells until their aggregated counts become sufficiently large to resist noise.
2. perturb the aggregated time series by adding noise to their largest Fourier coefficients
3. scale back with the (noisy) total number of visits of individual cells to get the individual time series

Naïve approach ($\epsilon=0.3$)



Our approach ($\epsilon=0.3$)



Limitations of These Schemes

- “only” provide aggregated statistics
 - Average values...
- Not useful for applications where an analyst necessarily needs raw individual-level data
 - synthetic data?
 - Work in progress: Generating anonymized mobility traces
 - By noising the data [Chen-ccs12]
 - DP-Where [ATT]
 - Generate a model from dataset
 - Noise/sanitize models
 - Regenerate traces from noisy models
 - Utility is usually limited!

Conclusion :

There are no universal solutions!

- **There are no “universal” anonymization solutions that fit all applications**
 - in order to get the best accuracy, they have to be customized to the application and the public characteristics of the dataset
 - specific context
 - specific utility/privacy tradeoff
 - Specific ADV models
 - Specific impacts..
- Anonymization is all about utility/efficiency trade-off!
 - Full-proof security is not always necessary (and probably impossible)!
 - It has to be performed with a PRA (Privacy Risk Analysis)

Conclusion :

Anonymization does not solve everything!

- ❑ Sanitization schemes protect against re-identification, **not inference!**
- ❑ You can learn and infer a lot from data
 - ❑ You can infer religion from Mobility data!
 - ❑ Interest from Google search requests
- ❑ You can learn and infer a lot from meta-data!
 - ❑ Who communicated with whom?
 - ❑ Is a user away/active?
- ❑ It is up to the society to decide what is acceptable or not!
 - ❑ By balancing the benefits with the risks*.

**Benefit-Risk Analysis for Big Data Projects, http://www.futureofprivacy.org/wp-content/uploads/FPF_DataBenefitAnalysis_FINAL.pdf*

BIG DATA

The Risks of Inference: The Target Case

- ❑ Target identified about 25 products that, when analyzed together, allowed him to assign each shopper a “pregnancy prediction” score.
- ❑ More important, he could also estimate her due date to within a small window
- ❑ Target could (and does) send coupons timed to very specific stages of her pregnancy.

See video: big data & Privacy:

- Bigdata-low.mp4 or <https://www.privacyinternational.org/node/572>



Active research areas

- Data anonymization of database records and other data structures (e.g., network graphs)
- Private communication (prevention of traffic analysis)
 - Anonymous and covert communication
- Crypto protocols
 - Privacy-enhanced authentication and identity management
 - Operations in the encrypted domain
 - Anonymous search and retrieval of information
 - Privacy-preserving biometric authentication
- Location privacy
- Ubiquitous environments
 - Constrained devices
 - Securing the physical link
- Social networks