Privacy Preserving Data Publishing

Benjamin Nguyen

INSA-CVL

M2 INIS

What is « anonymity » ?

- **Context : Personal Data Processing.**
- **Anonymity** = when the data can no longer be used to identify a person (*using reasonable effort*).
- **Anonymous data is not concerned by data protection texts.**
- No technique is suggested…
- Art. 29 WP prefers to speak about anonymization techniques.

→ WE WILL INVESTIGATE PRIVACY PRESERVING DATA PUBLISHING (PPDP) MODELS AND TECHNIQUES

Outline

- Attacks on statistical databases
- Anonymity and Privacy Preserving Data Publishing (PPDP)
- Data Partitionning
  - K-anonymity
  - L-diversity
- Continuous Releases
  - M-invarience
  - Sampling
- Data Perturbation
  - Local Perturbation
  - Input Perturbation
  - Statistical Perturbation
- Conclusion

Attaques contre les bases de données statistiques

- **Exemple :**
  - Relation Analyse(Patient, H/F, Age, Mutuelle, Leucocyte)

Patient	**H/F**	**Age**	**Mutuelle**	**Leucocyte**
Dupont	H	30	MMA	6000
Durand	F	25	LMDE	3000
Dulac	F	35	MMA	7000
Duval	H	45	IPECA	5500
Dubois	H	55	MGEN	3500
Dumont	H	38	MMA	7500
Dupré	F	32	IPECA	7200
Dupuis	F	50	MGEN	6800
Dufour	H	45	MAAF	4000
Dumas	H	40	Rempart	3800

Attaque contre les bases de données statistiques (suite)

- **Base de données statistique**
  - Base de données qui permet d'évaluer des requêtes qui dérivent des informations d'agrégation
  - Par exemple : des totaux, des moyennes
  - Mais pas des requêtes qui dérivent des informations particulières

- **Exemple :**
  - La requête « quelle est la moyenne du taux de leucocytes des patients ayant plus de 30 ans ? » est permise
  - La requête « quel est le taux de leucocytes de Dupont ? » est interdite

Attaque contre les bases de données statistiques (suite)

- **Exemple d'attaque simple :**
  - *U* veut découvrir le taux de Leucocyte de Dubois
  - *U* sait par ailleurs que Dubois est un adhérent masculin de la MGEN.

- **Requête 1**

```
SELECT COUNT ( Patient )
FROM   Analyse
       WHERE  H/F = 'H'
       AND    Mutuelle = 'MGEN' ;
```

Résultat : 1

- **Requête 2**

```
SELECT SUM ( Leucocyte )
       FROM   Analyse
       WHERE  H/F = 'H'
       AND    Mutuelle = 'MGEN' ;
```

Résultat : 3500

➔ Le système doit refuser de répondre à une requête pour laquelle la cardinalité du résultat est inférieure à une certaine borne *b*

Attaque contre les bases de données statistiques (suite)

- **Requête 3**

  SELECT COUNT (Patient)
  FROM Analyse

  Résultat : 10

- **Requête 4**

  SELECT COUNT (Patient)
  FROM Analyse
  WHERE NOT (H/F = 'H'
  AND Mutuelle = ‘MGEN’) ;

  Résultat: 9

- **Requête 5**

  SELECT SUM (Leucocyte)
  FROM Analyse

  Résultat : 54300

- **Requête 6**

  SELECT SUM (Leucocyte)
  FROM Analyse
  WHERE NOT (H/F = 'H'
  AND Mutuelle = ‘MGEN’) ;

  Résultat : 50800 ; 54300 – 50800 = 3500

Conséquence :

Le système doit aussi refuser de répondre à une requête pour laquelle la cardinalité du résultat est supérieure à $N - b$, où N est la cardinalité de la relation initiale

Attaque contre les bases de données statistiques (suite)

- **Problème :**
  - Mais limiter les requêtes à celles pour lesquelles le résultat a une cardinalité *c* telle que $b \leq c \leq N - b$ n'est pas suffisant pour éviter la compromission
  - Exemple : si $b = 2$, les requêtes auront une réponse si *c* est telle que $2 \leq c \leq 8$

- **Requête 7**

```
SELECT COUNT ( Patient )
      FROM   Analyse
      WHERE  H/F = 'H' ;
```

Résultat : 6

- **Requête 8**

```
SELECT COUNT ( Patient )
      FROM   Analyse
      WHERE H/F = 'H'
      AND      NOT (Mutuelle = ‘MGEN’) ;
```

Résultat : 5

- **Conséquence**
  - *U* peut déduire qu'il existe exactement un patient masculin qui a la MGEN comme mutuelle,
  - Il s’agit de Dubois, puisque *U* sait que cette description correspond à Dubois

Attaque contre les bases de données statistiques (suite)

- **Conséquence (suite)**
  - Le taux de Leucocyte de Dubois est facilement découvert de la façon suivante :

- **Requête 9**

```
SELECT SUM ( Leucocyte )
      FROM   Analyse
      WHERE  H/F = 'H' ;
```

Résultat : 30300

- **Requête 10**

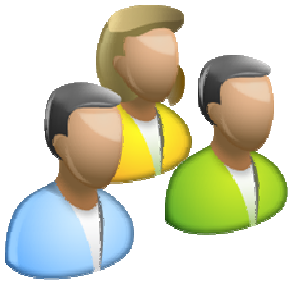
```
SELECT SUM ( Leucocyte )
      FROM   Analyse
      WHERE  H/F = 'H'
      AND    NOT (Mutuelle = 'MGEN') ;
```

Résultat : 26800 ; 30300 - 26800 = 3500

Les bases de données statistiques doivent donc limiter les requêtes…

Il faut trouver un moyen de publier des données tout en permettant un accès illimité aux données !

Privacy Preserving Data Publishing (PPDP)



Individuals

Raw data

Publisher *(trusted)*

Anonymized data
(or *sanitized …*)

Recipients *(no assumption of trust)*

PPDP Components

- **A privacy definition, which answers to:**
  What is « privacy »?

  →**DEBATABLE**
  (Many exist)

- **A utility metric, which answers to:**
  How to measure the utility of sanitized data or the information lost by a sanitization process?

  →**VERY GENERAL**
  (will in fact depend on the task)

- **A sanitization algorithm, which answers to:**
  How to enforce a privacy definition while minimizing a given utility metric?

  →**RESEARCH RESULTS**

Pseudonymization: A naïve privacy definition

SSN	Activity	Age	Diag
123	"M2 Stud."	22	Flu
456	"M1 Stud."	25	HIV
789	"L3 Stud."	20	Flu

Raw data

Pseudo	Activity	Age	Diag
ABC	"M2 Stud."	22	Flu
MNO	"M1 Stud"	25	HIV
XYZ	"L3 Stud."	20	Flu

Pseudonymized Data

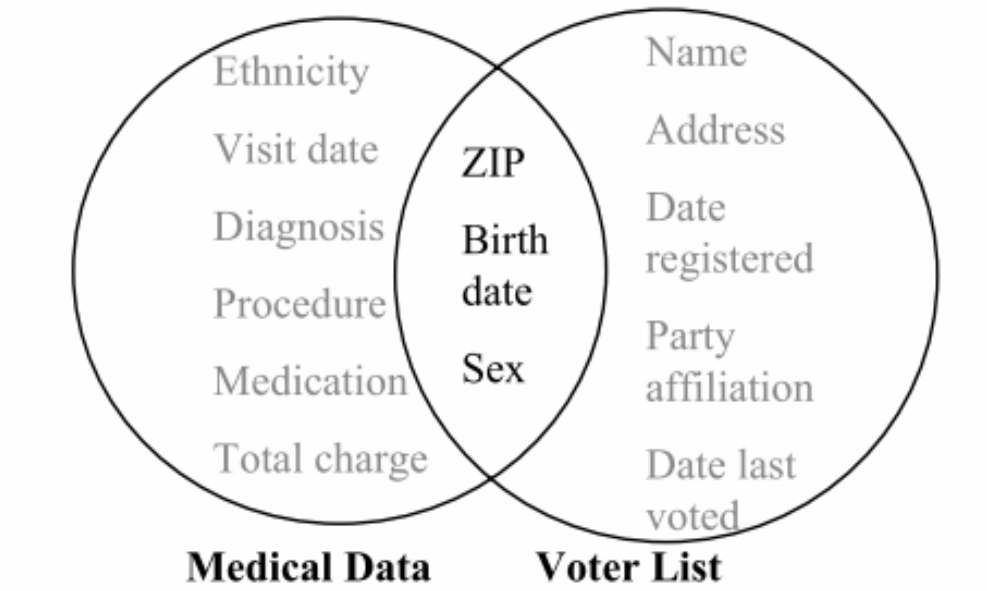
Individuals

Publisher *(trusted)*

Recipients *(no assumption of trust)*

Pseudonymization is not safe

- Sweeney [1] shows the existence of *quasi-identifiers*:
  - Medical data were « anonymized » and released;
  - A voter list was publicly available;
  - → Identification of medical records of Governor Weld by joining datasets on the *quasi-identifiers.*



  - In the US census of 1990: « 87% of the population in the US had **characteristics that likely made them unique** based only on {5-digit Zip, gender, date of birth} » [1].

Pseudonymization is not safe: cont'

In 2006, AOL™ released a list of web search queries [1]:

- 20 million search queries;
- issued by 658.000 unnamed users;

AnonID	Query	QueryTime
1326	*"holiday mansion houseboat"*	2006-03-29
1326	*"back to the future"*	2006-04-01
591476	*"english spanish translator"*	2006-03-20
591476	*"panama vacations"*	2006-03-20
591476	*"breast reduction"*	2006-03-23
591476	*"volunteer work at hospitals in brooklyn"*	2006-05-24
591476	...	...
591476	*"how to secretly poison your ex"*	2006-03-12

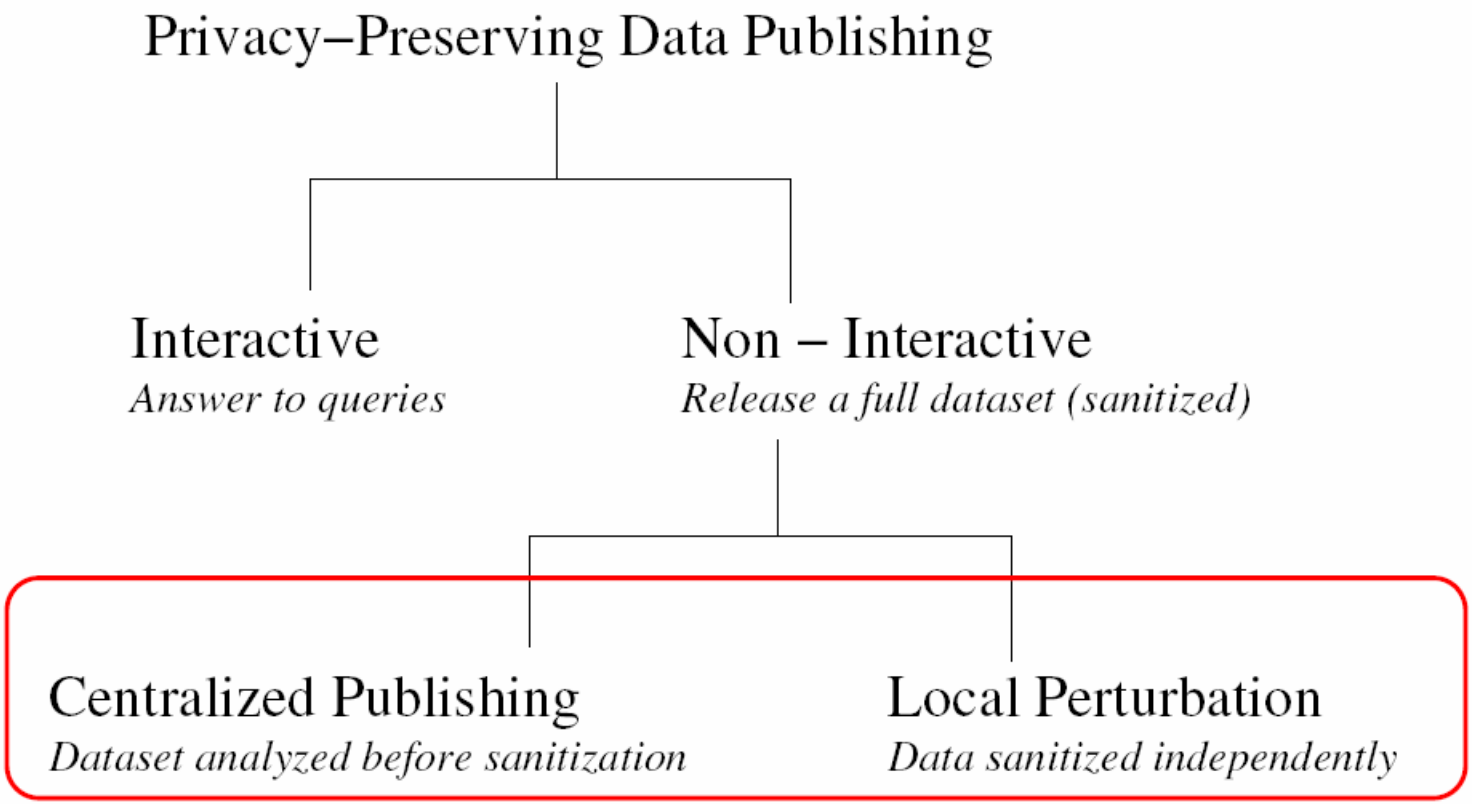
Pseudonymization is not safe: cont'

And especially:

AnonID	Query
4417749	people with last name *"Arnold"*
4417749	*"landscapers in Lilburn,Ga"*
4417749	*"60 single men"*
4417749	*"dog that urinates on everything"*
4417749	dog-related queries

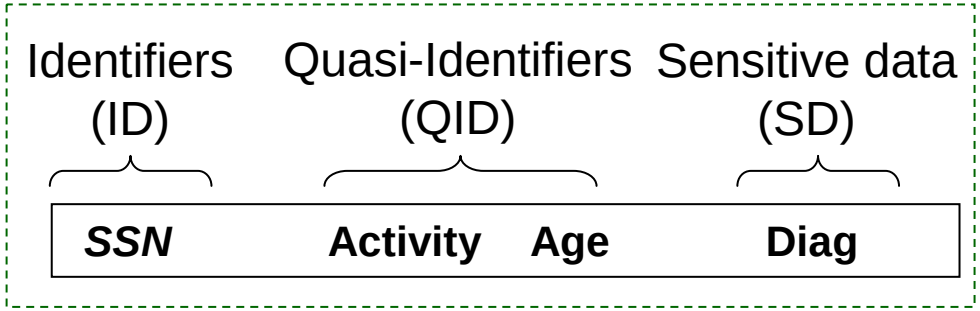
⇒ Few days after: Thelma Arnold is identified [2]. . . and AOL™ removes hastily the dataset from its website.

Classification of the approaches



Data Partitionning Family

Data classification



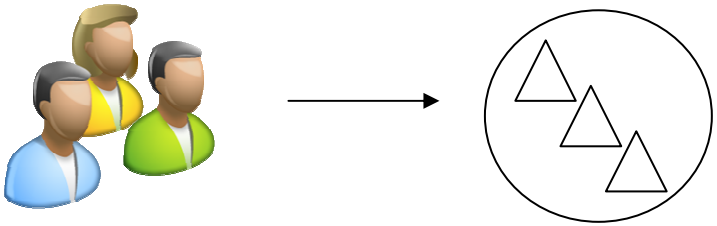
- For each tuple:
  - Identifiers must be removed;
  - The link between a quasi-identifier and its corresponding sensitive data must be *obfuscated* but remain *true* (see the *k*-anonymity section) ;

Questions

- On peut voir les attributs du QID comme les axes d'analyse des attributs du SD
  - Que seraient les attributs d'un SD typique? Leur cardinalité?

k-Anonymity [1]

Hide into the crowd

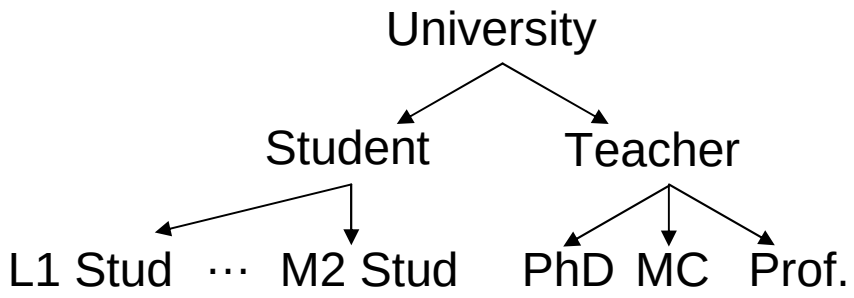


Intuition

- Make groups of *k* individuals, and associate each group to its corresponding group of sensitive records;
- A groups is called an *equivalence class* (or simply a *class);*

k-anonymity by Generalization (1)

- Rationale: Replace the QIDs by more general values such that they encompass at least *k* individuals.
- Value Generalization Hierarchies (VGH) define the « generalizes » relation:



VGH of attribute "Activity"
(example of a categorical VGH)

- Numerical VGH can be dynamically computed (see « Chapter 3: Sanitization algorithms »)

k-anonymity by Generalization (2)

- Back to our example, with k=3:

Name	**Activity**	**Age**	**Diag**
Sue	"M2 Stud."	22	Flu
Pat	"MC"	27	Cancer
Dan	"PhD"	26	Cancer
Bob	"M1 Stud."	21	HIV
Bill	"L3 Stud."	20	Flu
San	"PhD"	24	Cancer

Raw data

Activity	**Age**	**Diag**
"Student"	[20, 22]	Flu
"Student"	[20, 22]	HIV
"Student"	[20, 22]	Flu
"Teacher"	[24, 27]	Cancer
"Teacher"	[24, 27]	Cancer
"Teacher"	[24, 27]	Cancer

3-anonymous data
(by generalization)

k-anonymity by Bucketization [2]

- Rationale: form groups of *k* (QIDs, SDs), and split the groups in a QID table and a SD table;

Name	**Activity**	**Age**	**Diag**
Sue	"M2 Stud."	22	Flu
Pat	"MC"	27	Cancer
Dan	"PhD"	26	Cancer
Bob	"M1 Stud."	21	HIV
Bill	"L3 Stud."	20	Flu
San	"PhD"	24	Cancer

Raw data

Activity	**Age**	***Group***
"M2 Stud."	22	*G1*
"M1 Stud."	21	*G1*
"L3 Stud."	20	*G1*
"MC"	27	*G2*
"PhD"	26	*G2*
"PhD"	24	*G2*

QID Table

Group	**Diag**
G1	Flu, HIV, Flu
G2	Cancer * 3

SD Table

3-anonymous data
(by Bucketization)

k-anonymity guarantees that…

- An individual whose QID belongs to a class and who participated in the release can be associated to any of the *k* records of the group:
  - Eg, Sue can be associated to any of {Flu, HIV, Flu}

→ Record linkage probability = 1/k

Sue

Activity	Age	Diag
"Student"	[20, 22]	Flu
"Student"	[20, 22]	HIV
"Student"	[20, 22]	Flu

3-anonymous data
(by Generalization)

Questions

- Quid de l'utilité des classes suivantes?

"Student"	[20, 22]	→ {Flu, HIV, Flu}
"Teacher"	[24, 27]	→ {Cancer, Cancer, Cancer}

  – Et si elles se chevauchent:

"Student"	[20, 25]	→ {…}
"Teacher"	[24, 27]	→ {…}

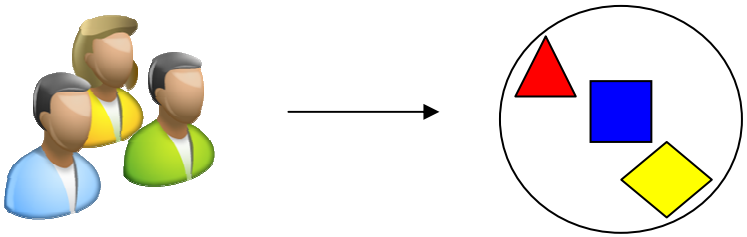
  – Quelles propriétés doivent être vérifiées pour qu'elles soient utilisables selon vous?

- Habituellement, faites-vous des croisements multi-sources? Comment?

- Quelle cohérence pour les données de sortie?
  – Cohérence forte comme ce que propose la généralisation?
  – Cohérence statistique (eg, préservation des counts)?

l-Diversity [3]

Diversify possibilities



Observation

- k-anonymity prevents record linkage but not attribute linkage.
- Example of attribute linkage:
  - Pat is bound to a class of 3 records;
  - But all records contain the value « Cancer »…

Pat

Activity	Age	Diag
"Teacher"	[24, 27]	Cancer
"Teacher"	[24, 27]	Cancer
"Teacher"	[24, 27]	Cancer

  - So Pat has « Cancer »…

Intuition

- Ensure that each *k*-anonymous group is diverse enough;
  - every class must be associated to at least *l* « well-represented » sensitive values;
  - « well-represented » can be instanciated in different ways;
- In practice, *l*-diversity builds on *k*-anonymity;

l-diversity by Generalization

- For example:

Name	**Activity**	**Age**	**Diag**
Sue	"M2 Stud."	22	Flu
Pat	"MC"	27	Cancer
Dan	"PhD"	26	Cancer
Bob	"M1 Stud."	21	HIV
Bill	"L3 Stud."	20	Flu
San	"PhD"	24	Cancer
John	"M2 Stud"	22	Cold
Jim	"M2 Stud"	23	Cancer

Raw data

Activity	**Age**	**Diag**
"Student"	[20, 23]	Flu
"Student"	[20, 23]	HIV
"Student"	[20, 23]	Cancer
"University"	[22, 24]	Flu
"University"	[22, 24]	Cold
"University"	[22, 24]	Cancer

(3-anonymous and 3-diverse) data (by generalization)

l-diversity guarantees that…

- An individual whose QID belongs to a class and who participated in the release can be associated to any of the *l* most present sensitive values with the same probability;
  - Eg, Bob can be associated to any of {Flu, HIV, Cancer} with the same probability;

→ Attribute linkage probability = 1/l

Bob

Activity	Age	Diag
"Student"	[20, 23]	Flu
"Student"	[20, 23]	HIV
"Student"	[20, 23]	Cancer

Questions

- Quid de l’utilité de classes l-diverses?
- Que pensez-vous du cas d’attaque par homogénéité des données sensibles?
  - Dépend de la valeur de k;
  - Dépend des attributs des données sensibles (cardinalité);

t-closeness [4]: extending *l*-diversity

- Intuition: Within each class, the distribution of sensitive values must be close to the global distribution by at most a factor *t*;

→ the adversary's posterior belief is the same as his prior belief (including the resulting global distribution);

- Example:

Non-Sensitive		Sensitive	
Age	Gender	Disease	Count
< 40	M	Flu	400
< 40	M	Cancer	200
≥ 40	M	Flu	400
≥ 40	M	Cancer	200
≥ 40	F	Flu	400
≥ 40	F	Cancer	200

Questions

- Quid de l’utilité des classes respectant la t-closeness?

Data Partitionning Family

Continuous release

Problem

- The dataset is continuously evolving by insertions and deletions;
- Releasing independently *k*-anonymous or *l*-diverse versions of the dataset may open privacy breaches;
- For example:

	Name	**Activity**	**Age**	**Diag**
t1	***Bob***	**"M1 Stud."**	**21**	**HIV**
	Bill	"L3 Stud."	20	Flu
	Jim	"M2 Stud"	23	Cancer
t2	***Bob***	**"M1 Stud."**	**21**	**HIV**
	Helen	"L1 Stud."	18	Cold
	Jules	"L1 Stud"	19	Dysp.

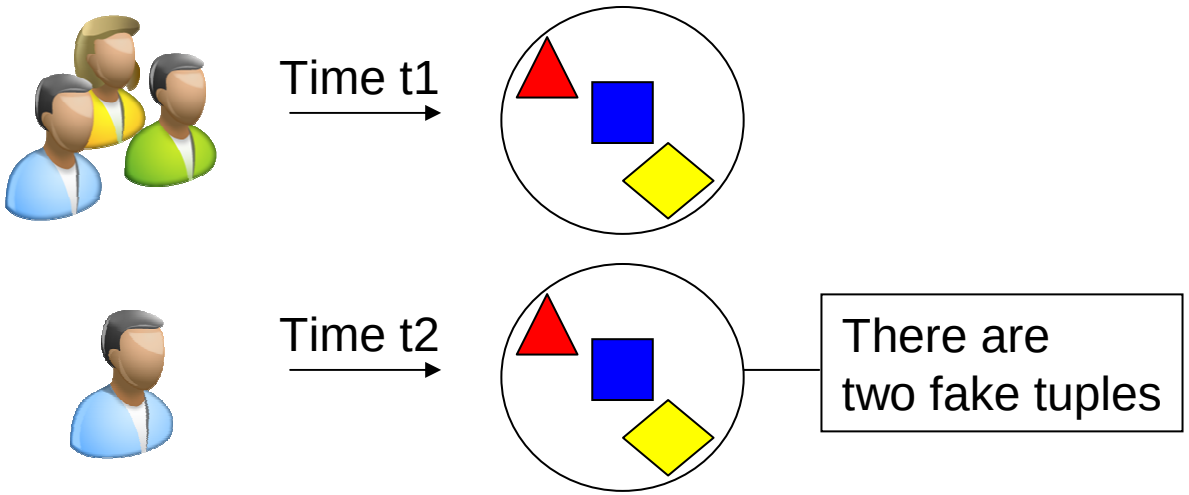
Activity	**Age**	**Diag**
"Student"	[20, 23]	Flu
"Student"	[20, 23]	**HIV**
"Student"	[20, 23]	Cancer
"Student"	[19, 21]	**HIV**
"Student"	[19, 21]	Cold
"Student"	[19, 21]	Dysp

Questions

- Quels sont les cadres applicatifs du continuous release?
  - Suivi individuel de chaque dossier?
  - Suivi d’une population?
  - …?

m-Invariance [5]

Never change



Intuition

- The set of sensitive data associated to a QID must be invariant.

	Name	Activity	Age	Diag
t1	***Bob***	**"M1 Stud."**	**21**	**HIV**
	Bill	"L3 Stud."	20	Flu
	Jim	"M2 Stud"	23	Cancer

t2	***Bob***	**"M1 Stud."**	**21**	**HIV**
	Helen	"L1 Stud."	18	Cold
	Jules	"L1 Stud"	19	Dysp.
	Gary	"PhD"	24	Gast.

Activity	Age	Diag
"Student"	[20, 23]	Flu
"Student"	[20, 23]	**HIV**
"Student"	[20, 23]	Cancer

Activity	Age	Diag
"Student"	[20, 23]	**HIV**
"Student"	[20, 23]	Flu
"Student"	[20, 23]	Cancer
"University"	[18, 24]	Cold
"University"	[18, 24]	Dysp.
"University"	[18, 24]	Gast.

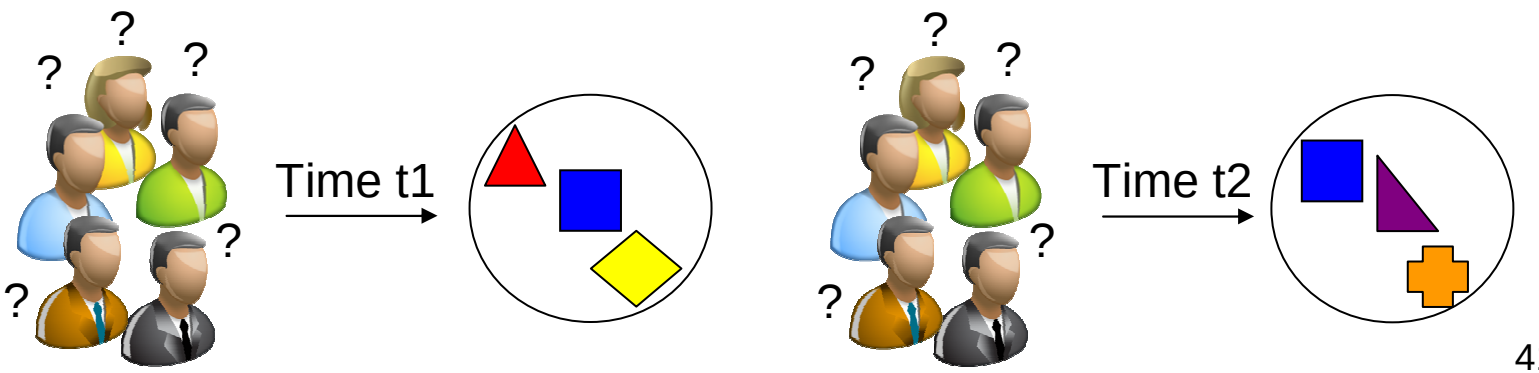
2 fake tuples

Questions

- Quid de l’utilité de ces classes se dégradant au cours du temps?

Directions for a sampling-based model

Introduce doubts



Intuition

- Assign a participation probability $P_{participation}$ to each individual into each release:
  - Pros:
    - Make anyone unable to state with certainty who participated in which release;
    - There is no constraint on data, so no need to introduce, eg, fake tuples as in m-invariance;
  - Cons:
    - Each release considers a sample of the population under study;
    - The samples considered by two releases are very likely to be different;

Questions

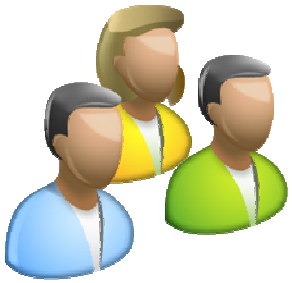
- Travaillez vous par échantillonage?
- Avez-vous des idées des tailles respectives de la population et de l’échantillon pour qu’il soit représentatif?

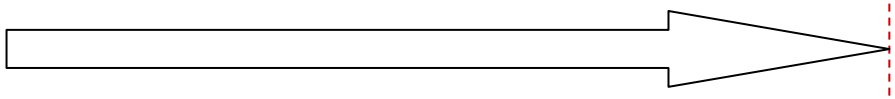
- Quid de l’utilité des données?

Data Perturbation Family

Local Perturbation

Local Perturbation





Anonymized data
(or *sanitized …*)

Individuals

Recipients *(no assumption of trust)*

Basic mechanism [6]

- Survey context;
- **Sensitive question:** Have you ever driven intoxicated?
- **Response:** truthful with probability *p* , lie with probability *(1-p);*
- Estimator:
  - Let π be the fraction of the population for which the true response is « *Yes* »
  - Expected proportion of « Yes »:
    $P(Yes) = (\pi * p) + (1 - \pi)*(1 - p)$
    → $\pi = [P(Yes) - (1 - p)] / (2p - 1)$
  - If m/n individuals answered « yes », π_{est} estimates π:
    $\pi_{est} = [m/n - (1 - p)] / (2p - 1)$

Extended mechanism [7]

- **Server: defines queries:**
  - A conjunction of values (eg, people that have « HIV+ = true » and « aids = false»);
  - And wants to know the fraction of individuals that agree with the conjunction;

Extended mechanism [7] cont'

- **Individuals: each one receives the values of each conjunction :**
  - Eg : the conjunction contains « HIV+ » and « aids »;
  - Generates the vector of all the possible answers (« HIV+ = true » and « aids = true », « HIV+ = true » and « aids = false », …) with his answer set to 1:

0	0	1	0

  - And flips each element of the vector with probability p;

1	0	1	0

Flipped (probability *p*) Not flipped (probability *(1 – p)*)

Extended mechanism [7] cont'

- **Server: receives the pertubed vectors:**
  - Estimate the count result:
    - r_{pert} = number of perturbed vectors that agree with the conjunction;
    - $r_{est} = (r_{pert} – p)/(1 – 2p)$;
  - The true result r is proven to be « not too far away » from r_{est} ;

Questions

- De quel ordre est le nombre typique:
  - D’attributs dans une requête?
  - De valeurs possibles par attributs?
  - De requêtes d’une étude épidémiologique?
- Les données auxquelles vous avez accès contiennent-elles déjà des erreurs non intentionnelles?
- De quels estimateurs statistiques avez-vous besoin?

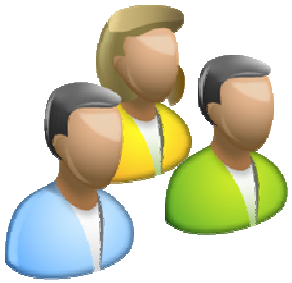
Privacy Guarantees (informally)

- Local perturbation mechanisms are proven to satisfy **(α, β)-privacy** [8] through **γ-amplification** [8]:
  - **(α, β)-privacy:** the gain of the adversarial probabilistic knowledge about each input tuple is bounded by the probabilities α and β;
  - **γ-amplification:** Any output tuple can be the sanitization of any input tuple with roughly the same probability (which depends on α and β);

Data Perturbation Family

Input Perturbation

Input Perturbation



Individuals

Raw data

Publisher *(trusted)*

Anonymized data
(or *sanitized …*)

Recipients *(no assumption of trust)*

Matrix masking [9]

- Huge amount of work from the statistical literature;
- General framework is called « matrix masking » [9]:
  - Z is an (n * p) data matrix
  - Z is perturbed: Z = AZB + C;
    - A: operates on the rows (delete/add);
    - B: operates on the columns (delete/add);
    - C: add noise to each cell;
  - Z is released;

The Post RAndomization Method [10] as an example

- Inspired from local perturbation techniques;
- PRAM perturbs randomly the entries of the database:
  - Suppose a single attribute in the db, taking value in 1…K;
  - Let p_{kl} denote the probability that value *k* be changed to value *l.*
  - Let P = $\{p_{kl}\}$ a K*K Markov matrix with p_{kl} as its (k, l)th entry
  - Process:
    - **Publisher:** Deliver both the data perturbed wrt P, and P;
    - **Recipient:** Compute an estimator of the frequencies: $T_{est} = (P^{-1})^T T_{pert}$

The Post RAndomization Method [10] as an example (cont')

- More:
  - generalizes to the multivariate case;
  - Impossibility to join the perturbed data with external data. But we can get estimates of the cross-tabulation frequencies;
  - Tailor P such that $T_{pert} = T_{est}$.
  - Privacy guarantees: identical to that of local perturbation;

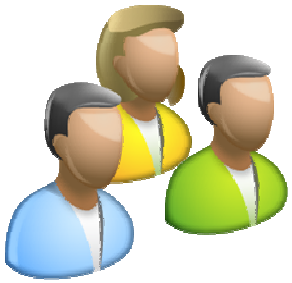
Data swapping [12]

- Sensitive values could also be swapped between records…
- Details are in [12]

Data Perturbation Family

Statistical Perturbation

Statistics Perturbation (interactive setting)



Individuals

Raw data

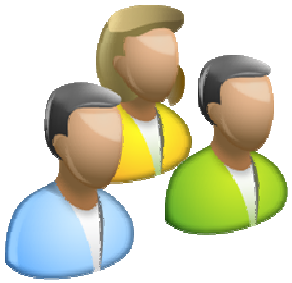
Publisher *(trusted)*

Statistical queries (eg, counts)

Perturbed answers

Recipients *(no assumption of trust)*

Statistics Perturbation (non-interactive setting)



Individuals

Raw data

Publisher *(trusted)*

Sanitized **statistics**

Recipients *(no assumption of trust)*

Achieving differential privacy [11] (interactive setting)

- Receive a statistical query (eg, a count): Q_1;
- Compute its *sensitivity* s_1 (depends on the output of the current query, and on the previous queries);
- Draw a sample η_1 from a Laplace distribution parameterized by s_1 (among others);
- Output the perturbed statistics: $Q_1 + \eta_1$;

- The error magnitude is low;
- The total number of queries is bounded;

Privacy guarantees (informally)

- Differential privacy [11]:
  - For any individual i, the sanitized outputs with and without including i's data are nearly indistinguishable;

  → Minimize the increased risk to an individual incurred by joining or leaving the database;

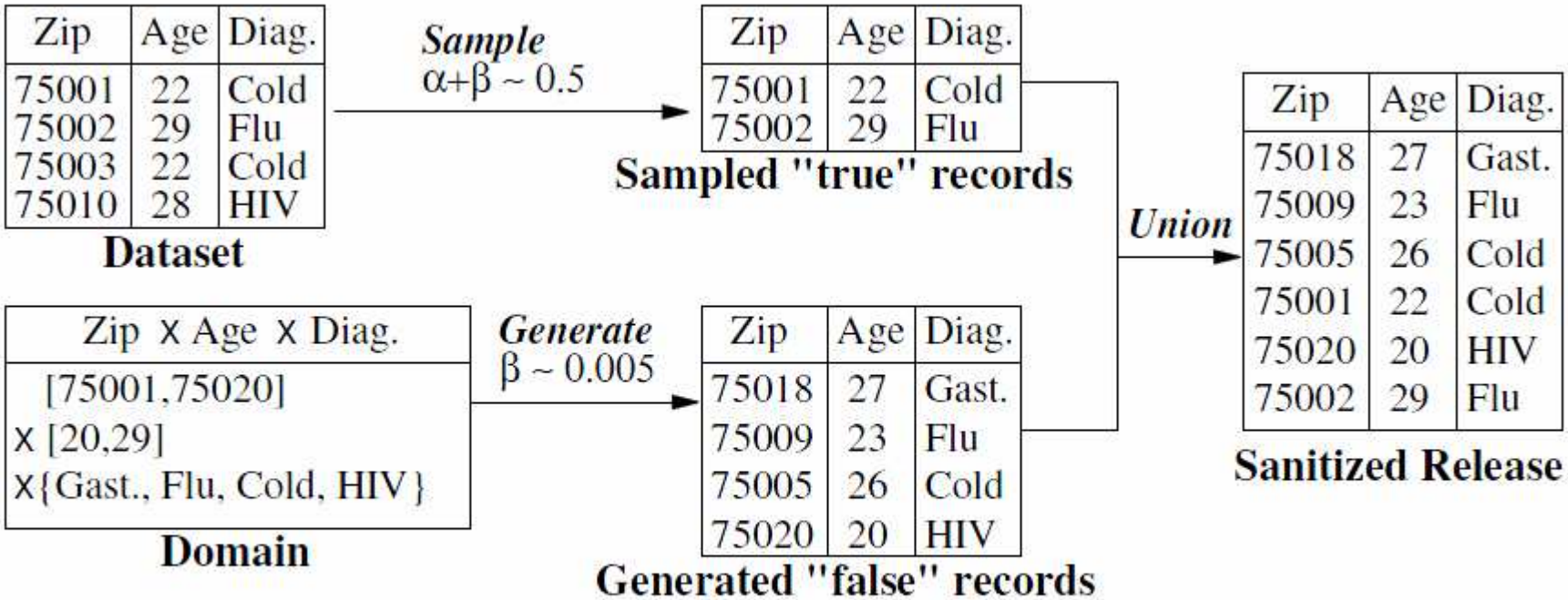
Differential Privacy

An algorithm A satisfies ε-differential privacy if

- For every pair of neighboring tables D_1, D_2(differ in one individual)
- For every output Ω

$$\Pr[A(D_1) = \Omega] \leq e^{\varepsilon} \Pr[A(D_2)=\Omega]$$

α, β – algorithm [14]



We can compute aggregate values such as COUNTs based on the estimator :

$$Q_{Cold} = (n_{sanitized} - \underbrace{\beta.n_{Domain}}_{=200*0.005=1}) / \alpha$$

=2 =0.5

Questions

- « Deviner » les statistiques d'intérêt sans « voir » les données est réaliste?

Conclusion

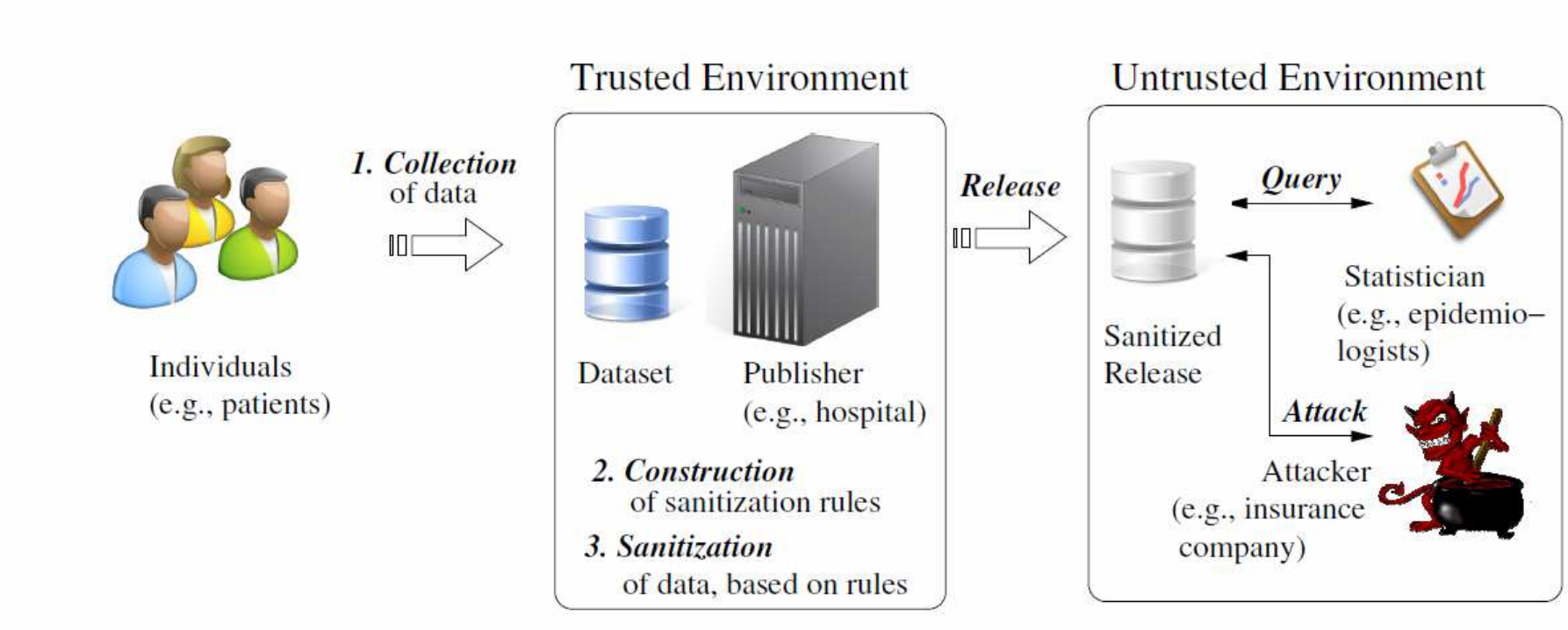
PPDP

- Many anonymity models exist.
- All are implemented, some are commercial products.

→ *Anonymization is possible and should be done.*

- « Utility » of anonymous data is hard to evaluate, what is important is to correctly anonymize with regards to a specific model and risk.

Traditional PPDP Process

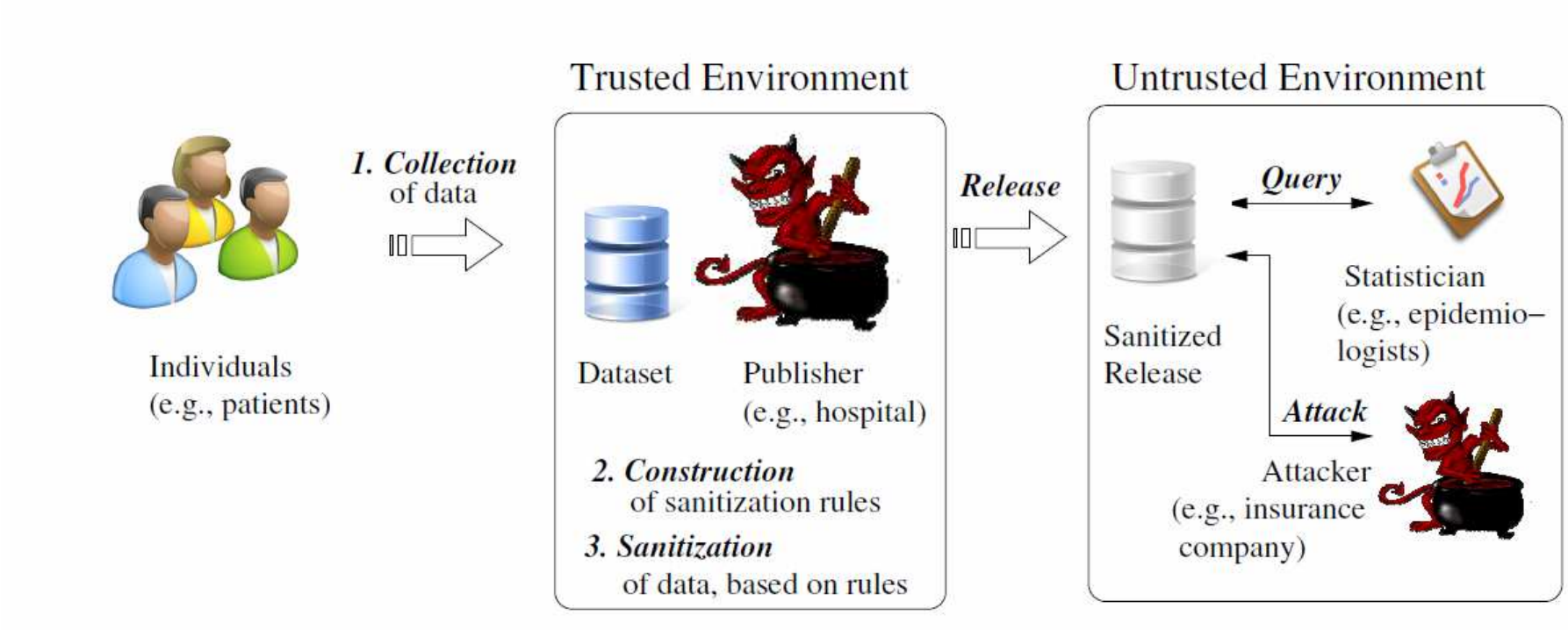


Individuals :
Private Data

Publisher
(trusted)

Recipients
(no trust assumption)
→Privacy Models
K-anon, L-div, Dif. Priv.

Traditional PPDP Process

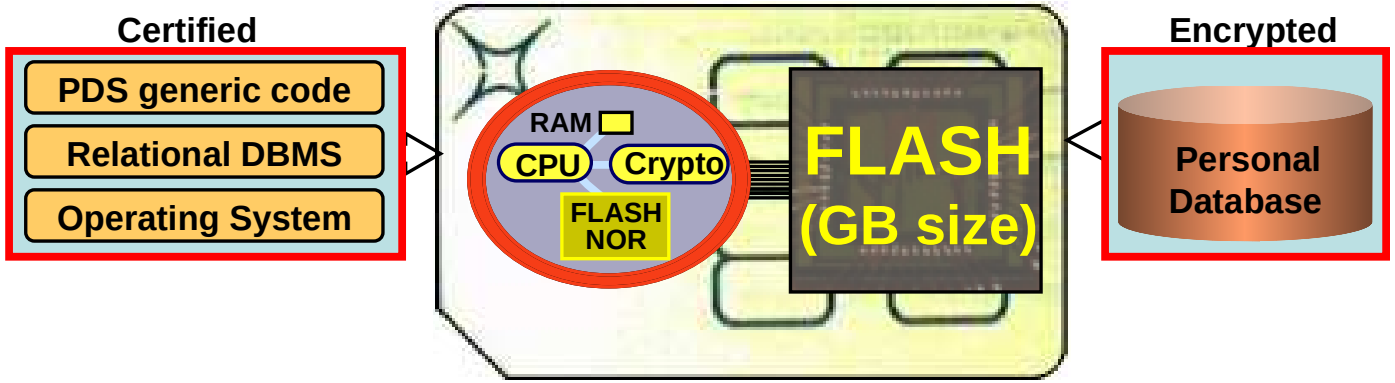


Individuals : Private Data

Publisher *(UNTRUSTED)*
→Secure Computation Needed

Recipients
(no trust assumption)
→Privacy Models
K-anon, L-div, Dif. Priv.

Introducing the Secure Personal Data Server



TAMPER RESISTANT
& SECURE !!

Simple Example :
K-anonymity & Mondrian Algorithm (LeFevre'06)

in cleartext

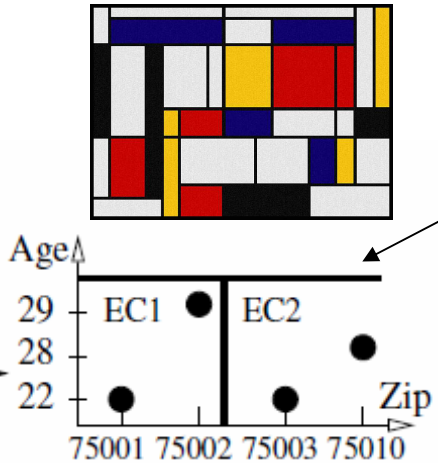
Executed on a trusted central server !

QI SD

Zip	Age	Diag.
75001	22	Cold
75002	29	Flu
75003	22	Cold
75010	28	HIV

Dataset

Form the Eq. Classes



2–anonymous Eq. Classes

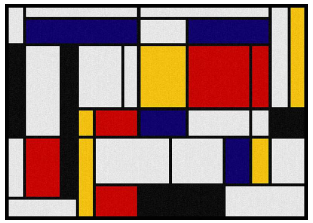
	Zip	Age	Diag.
EC1	[75001, 75002]	[22, 29]	Cold Flu
EC2	[75003, 75010]	[22, 29]	HIV Cold

Sanitized Release

1) Collection Phase 2) Construction Phase 3) Sanitization Phase

Simple Example :
K-anonymity & Mondrian Algorithm (PST'11 award)

encrypted by PDS



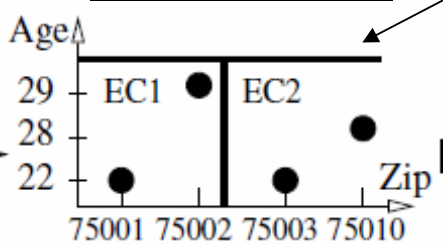
Executed on a *semi-honest* central server !

QI SD

Zip	Age	Diag.
75001	22	Cold
75002	29	Flu
75003	22	Cold
75010	28	HIV

Dataset

Form the Eq. Classes

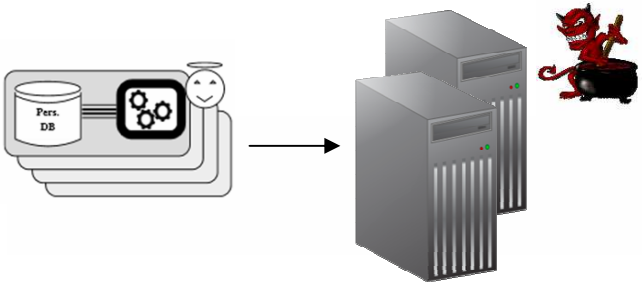


2–anonymous Eq. Classes

	Zip	Age	Diag.
EC1	[75001, 75002]	[22, 29]	Cold Flu
EC2	[75003, 75010]	[22, 29]	HIV Cold

Sanitized Release

1) Collection Phase



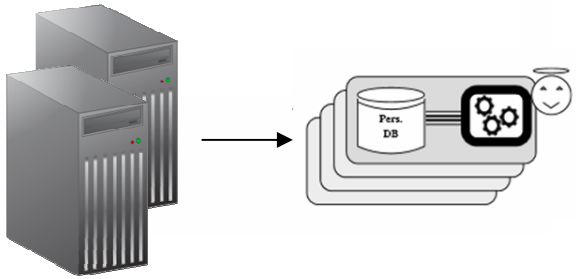
$t_i = (QI_i, E(SD_i))$

2) Construction Phase

$QI_i \rightarrow EC_j$

$t_i = (EC_j, E(SD_i))$

3) Sanitization Phase



$t_i = (EC_j, SD_i)$

Parallelizing and securing general PPDP (DAPD'13)

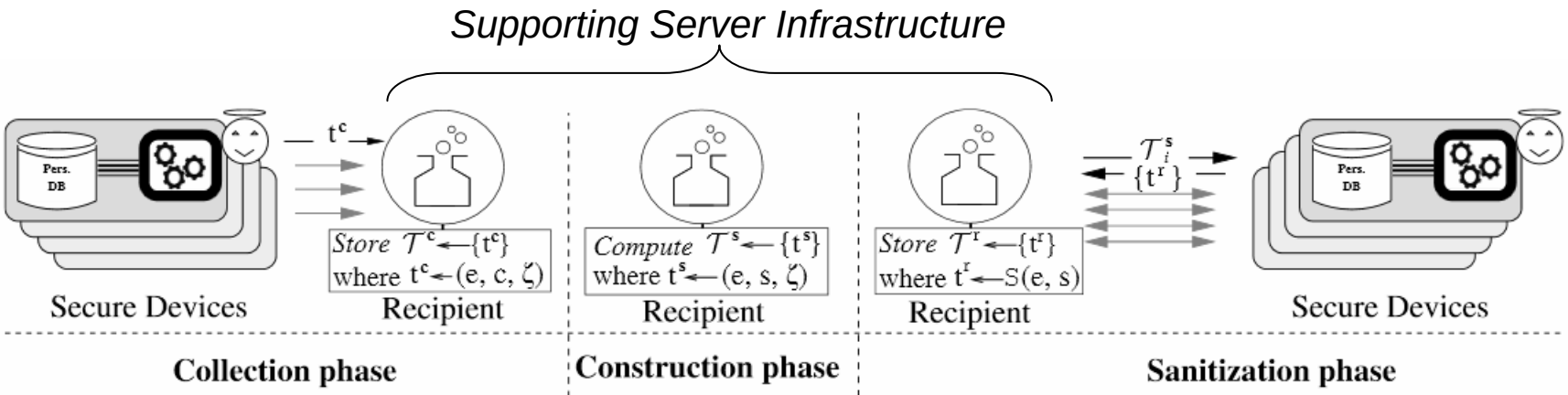
Collection phase is naturally parallel !

Construction phase is *algorithm dependant*. To remain as generic as possible, this task is delegated to the central server, while disclosing an amount of information compatible with the privacy requirements.

→*BREAK UP THE TUPLES !*

- Encrypted Data (e)
- Construction Information (c) / Sanitization information (s)
- Safety Information (ζ)

Sanitization is parallelized on the tokens by sending them batches of information.



Achievement : We have proposed and proven the security of a set of Generic Primitives to manage current privacy models

Malicious adversaries (ISPEC'11)

PB : Attacks launched in the case of covert adversaries can be the *deletion* and *copy* of tuples (malicious SSI) and *creation* of tuples (malicious Token)

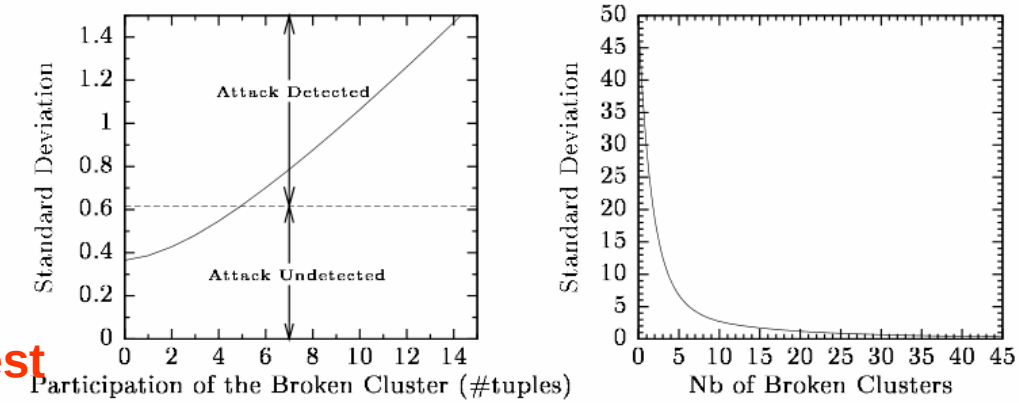
John Doe : [75010, 28, ?]

Inject k-1 tuples of the form [75010, 28, Flu]

[75010, 28, Flu], [75010, 28, Flu], … [75010, 28, HIV]

Classical prevention against malicious tokens is to CLUSTER.
We must also prevent impact of a broken cluster on others…

In the case of covert adversaries, detection is probabilistic.



(a) Standard Deviation Sensitivity wrt the Participation of the Broken Cluster

(b) Standard Deviation Sensitivity wrt the number of Broken Clusters

→Once active attacks are prevented, we are in the case of Semi-honest Adversary

→These counter measures are generic

Fin

Et vous?

Questions

- Collecte des données:
  - Qui collecte?
  - A qui?
  - Quelles sont les données collectées? Socio-médicales? Autres?
  - Par quels moyens?
  - Quid des études longitudinales?
    - Objectifs,
    - fréquence de collecte,
    - taille des cohortes,
    - etc.?
- Anonymisation des données
  - Qui anonymise?
  - Quels sont les procédés?
  - Comment sont publiées les données anonymisées?
  - A qui?
  - Dans quels cas les propriétés suivantes sont-elles requises?
    - **Les données anonymisées sont « vraies »:** par exemple, la ligne <30 ans, 75001, Cancer> implique qu'un participant de l'étude a bien ces valeurs;
    - **Les données reflètent globalement la vérité:** les lignes de données sont fausses mais la distribution globale des données anonymisées est celle des données de base, ie, conservation de certains invariants statistiques.

Encore merci!

References

[1] L. Sweeney. k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 10(5), 2002.

[2] Xiaokui Xiao , Yufei Tao, Anatomy: simple and effective privacy preservation, Proceedings of the 32nd international conference on Very large data bases, September 12-15, 2006, Seoul, Korea.

[3] Ashwin Machanavajjhala , Daniel Kifer , Johannes Gehrke , Muthuramakrishnan Venkitasubramaniam, L-diversity: Privacy beyond k-anonymity, ACM Transactions on Knowledge Discovery from Data (TKDD), v.1 n.1, p.3-es, March 2007.

[4] Ninghui Li, Tiancheng Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106--115, April 2007.

[5] Xiaokui Xiao , Yufei Tao, M-invariance: towards privacy preserving re-publication of dynamic datasets, Proceedings of the 2007 ACM SIGMOD international conference on Management of data, June 11-14, 2007, Beijing, China

[6] S. L. Warner, “Randomized Response: A survey technique for eliminating evasive answer bias,” Journal of the American Statistical Association, 1965.

[7] Nina Mishra , Mark Sandler, Privacy via pseudorandom sketches, Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, June 26-28, 2006, Chicago, IL, USA

References

[8] Alexandre Evfimievski , Johannes Gehrke , Ramakrishnan Srikant, Limiting privacy breaches in privacy preserving data mining, Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, p.211-222, June 09-11, 2003, San Diego, California.

[9] Duncan G.T., Jabine T.B., and De Wolf V.A. (eds.). Private Lives and Public Policies. Report of the Committee on National Statistics' Panel on Confidentiality and Data Access. National Academy Press, WA, USA, 1993.

[10] J. Gouweleeuw, P. Kooiman, L. C. R. J. Willenborg, and P.-P. de Wolf, "The Post Randomisation Method for Protecting Microdata," QUESTIIO, vol. 22, no. 1, 1998.

[11] C. Dwork, A Firm Foundation for Private Data Analysis, *To appear in* Communications of the ACM, 2010.

[12] S. E. Fienberg and J. McIntyre, "Data swapping: Variations on a theme by Dalenius and Reiss," in *Privacy in Statistical Databases*, pp. 14-29, 2004.

[13] Bee-Chung Chen, Daniel Kifer, Kristen LeFevre and Ashwin Machanavajjhala, Privacy-Preserving Data Publishing, In *Foundations and Trends in Databases,* vol.2, issue 1–2 , January 2009.

[14] Vibhor Rastogi, Dan Suciu and Sungho Hong, *The Boundary between privacy and utility in data publishing*, in *VLDB 2007*

[15] Article 29 Data Protection Working Party, *Opinion 05/2014 on Anonymisation Techniques,* april 2014