Privacy-Preserving Data Publishing

Where are we now ?

Talk @ Séminaire DIT - ENS Rennes

Tristan Allard
Druid team
Univ. Rennes 1 / Irisa lab.

January 24, 2017

Introduction

For l-Diversity and ϵ-Differential Privacy, two seminal privacy models !

Progress of the Talk

Progress of the Talk

Once upon a time in the very early 2000's I

Once upon a time in the very early 2000's II

- Around 360M Internet users[1]: ∼100M US, ∼100M EU, ∼100M Asia
- ADSL is spreading (against 56K modems)
- RAM: 64MB at ∼70$[2]
- HD: 40GB at ∼250$[2]
- First USB flash drive commercialized[3] (8MB)
- *"1999: The release of Oracle8i aimed to provide a database inter-operating better with the Internet (the i in the name stands for 'Internet')."*[4]
- Google.com is 3 years old and Adwords is launched (350 users) [5]

---

[1] http://www.internetworldstats.com/

[2] http://www.statisticbrain.com/average-historic-price-of-ram/

[3] https://en.wikipedia.org/wiki/USB_flash_drive

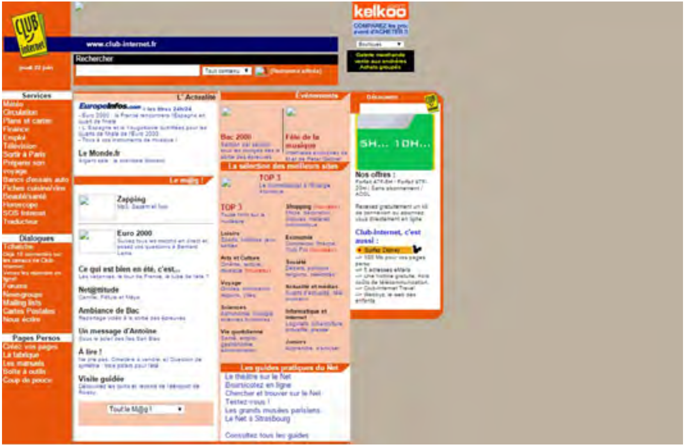
[4] https://en.wikipedia.org/wiki/Oracle_Database

[5] https://www.google.com/about/company/history/

From the archives I

From the archives II

From the archives III



From the archives IV

(re)Birth of a Problem (PPDP)

Governor Weld's Case I

In 2002, Sweeney accessed two datasets [46]:

- The Massachussets Group Insurance Commission (GIC):
  - collected health and **demographic data** of 135 000 state employees and families
  - produced a copy of the data for research purposes
  - Believed to be safe: names and social security numbers had been removed
- The voter list of Cambridge Massachussets (two diskettes, $20): **demographic data** and names;

Governor Weld's Case II

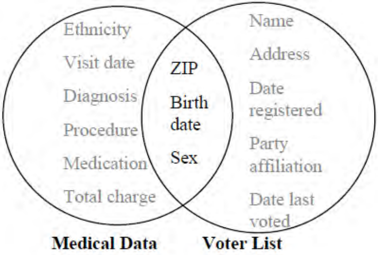


Figure: `Medical JOIN Voter ON (zip, DoB, sex)`

A straightforward disclosure

- Governor Weld lived in Cambridge and was part of the GIC dataset;
- In the voter list: 6 individuals had his birthdate, 3 of them were men, only one had Weld's zipcode;

Pseudonymity is not Enough

Publishing data while only removing direct identifiers, e.g., name, address, from data (aka *pseudonymity*) may be harmful not only for Governor Weld !

Simple Demographic Data is Identifying for Many Persons

The majority of the US population is unique wrt {zip code, DoB, sex} [45, 22].

k-Anonymity : Assumptions I

- Considers that individuals' data is made of :
  - Identifying attributes, or **ID**: **identify uniquely** each individual (e.g., $\langle$`SSN`$\rangle$);
  - Quasi-Identifying attributes, or **QID**: **may identify uniquely** some individuals (*e.g.*, $\langle$ `Zip`, `DoB`$\rangle$);
  - Sensitive attributes, or **SD**: sensitive data, e.g., $\langle$ `Disease` $\rangle$;

k-Anonymity : Assumptions II

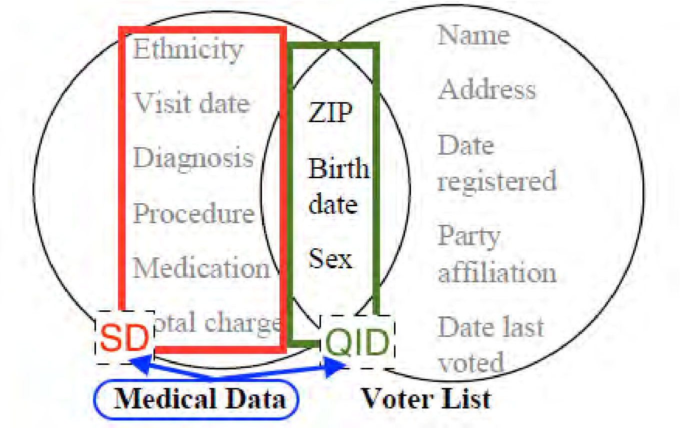


Figure: Quasi-identifiers and sensitive data in Gov. Weld's case

k-Anonymity: the Model I

Warning

We consider in this talk that each individual has a single record in the DB.

k-Anonymity: the Model II

A release is k-anonymous [46] if:

- It does not contain any direct identifier
- The QID of each record has been made indistinguishable from at least $(k-1)$ others

$\Rightarrow$ Each sensitive data is within a group that corresponds to at least k QID.

k-Anonymity: the Model III

Name	Zip	Age	Dis.
Bob	75001	22	Cold
Bill	75002	29	Flu
Don	75003	22	Cold
Sue	75010	28	HIV

Table: Raw data (*e.g.*, GIC medical data).

Zip	Age	Dis.
[75001, 75002]	[22, 29]	Cold
[75001, 75002]	[22, 29]	Flu
[75003, 75010]	[22, 29]	Cold
[75003, 75010]	[22, 29]	HIV

Table: A possible 2-Anonymous Release of the raw data.

k-ANONYMITY: the Model IV

Name	Zip	Age
Bob	75001	22

Zip	Age	Dis.
[75001, 75002]	[22, 29]	Cold
[75001, 75002]	[22, 29]	Flu
[75003, 75010]	[22, 29]	Cold
[75003, 75010]	[22, 29]	HIV

Table: Left: External knowledge made of a known QID (*e.g.*, voter list). Right: A possible 2-Anonymous release of the raw data.

$\Rightarrow$ Joins on QID are now ambiguous: what is Bob's disease?

k-Anonymity: the Model V

Vocabulary

- **Equivalence class**: A group of records indistinguishable wrt their QID
- **Sanitized release**: the set of equivalence classes finally published

Mondrian : A Simple Algorithm for Achieving k-Anonymity I

- **Goal**: form equivalence classes that span at least k similar QID values
- **How?** Greedily !
  - Starts with one *partition* of the dataset containing all the records
  - Recursively partitions it into smaller and smaller partitions
  - Finally replace the QID value of each record by the range of its partition

Mondrian : A Simple Algorithm for Achieving k-Anonymity II

Algorithm 1: `MondrianAnonymize`

input : A partition $\mathcal{P}$ to split

output: A set of partitions, each containing between k and $2k-1$ tuples

if *no allowable multidimensional cut for partition* **then return** $\mathcal{P}$;
else
  $dim \leftarrow$ `chooseDimension()`;
  $fs \leftarrow$ `frequencySet(`$\mathcal{P}$, dim`)`;
  $splitVal \leftarrow$ `findMedian(`fs`)`;
  $\mathcal{L} \leftarrow \{t \in \mathcal{P} : t.dim \leq splitVal\}$;
  $\mathcal{R} \leftarrow \{t \in \mathcal{P} : t.dim > splitVal\}$;
  return `MondrianAnonymize(`$\mathcal{L}$`)` $\cup$ `MondrianAnonymize(`$\mathcal{R}$`)`

Mondrian : A Simple Algorithm for Achieving k-Anonymity III

`MondrianAnonymize` internal calls:

- `chooseDimension`: choose the dimension in which to split (usually the widest one);
- `frequencySet`: set of unique values taken by the tuples for the chosen dimension, each paired with the number of times it appears;
- `findMedian`: find the median;

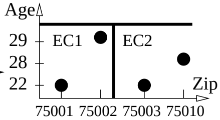
MONDRIAN Illustrated

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

QI: Zip, Age; SD: Diag.

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

In this example, we want 2-ANONYMITY (at least two records per class).

Mondrian, for Real I

Actually, Mr Mondrian was a painter !

Figure: Composition en rouge, jaune, bleu et noir. Mondrian. 1926

Mondrian, for Real II

And a `MondrianAnonymize` partitioning may look like this :

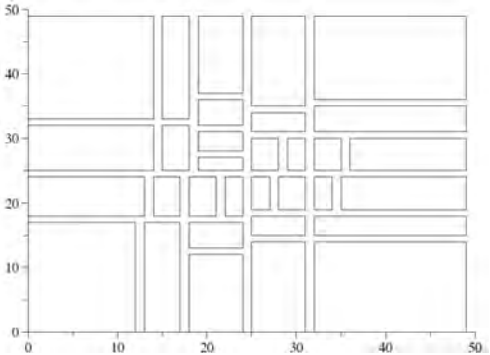


Figure: Example of a Mondrian partitioning [34] (synthetic data, 1000 tuples, k=25, normal distribution).

Components of a Privacy-preserving Data Publishing Solution

Three essential components exhibited by the k-Anonymity research track:

1. **Privacy model**: What does it mean for the data released to be privacy-preserving? Ex.: k-Anonymity.
2. **Privacy algorithm**: How to produce the privacy-preserving dataset to be released? Ex.: Mondrian.
3. **Utility metric**: How much useful is the released data? Ex.: low number of generalizations.

Pseudonymity does not work $\Rightarrow$ Which component(s) does it miss?

Progress of the Talk

Waiting for the Next Scandal

During a few years :

- Academics focus on the algorithmics aspects of k-Anonymity
- And pseudonymity fuels another scandal...

Thelma Arnold's Case I

In 2006, AOL releases a list of web search queries [5]:

- 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

Thelma Arnold's Case II

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

$\Rightarrow$ A few days after: Thelma Arnold is identified [6]... and AOL removes hastily the dataset from its website.

Call for Another Model

- On the same year, Machanavajjhala et al critically analyze the k-Anonymity guarantees
- **Limits of the adversarial model** are identified, an alternative model, called l-Diversity, is proposed

Progress of the Talk

Non-Informative Paradigm: Partition-Based Models and Algorithms
- k-Anonymity VS Pseudonymity
- Larval Period
- l-Diversity
- Endless Cycle

Differential Privacy Paradigm : Models, Algorithms, and Novelties

Conclusion

References

Some Defects of k-Anonymity

Name	**Zip**	**Age**
Bob	75001	22

Zip	**Age**	**Dis.**
[75001, 75002]	[22, 29]	Cold
[75001, 75002]	[22, 29]	Flu
[75003, 75010]	[22, 29]	Cold
[75003, 75010]	[22, 29]	HIV

Table: Attack considered by k-Anonymity. Left: External knowledge made of a known QID (*e.g.*, voter list). Right: A possible 2-Anonymous release.

1. **Homogeneity**: What if all the SD of the QI of an equivalence class are identical?
2. **Background knowledge**: What if the adversary knows that his victim is more or less likely to have a given sensitive data?

$\Rightarrow$ Motivate the l-Diversity model

Foundation: the Bayes-Optimal Privacy Model I

Founding intuition

Background knowledge about SD should be **expressed** and **taken into account** by the privacy model.

The Bayes-Optimal Privacy model [37] is an early attempt to this end (2006):

- **Background knowledge**: joint distribution between QI and SD
- **Prior belief**: given a targeted QI q and a SD s, probability of s given q
- **Posterior belief**: given a targeted QI q, a SD s, **and the sanitized release** $\mathcal{V}$, probability of s given q and $\mathcal{V}$
- **Privacy breach**: if *distance*(*posterior belief*, *prior belief*) $> \theta$ (too much gain in knowledge)

Foundation: the BAYES-OPTIMAL PRIVACY Model II

The intuition behind THIS definition of a privacy breach is **a way to envision privacy** (also called a *paradigm* in these slides) !

Paradigm#1: **Uninformative Principle** [37]

A privacy breach occurs when the *prior belief* of the adversary differs *significantly* from his *posterior belief*.

"If the **release of the statistics S** *make it possible to determine the value D_k* **more accurately** *than is possible* **without access to S**, *disclosure has taken place (...)"*
Dalenius 1977 [12]

Bayes-Optimal Privacy : Impractical

If Bayes-Optimal Privacy were practical, it could permit to check that releases do not allow significant knowledge gains...

But :

- Obtaining the joint distribution f that represents the adversarial background knowledge ?
- What if there are several adversaries ?
- What about other kinds of knowledge ?
- Cost of checking all the possible (q, s) pair !

l-Diversity I

l-Diversity: a simple and easy-to-check condition for protecting against **SD homogeneity** and **adversarial negation statements**.

l-Diversity II

l-Diversity

An l-diverse equivalence class contains at least l *well-represented* sensitive values.

l-Diversity III

"Well-represented" can be instantiated in many ways, among which:

- Naive l-Diversity : at least l distinct values appear ;
- Entropy l-Diversity: the entropy of the set of SD in each equivalence class should be at least $\log l$;
- Recursive (c, l)-Diversity: if the most frequent SD in a class is not much more frequent than the other SD of the class
- `(Put your idea here)`-Diversity

Progress of the Talk

The Family of Partition-Based Models and Algorithms

Many followers, based on producing equivalence classes by generalizing the QID.

Gave rise to the family of partition-based approaches :

1. Remove the ID attribute(s)
2. Form groups of records (partitions) according to the values of QID and SD of the actual records
3. And finally disclose information (statistics such as min/max) at the group level.

Weaknesses

- Proposal (year n) $\rightarrow$ Attack or limit + fix (year $n+1$)
- Various severe attacks/limits exist:
  - **No composability**: intersecting the respective sets of QID and of SD of two non-disjoint k-Anonymous releases may break k-Anonymity [50]
  - **Leaks in the execution sequences** (for optimality) : execution sequence depends on data $\Rightarrow$ minimality attacks [48]
  - **Naive adversarial reasonning models** : adversarial correlections between the QID and SD values of an equivalence class ignore the other classes $\Rightarrow$ Model the correlations between QID and SD values, in all the classes, by a bayesian network with probabilistic parameters (*aka* deFinetti attacks) [28]
  - **Numerous possible types of background knowledge** : negation statements [37], distribution of SD in the dataset [35], joint distribution between QID and SD [36, 37], logical sentences [11, 38], etc.

$\Rightarrow$ Is pursuing this cycle worth ?

RIP Partition-Based Approaches ?

Today in 2017 :

- Partition-based approaches have been shown to suffer from many flaws
- Strong interest decrease from academics
- *Differential privacy* and models inspired from it take the lead (see after)
- But. . .
  "Nous sommes en 50 avant Jésus-Christ. Toute la Gaule est occupée par les Romains. . . Toute ? Non ! Car un village peuplé d'irréductibles Gaulois résiste encore et toujours à l'envahisseur."

Progress of the Talk

Progress of the Talk

Introduction

- In parallel, an alternative research track is being followed
- Slightly different context: answer interactively to agregate queries (release statistics)

Uninformative Paradigm: "Wrong View"

- Uninformation : the opposite goal of data publishing ![6]
- The comparison between prior/posterior beliefs is hazardous:
  - Hard to know what the adversary knows or will know
    $\Rightarrow$ Random guesses.
  - Dalenius' desiderata is utopic : any learning can lead to a high knowledge gain, even if the *background knowledge is useless* without the DB, and even if the victim(s) *does not participate in the release*.
    Ex : Local DB: salaries (secret), objective: release average, auxiliary knowledge: "Bob's salary is 10% less than the DB average.".

[6] For example, learning that "Beer + Donuts = Diaper"
http://www.florent-masseglia.info/biere-et-couches-un-exemple-mythique-du-data-mining/

Differential Privacy Paradigm

- Global trends are not private and must be learnt
- Privacy is about each individual value, i.e., **each individual contribution** to the global trend

Paradigm#2: **Differential Privacy Paradigm**

A function `f` satisfies differential privacy iif: the possible impact of any individual on its result (its possible outputs) is limited.

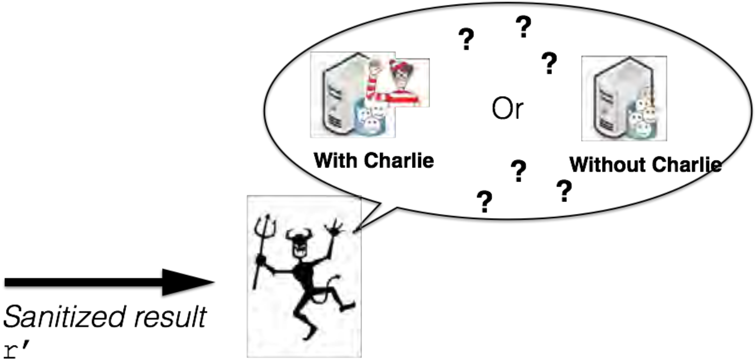
Differential Privacy Paradigm

- Global trends are not private and must be learnt
- Privacy is about each individual value, i.e., **each individual contribution** to the global trend

Paradigm#2: **Differential Privacy Paradigm**

A function `f` satisfies differential privacy iif: the possible impact of any individual on its result (its possible outputs) is limited.

Intuitions



Intuitions

Pr [f ()] ≈ Pr [f ()]

Figure: Limited impact of any possible Charlie

Intuitions

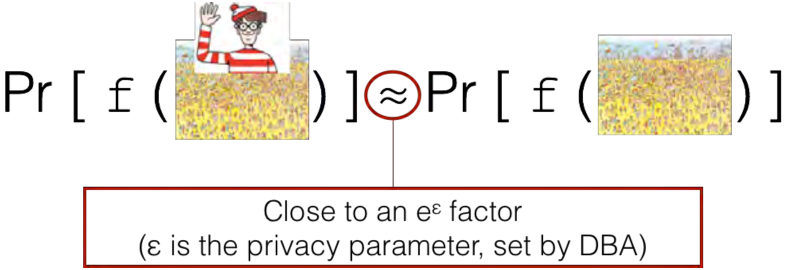


Figure: Limited impact of any possible Charlie

Initial Model

ϵ-differential privacy (from [14])

A **random function** $\mathtt{f}$ satisfies ϵ-differential privacy iff: **For all** $\mathcal{D}$ and $\mathcal{D}'$ **differing in at most one record**, and for any possible output $\mathcal{S}$ of $\mathtt{f}$, then it is true that:
$\mathtt{Pr}[\mathtt{f}(\mathcal{D}) = \mathcal{S}] \leq e^{\epsilon} \times \mathtt{Pr}[\mathtt{f}(\mathcal{D}') = \mathcal{S}]$

Initial Model

ϵ-differential privacy (from [14])

A **random function** $\texttt{f}$ satisfies ϵ-differential privacy iff: **For all** $\mathcal{D}$ and $\mathcal{D}'$ **differing in at most one record**, and for any possible output $\mathcal{S}$ of $\texttt{f}$, then it is true that:
$\texttt{Pr}[\texttt{f}(\mathcal{D}) = \mathcal{S}] \leq e^{\epsilon} \times \texttt{Pr}[\texttt{f}(\mathcal{D}') = \mathcal{S}]$

- $\texttt{f}$: here, an agregate query perturbed by adding random noise to its output
- "For all $\mathcal{D}$ and $\mathcal{D}'$": all possible datasets
- "$\mathcal{D}$ and $\mathcal{D}'$ differing in at most one record": here, $\mathcal{D}$ is $\mathcal{D}'$ with one tuple more or one tuple less (variant: one tuple with different values). Called *neighboring datasets*
- ϵ : the privacy parameter, public, common values: 0.01, 0.1, $\ln 2$, $\ln 3$
- $e^{\epsilon} \times \texttt{Pr}[\ldots]$: if one side is zero, the other must be zero too

Query Sensitivity

Different individuals, different impacts. . .

Query Sensitivity

Different individuals, different impacts...

- Presence/absence of an individual on the result of a `COUNT`: at worst +/- 1
- Presence/absence of an individual on the result of a `SUM`: $\mathtt{max}(|domain_{min}|, |domain_{max}|)$

Quantification of the worst-case impact of any possible individual on the output of the query `f`: called *query sensitivity*, and denoted $S_{\mathtt{f}}$.

Query Sensitivity

Different individuals, different impacts...

- Presence/absence of an individual on the result of a `COUNT`: at worst +/- 1
- Presence/absence of an individual on the result of a `SUM`: $\mathtt{max}(|domain_{min}|, |domain_{max}|)$

Quantification of the worst-case impact of any possible individual on the output of the query `f`: called *query sensitivity*, and denoted $S_{\mathtt{f}}$.

In general: $S_{\mathtt{f}} = \mathtt{max}_{\mathcal{D},\mathcal{D}'} ||\mathtt{f}(\mathcal{D}) - \mathtt{f}(\mathcal{D}')||_1$ where $\mathcal{D}$ and $\mathcal{D}'$ are two neighboring datasets.

Laplace Mechanism

A - "Excellent, but how to achieve differential privacy ?"
B - "Just add random noise to each query output, he said !"
A - "But from which distribution ? Uniform ? Gaussian ? Gamma ? Poisson ? . . . ? Any ?"

Laplace Mechanism

Given `f` and ϵ, adding a random variable sampled from a Laplace distribution with mean 0 and scale factor $S_{\mathtt{f}}/\epsilon$ satisfies ϵ-differential privacy [16] (easy to see).

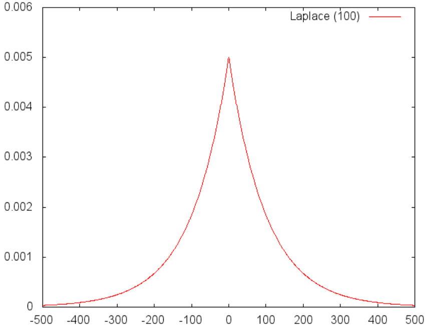


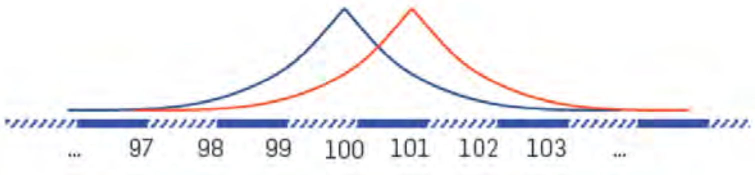
Figure: Laplace (0, 1/0.01)

Laplace Mechanism

Given $\mathtt{f}$ and ϵ, adding a random variable sampled from a Laplace distribution with mean 0 and scale factor $S_{\mathtt{f}}/\epsilon$ satisfies ϵ-differential privacy [16] (easy to see).

Assume that the COUNT when Bob participates to the dataset is $r = 101$:

- In red, distribution of perturbed outputs ($r' = r + n$) when Bob is in
- In blue, *idem* when Bob is out

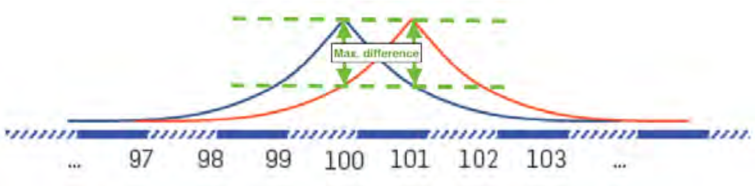


Laplace Mechanism

Given $\mathtt{f}$ and ϵ, adding a random variable sampled from a Laplace distribution with mean 0 and scale factor $S_{\mathtt{f}}/\epsilon$ satisfies ϵ-differential privacy [16] (easy to see).

Assume that the COUNT when Bob participates to the dataset is $r = 101$:

- In red, distribution of perturbed outputs ($r' = r + n$) when Bob is in
- In blue, *idem* when Bob is out



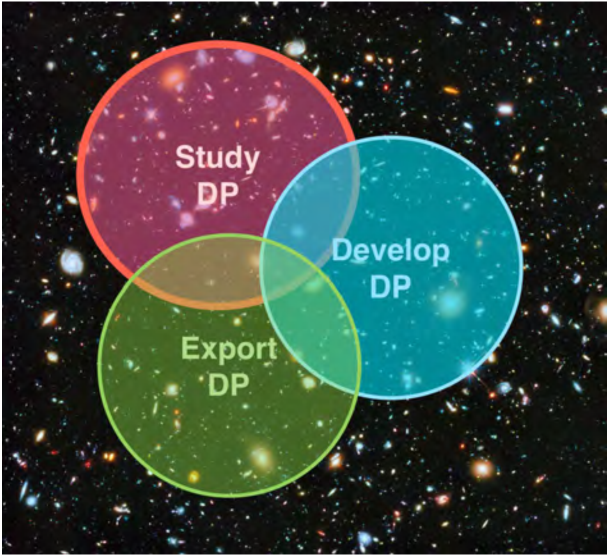
Nice Properties

- **Self-composability** : composing the outputs of two independant releases sanitized by differentially-private function(s) satisfies differential privacy :
  - Where $\epsilon_{final} = \sum \epsilon_i$ if input datasets are **not** disjoint
  - Or $\epsilon_{final} = \max \epsilon_i$ otherwise
- **No breach from post-processing** :
  - (*Laplace mechanism is independent from data*)
  - Any function applied to a differentially-private input produces a differentially-private output

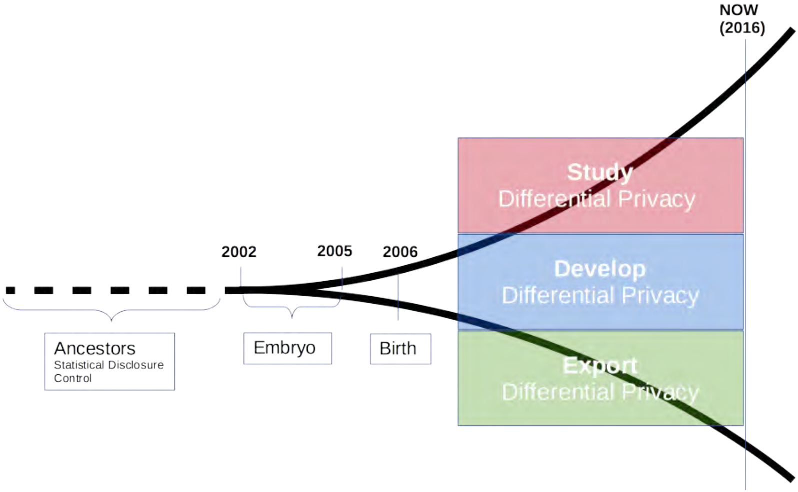
A non exact statement hides in this slide, can you find it ?

Progress of the Talk

Constellations



Constellations



Ancestors: [1].
Embryo : [8, 20].
Birth: [14, 16].

"Inventaire, à la Prévert ?"

- **Study:**
  - Assumptions (dataset and attacker) go explicit [30]
  - Relationships between models and paradigms [43, 29, 31]
  - Algorithmic hardness: *e.g.*, [19]
  - Less noise, more queries: *e.g.*, [23, 25, 49]
  - *etc.*
- **Develop:**
  - Distributed time-series: *e.g.*, [42]
  - Graphs: *e.g.*, [27, 41, 24]
  - Data cubes: *e.g.*, [13, 51]
  - Streaming data and pan-privacy: *e.g.*, [15, 17, 10, 40, 18]
  - *etc.*
- **Export:**
  - Relax secure multi-party computation algorithms: *e.g.*, [3, 9, 26, 32]
  - Use differentially private data structures for processing queries over encrypted data [*coming soon...*]
  - *etc.* ?

Disclaimer : My apologies for all the omissions !

Progress of the Talk

Relaxing Secure Multi-Party Computation Algorithms

Traditional secure multi-party computation (SMC) :

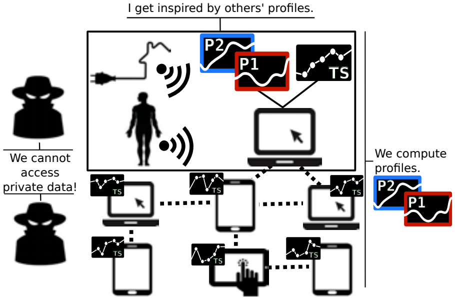
- How to compute `f` on n datasets $\mathcal{D}_1, \ldots, \mathcal{D}_n$ each stored on a distinct party such that (1) parties learn the result and (2) nothing else ?
- Solutions are usually based on complex cryptographic primitives. May be realistic when :
  1. n is small and
  2. do not connect/disconnect arbitrarily and
  3. $\mathcal{D}_i$ are small

And when the above conjunction does not hold ?
$\Rightarrow$ Relax the security model (point (2)) in order to allow the disclosure of differentially private information !

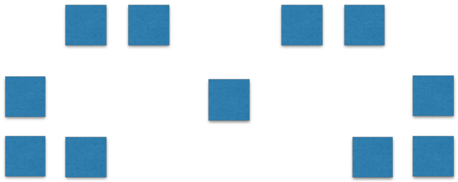
A Recent Illustration : Chiaroscuro [3, 4]

The problem :

- Compute representative **profiles** of **personal time-series distributed** in the **personal devices** of **large populations** of individuals ($\sim$ million) :
  - n is large,
  - each individual connects and disconnects arbitrarily,
  - and `f` is the k-Means algorithm



Centralized k-Means, Intuitively

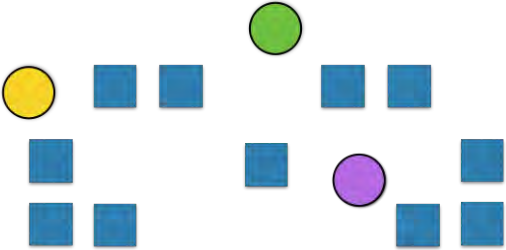


Data

Centralized k-Means, Intuitively

Choose k initial centroids at random ←

1. Assignment
2. Computation
3. Convergence

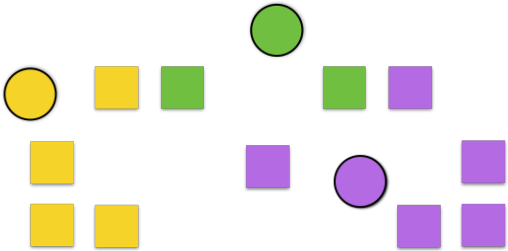


Choose k initial centroids at random

Centralized k-Means, Intuitively

Choose k initial centroids at random

1. Assignment ←
2. Computation
3. Convergence

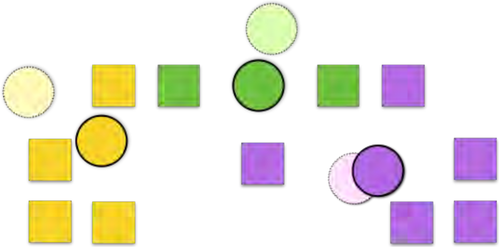


1. Assign each data point to the **closest** centroid (use, *e.g.*, euclidean distance)

Centralized k-Means, Intuitively

Choose k initial centroids at random

1. Assignment
2. **Computation** ←
3. Convergence

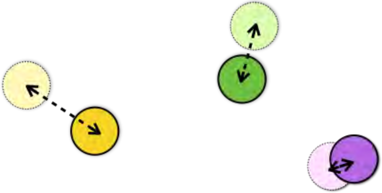


2. Compute the **barycenter** (*mean*) of each cluster. These means become *candidate centroids*.

Centralized k-Means, Intuitively

Choose k initial centroids at random

1. Assignment
2. Computation
3. **Convergence** ←

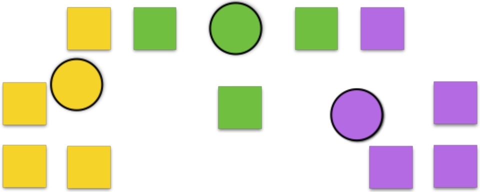


3. **Compare the distance** between the centroids and the means with a given threshold.

Centralized *k*-Means, Intuitively

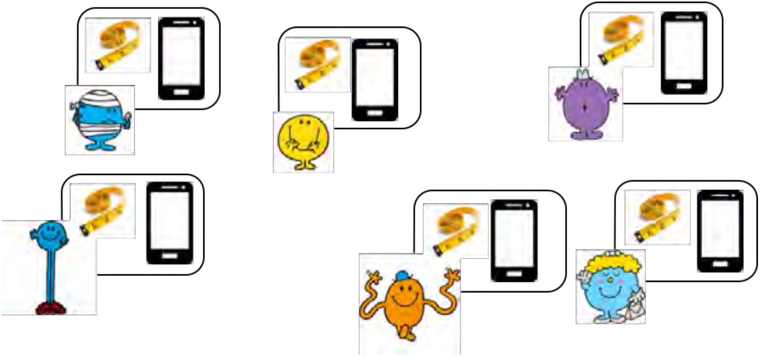
Choose k initial centroids at random

1. Assignment ←
2. Computation
3. Convergence



Etc until centroids converge.

Recall

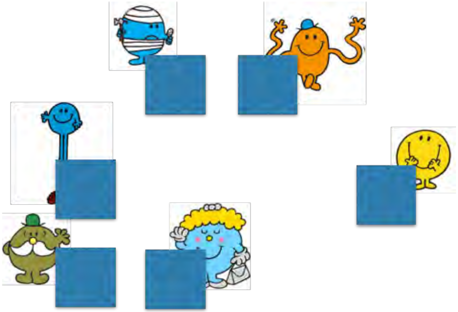


Avoid Reinventing the Wheel

Ingredients :

- **How to distribute computation ?**
  $\Rightarrow$ Adapt gossip algorithms (repeated point-to-point exchanges between participants)
- **How to preserve privacy ?**
  $\Rightarrow$ Encrypt : *additively-homomorphic* encryption and *threshold*-based decryption
  $\Rightarrow$ Perturb : *differential privacy* - a probabilistic variant - and distributed sum of *noise-shares*

k-Means with Chiaroscuro



Participants

k-Means with Chiaroscuro

Bootstrap
Get parameters (clustering, gossip, privacy) ***including initial centroids***

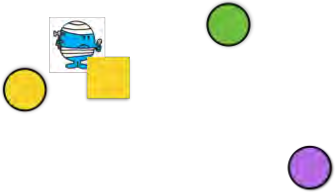
1. Assignment
2. Computation
3. Convergence

Participant #i

k-Means with Chiaroscuro

Bootstrap
Get parameters (clustering, gossip, privacy) including initial centroids

1. Assignment ←
2. Computation
3. Convergence

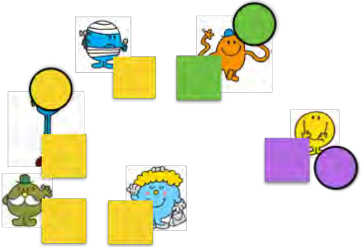


Participant #i

k-Means with Chiaroscuro

Bootstrap
Get parameters (clustering, gossip, privacy) including initial centroids

1. Assignment ←
2. Computation
3. Convergence

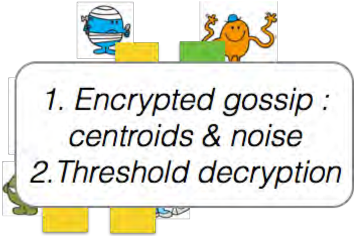


Participants

k-Means with Chiaroscuro

Bootstrap
Get parameters (clustering, gossip, privacy) including initial centroids

1. Assignment
2. **Computation** ⟵
3. Convergence

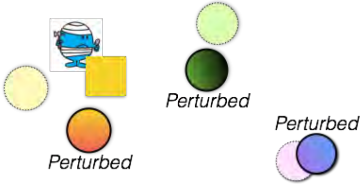


Participants

k-Means with Chiaroscuro

Bootstrap
Get parameters (clustering, gossip, privacy) including initial centroids

1. Assignment
2. **Computation** ←
3. Convergence

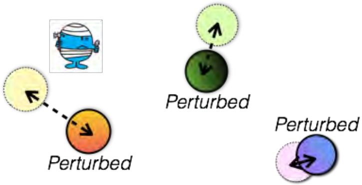


Participant #i

k-Means with Chiaroscuro

Bootstrap
Get parameters (clustering, gossip, privacy) including initial centroids

1. Assignment
2. Computation
3. **Convergence** ←
(& other termination criteria: max. number of iterations, quality monitoring)



Participant #i

Results

- Correct (similar to non-encrypted gossip computation)
- Secure against honest-but-curious participants **modulo differentially private disclosures**
- Experimental evaluations of quality (inertia of clusters) and performances (CPU cost, network cost, and latency) : affordable approach

Progress of the Talk

Conclusion

Privacy-preserving data publishing, where are we now ?

- A decade has passed and natural selection has left alive few approaches
- Severe flaws within partition-based approaches, hard to fix *a posteriori*
- In the meantime, differential privacy has born, grown, and is now expanding - *i.e.*, studied, developped, and exported

Progress of the Talk

[1] N. R. Adam and J. C. Worthmann.
Security-control methods for statistical databases: a comparative study.
ACM Comput. Surv., 21(4):515–556, Dec. 1989.

[2] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu.
Anonymizing tables.
In *Proceedings of the 10th International Conference on Database Theory*, ICDT'05, pages 246–258, Berlin, Heidelberg, 2005. Springer-Verlag.

[3] T. Allard, G. Hébrail, F. Masseglia, and E. Pacitti.
Chiaroscuro: Transparency and privacy for massive personal time-series clustering.
In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 779–794, New York, NY, USA, 2015. ACM.

[4] T. Allard, G. Hébrail, F. Masseglia, and E. Pacitti.

A new privacy-preserving solution for clustering massively distributed personal times-series.
In *Proceedings of the 32nd International Conference on Data Engineering*, ICDE '16, 2016.

[5] M. Arrington.
AOL Proudly Releases Massive Amounts of Private Data.
TechCrunch, 6th of August 2006.

[6] M. Barbaro and T. J. Zeller.
A Face Is Exposed for AOL Searcher No. 4417749.
The New York Times, 9th of August 2006.

[7] R. J. Bayardo and R. Agrawal.
Data privacy through optimal k-anonymization.
In *Proceedings of the 21st International Conference on Data Engineering*, ICDE '05, pages 217–228, Washington, DC, USA, 2005. IEEE Computer Society.

[8] A. Blum, C. Dwork, F. McSherry, and K. Nissim.
Practical privacy: the SuLQ framework.

In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '05, pages 128–138, New York, NY, USA, 2005. ACM.

[9] J. Cao, F. Rao, E. Bertino, and M. Kantarcioglu.
A hybrid private record linkage scheme: Separating differentially private synopses from matching records.
In *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*, pages 1011–1022, 2015.

[10] T.-H. H. Chan, E. Shi, and D. Song.
Private and Continual Release of Statistics.
ACM Trans. Inf. Syst. Secur., 14(3):26:1–26:24, Nov. 2011.

[11] B.-C. Chen, K. LeFevre, and R. Ramakrishnan.
Privacy skyline: privacy with multidimensional adversarial knowledge.

In *Proceedings of the 33rd international conference on Very large data bases*, VLDB '07, pages 770–781. VLDB Endowment, 2007.

[12] T. Dalenius.
Towards a methodology for statistical disclosure control.
Statistik Tidskrift, 15(5):429–444, 1977.

[13] B. Ding, M. Winslett, J. Han, and Z. Li.
Differentially Private Data Cubes: Optimizing Noise Sources and Consistency.
In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 217–228, New York, NY, USA, 2011. ACM.

[14] C. Dwork.
Differential privacy.
In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ICALP'06, pages 1–12, Berlin, Heidelberg, 2006. Springer-Verlag.

[15] C. Dwork.
Differential Privacy in New Settings.
In *Proceedings of the Twenty-first Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10, pages 174–183, Philadelphia, PA, USA, 2010. Society for Industrial and Applied Mathematics.

[16] C. Dwork, F. McSherry, K. Nissim, and A. Smith.
Calibrating noise to sensitivity in private data analysis.
In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag.

[17] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum.
Differential privacy under continual observation.
In *Proceedings of the 42nd ACM symposium on Theory of computing*, STOC '10, pages 715–724, New York, NY, USA, 2010. ACM.

[18] C. Dwork, M. Naor, T. Pitassi, G. N. Rothblum, and S. Yekhanin.

Pan-private streaming algorithms.
In *Innovations in Computer Science - ICS 2010, Tsinghua University, Beijing, China, January 5-7, 2010. Proceedings*, pages 66–80, 2010.

[19] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, and S. Vadhan.
On the complexity of differentially private data release: Efficient algorithms and hardness results.
In *Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing*, STOC '09, pages 381–390, New York, NY, USA, 2009. ACM.

[20] C. Dwork and K. Nissim.
Privacy-preserving datamining on vertically partitioned databases.
In *Advances in Cryptology: Proceedings of Crypto*, pages 528–544, 2004.

[21] B. C. M. Fung, K. Wang, and P. S. Yu.

Top-down specialization for information and privacy preservation.
In *Proceedings of the 21st International Conference on Data Engineering*, ICDE '05, pages 205–216, Washington, DC, USA, 2005. IEEE Computer Society.

[22] P. Golle.
Revisiting the uniqueness of simple demographics in the us population.
In *Proceedings of the 5th ACM workshop on Privacy in electronic society*, WPES '06, pages 77–80, New York, NY, USA, 2006. ACM.

[23] M. Hardt and K. Talwar.
On the geometry of differential privacy.
In *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, STOC '10, pages 705–714, New York, NY, USA, 2010. ACM.

[24] M. Hay, C. Li, G. Miklau, and D. Jensen.

Accurate estimation of the degree distribution of private networks.
In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, ICDM '09, pages 169–178, Washington, DC, USA, 2009. IEEE Computer Society.

[25] M. Hay, V. Rastogi, G. Miklau, and D. Suciu.
Boosting the accuracy of differentially private histograms through consistency.
Proc. VLDB Endow., 3(1-2):1021–1032, Sept. 2010.

[26] A. Inan, M. Kantarcioglu, G. Ghinita, and E. Bertino.
Private record matching using differential privacy.
In *Proceedings of the 13th International Conference on Extending Database Technology*, EDBT '10, pages 123–134, New York, NY, USA, 2010. ACM.

[27] V. Karwa, S. Raskhodnikova, A. Smith, and G. Yaroslavtsev.
Private analysis of graph structure.
ACM Trans. Database Syst., 39(3):22:1–22:33, Oct. 2014.

[28] D. Kifer.

Attacks on privacy and deFinetti's theorem.
In *Proceedings of the 35th SIGMOD international conference on Management of data*, SIGMOD '09, pages 127–138, New York, NY, USA, 2009. ACM.

[29] D. Kifer and B.-R. Lin.
Towards an axiomatization of statistical privacy and utility.
In *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '10, pages 147–158, New York, NY, USA, 2010. ACM.

[30] D. Kifer and A. Machanavajjhala.
No free lunch in data privacy.
In *Proceedings of the 2011 international conference on Management of data*, SIGMOD '11, pages 193–204, New York, NY, USA, 2011. ACM.

[31] D. Kifer and A. Machanavajjhala.
A rigorous and customizable framework for privacy.

In *Proceedings of the 31st symposium on Principles of Database Systems*, PODS '12, pages 77–88, New York, NY, USA, 2012. ACM.

[32] M. Kuzu, M. Kantarcioglu, A. Inan, E. Bertino, E. Durham, and B. Malin.
Efficient privacy-aware record integration.
In *Proceedings of the 16th International Conference on Extending Database Technology*, EDBT '13, pages 167–178, New York, NY, USA, 2013. ACM.

[33] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan.
Incognito: Efficient full-domain k-anonymity.
In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, pages 49–60, New York, NY, USA, 2005. ACM.

[34] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan.
Mondrian multidimensional k-anonymity.

In *Proceedings of the 22nd International Conference on Data Engineering*, ICDE '06, pages 25–, Washington, DC, USA, 2006. IEEE Computer Society.

[35] N. Li, T. Li, and S. Venkatasubramanian.
t-closeness: Privacy beyond k-anonymity and l-diversity.
In *Proceedings of the 23rd IEEE International Conference on Data Engineering*, ICDE '07, pages 106–115, april 2007.

[36] A. Machanavajjhala, J. Gehrke, and M. Götz.
Data publishing against realistic adversaries.
PVLDB, 2(1):790–801, August 2009.

[37] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam.
ℓ-diversity: Privacy beyond κ-anonymity.
In *Proceedings of the 22nd IEEE International Conference on Data Engineering*, ICDE '06, pages 24–, Washington, DC, USA, 2006. IEEE Computer Society.

[38] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern.

Worst-case background knowledge for privacy-preserving data publishing.
In *Proceedings of the 23rd IEEE International Conference on Data Engineering*, pages 126–135, 2007.

[39] A. Meyerson and R. Williams.
On the complexity of optimal k-anonymity.
In *Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '04, pages 223–228, New York, NY, USA, 2004. ACM.

[40] D. Mir, S. Muthukrishnan, A. Nikolov, and R. N. Wright.
Pan-private algorithms via statistics on sketches.
In *Proceedings of the Thirtieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '11, pages 37–48, New York, NY, USA, 2011. ACM.

[41] V. Rastogi, M. Hay, G. Miklau, and D. Suciu.

Relationship privacy: Output perturbation for queries with joins.
In *Proceedings of the Twenty-eighth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '09, pages 107–116, New York, NY, USA, 2009. ACM.

[42] V. Rastogi and S. Nath.
Differentially Private Aggregation of Distributed Time-series with Transformation and Encryption.
In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 735–746, New York, NY, USA, 2010. ACM.

[43] V. Rastogi, D. Suciu, and S. Hong.
The boundary between privacy and utility in data publishing.
In *Proceedings of the 33rd international conference on Very large data bases*, VLDB '07, pages 531–542. VLDB Endowment, 2007.

[44] P. Samarati and L. Sweeney.

Generalizing data to provide anonymity when disclosing information (abstract).
In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, PODS '98, pages 188–, New York, NY, USA, 1998. ACM.

[45] L. Sweeney.
Uniqueness of simple demographics in the u.s. population (white paper).
Carnegie Mellon University, Laboratory for International Data Privacy, 2000.

[46] L. Sweeney.
k-anonymity: a model for protecting privacy.
Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 10(5):557–570, 2002.

[47] K. Wang, P. S. Yu, and S. Chakraborty.
Bottom-up generalization: A data mining solution to privacy protection.

In *Proceedings of the Fourth IEEE International Conference on Data Mining*, ICDM '04, pages 249–256, Washington, DC, USA, 2004. IEEE Computer Society.

[48] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei.
Minimality attack in privacy preserving data publishing.
In *Proceedings of the 33rd International Conference on Very Large Data Bases*, VLDB '07, pages 543–554. VLDB Endowment, 2007.

[49] X. Xiao, G. Bender, M. Hay, and J. Gehrke.
ireduct: Differential privacy with reduced relative errors.
In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 229–240, New York, NY, USA, 2011. ACM.

[50] X. Xiao and Y. Tao.
M-invariance: Towards privacy preserving re-publication of dynamic datasets.

In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 689–700, New York, NY, USA, 2007. ACM.

[51] Y. Xiao, J. J. Gardner, and L. Xiong.
DPCube: Releasing Differentially Private Data Cubes for Health Information.
In *ICDE*, pages 1305–1308, 2012.

Appendix

Achieving k-Anonymity

- The more general a value is, the more people correspond to it : "less people in Urrugne, than in Pays Basque, than in France."
- Based on generalizing/suppressing the values of the attributes of the QID (also called *recoding*)
- Numerical attribute : from values to ranges
- Categorical attribute: need a taxonomy (*e.g.*, Urrugne > Pays Basque > France),
- Output an optimal release, i.e., one that satisfies k-Anonymity with a minimal *number of generalizations*
  $\Rightarrow$ Shown to be hard [2, 39]
  $\Rightarrow$ Many alternative strategies/simplifications/heuristics (e.g., [2, 7, 21, 33, 44, 39, 47])
- Not the focus of this talk but lets have a quick look at one of them...

Formalizing the Bayes-Optimal Model I

- Background knowledge: joint distribution between quasi-identifiers and sensitive data : $f(s, q)$.

Prior belief

Given a target QI q (the victim) and a sensivite data s :

$$\alpha(q, s) = \mathtt{Pr}_f(s|q) = \frac{f(s, q)}{\sum_{s' \in SD} f(s', q)} \tag{1}$$

Formalizing the Bayes-Optimal Model II

- Let $\mathcal{V}$ be the sanitized release
- Let $q^\star$ be the QI of the equivalence class that contains q
- Let $n(q^\star, s)$ be the number of tuples $\langle q^\star, s\rangle$ in $\mathcal{V}$;
- Let $f(s|q^\star)$ be the conditional probability that s be associated to the QIs that have been generalized to $q^\star$;

Posterior belief

Given a target QI q, a sensitive data s, and the release $\mathcal{V}$:

$$\beta(q, s, \mathcal{V}) = \mathtt{Pr}(s|q \wedge \mathcal{V}) = \frac{n(q^\star, s)\frac{f(s|q)}{f(s|q^\star)}}{\sum_{s' \in SD} n(q^\star, s')\frac{f(s'|q)}{f(s'|q^\star)}} \tag{2}$$

(proof in [37])

Formalizing the Bayes-Optimal Model III

A sanitized release $\mathcal{V}$ satisfies BAYES-OPTIMAL PRIVACY if:

$$\forall q \in QI, s \in SD, \mathtt{abs}(\alpha(q, s) - \beta(q, s, \mathcal{V})) < \tau \tag{3}$$

where abs returns the absolute value of its argument and τ is the user-defined threshold over the adversarial knowledge gain.
Note: alternative definitions exist [37].

Example I

Let the adversary's background knowledge about Don be:

$f(\langle\ q_{Don}, Cold\rangle) = 0.1$	$\alpha(q_{Don}, Cold) = ??$
$f(\langle q_{Don}, Flu\rangle) = 0.01$	$\alpha(q_{Don}, Flu) = ??$
$f(\langle q_{Don}, HIV\rangle) = 0.14$	$\alpha(q_{Don}, HIV) = ??$

What is his prior belief about Don ?

Example II

Answer:

$f(\langle\ q_{Don}, Cold\rangle) = 0.1$	$\alpha(q_{Don}, Cold) = 0.1/0.25 = 0.4$
$f(\langle q_{Don}, Flu\rangle) = 0.01$	$\alpha(q_{Don}, Flu) = 0.01/0.25 = 0.04$
$f(\langle q_{Don}, HIV\rangle) = 0.14$	$\alpha(q_{Don}, HIV) = 0.14/0.25 = 0.56$

Example III

Let the adversary's background knowledge about any individual other than Don be:

$f(\langle\ q_i, Cold\rangle) = 0.083$	$\alpha(q_i, Cold) =??$
$f(\langle q_i, Flu\rangle) = 0.083$	$\alpha(q_i, Flu) =??$
$f(\langle q_i, HIV\rangle) = 0.083$	$\alpha(q_i, HIV) =??$

What is his prior belief about any other individual ?

Example IV

Answer:

$f(\langle\ q_i, Cold\rangle) = 0.083$	$\alpha(q_i, Cold) = 0.083/0.25 = 0.33$
$f(\langle q_i, Flu\rangle) = 0.083$	$\alpha(q_i, Flu) = 0.083/0.25 = 0.33$
$f(\langle q_i, HIV\rangle) = 0.083$	$\alpha(q_i, HIV) = 0.083/0.25 = 0.33$

Example V

Let $\mathcal{V}$ be the 2-anonymous release:

Zip	Age	Dis.
[75001, 75002]	[22, 29]	Cold
[75001, 75002]	[22, 29]	Flu
[75003, 75010]	[22, 29]	Cold
[75003, 75010]	[22, 29]	HIV

Recall that $q_{Don} = \langle 75003, 22 \rangle$ and is known by the adversary.

What is his posterior belief about Don ?

Example VI

Answer:

In the above release, $q^{\star}_{Don} = \langle[75003, 75010], [22, 29]\rangle$.

Then, the adversary's posterior belief about Don is:

$$\boxed{\begin{array}{l} \beta(q_{Don}, Flu, \mathcal{V}) = \frac{0 * \frac{0.04}{0.37}}{1.18} = 0 \\ \beta(q_{Don}, Cold, \mathcal{V}) = \frac{1 * \frac{0.4}{0.73}}{1.18} = 0.46 \\ \beta(q_{Don}, HIV, \mathcal{V}) = \frac{1 * \frac{0.56}{0.89}}{1.18} = 0.54 \end{array}}$$

Example VII

As a result:

Prior	**Posterior**
$\alpha(q_{Don}, Cold) = 0.4$	$\beta(q_{Don}, Cold, \mathcal{V}) = 0.46$
$\alpha(q_{Don}, Flu) = 0.04$	$\beta(q_{Don}, Flu, \mathcal{V}) = 0$
$\alpha(q_{Don}, HIV) = 0.56$	$\beta(q_{Don}, HIV, \mathcal{V}) = 0.54$

Is there a privacy breach ?

RECURSIVE (c, l)-DIVERSITY

For each class:

- Count the occurence of each sensitive value;
- and sort them by descending order.

Let r_1 be the first count, ..., r_m be the m^{th}.

Recursive (c, l) Diversity

An equivalence class satisfying RECURSIVE (c, l)-DIVERSITY satisfies: $r_1 < c(r_l + r_{l+1} + ... + r_m)$.
A release $\mathcal{V}$ satisfies RECURSIVE (c, l)-DIVERSITY if all its equivalence classes satisfy it.

Examples

What is the protection offered by the classes having the following counts?

r_1	100
r_2	6
r_3	5
r_4	3

Examples

What is the protection offered by the classes having the following counts?

r_1	100
r_2	6
r_3	5
r_4	3

r_1	7
r_2	6
r_3	5
r_4	3

Recursive (c, l) Diversity, bis I

Assume that the counts of Don's class are as follows:

r_1	7
r_2	6
$\boxed{r_3}$	5
r_4	3
r_5	1
r_6	1

$\Rightarrow$ Satisfies Recursive $(1, 3)$-Diversity.

Recursive (c, l) Diversity, bis II

The adversary knows that Don **does not** have flu.

If the count of flu is r_2:

$$\begin{array}{|c|c|} \hline r_1 & 7 \\ r_2 & 6 \\ \boxed{r_3} & 5 \\ r_4 & 3 \\ r_5 & 1 \\ r_6 & 1 \\ \hline \end{array} \Rightarrow \begin{array}{|c|c|} \hline r_1 & 7 \\ \boxed{r_2} & 5 \\ r_3 & 3 \\ r_4 & 1 \\ r_5 & 1 \\ \hline \end{array}$$

$\Rightarrow$ Satisfies RECURSIVE $(1,\ 2)$-DIVERSITY.

Recursive (c, l) Diversity, bis III

The adversary knows that Don **does not** have flu.

If the count of flu is r_6:

$$\begin{array}{|c|c|} \hline r_1 & 7 \\ r_2 & 6 \\ \boxed{r_3} & 5 \\ r_4 & 3 \\ r_5 & 1 \\ r_6 & 1 \\ \hline \end{array} \Rightarrow \begin{array}{|c|c|} \hline r_1 & 7 \\ r_2 & 6 \\ \boxed{r_3} & 5 \\ r_4 & 3 \\ r_5 & 1 \\ \hline \end{array}$$

$\Rightarrow$ Satisfies Recursive $(1, 3)$-Diversity.

Recursive (c, l) Diversity, bis IV

$\textsc{Recursive}$ (c, l)-$\textsc{Diversity}$ + 1 negation statement $\rightarrow$ What is the protection level at worst?

Limits of differential privacy

Even differential privacy has its limits ;)

But they are hard to grasp (underlying assumptions are most often only implicit). Actually, we have assumptions [30]:

- **About the dataset**.
  - *"Differential privacy works without any assumption about the dataset."* : Wrong
  - $\Rightarrow$ All tuples are considered independant !
- **About the attacker**.
  - *"Differential privacy works against arbitrary background knowledge."*: Wrong
  - $\Rightarrow$ Differential privacy does not compose with the deterministic release of marginal counts

Private Record Matching [26]

Context:

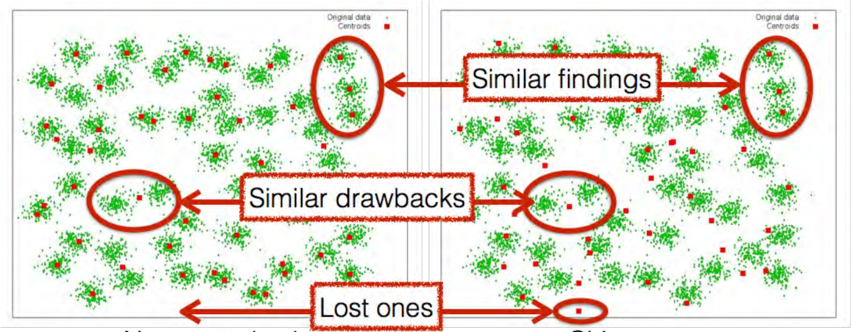
- Two mutually distrustful entities hold a DB
- They want to match their records (*i.e.*, join "close" records together)
- So that the non-matching records of each entity remain hidden to the other

Proposal :

- Overcome the efficiency limits of the Secure Multiparty Computation protocols (SMC)
- By disclosing differentially private information (relaxing the security definition):
  - Partition the records into regions (eg, age in [45, 50[)
  - Publish differentially private stats of each partition in order to identify those for which some records may match (eg, partitions [35, 48[and [45, 50[)
  - Match by a SMC the regions that have not been filtered out

Chiaroscuro and 2D Points

On a set of $750K$ 2D random points[7] distributed in 50 clusters :



Non-perturbed
(6th iteration)

Chiaroscuro
(6th iteration, GREEDY strategy, no smoothing)

[7] From : I. Kärkkäinen and P. Fränti, "Dynamic local search algorithm for the clustering problem", Research Report A-2002-6, available at `https://cs.joensuu.fi/sipu/datasets/`