Privacy-Preserving Data Publishing
Where are we now?

MDD Summer School

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

June 9, 2016
Introduction

For \( l \)-Diversity and \( \epsilon \)-Differential Privacy, two seminal privacy models!
Progress of the Talk

Non-Informative Paradigm: Partition-Based Models and Algorithms

Differential Privacy Paradigm: Models, Algorithms, and Novelties

Conclusion

References
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
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\(^1\): \(\sim\)100M US, \(\sim\)100M EU, \(\sim\)100M Asia
- ADSL is spreading (against 56K modems)
- RAM: 64MB at \(\sim\)70\(^{\text{\$}}\)\(^2\)
- HD: 40GB at \(\sim\)250\(^{\text{\$}}\)\(^2\)
- 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 |
From the archives II
From the archives III
From the archives IV

Enter your search terms...

Google Search  I'm Feeling Lucky
...or browse web pages by category.

Feeling lucky? Test your search skills with the Google Quiz

Search the Web on your Wireless Phone or PDA

©2000 Google Inc.  About | Search Tips | Google Buttons | Add Google to Your Site | Jobs@Google
(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: A join is enough: “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].
Consider that individuals’ data is made of:

- Identifying attributes, or **ID**: identify uniquely each individual (e.g., ⟨SSN⟩);
- Quasi-Identifying attributes, or **QID**: may identify uniquely some individuals (e.g., ⟨Zip, DoB⟩);
- Sensitive attributes, or **SD**: sensitive data, e.g., ⟨Disease⟩;
$k$-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.
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**

<table>
<thead>
<tr>
<th>Name</th>
<th>Zip</th>
<th>Age</th>
<th>Dis.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>75001</td>
<td>22</td>
<td>Cold</td>
</tr>
<tr>
<td>Bill</td>
<td>75002</td>
<td>29</td>
<td>Flu</td>
</tr>
<tr>
<td>Don</td>
<td>75003</td>
<td>22</td>
<td>Cold</td>
</tr>
<tr>
<td>Sue</td>
<td>75010</td>
<td>28</td>
<td>HIV</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Dis.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[75001, 75002]</td>
<td>[22, 29]</td>
<td>Cold</td>
</tr>
<tr>
<td>[75001, 75002]</td>
<td>[22, 29]</td>
<td>Flu</td>
</tr>
<tr>
<td>[75003, 75010]</td>
<td>[22, 29]</td>
<td>Cold</td>
</tr>
<tr>
<td>[75003, 75010]</td>
<td>[22, 29]</td>
<td>HIV</td>
</tr>
</tbody>
</table>

**Table:** A possible 2-Anonymous Release of the raw data.
**k-Anonymity: the Model IV**

<table>
<thead>
<tr>
<th>Name</th>
<th>Zip</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>75001</td>
<td>22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Dis.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[75001, 75002]</td>
<td>[22, 29]</td>
<td>Cold</td>
</tr>
<tr>
<td>[75001, 75002]</td>
<td>[22, 29]</td>
<td>Flu</td>
</tr>
<tr>
<td>[75003, 75010]</td>
<td>[22, 29]</td>
<td>Cold</td>
</tr>
<tr>
<td>[75003, 75010]</td>
<td>[22, 29]</td>
<td>HIV</td>
</tr>
</tbody>
</table>

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

⇒ 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
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
  ⇒ Shown to be hard [2, 39]
  ⇒ 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...
Mondrian for Example 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 for Example II

Algorithm 1: MondrianAnonymize

input : A partition $\mathcal{P}$ to split
output: A set of partitions, each containing between $k$ and $2k - 1$ tuples

1. if no allowable multidimensional cut for partition then return $\mathcal{P}$;
2. else
3. \hspace{1em} \textit{dim} $\leftarrow$ chooseDimension();
4. \hspace{1em} $fs$ $\leftarrow$ frequencySet($\mathcal{P}$, \textit{dim});
5. \hspace{1em} $splitVal$ $\leftarrow$ findMedian($fs$);
6. \hspace{1em} $\mathcal{L}$ $\leftarrow$ $\{t \in \mathcal{P} : t.\text{dim} \leq splitVal \}$;
7. \hspace{1em} $\mathcal{R}$ $\leftarrow$ $\{t \in \mathcal{P} : t.\text{dim} > splitVal \}$;
8. return MondrianAnonymize($\mathcal{L}$) $\cup$ MondrianAnonymize($\mathcal{R}$)
Mondrian for Example 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;
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

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
Waiting for the Next Scandal

During a few years:
- Academics focus on the algorithmics aspects of $k$-Anonymity
- And pseudonymity fuels another scandal...
In 2006, AOL releases a list of web search queries [5]:

- 20 million search queries
- issued by 658,000 unnamed users

<table>
<thead>
<tr>
<th>AnonID</th>
<th>Query</th>
<th>QueryTime</th>
</tr>
</thead>
<tbody>
<tr>
<td>1326</td>
<td>“holiday mansion houseboat”</td>
<td>2006-03-29</td>
</tr>
<tr>
<td>1326</td>
<td>“back to the future”</td>
<td>2006-04-01</td>
</tr>
<tr>
<td>591476</td>
<td>“english spanish translator”</td>
<td>2006-03-20</td>
</tr>
<tr>
<td>591476</td>
<td>“panama vacations”</td>
<td>2006-03-20</td>
</tr>
<tr>
<td>591476</td>
<td>“breast reduction”</td>
<td>2006-03-23</td>
</tr>
<tr>
<td>591476</td>
<td>“volunteer work at hospitals in brooklyn”</td>
<td>2006-05-24</td>
</tr>
<tr>
<td>591476</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>591476</td>
<td>“how to secretly poison your ex”</td>
<td>2006-03-12</td>
</tr>
</tbody>
</table>
Thelma Arnold’s Case II

And especially:

<table>
<thead>
<tr>
<th>AnonID</th>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>4417749</td>
<td>people with last name “Arnold”</td>
</tr>
<tr>
<td>4417749</td>
<td>“landscapers in Lilburn, Ga”</td>
</tr>
<tr>
<td>4417749</td>
<td>“60 single men”</td>
</tr>
<tr>
<td>4417749</td>
<td>“dog that urinates on everything”</td>
</tr>
<tr>
<td>4417749</td>
<td>dog-related queries</td>
</tr>
</tbody>
</table>

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

![Thelma Arnold with a dog]
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 \textit{k-Anonymity}

<table>
<thead>
<tr>
<th>Name</th>
<th>Zip</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>75001</td>
<td>22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Dis.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[75001, 75002]</td>
<td>[22, 29]</td>
<td>Cold</td>
</tr>
<tr>
<td>[75001, 75002]</td>
<td>[22, 29]</td>
<td>Flu</td>
</tr>
<tr>
<td>[75003, 75010]</td>
<td>[22, 29]</td>
<td>Cold</td>
</tr>
<tr>
<td>[75003, 75010]</td>
<td>[22, 29]</td>
<td>HIV</td>
</tr>
</tbody>
</table>

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
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 \( \text{distance}(\text{posterior belief}, \text{prior belief}) > \theta \) (too much gain in knowledge)
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 !
/-DIVERSITY:

/-DIVERSITY: a simple and easy-to-check condition for protecting against SD homogeneity and adversarial negation statements.
An $l$-diverse equivalence class contains at least $l$ well-represented sensitive values.
“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

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
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 2016 :

- 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 . . .

Progress of the Talk

Non-Informative Paradigm: Partition-Based Models and Algorithms

Differential Privacy Paradigm: Models, Algorithms, and Novelties

Conclusion

References
Progress of the Talk

Non-Informative Paradigm: Partition-Based Models and Algorithms

Differential Privacy Paradigm: Models, Algorithms, and Novelties
   Basics
   An Expanding Universe
   Focus on Export

Conclusion

References
Introduction

- In parallel, an alternative research track is being followed
- Slightly different context: answer interactively to aggregate queries (release statistics)
Uninformative Paradigm: “Wrong View”

- Uninformation: the opposite goal of data publishing!
- The comparison between prior/posterior beliefs is hazardous:
  - Hard to know what the adversary knows or will know ⇒ 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”
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

Sanitized result $r'$
Intuitions

\[ \Pr \left( f \left( \begin{array}{c} \text{\textbullet} \\ \text{\textbullet} \end{array} \right) \right) \approx \Pr \left( f \left( \begin{array}{c} \text{\textbullet} \\ \text{\textbullet} \end{array} \right) \right) \]

**Figure:** Limited impact of any possible Charlie
Intuitions

$$\Pr \left[ f \left( \cdot \right) \right] \approx \Pr \left[ f \left( \cdot \right) \right]$$

Close to an $e^\varepsilon$ factor
($\varepsilon$ is the privacy parameter, set by DBA)

**Figure:** Limited impact of any possible Charlie
Initial Model

$\epsilon$-differential privacy (from [14])

A random function $f$ satisfies $\epsilon$-differential privacy iff: For all $D$ and $D'$ differing in at most one record, and for any possible output $S$ of $f$, then it is true that:

$$\Pr[f(D) = S] \leq e^\epsilon \times \Pr[f(D') = S]$$
Initial Model

$\epsilon$-differential privacy (from [14])

A random function $f$ satisfies $\epsilon$-differential privacy iff: For all $\mathcal{D}$ and $\mathcal{D}'$ differing in at most one record, and for any possible output $S$ of $f$, then it is true that:

$$\Pr[f(\mathcal{D}) = S] \leq e^\epsilon \times \Pr[f(\mathcal{D}') = S]$$

- $\epsilon$: here, an aggregate 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
- $\epsilon$: the privacy parameter, public, common values: 0.01, 0.1, ln 2, ln 3
- $e^\epsilon \times \Pr[\ldots]$ : 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: 
  \[ \max(|domain_{min}|, |domain_{max}|) \]

Quantification of the worst-case impact of any possible individual on the output of the query \( f \): called query sensitivity, and denoted \( S_f \).
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:
  \[ \max(\text{domain}_{\text{min}}, \text{domain}_{\text{max}}) \]

Quantification of the worst-case impact of any possible individual on the output of the query \( f \): called *query sensitivity*, and denoted \( S_f \).

In general: \( S_f = \max_{\mathcal{D}, \mathcal{D}'} \| f(\mathcal{D}) - f(\mathcal{D}') \|_1 \) 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 !”
Laplace Mechanism

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

Figure: Laplace (0, 1/0.01)
Laplace Mechanism

Given $f$ and $\epsilon$, adding a random variable sampled from a Laplace distribution with mean 0 and scale factor $S_f/\epsilon$ satisfies $\epsilon$-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 $f$ and $\epsilon$, adding a random variable sampled from a Laplace distribution with mean 0 and scale factor $S_f/\epsilon$ satisfies $\epsilon$-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 independent releases sanitized by differentially-private function(s) satisfies differential privacy:
  - Where $\epsilon_{\text{final}} = \sum \epsilon_i$ if input datasets are not disjoint
  - Or $\epsilon_{\text{final}} = \max \epsilon_i$ 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
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
  ► ⇒ All tuples are considered independent!

► About the attacker.
  ► “Differential privacy works against arbitrary background knowledge.”: Wrong
  ► ⇒ Differential privacy does not compose with the deterministic release of marginal counts
Progress of the Talk

Non-Informative Paradigm: Partition-Based Models and Algorithms

Differential Privacy Paradigm: Models, Algorithms, and Novelties
  Basics
  An Expanding Universe
  Focus on Export

Conclusion

References
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

Non-Informative Paradigm: Partition-Based Models and Algorithms

Differential Privacy Paradigm: Models, Algorithms, and Novelties
  Basics
  An Expanding Universe
  Focus on Export

Conclusion

References
Traditional secure multi-party computation (SMC):

- How to compute $f$ on $n$ datasets $D_1, \ldots, D_n$ 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. $D_i$ 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 (~ million):
  - *n* is large,
  - each individual connects and disconnects arbitrarily,
  - and *f* is the *k*-Means algorithm

![Diagram showing data collection and processing](image-url)
Centralized $k$-Means, Intuitively
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?**
  ⇒ Adapt gossip algorithms (repeated point-to-point exchanges between participants)

- **How to preserve privacy?**
  ⇒ Encrypt: *additively-homomorphic* encryption and *threshold*-based decryption
  ⇒ 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

1. Encrypted gossip: centroids & noise
2. Threshold decryption
$k$-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

Non-Informative Paradigm: Partition-Based Models and Algorithms

Differential Privacy Paradigm: Models, Algorithms, and Novelties

Conclusion

References
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 \textit{a posteriori}
- In the meantime, differential privacy has born, grown, and is now expanding - \textit{i.e.}, studied, developed, and exported
Progress of the Talk

Non-Informative Paradigm: Partition-Based Models and Algorithms

Differential Privacy Paradigm: Models, Algorithms, and Novelties

Conclusion

References


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.

AOL Proudly Releases Massive Amounts of Private Data.
TechCrunch, 6th of August 2006.

A Face Is Exposed for AOL Searcher No. 4417749.

Data privacy through optimal k-anonymization.

Practical privacy: the SuLQ framework.


Pan-private streaming algorithms.


On the complexity of differentially private data release: Efficient algorithms and hardness results.

Privacy-preserving datamining on vertically partitioned databases.

Top-down specialization for information and privacy preservation.


Revisiting the uniqueness of simple demographics in the US population.

On the geometry of differential privacy.

Accurate estimation of the degree distribution of private networks.


Attacks on privacy and deFinetti’s theorem.

Towards an axiomatization of statistical privacy and utility.

No free lunch in data privacy.

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


[35] N. Li, T. Li, and S. Venkatasubramanian. 
$t$-closeness: Privacy beyond $k$-anonymity and $l$-diversity. 

Data publishing against realistic adversaries. 

$l$-diversity: Privacy beyond $\kappa$-anonymity. 

[38] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern.
Worst-case background knowledge for privacy-preserving data publishing.


On the complexity of optimal k-anonymity.

Pan-private algorithms via statistics on sketches.

Relationship privacy: Output perturbation for queries with joins.


Differentially Private Aggregation of Distributed Time-series with Transformation and Encryption.


The boundary between privacy and utility in data publishing.


Generalizing data to provide anonymity when disclosing information (abstract).


Uniqueness of simple demographics in the u.s. population (white paper).

k-anonymity: a model for protecting privacy.

Bottom-up generalization: A data mining solution to privacy protection.


DPCube: Releasing Differentially Private Data Cubes for Health Information.
Appendix
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 sensitive data $s$:

$$\alpha(q, s) = \Pr_f(s|q) = \frac{f(s, q)}{\sum_{s' \in SD} f(s', q)}$$  \hspace{1cm} (1)
Formalizing the Bayes-Optimal Model II

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

Posterior belief

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

$$
\beta(q, s, \mathcal{V}) = \Pr(s|q \land \mathcal{V}) = \frac{n(q^*, s) \frac{f(s|q)}{f(s|q^*)}}{\sum_{s' \in SD} n(q^*, s') \frac{f(s'|q)}{f(s'|q^*)}}
$$

(2)
(proof in [37])
A sanitized release $\mathcal{V}$ satisfies **Bayes-Optimal Privacy** if:

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

where \text{abs} returns the absolute value of its argument and $\tau$ is the user-defined threshold over the adversarial knowledge gain. Note: alternative definitions exist [37].
Example 1

Let the adversary’s background knowledge about Don be:

<table>
<thead>
<tr>
<th></th>
<th>f(⟨q_{Don}, Cold⟩) = 0.1</th>
<th>α(q_{Don}, Cold) =??</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>f(⟨q_{Don}, Flu⟩) = 0.01</td>
<td>α(q_{Don}, Flu) =??</td>
</tr>
<tr>
<td></td>
<td>f(⟨q_{Don}, HIV⟩) = 0.14</td>
<td>α(q_{Don}, HIV) =??</td>
</tr>
</tbody>
</table>

What is his prior belief about Don?
Example II

Answer:

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

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

<table>
<thead>
<tr>
<th>Condition</th>
<th>Value</th>
<th>Prior Belief</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(⟨ q_i, Cold ⟩)$</td>
<td>0.083</td>
<td>$\alpha(q_i, Cold) = ??$</td>
</tr>
<tr>
<td>$f(⟨ q_i, Flu ⟩)$</td>
<td>0.083</td>
<td>$\alpha(q_i, Flu) = ??$</td>
</tr>
<tr>
<td>$f(⟨ q_i, HIV ⟩)$</td>
<td>0.083</td>
<td>$\alpha(q_i, HIV) = ??$</td>
</tr>
</tbody>
</table>

What is his prior belief about any other individual?
Example IV

Answer:

| $f(\langle q_i, \text{Cold} \rangle)$ | 0.083 |
| $f(\langle q_i, \text{Flu} \rangle)$ | 0.083 |
| $f(\langle q_i, \text{HIV} \rangle)$ | 0.083 |
| $\alpha(q_i, \text{Cold})$ | $0.083/0.25 = 0.33$ |
| $\alpha(q_i, \text{Flu})$ | $0.083/0.25 = 0.33$ |
| $\alpha(q_i, \text{HIV})$ | $0.083/0.25 = 0.33$ |
Example V

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

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Dis.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[75001, 75002]</td>
<td>[22, 29]</td>
<td>Cold</td>
</tr>
<tr>
<td>[75001, 75002]</td>
<td>[22, 29]</td>
<td>Flu</td>
</tr>
<tr>
<td>[75003, 75010]</td>
<td>[22, 29]</td>
<td>Cold</td>
</tr>
<tr>
<td>[75003, 75010]</td>
<td>[22, 29]</td>
<td>HIV</td>
</tr>
</tbody>
</table>

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, \( q^*_\text{Don} = \langle [75003, 75010], [22, 29] \rangle \).

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

\[
\begin{align*}
\beta(q_{\text{Don}}, \text{Flu}, \mathcal{V}) &= \frac{0 \times \frac{0.04}{0.37}}{1.18} = 0 \\
\beta(q_{\text{Don}}, \text{Cold}, \mathcal{V}) &= \frac{1 \times \frac{0.4}{0.73}}{1.18} = 0.46 \\
\beta(q_{\text{Don}}, \text{HIV}, \mathcal{V}) &= \frac{1 \times \frac{0.56}{0.89}}{1.18} = 0.54
\end{align*}
\]
Example VII

As a result:

<table>
<thead>
<tr>
<th>Prior</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha(q_{Don}, Cold) = 0.4$</td>
<td>$\beta(q_{Don}, Cold, \mathcal{V}) = 0.46$</td>
</tr>
<tr>
<td>$\alpha(q_{Don}, Flu) = 0.04$</td>
<td>$\beta(q_{Don}, Flu, \mathcal{V}) = 0$</td>
</tr>
<tr>
<td>$\alpha(q_{Don}, HIV) = 0.56$</td>
<td>$\beta(q_{Don}, HIV, \mathcal{V}) = 0.54$</td>
</tr>
</tbody>
</table>

Is there a privacy breach?
**Recursive \((c, l)\)-Diversity**

For each class:

- Count the occurrence 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?

<table>
<thead>
<tr>
<th>r₁</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>r₂</td>
<td>6</td>
</tr>
<tr>
<td>r₃</td>
<td>5</td>
</tr>
<tr>
<td>r₄</td>
<td>3</td>
</tr>
</tbody>
</table>
Examples

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

\[
\begin{array}{c|c}
 r_1 & 100 \\
 r_2 & 6 \\
 r_3 & 5 \\
 r_4 & 3 \\
\end{array}
\quad
\begin{array}{c|c}
 r_1 & 7 \\
 r_2 & 6 \\
 r_3 & 5 \\
 r_4 & 3 \\
\end{array}
\]
Assume that the counts of Don’s class are as follows:

| \( r_1 \) | 7 |
| \( r_2 \) | 6 |
| \( r_3 \) | 5 |
| \( r_4 \) | 3 |
| \( r_5 \) | 1 |
| \( r_6 \) | 1 |

⇒ Satisfies **Recursive (1, 3)-Diversity**.
The adversary knows that Don does not have flu.

If the count of flu is $r_2$:

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

$\Rightarrow$ 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 \(r_6\):

\[
\begin{array}{c|c}
  r_1 & 7 \\
r_2 & 6 \\
r_3 & 5 \\
r_4 & 3 \\
r_5 & 1 \\
r_6 & 1 \\
\end{array}
\quad \Rightarrow 
\begin{array}{c|c}
  r_1 & 7 \\
r_2 & 6 \\
r_3 & 5 \\
r_4 & 3 \\
r_5 & 1 \\
\end{array}
\]

⇒ Satisfies Recursive \((1, 3)\)-Diversity.
Recursive \((c, l)\) Diversity, bis IV

**Recursive \((c, l)\)-Diversity** + 1 negation statement \(\rightarrow\) What is the protection level at worst?
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 (e.g., age in [45, 50])
  - Publish differentially private stats of each partition in order to identify those for which some records may match (e.g., 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\textsuperscript{7} distributed in 50 clusters: