Anonymization, privacy

Privacy-Preserving Data Publishing

Pierre Senellart

Mostly derived from material by from Tristan Allard, see
`http://www.ens-rennes.fr/medias/fichier/seminaire-ens-ppdp-2017-01-final_1485338538859-pdf`

18 February 2022

Plan

Plan

Data Publishing

- Extraordinary amount and quality of data captured and owned by companies, research institutions, government, individuals
- Huge number of applications could benefit of data sharing:
  - Scientific research
  - Social studies
  - Market research
- Potentially financial incentive: the data can be sold
- Some of these datasets have nothing to do about individual human beings (e.g., data on particle detection in a particle accelerator), but many contain personal information

Personal and Sensitive Data

- A dataset related to persons typically contains:

  Personally identifying information: Information that can be used (more or less easily) to identify a specific person: full names, email addresses, social security numbers, IP addresses, precise address, etc.

  Sensitive data: Data that is not publicly attached to a person, that this person may not want to be released (medical data, shopping history, and other consumer habits, personal communications, etc.)

  Other data: neither sensitive, not directly identifying (but see later)

- Depending on the context, some information (e.g., telephone number) may be personally identifying, sensitive, or neither
- Clearly, one does not want to release sensitive data attached to personally identifying information

First Idea: Pseudonymization

- Replace all information that is potentially personally identifying with pseudonyms (e.g., encrypted information, sequentially or randomly generated identifiers)
- Release the resulting dataset
- Usually, the institution that produces the dataset may want to keep the mapping between pseudonyms and personal information

The Problem with Pseudonymization

Seemingly benign, non personally identifying information, may end up being usable to reidentify pseudonyms, especially when combined.

Plan

Governor Weld's Case (1/3)

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;

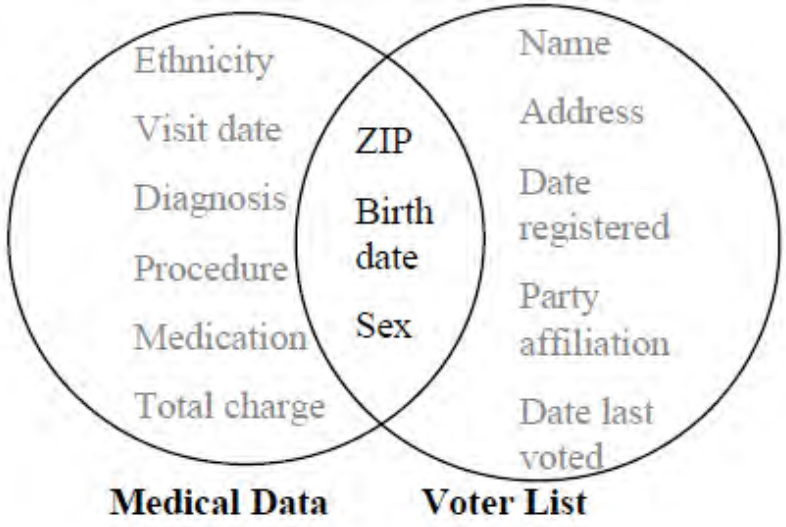


Figure: `Medical JOIN Voter ON (zip, DoB, sex)`

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;

Governor Weld's Case (3/3)

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].

AOL Query Set (1/2)

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

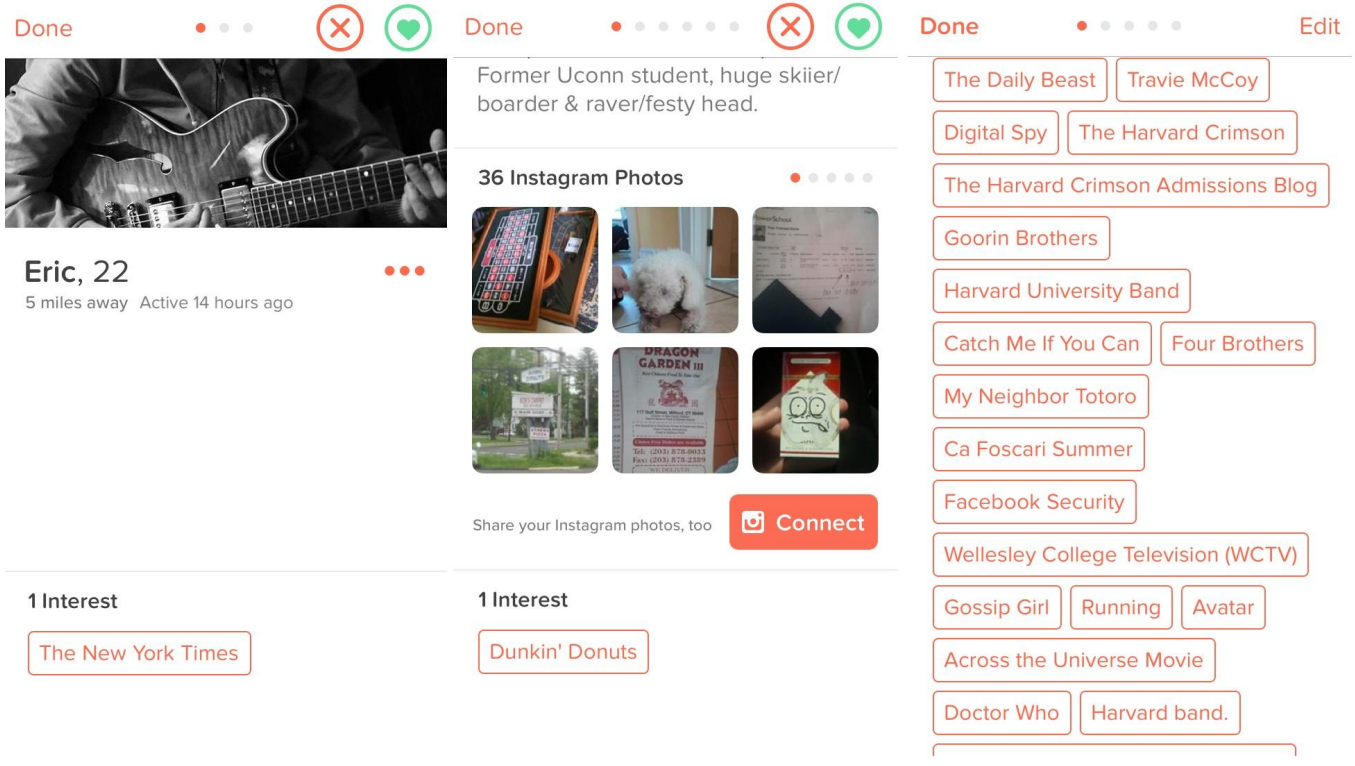
AOL Query Set (2/2)

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.

Tinder and Facebook



Tinder interests coming from Facebook. What could go wrong with this?

Plan

General Data Protection Regulation

- EU Regulation 2016/679
- Regulates the use of personal data in the EU
- Applies to every processing of personal data of an EU resident, even by companies not resident in the EU
- In France, completes the historical law on "Informatique et liberté" from 1978, which was revised to implement the GDPR in 2019; data protection office: CNIL

Article 4

For the purposes of this Regulation:

- *'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;*
- *(...)*
- *'pseudonymisation' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;*
- *(...)*

Recital 26

The principles of data protection should apply to any information concerning an identified or identifiable natural person.
Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person.
To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly.
To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.
The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.

Article 6

Processing shall be lawful only if and to the extent that at least one of the following applies:

- *the data subject has given consent to the processing of his or her personal data for one or more specific purposes;*
- *processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;*
- *processing is necessary for compliance with a legal obligation to which the controller is subject;*
- *processing is necessary in order to protect the vital interests of the data subject or of another natural person;*
- *processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;*
- *processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data (...)*

Plan

Model (1/2)

- 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 $\rangle$;

Model (2/2)

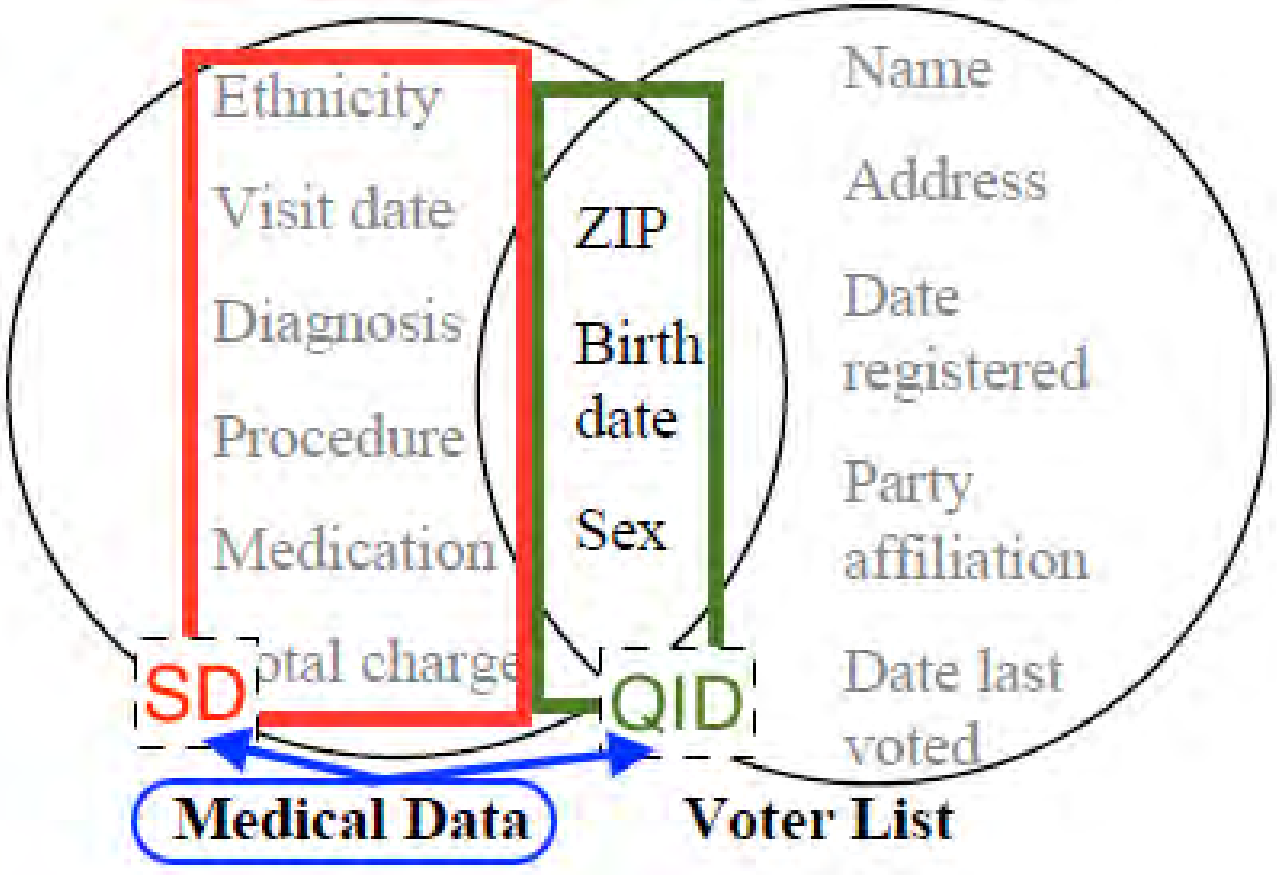


Figure: Quasi-identifiers and sensitive data in Gov. Weld's case

Plan

k-Anonymity Model (1/5)

Warning

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

k-Anonymity Model (2/5)

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 Model (3/5)

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 Model (4/5)

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 Model (5/5)

Vocabulary

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

Achieving k-Anonymity

- Generalizing attribute values:
  - Range for numerical attributes
  - Supercategory for categorical attributes
  - "Wildcard value"
- Deleting tuples
- Adding new tuples
- Changing attribute values
- Every such modification affects the utility of the dataset; also, truthfulness of modifications in question

Optimal k-Anonymization

Simple model: only modification allowed is replacing QID values with wildcards.

Theorem ([39])

Achieving k-anonymity with minimal wildcard replacements is NP-hard for $k \geq 3$.

Proof.

By reduction from k-dimensional perfect matching. □

Mondrian (1/6)

- **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 (2/6)

Algorithm 1: `MondrianAnonymize`

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

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

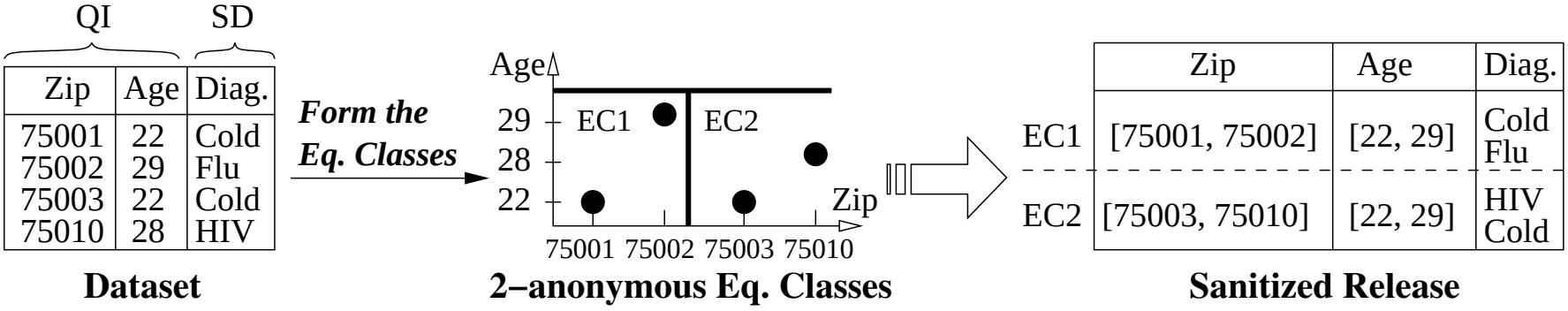
```
1 if no allowable multidimensional cut for partition then return P ;
2 else
3     dim ← chooseDimension();
4     fs ← frequencySet(P, dim);
5     splitVal ← findMedian(fs);
6     L ← {t ∈ P : t.dim ≤ splitVal };
7     R ← {t ∈ P : t.dim > splitVal };
8     return MondrianAnonymize(L) ∪ MondrianAnonymize(R)
```

Mondrian (3/6)

`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 (4/6)



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

Mondrian (5/6)

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

Mondrian (6/6)

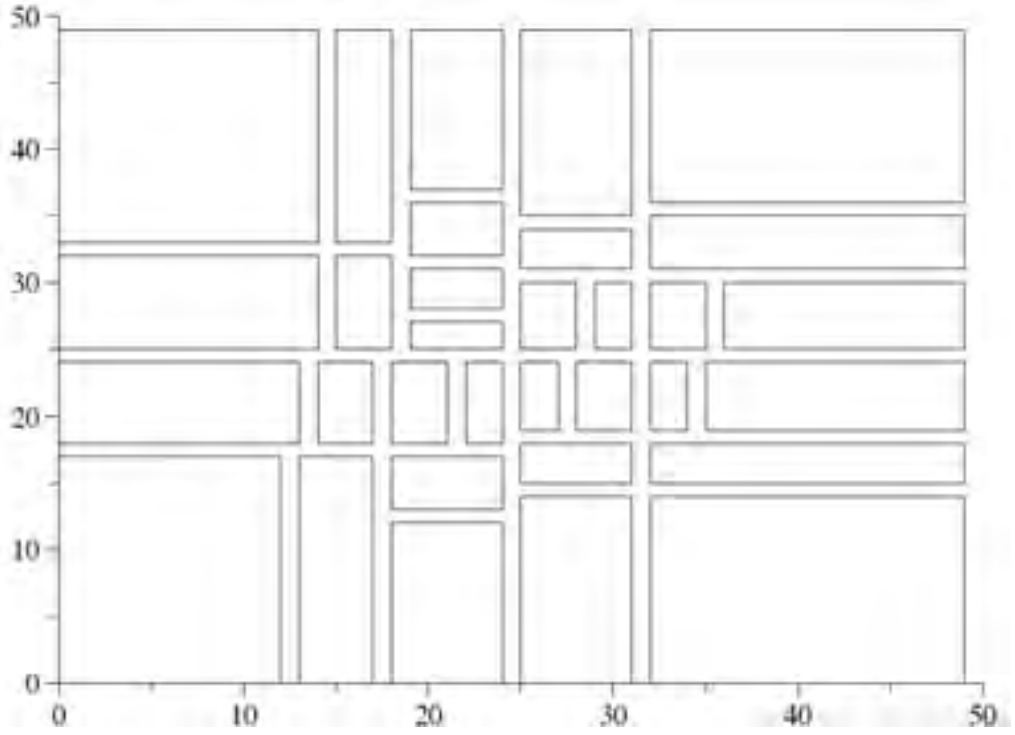


Figure: Example of a Mondrian partitioning [34] (synthetic data, 1000 tuples, k=25, normal distribution).

Privacy Paradigm

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?

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

Plan

Intuition (1/2)

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)

Intuition (2/2)

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]

Formalization (1/3)

- 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 :

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

Formalization (2/3)

- 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 $\mathcal{V}$:

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

(proof in [37])

Formalization (3/3)

A sanitized release $\mathcal{V}$ satisfies Bayes-Optimal Privacy if:

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

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 (1/6)

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 (2/6)

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 (3/6)

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 (4/6)

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 (5/6)

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 (6/6)

Answer:

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

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

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

Impractical Model

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 !

Plan

l-Diversity (1/3)

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

l-Diversity (2/3)

l-Diversity

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

l-Diversity (3/3)

"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

Beyond *l*-Diversity: Partition-Based Models

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 of Partition-Based Models

- 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.

Plan

In Brief

- (Moral and legal) necessity to not disclose any sensitive data when publishing datasets involving personal information
- Pseudonymization is not good enough; de-identification possible
- Much research about how to fix this, starting with k-anonymity
- But except maybe in very specific circumstances (perfect understanding of the background knowledge), the entire approach seems flawed (endless cycle of attacks and fixes, lack of composability, etc.)
- Giving up on privacy-preserving data publishing, and moving to privacy-preserving querying of datasets

[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.