Introduction to Privacy Models

104369: Privacitat de Dades i Seguretat

G. Navarro-Arribas

UAB -- DEIC

v. 20261014

©G. Navarro-Arribas – (2026)

Privacy Models

k-Anonymity

Differential Privacy

©G. Navarro-Arribas – (2026)

Privacy Models

k-Anonymity

Problems in k-anonymity

Differential Privacy

©G. Navarro-Arribas – (2026)

Privacy Models i

$\rightarrow$ A computational definition for privacy.

- Is this dataset private?
- How private?
- Is this procedure private enough?
- ...

Privacy Models ii

Examples:

- **Reidentification privacy**. Avoid finding a record in a database.
- **k-Anonymity**. A record indistinguishable with $k-1$ other records.
- **Secure multiparty computation**. Several parties want to compute a function of their databases, but only sharing the result.
- **Differential privacy**. The output of a query to a database should not depend (much) on whether a record is in the database or not.
- **Result privacy**. We want to avoid some results when an algorithm is applied to a database.

©G. Navarro-Arribas – (2026)

Privacy Models iii

- **Integral privacy**. Inference on the databases. E.g., changes have been applied to a database.
- **Homomorphic encryption**. We want to avoid access to raw data and partial computations.

©G. Navarro-Arribas – (2026)

Privacy Models

k-Anonymity

Problems in k-anonymity

Differential Privacy

k-Anonymity

(Samarati, Sweeney: 1998, 2001, 2002)

Definition

Let $RT(A_1, \ldots, A_n)$ be a table, and QI_{RT} be the quasi-identifier associated with it. RT is said to satisfy k-anonymity if and only if each sequence of values in $RT[QI_{RT}]$ appears with at least k occurrences in $RT[QI_{RT}]$.

Example:
k-anonymous table for $k = 2$, when $QI_{RT} = \{City, Age\}$.

City	Age	Illnes
Barcelona	30	Cancer
Barcelona	30	AIDS
Tarragona	60	AIDS
Tarragona	60	Cancer
Tarragona	60	Heart attack

©G. Navarro-Arribas – (2026)

Example: k-anonymity by generalization

City	Age	Salary
Guayaquil	23	600
Quito	48	500
Loja	20	734
Guayaquil	32	543
Cuenca	38	890
Quito	41	678
Portoviejo	21	399
Quevedo	44	943
Machala	31	632

©G. Navarro-Arribas – (2026)

Example: k-anonymity by generalization

City	Age	Salary
Guayaquil	23	600
Quito	48	500
Loja	20	734
Guayaquil	32	543
Cuenca	38	890
Quito	41	678
Portoviejo	21	399
Quevedo	44	943
Machala	31	632

$\longrightarrow$

$QI = \{City, Age\}$

k-anonymity for $k = 3$

$\longrightarrow$

Example: k-anonymity by generalization

City	Age	Salary
Guayaquil	23	600
Quito	48	500
Loja	20	734
Guayaquil	32	543
Cuenca	38	890
Quito	41	678
Portoviejo	21	399
Quevedo	44	943
Machala	31	632

$\longrightarrow$

$QI = \{City, Age\}$

k-anonymity for $k = 3$

$\longrightarrow$

City	Age	Salary
Ecuador	[20, 29]	600
Ecuador	[40, 49]	500
Ecuador	[20, 29]	734
Ecuador	[30, 39]	543
Ecuador	[30, 39]	890
Ecuador	[40, 49]	678
Ecuador	[20, 29]	399
Ecuador	[40, 49]	943
Ecuador	[30, 39]	632

k-Anonymity definitions

Definition 2

A table satisfies k-anonymity when it is partitioned into sets of at least k indistinguishable records.

- Discussion
  - k-Anonymity is not a protection procedure in itself but **a condition to be satisfied** by the protected data set.
  - Goal: avoid disclosure (identity disclosure / attribute disclosure).
  - Different levels of k lead to different protections.
  - k-anonymity through generalization and suppression: NP-Hard problem!

©G. Navarro-Arribas – (2026)

Privacy Models

k-Anonymity

Problems in k-anonymity

Differential Privacy

Attacks: homogeneity attack

City	Age	Illnes
Barcelona	30	Cancer
Barcelona	30	Cancer
Tarragona	60	AIDS
Tarragona	60	AIDS

©G. Navarro-Arribas – (2026)

Attacks: homogeneity attack

City	Age	Illnes
Barcelona	30	Cancer
Barcelona	30	Cancer
Tarragona	60	AIDS
Tarragona	60	AIDS

Homogeneity attack

When all indistinguishable records in a cluster (anonymity group) are also indistinguishable with respect to a confidential variable.
$\rightarrow$ attribute disclosure

Attacks: External knowledge attack

City	Age	Illnes
Barcelona	30	Cancer
Barcelona	30	Cancer
Tarragona	60	AIDS
Tarragona	60	AIDS
Tarragona	60	Heart attack

What if the attacker knows that person *X* from Tarragona had a *heart attack*.

©G. Navarro-Arribas – (2026)

Attacks: External knowledge attack

City	Age	Illnes
Barcelona	30	Cancer
Barcelona	30	Cancer
Tarragona	60	AIDS
Tarragona	60	AIDS
Tarragona	60	Heart attack

What if the attacker knows that person *X* from Tarragona had a *heart attack*.

External knowledge attack

Some information about a record is used to deduce information of the same or another record.

©G. Navarro-Arribas – (2026)

Beyond k-Anonymity: p-Sensitive k-anonymity

p-Sensitive k-anonymity (Truta, Vinay, 2006)

A data set is said to satisfy p-sensitive k-anonymity for $k > 1$ and $p \leq k$ if it satisfies k-anonymity and, for each group of records with the same combination of values for quasi-identifiers, the number of distinct values for each confidential value is at least p (within the same group)

l-Diversity

l-diversity (Machanavajjhala et al. 2006)

It forces l different categories in each set. However, in this case, categories should have to be well-represented.

- Different interpretation of *well-represented*:
  - Distinct values (p-sensitive).
  - Entropy $> \log(l)$.
  - ...

t-Closeness (Li, Li, Venkatasubramanian, 2007)

t-Closeness

The distribution of the attribute in any k-anonymous subset of the database is similar to the one of the full database.

- Similarity is defined in terms of the distance between the two distributions and such distance should be below a given threshold t.
- Low threshold makes the utility of the data doubtful: large information loss

©G. Navarro-Arribas – (2026)

Privacy Models

k-Anonymity

Problems in k-anonymity

Differential Privacy

Differential Privacy

Interactive vs. non-interactive scenario

We know the query (function) to apply to the database: f.
Motivation of Differential Privacy:

- the result of a query should not depend on the presence (or absence) of a particular individual,
- the impact of any individual in the output of the query is limited.

See real-world deployments according to Wikipedia.

Differential Privacy Definition

Differential Privacy

A function K_q for a query q gives ε-differential privacy if for all datasets D_1 and D_2 differing in at most one element, and all $S \subseteq Range(K_q)$,

$$\frac{Pr[K_q(D_1) \in S]}{Pr[K_q(D_2) \in S]} \leq e^{\varepsilon} \tag{1}$$

DP Example 1

name	*age*	*salary*
Ataúlfo	43	100
Sigerico	15	300
Walia	50	200
Teodoredo	51	100
Turismundo	16	500
Tedorico II	40	300
Eurico	14	700
Alarico II	49	300
Gesaleico	69	500
Amalarico	17	200

Query f: "number of records with age ≥ 18" $\Longrightarrow f(D) = 6$.
Is f safe (privacy-preserving)?

DP Example 2

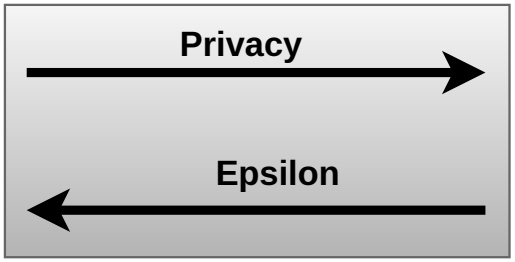
name	*age*	*salary*
Ataúlfo	43	100
Sigerico	**15**	**300**
Walia	50	200
Teodoredo	51	100
Turismundo	16	500
Tedorico II	40	300
Eurico	14	700
Alarico II	49	300
Gesaleico	69	500
Amalarico	17	200

name	*age*	*salary*
Ataúlfo	43	100
Walia	50	200
Teodoredo	51	100
Turismundo	16	500
Tedorico II	40	300
Eurico	14	700
Alarico II	49	300
Gesaleico	69	500
Amalarico	17	200

What about f now?

©G. Navarro-Arribas – (2026)

Meaning of epsilon



- e^{ε}: bound ratio of probabilities.
- $\varepsilon = 0 \rightarrow$ total privacy, but ...

A very simplistic example

Supose we want to prevent the previous problem by defining:

$$K_f(D) = f(D) + r$$

where $r \sim \mathcal{U}(-2, 2)$ (r is taken from a uniform distribution over the interval $[-2, 2]$). So e.g:

- If $K_f(D_1) = 7$, $K_f(D_2) = 5 \Rightarrow f(D_1) = ?, f(D_2) = ?$.

ε as a privacy budget

Each query *expends from* the privacy budget ε.

Example

Supose $K_f(D_1) = 5$, we repeat the query and now $K_f(D_1) = 7, \ldots$

- Can the level of privacy be affected by the number of queries?
- This is formalized as the **composition theorems** in DP.
- Noise is commonly obtained from a Laplace distribution.

DP composition theorems

Given some randomized queries $Kf_1, \ldots, Kf_n$, each one providing ε-differential privacy for $\varepsilon_1, \ldots, \varepsilon_n$:

Sequential composition

$Kf_1, \ldots, Kf_n$ with $\varepsilon_1, \ldots, \varepsilon_n$ applied to D provide $\varepsilon = \sum_{i=1}^{n} \varepsilon_i$.

Parallel composition

$Kf_1, \ldots, Kf_n$ with $\varepsilon_1, \ldots, \varepsilon_n$ applied to disjoint $D1, \ldots, D_n$ provide $\varepsilon = \max_{i=1}^{n} \varepsilon_i$.

References i

[1] Vicenç Torra.
Guide to Data Privacy: Models, Technologies, Solutions.
Undergraduate Topics in Computer Science. Springer International Publishing, Cham, 2022.