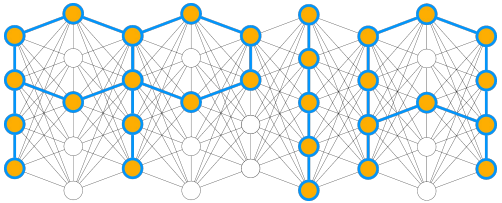


k-ANONYMITY, ℓ-DIVERSITY, t-CLOSENESS

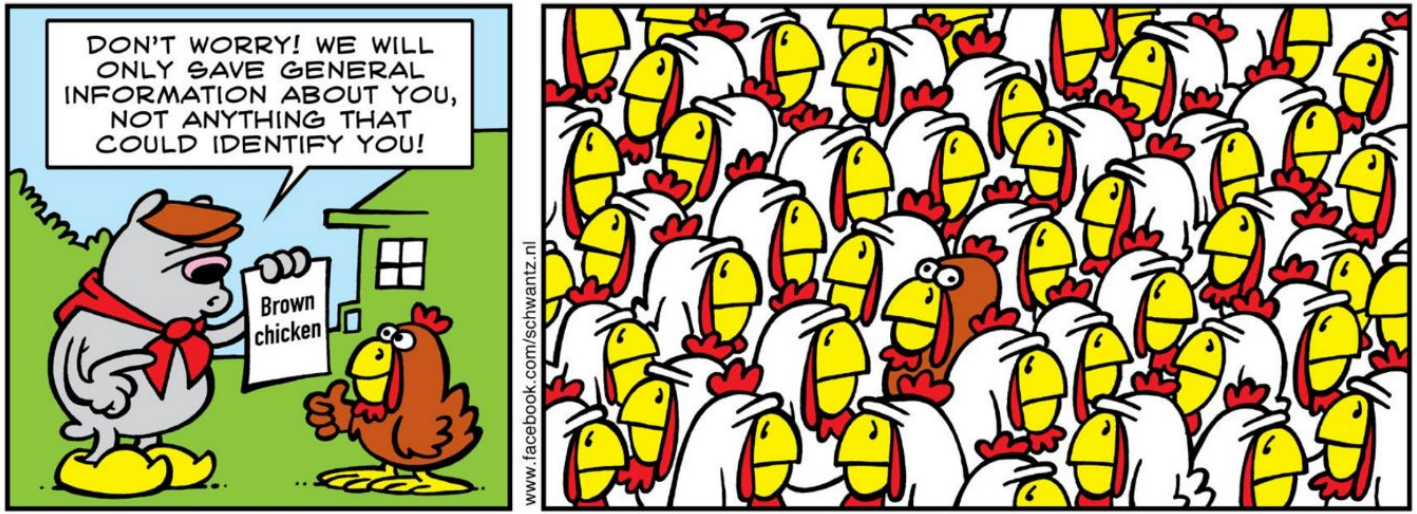
PRIVACY PRESERVING INFORMATION ACCESS

PhD in Information Engineering
A.Y. 2025/2026

GUGLIELMO FAGGIOLI
Intelligent Interactive Information Access (IIIA) Hub
Department of Information Engineering
University of Padua

k-ANONYMITY

IS DATA MINIMIZATION ENOUGH?



IDENTITY LINKING ATTACK

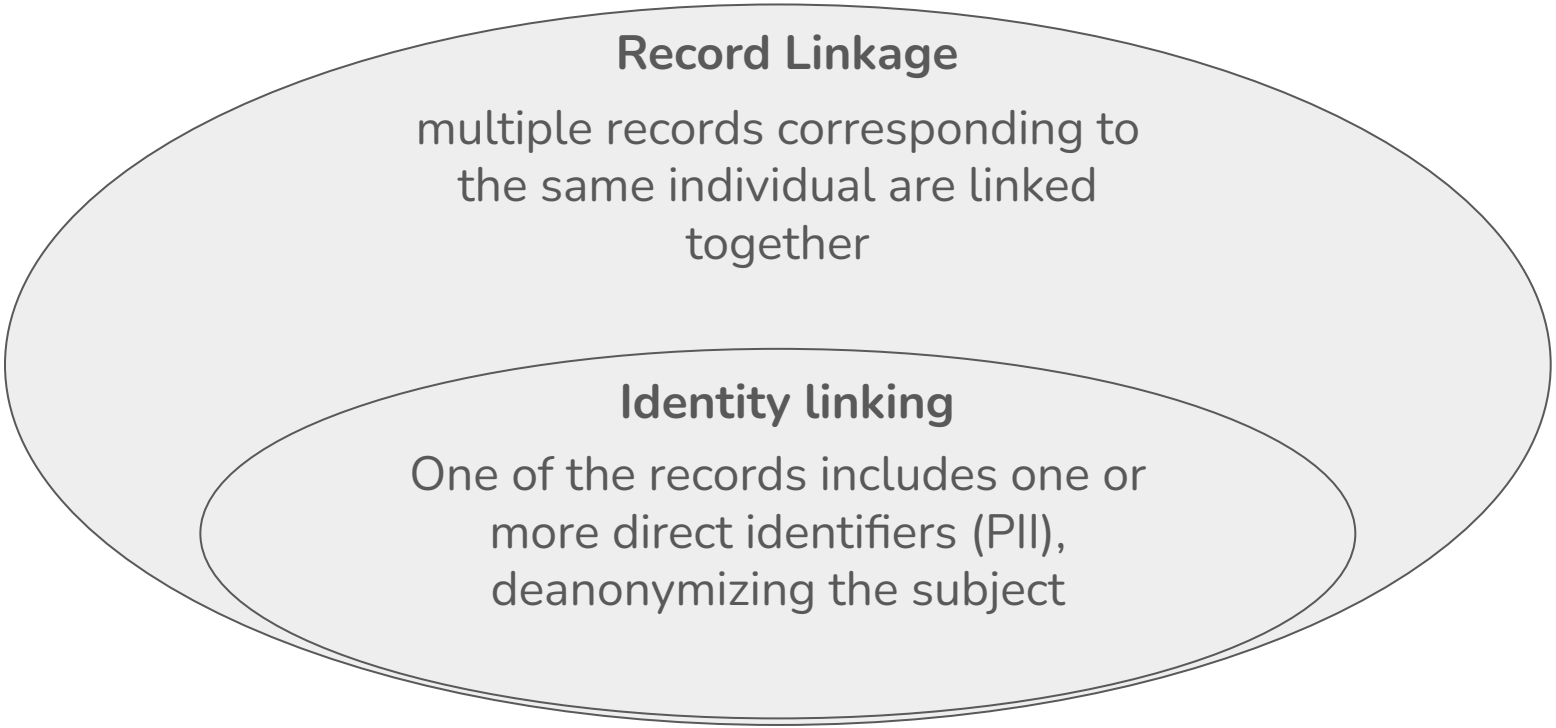
De-identification and sanitization (removing identifiers) do not grant anonymity.

Problems arise due to:

- large amount of information: publicly accessible data sources, social networks
- high computational power to link data

These factors enable the **identity linking attack**: multiple data sources are fused to aggregate information about the same person and de-anonymize data.

RECORD LINKAGE VS IDENTITY LINKING



IDENTITY LINKING ATTACK: AN EXAMPLE

SSN	Name	Race	DoB	Sex	ZIP	Marital Status	Income
1230044954330	Yu Zhong	asian	64/04/12	F	35138	divorced	33.000
1230082394331	Tao Jiang	asian	64/09/13	F	35142	divorced	54.000
2934944954322	Tang Hanying	asian	64/04/15	F	35148	married	22.000
1230004449530	Djimon Igwe	black	63/03/15	M	35138	married	11.000
3823893498549	Ashton Katy	black	63/02/18	M	35138	married	178.000
7938458593247	Alyssa Bryce	black	64/09/27	F	35137	single	23.000
9584935832839	Jean Harmon	white	64/09/27	F	35137	single	23.000
7852438634549	Caitlyn Em	white	64/09/27	F	35141	single	23.000
3582347528778	Louise Wayland	white	64/09/27	F	35141	widow	56.000

IDENTITY LINKING ATTACK: AN EXAMPLE

SSN	Name	Race	DoB	Sex	ZIP	Marital Status	Income
[redacted]	[redacted]	asian	64/04/12	F	35138	divorced	33.000
[redacted]	[redacted]	asian	64/09/13	F	35142	divorced	54.000
[redacted]	[redacted]	asian	64/04/15	F	35148	married	22.000
[redacted]	[redacted]	black	63/03/15	M	35138	married	11.000
[redacted]	[redacted]	black	63/02/18	M	35138	married	178.000
[redacted]	[redacted]	black	64/09/27	F	35137	single	23.000
[redacted]	[redacted]	white	64/09/27	F	35137	single	23.000
[redacted]	[redacted]	white	64/09/27	F	35141	single	23.000
[redacted]	[redacted]	white	64/09/27	F	35141	widow	56.000

What if we knew Louise Wayland, who was born on 64/09/27, lives in Padova and is a widow?

IDENTITY LINKING ATTACK: AN EXAMPLE

SSN	Name	Race	DoB	Sex	ZIP	Marital Status	Income
[redacted]	[redacted]	asian	64/04/12	F	35138	divorced	33.000
[redacted]	[redacted]	asian	64/09/13	F	35142	divorced	54.000
[redacted]	[redacted]	asian	64/04/15	F	35148	married	22.000
[redacted]	[redacted]	black	63/03/15	M	35138	married	11.000
[redacted]	[redacted]	black	63/02/18	M	35138	married	178.000
[redacted]	[redacted]	black	64/09/27	F	35137	single	23.000
[redacted]	[redacted]	white	64/09/27	F	35137	single	23.000
[redacted]	[redacted]	white	64/09/27	F	35141	single	23.000
[redacted]	[redacted]	white	64/09/27	F	35141	widow	56.000

What if we knew Louise Wayland, who was born on 64/09/27, lives in Padova and is a widow?

QUASI-IDENTIFIERS

quasi-identifiers are set of attributes included in the private table, which can be also externally available and therefore exploitable for linking.

In our example, Race, DoB, Sex, ZIP and Marital Status are quasi-identifiers: they do not allow to recognize univocally the person, but knowing them enables the **identity linking attack**.

k-ANONYMITY REQUIREMENT

k-anonymity requirement: Each release of data must be such that **every combination of values of quasi-identifiers** can be indistinctly matched to **at least k respondents**.

k-ANONYMITY

k-anonymity: Let $T(A_1, \ldots, A_m)$ be a table, and QI be set of quasi-identifiers associated with it. T is said to satisfy k-anonymity with respect to QI iff each sequence of values in T[QI] appears at least with k occurrences in T[QI].

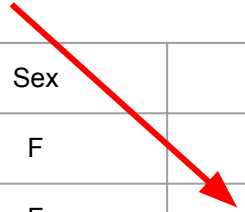
k-ANONYMITY

In other words, in your k-anonymized data, at least k tuples should be identical for what concerns the quasi-identifiers.

The set of tuples having the same quasi-identifiers is called **equivalence class** or **q-block**.

k-ANONYMITY IN OUR EXAMPLE

if we want to implement k-anonymity with k=2, there must not be less then 2 tuples with the same quasi-identifiers combination



Race	DoB	Sex	ZIP	Marital Status
asian	64/04/12	F	35138	divorced
asian	64/09/13	F	35142	divorced
asian	64/04/15	F	35148	married
black	63/03/15	M	35138	married
black	63/02/18	M	35138	married
black	64/09/27	F	35137	single
white	64/09/27	F	35137	single
white	64/09/27	F	35141	single
white	64/09/27	F	35141	widow

k-ANONYMITY IN OUR EXAMPLE

if we want to implement k-anonymity with k=2, there must not be less then 2 tuples with the same quasi-identifiers combination

Race	DoB	Sex	ZIP	Marital Status
asian	64/04/12	F	351**	divorced
asian	64/09/13	F	351**	divorced
asian	64/04/15	F	351**	married
black	63/03/15	M	351**	married
black	63/02/18	M	351**	married
black	64/09/27	F	351**	single
white	64/09/27	F	351**	single
white	64/09/27	F	351**	single
white	64/09/27	F	351**	widow

k-ANONYMITY IN OUR EXAMPLE

if we want to implement k-anonymity with k=2, there must not be less then 2 tuples with the same quasi-identifiers combination

Race	DoB	Sex	ZIP	Marital Status
asian	64/04/12	F	351**	divorced
asian	64/09/13	F	351**	divorced
asian	64/04/15	F	351**	married
black	63/03/15	M	351**	married
black	63/02/18	M	351**	married
black	64/09/27	F	351**	single
white	64/09/27	F	351**	single
white	64/09/27	F	351**	single
white	64/09/27	F	351**	widow

k-ANONYMITY IN OUR EXAMPLE

if we want to implement k-anonymity with k=2, there must not be less then 2 tuples with the same quasi-identifiers combination

Race	DoB	Sex	ZIP	Marital Status
asian	64/04/12	F	351**	been_married
asian	64/09/13	F	351**	been_married
asian	64/04/15	F	351**	been_married
black	63/03/15	M	351**	been_married
black	63/02/18	M	351**	been_married
black	64/09/27	F	351**	single
white	64/09/27	F	351**	single
white	64/09/27	F	351**	single
white	64/09/27	F	351**	been_married

k-ANONYMITY IN OUR EXAMPLE

if we want to implement k-anonymity with k=2, there must not be less then 2 tuples with the same quasi-identifiers combination

Race	DoB	Sex	ZIP	Marital Status
person	64/04/12	F	351**	been_married
person	64/09/13	F	351**	been_married
person	64/04/15	F	351**	been_married
person	63/03/15	M	351**	been_married
person	63/02/18	M	351**	been_married
person	64/09/27	F	351**	single
person	64/09/27	F	351**	single
person	64/09/27	F	351**	single
person	64/09/27	F	351**	been_married

k-ANONYMITY IN OUR EXAMPLE

if we want to implement k-anonymity with k=2, there must not be less then 2 tuples with the same quasi-identifiers combination

Race	DoB	Sex	ZIP	Marital Status
person	64/04/**	F	351**	been_married
person	64/09/**	F	351**	been_married
person	64/04/**	F	351**	been_married
person	63/03/**	M	351**	been_married
person	63/02/**	M	351**	been_married
person	64/09/**	F	351**	single
person	64/09/**	F	351**	single
person	64/09/**	F	351**	single
person	64/09/**	F	351**	been_married

k-ANONYMITY IN OUR EXAMPLE

if we want to implement k-anonymity with k=2, there must not be less then 2 tuples with the same quasi-identifiers combination

Race	DoB	Sex	ZIP	Marital Status
person	64/04/**	F	351**	been_married
person	64/09/**	F	351**	been_married
person	64/04/**	F	351**	been_married
person	63/03/**	M	351**	been_married
person	63/02/**	M	351**	been_married
person	64/09/**	F	351**	single
person	64/09/**	F	351**	single
person	64/09/**	F	351**	single
person	64/09/**	F	351**	been_married

k-ANONYMITY IN OUR EXAMPLE

if we want to implement k-anonymity with k=2, there must not be less then 2 tuples with the same quasi-identifiers combination

Race	DoB	Sex	ZIP	Marital Status
person	64/**/**	F	351**	been_married
person	64/**/**	F	351**	been_married
person	64/**/**	F	351**	been_married
person	63/**/**	M	351**	been_married
person	63/**/**	M	351**	been_married
person	64/**/**	F	351**	single
person	64/**/**	F	351**	single
person	64/**/**	F	351**	single
person	64/**/**	F	351**	been_married

IMPLEMENTING k-ANONYMITY: GENERALIZATION

To achieve k-anonymity maintaining truthfulness of our data we exploit **generalization**: substitute the quasi-identifiers with a generalization of them.

The generalization relies on the concept of *domain* (the possible values for a certain attribute) and *generalized domains*.

A generalized domain contains multiple domains.

GENERALIZATION RELATIONSHIP

The generalization relationship $\leq_D$ define a partial ordering between domains such that:

Condition 1: $\forall\ D_i, D_j, D_z \in$ **Dom**: $D_i \leq_D D_j\ , D_i \leq_D D_z \Rightarrow D_j \leq_D D_z \vee D_z \leq_D D_j$

Condition 2: all maximal elements of **Dom** are singleton.

Where **Dom** is a set of domains.

GENERALIZATION RELATIONSHIP

The generalization relationship $\leq_D$ define a partial ordering between domains such that:

Condition 1: $\forall\ D_i, D_j, D_z \in \textbf{Dom}: D_i \leq_D D_j\ , D_i \leq_D D_z \Rightarrow D_j \leq_D D_z \vee D_z \leq_D D_j$

Condition 2: all maximal elements of **Dom** are singleton.

The generalization relationship is deterministic

Where **Dom** is a set of domains.

GENERALIZATION RELATIONSHIP

The generalization relationship $\leq_D$ define a partial ordering between domains such that:

Condition 1: $\forall D_i, D_j, D_z \in \textbf{Dom}: D_i \leq_D D_j, D_i \leq_D D_z \Rightarrow D_j \leq_D D_z \vee D_z \leq_D D_j$

Condition 2: all maximal elements of **Dom** are singleton.

Where Dom is a set of domains

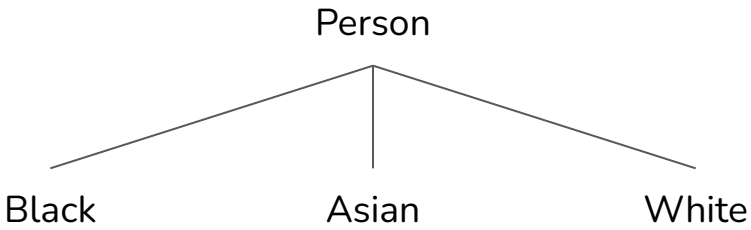
All values in each domain can always be generalized to a single value.

GENERALIZATION RELATIONSHIP

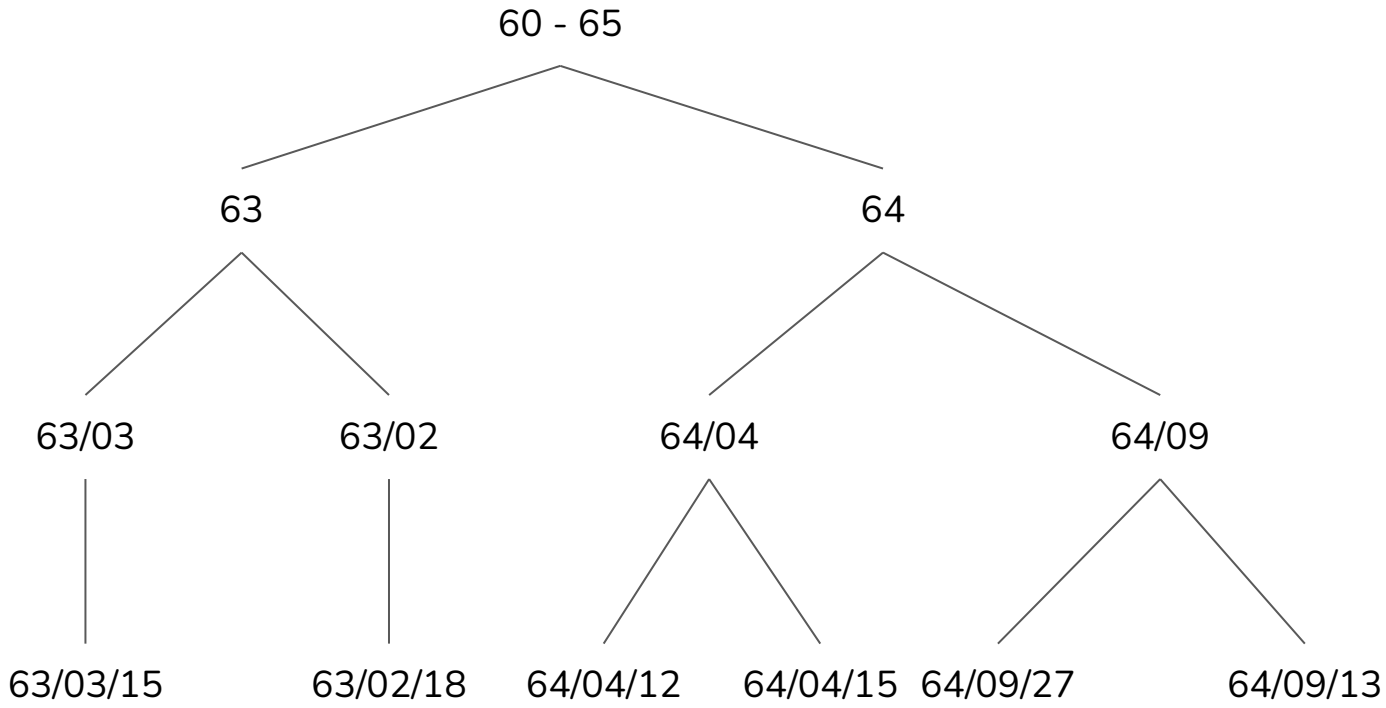
Conditions C1 and C2 allow to define a totally ordered hierarchy of domains, called **domain generalization hierarchy** (DGH_D).

A **value generalization relationship**, denoted $\leq_V$ associates with each value in domain D_i a unique value in domain D_j, direct generalization of D_i. The value generalization relationship implies the existence, for each domain D, of a **Value Generalization Hierarchy** (VGH_D).

GENERALIZATION RELATIONSHIP - RACE



GENERALIZATION RELATIONSHIP - DATE OF BIRTH



WHAT TO GENERALIZE?

Generalization can be applied at the level of:

- Attribute (AG)
- Cell (CG): generalized table may contain values at different generalization levels.

Should the sensitive attribute be generalized? Yes, if it is also a quasi-identifier, but you rapidly increase the risk of making data less useful.

k-ANONYMITY IN OUR EXAMPLE

if we want to implement k-anonymity with k=2, there must not be less then 2 tuples with the same quasi-identifiers combination

Race	DoB	Sex	ZIP	Marital Status
person	64/04/**	F	351**	been_married
person	64/09/**	F	351**	been_married
person	64/04/**	F	351**	been_married
person	63/**/**	M	351**	been_married
person	63/**/**	M	351**	been_married
person	64/09/**	F	351**	single
person	64/09/**	F	351**	single
person	64/09/**	F	351**	single
person	64/09/**	F	351**	been_married

SUPPRESSION

Outliers might make the generalization challenging: it might be necessary to use highly generalized values to allow extremely rare cases to comply with the k-anonymity requirements.

In such cases it might be more beneficial to suppress outlier information to reduce the amount of generalization required to respect the k-anonymity.

WHAT TO SUPPRESS?

Suppression can be applied at the level of:

- Tuple (TS)
- Attribute (AS)
- Cell (CS)

Should the sensitive attribute be suppressed?

k-ANONYMITY IN OUR EXAMPLE

if we want to implement k-anonymity with k=2, there must not be less then 2 tuples with the same quasi-identifiers combination

Race	DoB	Sex	ZIP	Marital Status
person	64/04/**	F	351**	been_married
person	64/09/**	F	351**	been_married
person	64/04/**	F	351**	been_married
~~person~~	~~63/03/**~~	~~M~~	~~351**~~	~~been_married~~
~~person~~	~~63/02/**~~	~~M~~	~~351**~~	~~been_married~~
person	64/09/**	F	351**	single
person	64/09/**	F	351**	single
person	64/09/**	F	351**	single
person	64/09/**	F	351**	been_married

k-ANONYMITY IN OUR EXAMPLE

if we want to implement k-anonymity with k=2, there must not be less then 2 tuples with the same quasi-identifiers combination

Race	DoB	Sex	ZIP	Marital Status
person	64/04/**	F	351**	been_married
person	64/09/**	F	351**	been_married
person	64/04/**	F	351**	been_married
person	63/03/**	M	351**	been_married
person	63/02/**	M	351**	been_married
person	64/09/**	F	351**	single
person	64/09/**	F	351**	single
person	64/09/**	F	351**	single
person	64/09/**	F	351**	been_married

GENERALIZED TABLE WITH SUPPRESSION

Generalized table - with suppression: Let T_i and T_j be two tables defined on the same set of attributes. Table T_j is said to be a generalization (with tuple suppression) of table T_i, denoted $T_i \preccurlyeq T_j$, if:

1. $|T_j| \leq |T_i|$;
2. the domain $dom(A, T_j)$ of each attribute A in T_j is equal to, or a generalization of, the domain $dom(A, T_i)$ of attribute A in T_i;
3. it is possible to define an injective function associating each tuple t_j in T_j with a tuple t_i in T_i, such that the value of each attribute in t_j is equal to, or a generalization of, the value of the corresponding attribute in t_i.

DISTANCE VECTOR

Distance vector: Let $T_i(A_1, ..., A_n)$ and $T_j(A_1, ..., A_n)$ be two tables such that $T_i \preccurlyeq T_j$.

The **distance vector** of Tj from T_i is the vector $DV_{i,j} = [d_1, ..., d_n]$, where each d_z, z = 1, . . . , n, is the length of the unique path between dom(Az, T_i) and dom(Az, T_j) in the domain generalization hierarchy DGH_{Dz}.

k-MINIMAL GENERALIZATION

k-minimal generalization - with suppression: Let T_i and T_j be two tables such that $T_i \leqslant T_j$, and let **MaxSup** be the specified threshold of acceptable suppression. T_j is said to be a k-minimal generalization of table T_i iff:

1. T_j satisfies k-anonymity enforcing minimal required suppression: T_j satisfies k-anonymity and $\forall T_z : T_i \leqslant T_z$, $DV_{i,z} = DV_{i,j}$, T_z satisfies k-anonymity $\Rightarrow |T_j| \geq |T_z|$
2. $|T_i| - |T_j| \leq$ **MaxSup**
3. $\forall T_z : T_i \leqslant T_z$ and T_z satisfies conditions 1 and 2 $\Rightarrow \neg(DV_{i,z} < DV_{i,j})$.

PREFERRED GENERALIZATION-SUPPRESSION CRITERIA

minimum absolute distance: use generalizations with the smallest total number of generalization steps;

minimum relative distance: use generalizations that minimizes the total number of relative steps (a step is made relative by dividing it over the height of the domain hierarchy to which it refers);

maximum distribution prefers the generalizations with the greatest number of distinct tuples;

minimum suppression prefers the generalizations that suppresses less tuples.

POSSIBLE APPROACHES

		Suppression			
		Tuple	Attribute	Cell	None
Generalization	Attribute	AG_TS	AG_AS	AG_CS	AG_
	Cell	CG_TS	CG_AS	CG_CS	CG_
	None	_TS	_AS	_CS	

BEST K

There isn’t a k that works in every scenario.

Obviously, 1 and n represent extreme scenarios were privacy is absent or complete.

1/k is the probability of being discovered - if you use k=2, then there is 50% chance that your data might be used in an identification attack.

5 might be already a satisfactory k, but it highly depends on your setup.

BEST K

As a rule of thumb:

Type of data	minimum k
Low sensitivity data Public Library Visitors, Survey on TV Preferences	3 to 5
Medium sensitivity data Employment Survey, Consumer Purchase Behavior	5 to 10
High sensitivity data Medical records, Financial transactions	10 to 20

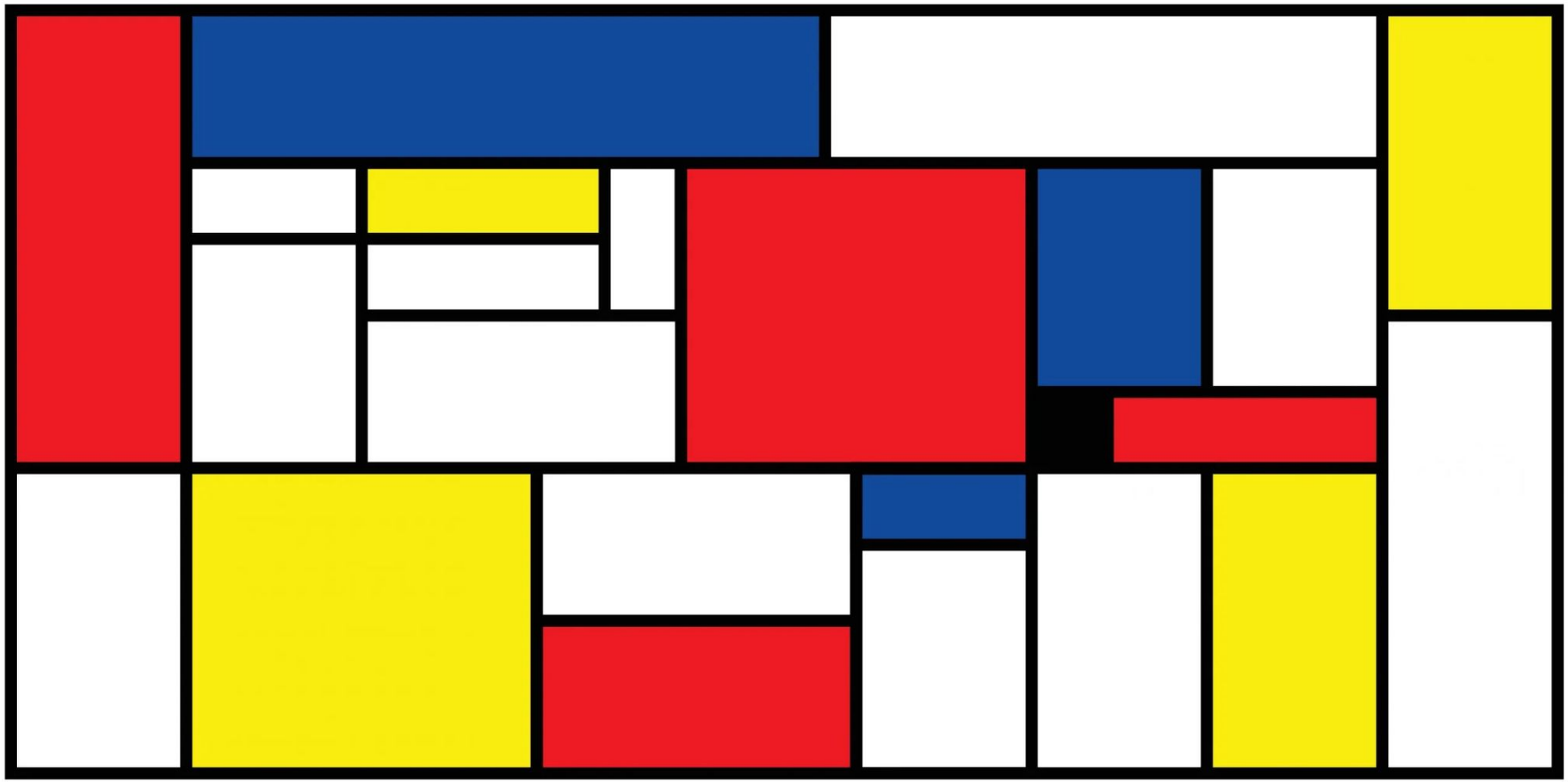
BEST K

As a rule of thumb:

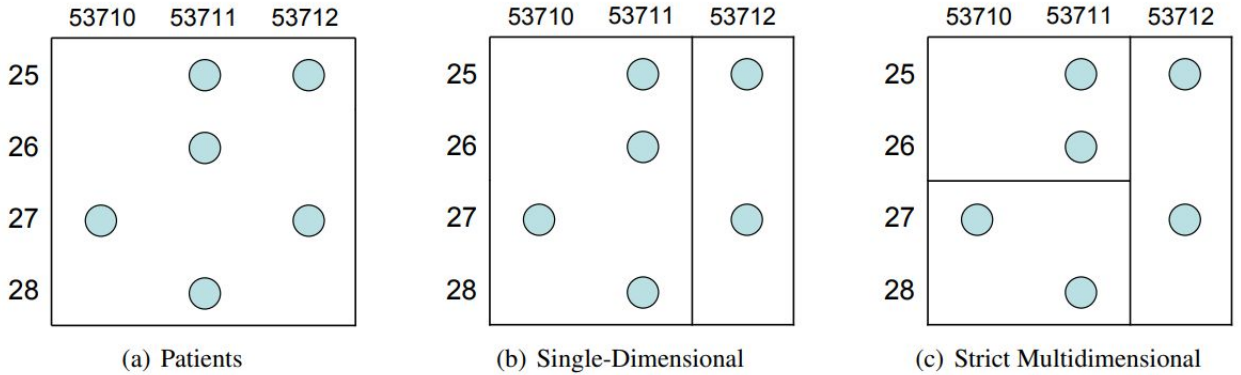
Type of data	minimum k
Low sensitivity data Public Library Visitors, Survey on TV Preferences	3 to 5
Medium sensitivity data Employment Survey, Consumer Purchase Behavior	5 to 10
High sensitivity data Medical records, Financial transactions	10 to 20

START BY FIXING THE TOLERABLE RISK, CHOOSE THE MINIMUM K THAT ALLOWS YOU TO ACHIEVE IT.

HOW TO ACHIEVE K-ANONYMITY: MONDRIAN



HOW TO ACHIEVE K-ANONYMITY: MONDRIAN



Given your points in the space, construct the mondrian tessellation.

Use some heuristics (e.g., Entropy), to determine where to partition the space; backtrack if the final solution does not satisfy the required k-anonymity.

HOMOGENEITY ATTACK

What if data are too homogeneous?

We are interested to identify the disease of a female person, born in 1964, which is single.

HOMOGENEITY ATTACK

What if data are too homogeneous?

Race	DoB	Sex	ZIP	Marital Status	Income
person	64/**/**	F	351**	been_married	33.000
person	64/**/**	F	351**	been_married	54.000
person	64/**/**	F	351**	been_married	22.000
person	63/**/**	M	351**	been_married	11.000
person	63/**/**	M	351**	been_married	178.000
person	64/**/**	F	351**	single	23.000
person	64/**/**	F	351**	single	23.000
person	64/**/**	F	351**	single	23.000
person	64/**/**	F	351**	been_married	56.000

HOMOGENEITY ATTACK

What if data are too homogeneous?

Race	DoB	Sex	ZIP	Marital Status	Income
person	64/**/**	F	351**	been_married	33.000
person	64/**/**	F	351**	been_married	54.000
person	64/**/**	F	351**	been_married	22.000
person	63/**/**	M	351**	been_married	11.000
person	63/**/**	M	351**	been_married	178.000
person	64/**/**	F	351**	single	23.000
person	64/**/**	F	351**	single	23.000
person	64/**/**	F	351**	single	23.000
person	64/**/**	F	351**	been_married	56.000

BACKGROUND ATTACK

What if the malicious user has some background knowledge?

We are interested in a male - the subject drives a very expensive car.

BACKGROUND ATTACK

What if the malicious user has some background knowledge?

Race	DoB	Sex	ZIP	Marital Status	Income
person	64/**/**	F	351**	been_married	33.000
person	64/**/**	F	351**	been_married	54.000
person	64/**/**	F	351**	been_married	22.000
person	63/**/**	M	351**	been_married	11.000
person	63/**/**	M	351**	been_married	178.000
person	64/**/**	F	351**	single	23.000
person	64/**/**	F	351**	single	23.000
person	64/**/**	F	351**	single	23.000
person	64/**/**	F	351**	been_married	56.000

BACKGROUND ATTACK

What if the malicious user has some background knowledge?

Race	DoB	Sex	ZIP	Marital Status	Income
person	64/**/**	F	351**	been_married	33.000
person	64/**/**	F	351**	been_married	54.000
person	64/**/**	F	351**	been_married	22.000
~~person~~	~~63/**/**~~	~~M~~	~~351**~~	~~been_married~~	~~11.000~~
person	63/**/**	M	351**	been_married	178.000
person	64/**/**	F	351**	single	23.000
person	64/**/**	F	351**	single	23.000
person	64/**/**	F	351**	single	23.000
person	64/**/**	F	351**	been_married	56.000

ℓ-DIVERSITY

UNINFORMATIVE DISCLOSURE PRINCIPLE

Uninformative Principle:
The published table should provide the adversary with **little additional information** beyond the background knowledge.
In other words, there should **not be a large difference between the prior and posterior beliefs**.

DISTINCT ℓ-DIVERSITY

Distinct ℓ-diversity: The simplest understanding of “well represented” would be to ensure there are **at least ℓ distinct values** for the sensitive attribute **in each equivalence class**.

Distinct ℓ-diversity is vulnerable to too unbalanced classes: even though there are ℓ distinct values, there might be one class appearing much more frequently than the others.

ENTROPY

The **entropy** is a measure of “disorder” of a given set.

Called $s \in S$ the possible values that a certain attribute can take and $p(s, E)$ the probability that, by taking random element of a set E, it is exactly s, we define the entropy as:

$$\text{Entropy}(E) = -\sum_{s \in S} p(s, E) \cdot \log p(s, E)$$

ENTROPY ℓ-DIVERSITY

Entropy ℓ-Diversity: a q-block is Entropy ℓ-Diverse if its entropy is bigger than log(ℓ).

A table is Entropy ℓ-Diverse if, for every q-block is Entropy ℓ-Diverse.

This approach is very conservative, granting high diversity, but poorly applicable in practice.

COMPUTING ℓ-DIVERSITY

STEP 1 - Identify q-blocks: count how many tuples share the same set of quasi-identifiers.

COMPUTING ℓ-DIVERSITY

STEP 2 - look at the distribution of the sensitive attribute(s) in each equivalence class

income	<10k	10k	20k	30k	<30k	TOT
q-block$_1$	50	0	30	50	10	140
q-block$_2$	110	5	5	5	5	130
q-block$_3$	20	50	40	10	20	140
q-block$_4$	0	0	10	70	40	120
q-block$_5$	40	10	80	20	60	210

COMPUTING ℓ-DIVERSITY

STEP 3 - for **distinct ℓ diversity**: count the number of distinct values for the sensitive attribute

income	<10k	10k	20k	30k	<30k	TOT
q-block$_1$	50	0	30	50	10	140
q-block$_2$	110	5	5	5	5	130
q-block$_3$	20	50	40	10	20	140
q-block$_4$	0	0	10	70	40	120
q-block$_5$	40	10	80	20	60	210

COMPUTING ℓ-DIVERSITY

STEP 3 - for **distinct ℓ diversity**: count the number of distinct values for the sensitive attribute

income	<10k	10k	20k	30k	<30k	TOT	
q-block$_1$	50	0	30	50	10	140	4
q-block$_2$	110	5	5	5	5	130	5
q-block$_3$	20	50	40	10	20	140	5
q-block$_4$	0	0	10	70	40	120	3
q-block$_5$	40	10	80	20	60	210	5

COMPUTING ℓ-DIVERSITY

STEP 3.0 - for **entropy ℓ diversity**: convert occurrences in probabilities (frequencies)

income	<10k	10k	20k	30k	<30k	TOT
q-block$_1$	.36	0	.21	.36	.07	1
q-block$_2$	.84	.04	.04	.04	.04	1
q-block$_3$	.14	.36	.29	.07	.14	1
q-block$_4$	0	0	.08	.59	.33	1
q-block$_5$	.19	.05	.38	.09	.29	1

COMPUTING ℓ-DIVERSITY

STEP 3.1 - for **entropy ℓ diversity**: compute logs

income	<10k		10k		20k		30k		<30k		TOT
q-block$_1$	.36	-1.0	0	0	.21	-1.6	.36	-1.0	.07	-2.7	1
q-block$_2$	.84	-.17	.04	-3.2	.04	-3.2	.04	-3.2	.04	-3.2	1
q-block$_3$	.14	-2.0	.36	-1.0	.29	-1.2	.07	-2.7	.14	-2.0	1
q-block$_4$	0	0	0	0	.08	-2.5	.59	-.53	.33	-1.1	1
q-block$_5$	.19	-1.7	.05	-3.0	.38	-.97	.09	-2.4	.29	-1.2	1

COMPUTING ℓ-DIVERSITY

STEP 3.2 - for **entropy ℓ diversity**: compute products and sum

income	<10k	10k	20k	30k	<30k	Entropy
q-block$_1$	-.37	0	-.33	-.37	-.19	1.25
q-block$_2$	-.15	-.13	-.13	-.13	-.13	0.66
q-block$_3$	-.28	-.37	-.36	-.19	-.28	1.46
q-block$_4$	0	0	-.20	-.31	-.37	0.88
q-block$_5$	-.32	-.15	-.37	-.22	-.36	1.41

ENFORCING ℓ-DIVERSITY

The procedure is substantially equal to those available for the k-anonymity (suppression and generalization), with the difference that our search for the minimum stops when we have satisfied the ℓ-diversity requirement rather than the simpler k-Anonymity.

LIMITATIONS: COMPLEXITY

ℓ-diversity is extremely stringent: it might be very hard to achieve.

Consider a situation with one sensitive attribute – a test for a certain virus – with 2 values: positive and negative. Negatives are much more present (99%).

If we have 10000 records, then to obtain a 2-diverse distinct table, we can have at most 10000*1% = 100 equivalence classes (each class should at least contain 1 positive example).

LIMITATIONS: SKEWNESS ATTACK

Consider the previous example: if we have a class with equal negatives and positives, the proportion of incidence of the virus in that equivalence class will be very skewed compared to the general population.

Similarly, if we have an extremely skewed distribution toward the positives, then we might end up considering more likely a positive incidence of the virus simply due to the anonymization of our data.

LIMITATIONS: SIMILARITY ATTACK

When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information.

601**	2*	3k
601**	2*	4k
601**	2*	5k
6012*	4*	6k
6012*	4*	11k
6012*	4*	8k
601**	3*	7k
601**	3*	9k
601**	3*	10k

LIMITATIONS: SIMILARITY ATTACK

When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information.

The table satisfies distinct and entropy 3-diversity.

Knowing that Bob (23 years, living in Ancona - cap 60121) is in our dataset, is he rich?

601**	2*	3k
601**	2*	4k
601**	2*	5k
6012*	4*	6k
6012*	4*	11k
6012*	4*	8k
601**	3*	7k
601**	3*	9k
601**	3*	10k

LIMITATIONS: SIMILARITY ATTACK

When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information.

The table satisfies distinct and entropy 3-diversity.

Knowing that Bob (23 years, living in Ancona - cap 60121) is in our dataset, we also know that he has a relative low income.

601**	2*	3k
601**	2*	4k
601**	2*	5k
6012*	4*	6k
6012*	4*	11k
6012*	4*	8k
601**	3*	7k
601**	3*	9k
601**	3*	10k

t-CLOSENESS

t-CLOSENESS

The t-closeness Principle: An equivalence class is said to have t-closeness if the **distance** between the **distribution of a sensitive attribute** in an equivalence class and the **distribution of the attribute in the whole table is no more than a threshold t**.

A table is said to have t-closeness if all equivalence classes have t-closeness.

PROPERTIES OF THE t-CLOSENESS

Generalization Property Let T be a table, and let A and B be two generalizations on T such that A is more general than B. If T satisfies t-closeness using B, then T also satisfies t-closeness using A.

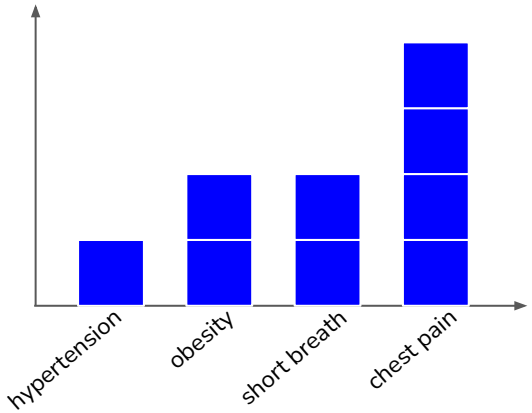
Subset Property Let T be a table and let C be a set of attributes in T. If T satisfies t-closeness with respect to C, then T also satisfies t-closeness with respect to any set of attributes D such that $D \subset C$.

EARTH MOVER’S DISTANCE (EMD)

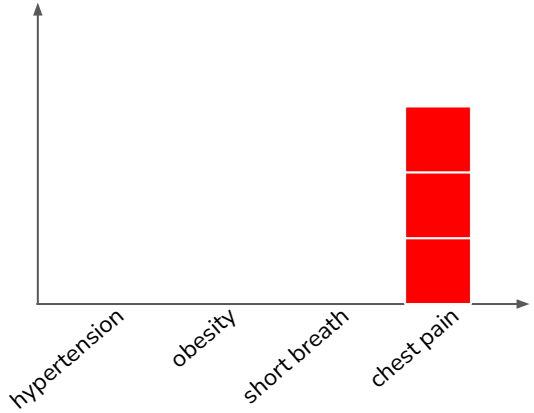
the **Earth Mover’s Distance** is a measure that allows to quantify how much two distributions differ.

If I have two heaps of dirt (distributions) how much dirt (and how far) should I move, to make the two heaps look the same?

EARTH MOVER'S DISTANCE (EMD)



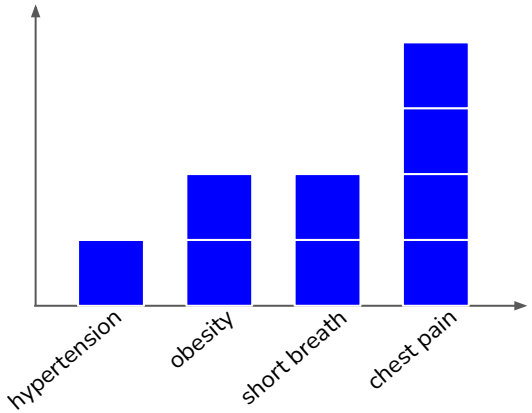
ENTIRE DATASET



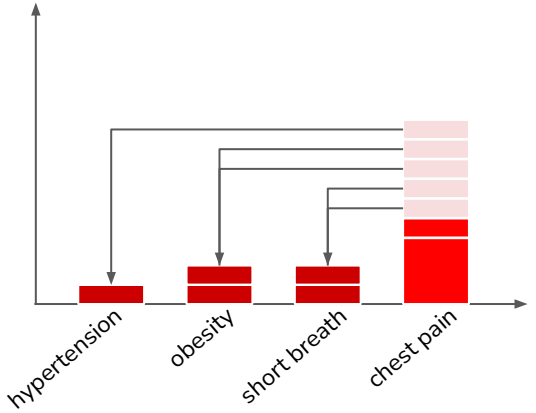
EQUIVALENCE CLASS

(person, 64/**/**,F, 351**, been_married)

EARTH MOVER'S DISTANCE (EMD)



ENTIRE DATASET



EQUIVALENCE CLASS

(person, 64/**/**,F, 351**, been_married)

EARTH MOVER'S DISTANCE (EMD): CATEGORICAL ATTRIBUTES

In the case of categorical attributes, we assume each value to be at distance 1. Called P and Q the distributions for the entire dataset and for the q-block, EMD is:

$$D[P, Q] = \frac{1}{2}\sum_{i=1}^{|S|}|p_i - q_i|$$

EARTH MOVER'S DISTANCE (EMD): NUMERICAL ATTRIBUTES

Discrete values:

called $r_i = p_i - q_i$

$$D[P, Q] = \frac{1}{|S| - 1} \sum_{i=1}^{|S|} \left| \sum_{j=1}^{i} r_j \right|$$

Continuous values:

$$D[P, Q] = \int_S |P(x) - Q(x)| dx$$

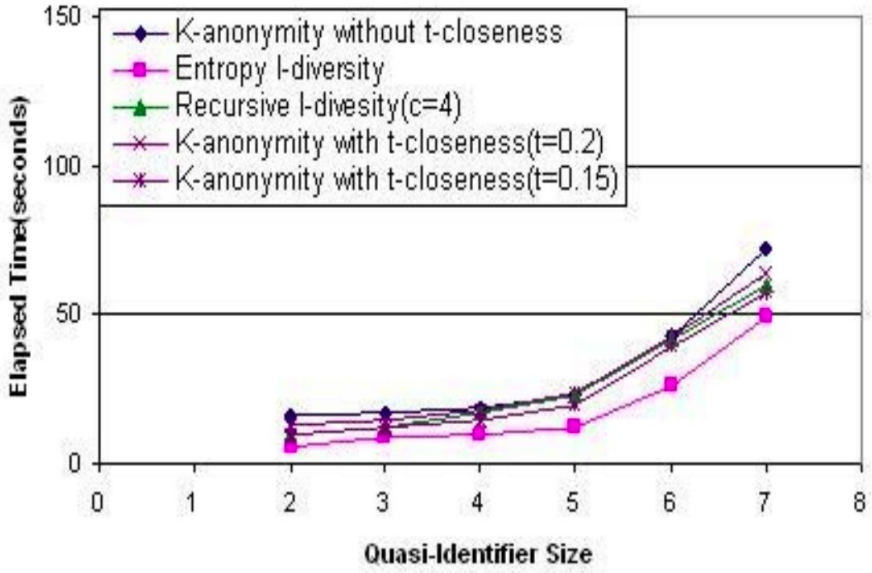
FINDING A t-CLOSE TABLE

Akin to k-anonymity and ℓ-diversity, we keep going on generalizing and suppressing until we have achieved the required level of t-closeness.

EVALUATING THE QUALITY OF OUR ANONYMIZATION TECHNIQUES

EFFICIENCY

Implementing k-anonymity, l-diversity, and t-closeness is a NP hard task. The first aspect that you should consider is the efficiency of the anonymization.



DATA QUALITY: EQUIVALENCE CLASSES SIZE

Besides observing the speed of the anonymization, we are also interested in measuring the risk of re-identification.

The first strategy consists in computing the **average dimension of the equivalence classes** in the anonymized table.

The bigger q-blocks are, the less likely it is that the user is de-anonymized (but also the more information we have lost).

DATA QUALITY: DISCERNIBILITY

We call t the records of a table T and E_t its equivalence class. We define the discernibility penalty as:

$$\text{penalty}(t) = \begin{cases} |T|, \text{ if t is suppressed} \\ \\ |E_t|, \text{ else} \end{cases}$$

the **discernibility** is defined as follows:

$$\text{discernibility}(T) = \sum_{t \in T} \text{penalty}(t)$$

K-ANONYMITY AND L-DIVERSITY IN PRACTICE

https://anjana.readthedocs.io/en/latest/intro.html

https://github.com/kaylode/k-anonymity

A commercial solution:

https://cloud.google.com/dlp/docs/compute-k-anonymity