

FATE P01, P02

Data Anonymization

Ashwin S they/them

ashwin.singh01@estudiant.upf.edu



**Universitat
Pompeu Fabra**
Barcelona

Fairness, Accountability, Transparency
and Ethics of Data Processing (FATE)

Programming Sessions



1. **6 Labs = 3 Modules x 2 Sessions**

2. **Grading Policy**

M1: Data Anonymization	(40%)	Submit by 11:59 PM, Feb 03, 2025
------------------------	-------	----------------------------------

M2: Algorithmic Fairness I	(20%)	Submit by 11:59 PM, Feb 20, 2025
----------------------------	-------	----------------------------------

M3: Algorithmic Fairness II	(40%)	Submit by 11:59 PM, Mar 07, 2025
-----------------------------	-------	----------------------------------

You must attend at least one session of each module to be eligible for grading.

3. **Queries / Discussion**

Please post on the Aula Global Forum for everyone's benefit.

4. **Email:** ashwin.singh01@estudiant.upf.edu

Overview



1. Attributes in Microdata
2. k -Anonymity
3. Methods for k -Anonymity
 - a. Global Recoding (Generalization)
 - b. Micro-Aggregation
4. ℓ -Diversity [New!]
5. Local Suppression for ℓ -Diversity

Attributes In Microdata



1. Identifiers

Can unambiguously identify a person (passport, DNI/NIE, email address, etc.)

2. Quasi-Identifiers

Can sometimes identify the person when combined with other quasi-identifiers.

(**gender** <-> zip-code <-> age)

3. Confidential Attributes

Sensitive information about a person (salary, ethnicity, **gender**, etc.)

4. Non-Confidential Attributes

Whatever remains...

k-Anonymity

D1: $x_1, x_2 \in X$ belong to the same **equivalence class** if \forall quasi-identifiers $q \in Q, x_1(q) = x_2(q)$.

D2: X satisfies **k-anonymity** when there exist at least k elements in each **equivalence class**.

Gender	Zip-Code	Salary	Equivalence Class
man	08010	<5000€	C1
man	08010	<5000€	C1
woman	08022	>5000€	C2
woman	08022	>5000€	C2

Global Recoding (Generalization)

Gender	Zip-Code	Generalized Zip-Code	Salary	Equivalence Class
man	08011	0801*	<5000€	C1
man	08012	0801*	<5000€	C1
woman	08020	0802*	>5000€	C2
woman	08022	0802*	>5000€	C2

Remember

All instances in the database are modified during generalization.

(which is why it is **global**)

It is non-perturbative. Why ?

Quasi-Identifiers	Confidential Attribute
-------------------	------------------------

Micro-Aggregation

Company	Workers	Aggregated Workers	Profit	Equivalence Class
A	44	41	+1000€	C1
B	25	27	+500€	C2
C	39	41	+750€	C1
D	30	27	+300€	C2

Discuss

Why do we take the **mean**?

Is it **perturbative** OR
non-perturbative?

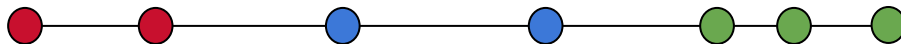
Identifier	Quasi-Identifiers	Confidential Attribute
------------	-------------------	------------------------

Micro-Aggregation ► Intuition

Univariate Case

Sort data by the continuous quasi-identifier

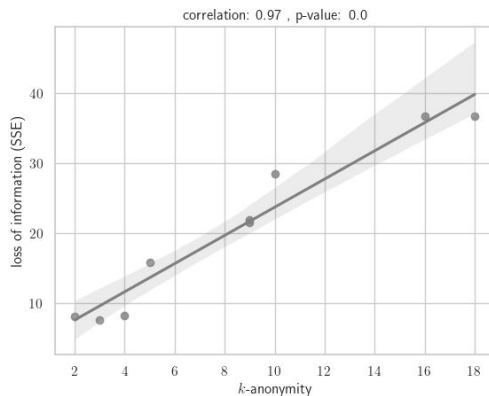
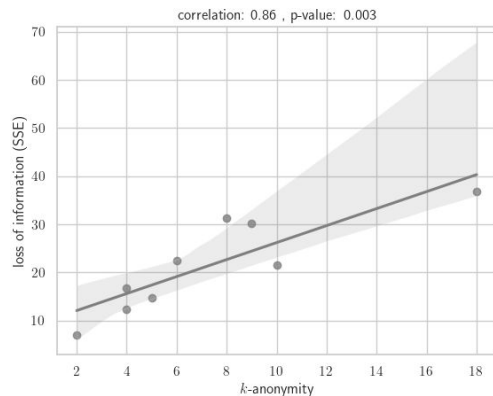
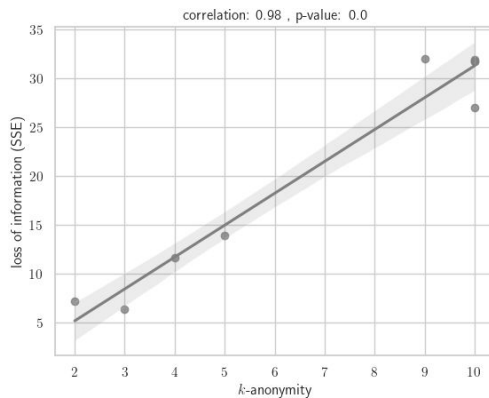
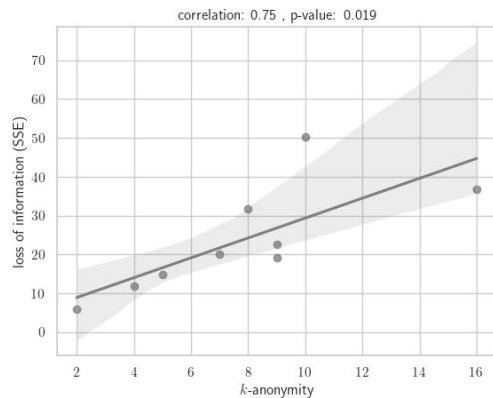
Assign first k items class 1, next k items class 2, and so on... (*not always optimal)



Multivariate Case

k -Partition Problem with Minimal Information Loss [NP-Hard]

Micro-Aggregation ► Loss of Information



Plots generated across different runs of **Gaussian Mixture Model** to cluster and micro-aggregate data to achieve k -anonymity.

Loss of Information

$$SSE = \sum_{c \in C} \sum_{q \in Q_c} (q - \bar{q}_c)'(q - \bar{q}_c)$$

C is the set of equivalence classes

Q_c is the set of quasi-identifier vectors in equivalence class c

A dataset X satisfies **ℓ-diversity** when there exist at least ℓ distinct values of the **confidential attribute** in each **equivalence class**.

Zip-Code	Profession	Equivalence Class
0801*	Software Engineer	C1
0801*	Software Engineer	C1
0802*	Architect	C2
0802*	Doctor	C2

What is the problem here?

ℓ -Diversity ► Local Suppression

A dataset X satisfies **ℓ -diversity** when there exist at least ℓ distinct values of the **confidential attribute** in each **equivalence class**.

Zip-Code	Profession	Equivalence Class
0801*	Software Engineer	C1
0801*	Engineer	C1
0802*	Architect	C2
0802*	Doctor	C2

What is the problem now?