Data Protection and Privacy

University of Genoa

Lesson 2: Multidimensional data (1)

Gaspare Ferraro <ferraro@gaspa.re>

Classification of Privacy Preserving Methods

Preamble: Data anonymization methods should focus on *semantics* of data and not on the *syntax*

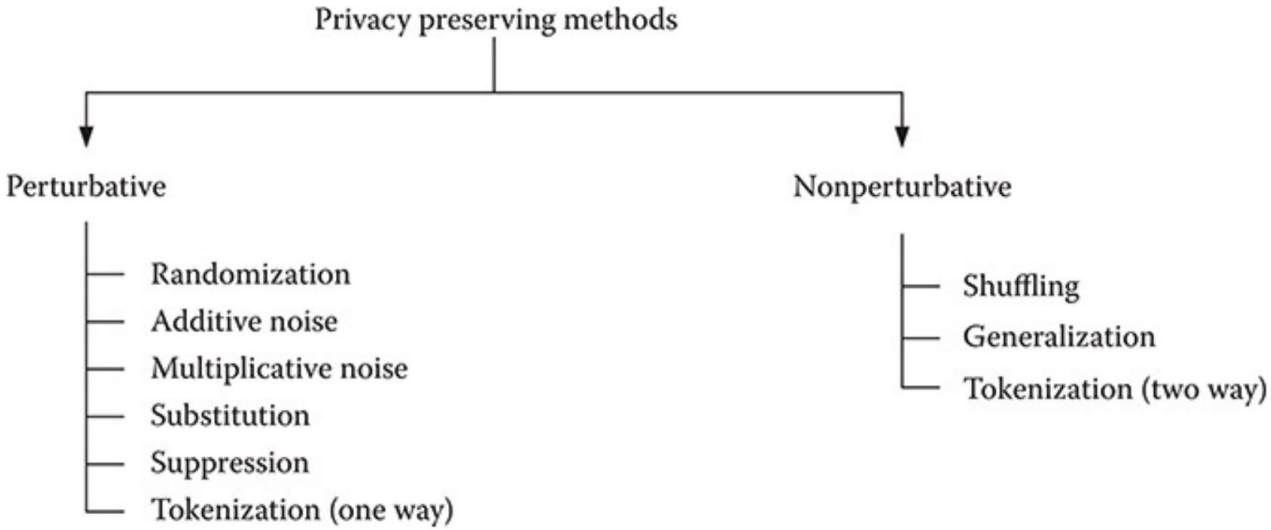
Syntax → grammar

Semantics → meaning

Principle: Understand the semantics of data in the context of the application as to apply proper anonymization techniques

Classification of Privacy Preserving Methods

Perturbative techniques are generally referred as **masking**
Non-perturbative techniques are generally referred as **anonymization**



A recap

- EI allow to directly identify the user, often contain the primary key
- QI could allow to indirectly identify the user if combined with external knowledge
- SD should be not anonymized to maximize utility (in general)

EI		QI				SD			
ID	Name	Gender	Age	Address	Zip	Basic	HRA	Med	All
12345	John	M	25	1, 4th St.	560001	10,000	5,000	1000	6,000
56789	Harry	M	36	358, A dr.	560068	20,000	10,000	1000	12,000
52131	Hari	M	21	3, Stone Ct	560055	12,000	6,000	1000	7,200
85438	Mary	F	28	51, Elm st.	560003	16,000	8,000	1000	9,600
91281	Srini	M	40	9, Ode Rd	560001	14,000	7,000	1000	8,400
11253	Chan	M	35	3, 9th Ave	560051	8,000	4,000	1000	4,800

Protecting Explicit Identifiers (EI)

Requirements:

1. Referential integrity
2. Consistency across tables and databases

One-way tokenization: $x \rightarrow h(x)$ with h(x) a one-way function (not reversible)

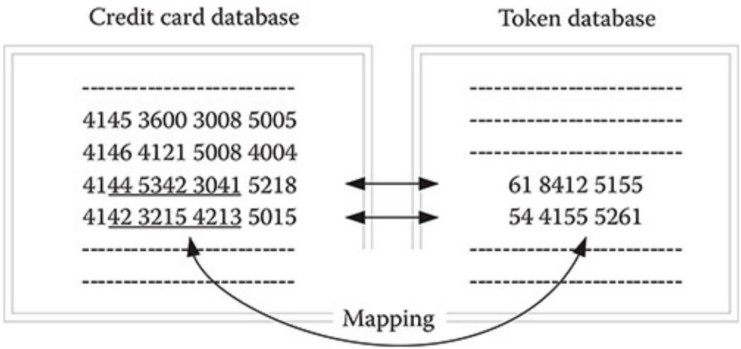
Tokenizing ID and Substituting Name

ID	Name	ID	Name
(a) Original EI		(b) Masked EI	
12345	John	40011	Jack
56789	Harry	81100	Sammy
52131	Hari	62410	Mark
85438	Mary	79820	Jane
91281	Srini	14532	Singh
11253	Chan	22244	Khan

Protecting Explicit Identifiers (EI) - 2

Tokenization:

- A form of randomization, but more secure
- It preserves the format of data
- Token value has no relation with the original data (loss of semantics)
- One-way vs Two-way (reversible, non-perturbative)



Example of Two-way tokenization

Protecting Explicit Identifiers (EI) - 3

Protecting names: **substitution**

- It requires a look-up table

First Name	Last Name	Gender
Mark	Anthony	M
Lina	Roy	F
Larry	Rowe	M
Roy	Fred	M
Lara	Dow	F

(a)

(b)

David	Anderson	M
Jane	Croft	F
Clive	Richards	M
Prill	James	M
Mary	Thomas	F

(c)

Protecting Quasi-Identifiers (QI)

Record linkage: is the task of finding records in a data set that refer to the same entity across different data sources.

ID	First Name	Last Name	Gender	Address	DOB	Zip	Disease
—	—	—	—	—	—	—	—
12432			M	MA	21/02/1946	01880	Cancer
—	—	—	—	—	—	—	—

(a)

Voter ID	First Name	Last Name	Gender	Address	DOB	Zip
—	—	—	—	—	—	—
893423		Weld	M	MA	21/02/1946	01880
—	—	—	—	—	—	—

(b)

- QI attributes are **categorical**, they can have two or more categories, but without any intrinsic ordering to the categories.
- The finite range of these categorical values needs to be considered while coming up with the anonymization approach

Challenges in Protecting QI

Aspects to deal with:

1. The analytical utility of QI needs to be preserved.
2. Correlation of QI attributes with SD needs to be maintained to support the utility of anonymized data.

Challenges:

1. High dimensionality → it becomes difficult to define a clear boundary between QI and SD.
2. Background knowledge of the adversary → unknown, assumptions should be made.
3. Availability of external knowledge → increasing
4. Correlation with SD to ensure utility (see next slide)
5. Maintaining analytical utility → the anonymized QI attributes should support all the different queries that the original data set supported.

Challenges in Protecting QI - 2

Correlation and Anonymization: e.g. How many employees have a Doctorate?

ID	Gender	Day	Month	Year	Address	City	Zip Code	Education	Years of Experience	Salary
1	M	—	—	1968	512, -------	BLR	560002	Doctorate	20	34,000
2	M	—	—	1970	115, -------	BLR	560001	Postgraduate	19	24,000
3	M	—	—	1967	188, -------	BLR	560033	Doctorate	22	36,000
4	F	—	—	1985	157, -------	BLR	560004	Graduate	10	14,000
5	F	—	—	1982	121, -------	BLR	560068	Postgraduate	12	16,000
6	M	—	—	1970	610, -------	BLR	560001	Postgraduate	18	22,000

Cannot say

perturbative technique on dataset

ID	Gender	Day	Month	Year	Address	City	Zip Code	Education	Years of Experience	Salary
1	M	—	—	1968	512, -------	BLR	560001	Graduate	20	34,000
2	M	—	—	1970	115, -------	BLR	560004	Graduate	19	24,000
3	M	—	—	1967	188, -------	BLR	560068	Graduate	22	36,000
4	F	—	—	1985	157, -------	BLR	560001	Graduate	10	14,000
5	F	—	—	1982	121, -------	BLR	560033	Graduate	12	16,000
6	M	—	—	1970	610, -------	BLR	560002	Graduate	18	22,000

Protecting Sensitive Data (SD)

SD should be not anonymized to preserve utility, but in some cases they can be used for re-identification. Consider the following example with random perturbation.

Base Salary	Allowance	Medicals	Perks	Total
10,000	5000	1000	6000	22,000
12,000	6000	1000	7200	26,200
9,000	4500	1000	5000	19,000
14,000	7000	1000	8400	30,400
13,000	6500	1000	7800	28,300
11,000	5500	1000	6600	24,100
15,000	7500	1000	9000	32,500
10,500	5250	1000	6300	23,050
12,500	6250	1000	7500	27,250
9,500	4750	1000	5700	20,950

Base Salary	Allowance	Medicals	Perks	Total
10,500	5250	1000	6300	23,050
12,800	6400	1000	7680	27,880
9,760	4880	1000	5856	21,496
11,950	5975	1000	7170	26,095
14,000	7000	1000	8400	30,400
10,250	5125	1000	6150	22,525
13,830	6915	1000	8298	30,043
10,500	5250	1000	6300	23,050
12,200	6100	1000	7320	26,620
10,700	5350	1000	6420	23,470

The mean and covariance of both tables are the same.

Group-based Anonymization: K-Anonymity

Record linkage: As most QI attributes are also present in external data sources, such as a voters database, the anonymization technique should prevent the linking of a record owners QI attribute to these external data sources.

Utility of the transformed data: Naive perturbation of QI attributes renders the data unusable. Non-perturbative techniques, such as generalization, preserve the truth in the data table.

Protection of outlier records: It is difficult to mask outlier records. When techniques such as additive noise are used to transform the data, outlier values still show up. For example, when the distribution (statistical) is computed, one cannot hide the net worth of Warren Buffet or Bill Gates!

The correlation/association between QI and SD must be preserved and protected.

K-Anonymization

Consider the following example: how to prevent the record linkage and preserve utility?

ID	Gender	Day	Month	Year	Address	City	Zip Code	Education	Years of Experience	Salary
1	M	15	07	1973		BLR	560001	Doctorate	20	35,000
2	M	20	11	1975		BLR	560045	Masters	17	28,000
3	F	12	12	1977		BLR	560033	Graduate	18	15,000
4	F	08	07	1974		BLR	560041	Doctorate	20	38,000
5	F	17	06	1985		BLR	560003	Graduate	12	10,000
6	M	05	07	1980		BLR	560002	Graduate	10	9,000
7	F	01	02	1977		BLR	560044	Masters	15	18,000
8	M	03	01	1978		BLR	560001	Masters	18	22,000
9	M	10	11	1980		BLR	560042	Graduate	20	15,000
10	F	18	12	1982		BLR	560031	Doctorate	15	32,000
11	M	22	10	1980		BLR	560035	Masters	12	14,000
12	M	25	11	1979		BLR	560033	Masters	14	16,000

Name	Gender	Date of Birth	Address	City	Zip
Hari	M	05/07/1980		Bangalore	560002

K-Anonymization is a technique for preserving individual identification by transforming the record set so that **each record of a table identical to at least k-1 other records**.

K-Anonymization

K-anonymization is granted by **generalizing** and **suppressing** the value of attributes.

Generalization: technique of replace more specific values with generic and semantically similar values. It can be applied at cell or tuple or attribute levels.

Education	Education (4-Anonymous)
Doctorate	Grad school
Masters	Grad school
Bachelors	Bachelors
Doctorate	Grad school
Bachelors	Bachelors
Bachelors	Bachelors
Masters	Masters
Masters	Masters
Bachelors	Bachelors
Doctorate	Grad school
Masters	Masters
Masters	Masters

K-Anonymization

le anonimizzazioni che abbiamo non perdono tante informazioni

-togliere l'ultima cifra del salario
-education piu generica
-zipcode tolto ultime cifre non necessarie

ID	Gender	Day	Month	Year	Address	City	Zip Code	Education	Years of Experience	Salary
1	Any Sex	—	—	1973	—	BLR	560010	Any_Degree	20	35,000
2	Any Sex	—	—	1975	—	BLR	560050	Any_Degree	17	28,000
3	Any Sex	—	—	1977	—	BLR	560040	Any_Degree	18	15,000
4	Any Sex	—	—	1974	—	BLR	560040	Any_Degree	20	38,000
5	Any Sex	—	—	1985	—	BLR	560010	Any_Degree	12	10,000
6	Any Sex	—	—	1980	—	BLR	560010	Any_Degree	10	9,000
7	Any Sex	—	—	1977	—	BLR	560050	Any_Degree	15	18,000
8	Any Sex	—	—	1978	—	BLR	560000	Any_Degree	18	22,000
9	Any Sex	—	—	1980	—	BLR	560030	Any_Degree	20	15,000
10	Any Sex	—	—	1982	—	BLR	560030	Any_Degree	15	32,000
11	Any Sex	—	—	1980	—	BLR	560040	Any_Degree	12	14,000
12	Any Sex	—	—	1979	—	BLR	560040	Any_Degree	14	16,000

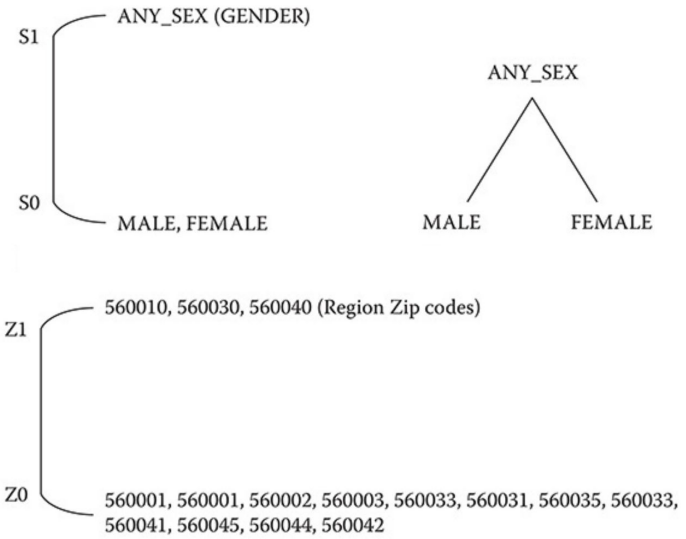
4-anonymous salary table.

Data Generalization: how to

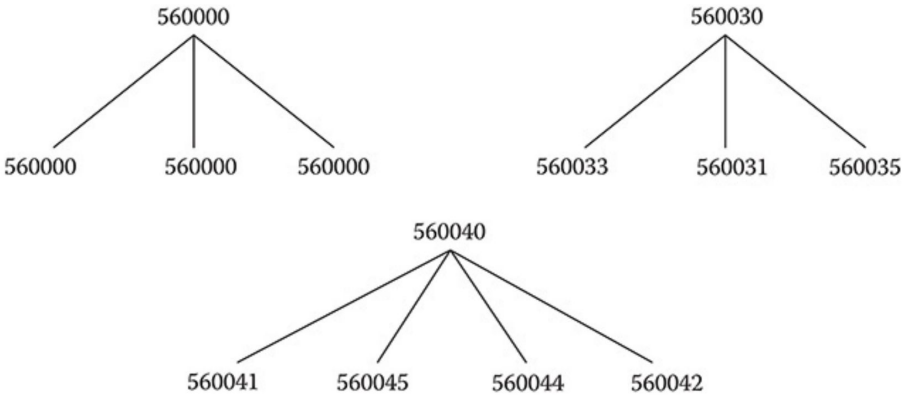
Generalization uses the concept of domain generalization and value generalization.

The value generalization hierarchy associates a value in domain D_i to a unique value in the general domain D_j.

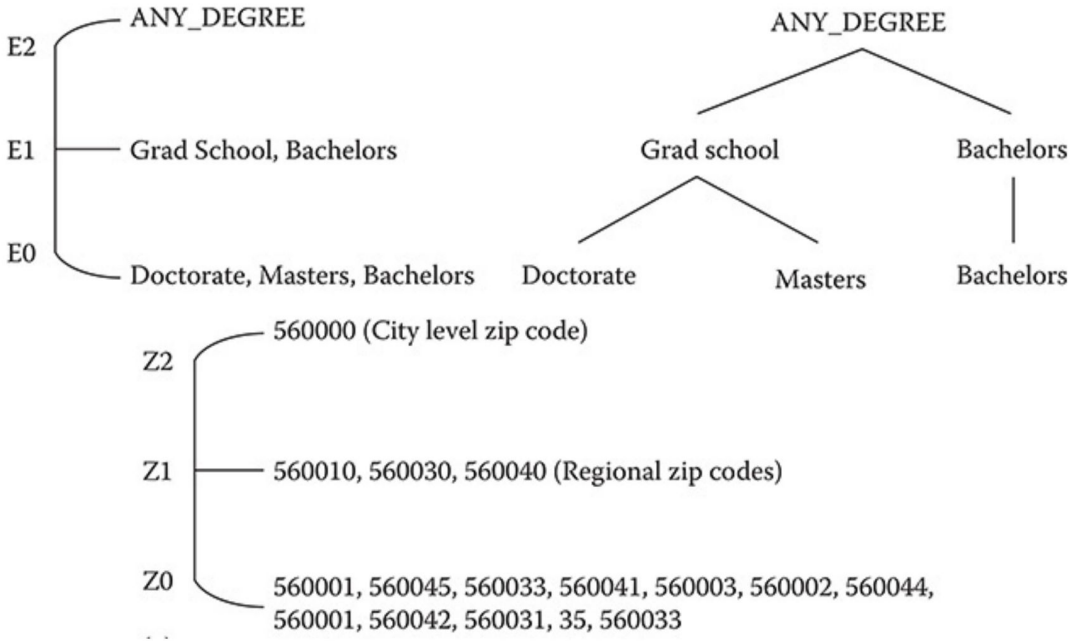
Example of attributes generalizations:



Data Generalization: how to



Data Generalization: how to



Data Generalization: how to

Gender	Zip Code	Education
M	560001	Doctorate
M	560045	Masters
F	560033	Bachelors
F	560041	Doctorate
F	560003	Bachelors
M	560002	Masters
F	560044	Masters
M	560001	Bachelors
M	560042	Doctorate
F	560031	Masters
M	560035	Masters
M	560033	Doctorate

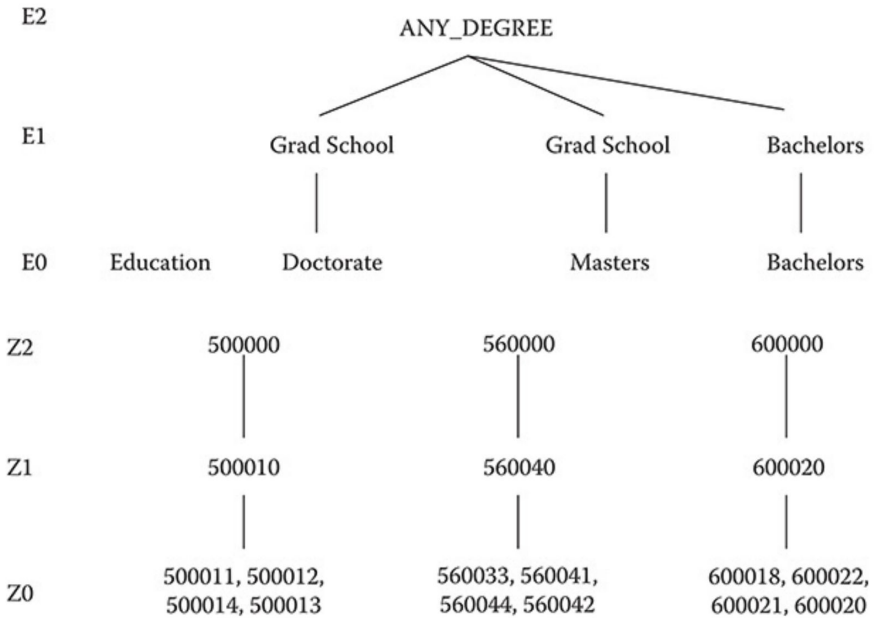
Gender	Zip Code	Education
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE

Example of full domain generalization.

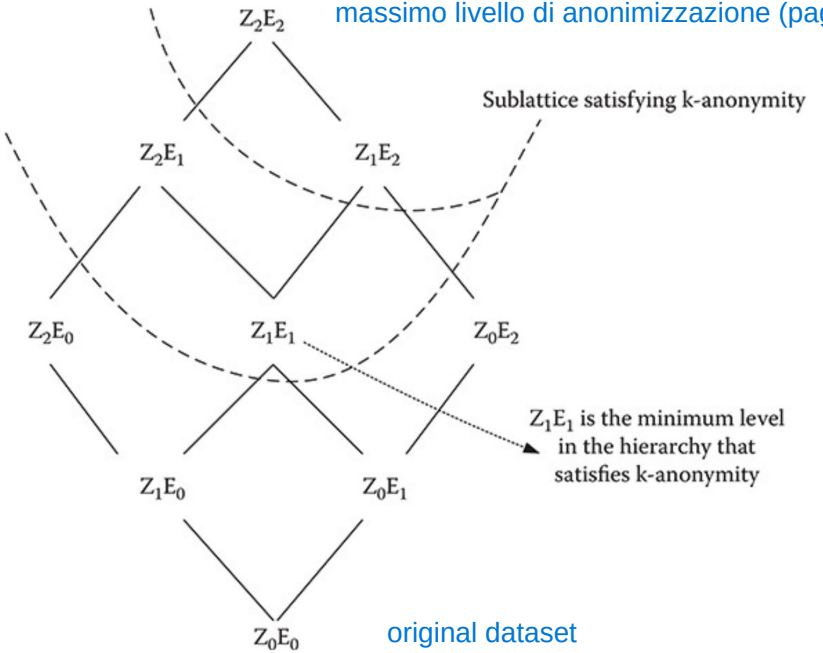
Implementing k-Anonymity: Samarati's approach

- Domain Generalization Hierarchy
- k-anonymity is calculated through the AG-TS technique (attributes-tuples)
- Goal: minimum level of generalization that satisfies k-anonymity
- Output: the node <Qi, Qj> that is closest to the most specific node in the lattice structure that satisfies k-anonymity.

k-anonymity: example



k-anonymity: example -2



massimo livello di anonimizzazione (pagina prima)

original dataset

Selecting the value of k

Optimal k

$$k = f(P_R, U_R, C_R, G_L, C)$$

- P_R is the privacy requirement of the data owner
- U_R is the utility requirement of users of anonymized data
- C_R is the compliance requirement of privacy of data
- G_L is the generalization level
- C refers to the constraints

tutte le info necessarie alla scelta del grado di anonimiz