FERRARO

MULTIDIMENSIONAL DATA (DATABASE/RELATIONAL DATA)

[k-Anonimato e metodi basati su cluster per la privacy | ~elf11.github.io]

It's important to remember that we have four types of attributes:

- Explicit identifiers (EI): attributes that identify a customer (also called record owner) directly. These include attributes like social security number (SSN), insurance ID, and name. EI allow to directly identify the user, often contain the primary key. EI by default are completely masked (perturbed).
- Quasi-identifiers (QI): attributes that include geographic and demographic information, phone numbers, age and e-mail IDs. Quasi-identifiers are also defined as those attributes that are publicly available, for example, a voters database. QI could allow to indirectly identify the user if combined with external knowledge. QIs are anonymized.
- Sensitive data (SD): attributes that contain confidential information about the record owner, such as health issues, financial status, and salary, which cannot be compromised at any cost. SD should be not anonymized to maximize utility (in general). SD are left as is to enable analysis.
- Nonsensitive data (NSD): data that are not sensitive for the given context.

Explicit Identifiers		Quasi-Identifiers				
ID	First Name	DOB	Gender	Address	Zip Code	Phone
1	Ravi	1970	Male	Fourth Street	66001	92345-67567
2	Hari	1975	Male	Queen Street	66011	98769-66610
3	John	1978	Male	Penn Street	66003	97867-00055
4	Amy	1980	Female	Ben Street	66066	98123-98765

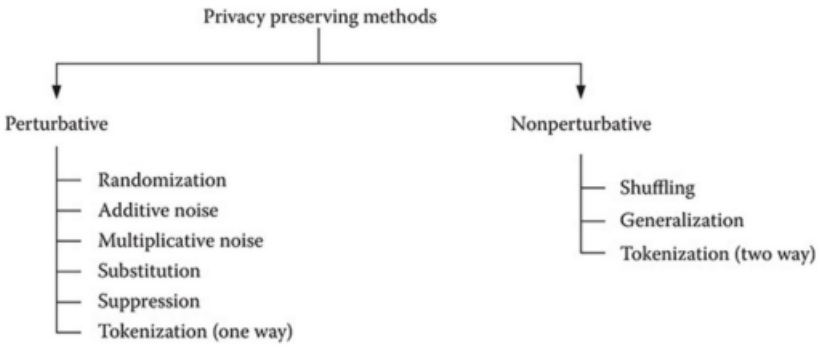
	Sensitive Data				
ID	Account Number	Account Type	Account Balance	Credit Limit	Nonsensitive Data
1	12345	Savings	10,000	20,000	
2	23456	Checking	5,000	15,000	
3	45678	Savings	15,000	30,000	
4	76543	Savings	17,000	25,000	

Remember that each record or row is independent of others.

When we talk about the "privacy preserving methods" we need to know that the data anonymization methods should focus on semantics of data (which is the meaning) and not on the syntax (which is grammar, so how to put together words) → understand the semantics of data in the context of the application ofc.

There are:

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



EI (it's easy to anonymize them because they're unique)

Requirements:

- Referential integrity(?).
- Consistency across tables and databases.

How can we protect EI? We can use 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). The first technique is used when we don't want to return back after anonymization. We can use both for ID and name.

If we want to protect names, we can use substitution that requires a look-up table (It is used to replace explicit identifiers, such as people's names, with anonymized identifiers or pseudonyms. This helps protect the identity of people in the data without compromising the usefulness of the information, while still allowing the anonymous identifier to be matched to the real name when necessary using the matching table).

QI

Record linkage is the task of finding records (values of attributes, all the rows) in a data set that refer to the same entity across different data sources.

QI attributes are categorical, they can have two or more categories(?), but without any intrinsic ordering to the categories. It's important to preserve the analytical utility of QI and the correlation between QI and SD.

What are the problem?

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

Example:

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

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

The first table is the original table and we can say how many employees have a doctorate, if we apply a perturbative technique on zip code and education we can't say after only seeing this table.

SD

SD should be not anonymized to preserve utility, but in some cases, they can be used for re-identification (new identification, is it a good thing?). 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

I want to protect:

- 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 is a technique for preserving individual identification by transforming the record set (equivalence class) so that each record of a table identical to at least k-1 other records. K-anonymization is granted by generalizing and suppressing the value of attributes (I can eliminate some or all values of one or more attributes). Generalization is a technique of replace more specific values with generic (range) and semantically similar values. It can be applied at cell or tuple or attribute levels. I can use suppression when generalize is too stranger (because I lose information).

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

Nel caso seguente, ogni salary non dovrebbe comparire almeno tre volte? No, perché posso applicare k-anonymity solo sui QI. Si, ma allora che senso ha?:

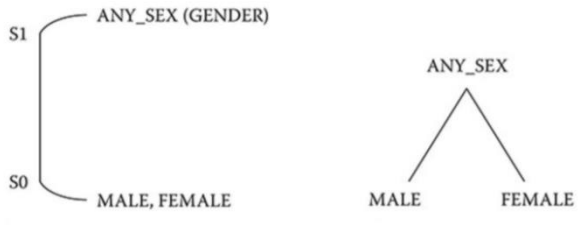
ID	Gender	Day	Month	Year	Address	City	Zip Code	Education	Years of Experience	Salary
1	M	15	07	1973		BLR	560001	Doctorate	20	35,000
2	M	20	11	1975		BLR	560045	Masters	17	28,000
3	F	12	12	1977		BLR	560033	Graduate	18	15,000
4	F	08	07	1974		BLR	560041	Doctorate	20	38,000
5	F	17	06	1985		BLR	560003	Graduate	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

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

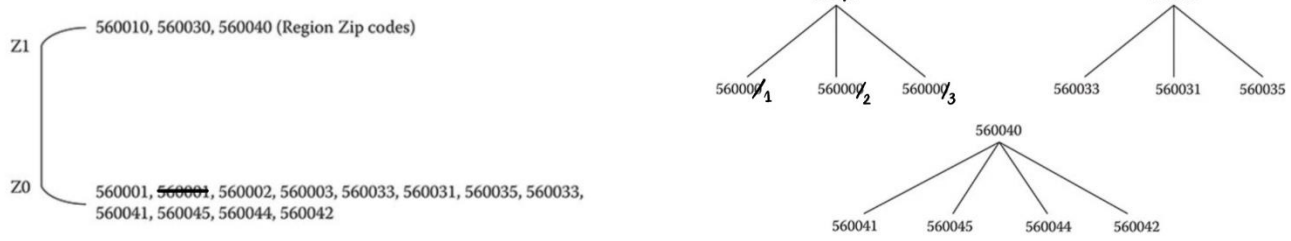
4-anonymous salary table.

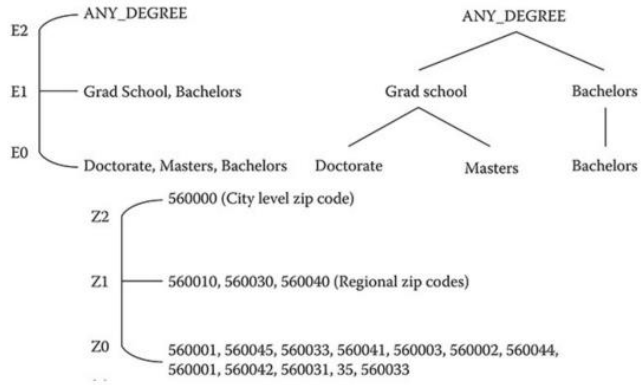
DATA GENERALIZATION

Generalization uses the concept of domain generalization and value generalization. The value generalization hierarchy associates a value in domain Di to a unique value in the general domain Dj.



Non dovrebbe esserci anche 560050?:





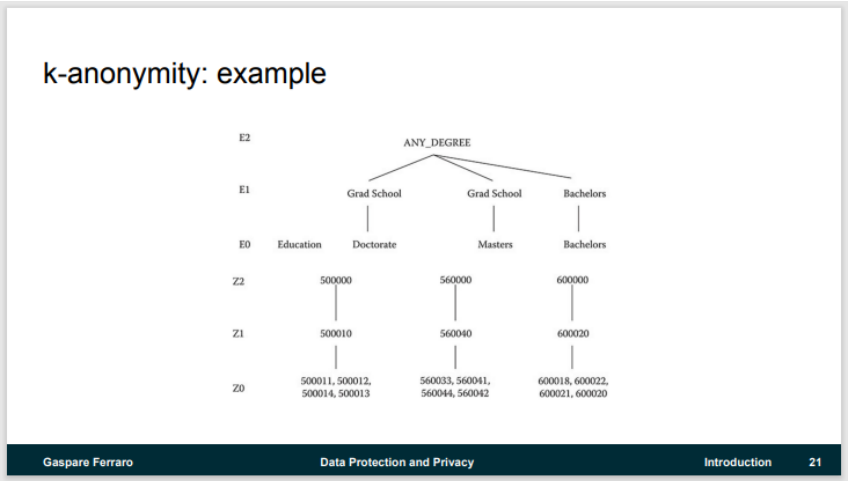
Gender	Zip Code	Education
M	560001	Doctorate
M	560045	Masters
F	560033	Bachelors
F	560041	Doctorate
F	560003	Bachelors
M	560002	Masters
F	560044	Masters
M	560001	Bachelors
M	560042	Doctorate
F	560031	Masters
M	560035	Masters
M	560033	Doctorate

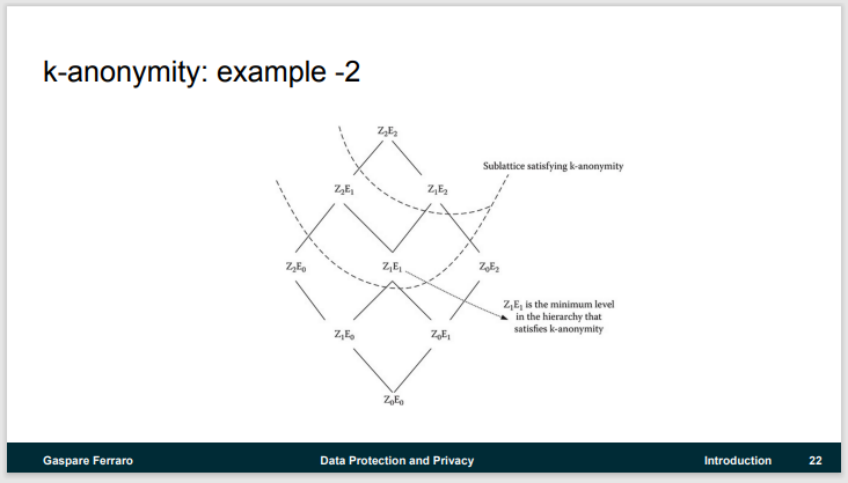
Gender	Zip Code	Education
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000	ANY_DEGREE
ANY_SEX	560000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

Example of full domain generalization.

IMPLEMENTING K-ANONYMITY: SAMARATI'S APPROACH

NON HO CAPITO:





The optimal k is:

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

where:

- 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.

Problems:

- Provable privacy.

  Fundamental concepts:

  - Equivalence class: In k-anonymity, k tuples form an equivalence class where records within a class are indistinguishable from each other with regards to sensitive attributes. This means that within a class, sensitive attributes have the same values, ensuring a certain level of anonymization.
  - Identity disclosure: Occurs when an adversary is able to link the sensitive information in the anonymized table to an external database that can identify a specific person. This link may compromise your privacy.
  - Anonymization and probability: Anonymization is designed (progettata) to prevent identity revelation. The probability of linking an individual to an external record based on QIs must

be reduced to a maximum of 1/k, where k is the minimum number of records or tuples that share the same set of QIs.

In essence, “provable privacy” in the context of k-anonymity refers to the ability to demonstrate that the implemented anonymization technique can protect data against individual identification and privacy-compromising attacks, reducing the probability between anonymized records and external databases.

- Efficiency and performance of the algorithm

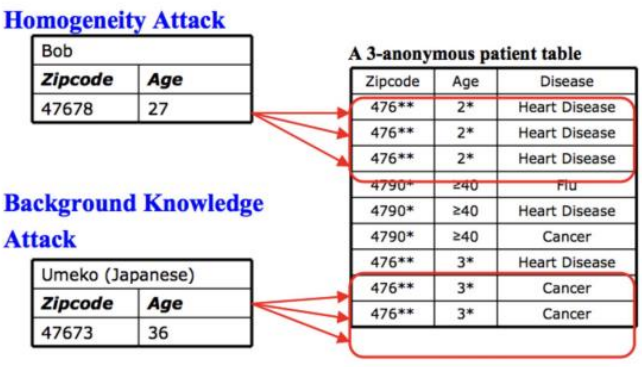
  It’s NP-hard.

- Scalability
  - Horizontal scalability: In the context of data anonymization, it could mean the ability to extend data protection to an increasing number of attributes without compromising performance or security.
  - Vertical scalability: In the context of data anonymization, it could mean the ability to handle an increasing number of records or tuples in the table without compromising the efficiency of the anonymization process.
  - Problems with the high dimensionality of quasi-identifiers (QIs): QIs are attributes or combinations of attributes that could, if not protected, allow the identification of individuals. The high dimensionality of QIs can cause problems in data anonymization. With many sensitive or quasi-identifying attributes, it can become more challenging to effectively protect anonymity without compromising the usefulness of the data.

  In short, scalability in this context is about the ability to handle large volumes of data and workloads, while issues related to the high dimensionality of quasi-identifiers highlight the challenges in protecting anonymity when dealing with multiple sensitive or quasi-sensitive attributes. -identifiers within a dataset.

- Robustness

  K-anonymity suffers from 3 kinds of attacks:
  - Homogeneity Attack: it leverages (sfrutta) the case where all the values for a sensitive value within a set of k records are identical. In such cases, even though the data has been k-anonymized, the sensitive value for the set of k records may be exactly predicted.
  - Background Knowledge Attack: it leverages an association between one or more quasi-identifier attributes with the sensitive attribute to reduce the set of possible values for the sensitive attribute.
  - Complementary Release Attack: different releases of the same private table can be linked together to compromise k-anonymity.



In the first case, one is Bob for sure, in the second one no because we have one record free.

K-anonymity does not provide privacy if:

- SD is an equivalence class lack diversity.
- The attacker has a background knowledge.

- Utility/data quality

Drawbacks (svantaggi):

- It is not robust enough to prevent homogeneity attacks.
- Optimal k-anonymization is NP-hard: it is computationally very hard to solve.
- It has a problem with high-dimensional data and large record sizes.
- It is difficult to balance or optimize privacy versus utility, as higher levels of k provide high privacy and low utility and vice versa.
- No scientific method is available to determine an optimal value of k.
- The use of suppression leads to high information loss or low utility and using only generalization leads to a highly generic table having very low utility.

Advantages:

- No additional noise is added to the original data.
- All of the tuples in the anonymized data remain trustworthy.

L-DIVERSITY

l-diversity is an extension of the k-anonymity model and adds the promotion of intra-group diversity for sensitive values in the anonymization mechanism. A table is said to have l-diversity if every equivalence class of the table has l-diversity.

K-anonymity aims to protect the identity of individuals by ensuring that (assicurandosi che), within a group or equivalence class, there are at least k records that are indistinguishable with respect to sensitive attributes. However, one of the critical issues of k-anonymity is that, within these classes, the values of sensitive attributes can be homogeneous, they present little variety or diversity.

l-diversity aims to address (affrontare) this problem by introducing greater diversity within each k-anonymity equivalence class. This means that, in addition to ensuring that there are at least k indistinguishable records, it is also required that there is variety (diversity) in the sensitive attribute values within the class.

So, the l-diversity idea is:

- It focuses on sensitive attributes, assuming in blocks where QI are assumed identical (q* - block).
- Since in a (q* - block) the SD and QI are identical then the homogeneity attack could be performed, the l-diversity aims at modifying the SD attributes in each (q* - block) to make them “diverse”.
- A (q* - block) is l-diverse if each sensitive attribute in each (q* - block) is “well-represented” by at least l diverse values.

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

Original table (top), 4-Anonymous Table (bottom-left), 3-Diverse Table (bottom-right)

On the right I have l-diverse values in each group (block or class od equivalence) of SD.

l-diversity modeling is based on probability and generalization. When we talk about probability, we have to remember Bayes' Theorem (principio probabilistico che descrive come la probabilità di un evento condizionato possa essere calcolata conoscendo la probabilità di altri eventi correlati. Noi lo applichiamo per modellare come l'avversario potrebbe cercare di ricavare informazioni sensibili da dati anonimizzati):

$$P(A|B) = (P(B|A)\ P(A)) / P(B)$$

where:

- P(A) and P(B) are the probabilities of observing A and B without regards to each other.
- P(A|B) is a conditional probability, is the probability of observing event A given that B is true.
- P(B|A) is a conditional probability, is the probability of observing event B given that A is true.

Assumptions:

- A table T is a random sample of dimension n from a large population Ω.
- There are a single non-sensitive attribute Q and a single sensitive attribute S in T.
- f is the distribution of Q and S in the population Ω.
- Attacker's assumptions (worst case), knows the distribution f, the fact that Bob corresponds to a record $t \in T$ that has been generalized in a record $t^* \in T^*$ and Bob's non-sensitive attribute (she knows that t[Q] = q, where t[Q] indicates the projection).

Adversary's background knowledge:

- Instance-level: the adversary has some information on individuals contained in the table.
- Demographic: the adversary has partial knowledge about the distribution of sensitive and non-sensitive attributes in the population.

Prior and Posterior knowledge of the attacker:

- Prior knowledge (before data release): $\alpha(q,s) = P_f(t[S] = s, t[Q] = q)$ -> probabilità di trovare una combinazione di valori specifici di attributi sensibili e non sensibili.
- Posterior knowledge (after the release of table T*): $\beta(q, s, T^*) = P_f(t[S] = s, t[Q] = q \wedge \exists\ t^* \in T^*, t \rightarrow t^*)$ -> probabilità di trovare quella combinazione di valori dopo il rilascio dei dati anonimizzati.

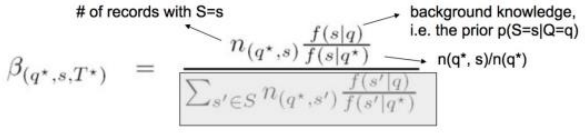
These concepts show how analysis of adversary probabilities and knowledge can be used to evaluate (valutare) how an anonymization is good to protecting sensitive data from attacks that attempt to reconstruct individual information from anonymized data (questi concetti mostrano come l'analisi delle probabilità e delle conoscenze dell'avversario possa essere utilizzata per valutare quanto un'anonimizzazione sia buono nel proteggere i dati sensibili da attacchi che tentano di ricostruire informazioni individuali da dati anonimizzati).

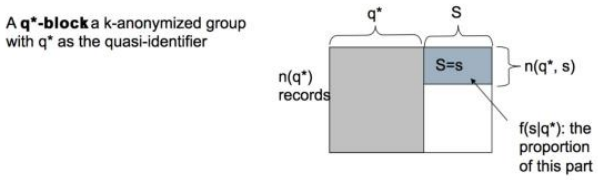
Positive and Negative disclosure (divulgazione):

- Positive disclosure: if the adversary can correctly identify the value of a sensitive attribute with high probability → prior knowledge is small, posterior knowledge is large.
- Negative disclosure: if the adversary can eliminate some possible values of the sensitive attributes → prior knowledge is large, posterior knowledge is small.
- Uninformative Principle: The published data should provide the adversary with additional little information beyond the background knowledge α(q,s) ≈ β(q, s, T*). Uninformative Principle states that published data should provide (dovrebbe fornire) the adversary with a minimum amount of information additional to the baseline knowledge the adversary possesses before the data is released. It seeks (si cerca di) to minimize the amount of new information the adversary can gain (ottenere) from the anonymized data (T*) relative to his baseline knowledge (α(q,s)), which represents the information he had available before releasing the data. The principle suggests that the posteriori knowledge of the adversary (β(q, s, T*)) obtained after the data release should be approximately similar to the a priori knowledge of the adversary (α(q,s)), the information that the adversary possessed before obtaining the anonymized data. In other words, anonymized data should not reveal significantly more information than the adversary already knew before the data was released.

Theorem 3.1 *Let q be a value of the nonsensitive attribute Q in the base table T; let $q^\star$ be the generalized value of q in the published table $T^\star$; let s be a possible value of the sensitive attribute; let $n_{(q^\star,s')}$ be the number of tuples $t^\star \in T^\star$ where $t^\star[Q] = q^\star$ and $t^\star[S] = s'$; and let $f(s' \mid q^\star)$ be the conditional probability of the sensitive attribute conditioned on the fact that the nonsensitive attribute Q can be generalized to $q^\star$. Then the following relationship holds:*

$$\beta_{(q,s,T^\star)} = \frac{n_{(q^\star,s)} \frac{f(s|q)}{f(s|q^\star)}}{\sum_{s' \in S} n_{(q^\star,s')} \frac{f(s'|q)}{f(s'|q^\star)}} \qquad (1)$$





Derived from the relationship between observed belief and privacy disclosure (positive)

- Extreme situation: $\beta(q,s,T^*) \approx 1$ => positive disclosure

$$\beta_{(q^*,s,T^*)} = \frac{n_{(q^*,s)}\frac{f(s|q)}{f(s|q^*)}}{\sum_{s}\in S\, n_{(q^*,s')}\frac{f(s'|q)}{f(s'|q^*)}}$$

Minimize the contribution of other items, and make

$$n_{(q^*,s)}\frac{f(s|q)}{f(s|q^*)} \approx \sum_{s'\in S} n_{(q^*,s')}\frac{f(s'|q)}{f(s'|q^*)}$$

Possibility 1. n(q*, s') << n(q*, s) => Lack of diversity

Possibility 2. Strong background knowledge helps to eliminate other items

- Knowledge: except one s, other s' are not likely true while Q=q => f(s'|q) ≈ 0

=> $\forall s' \neq s,\quad n_{(q^*,s')}\frac{f(s'|q)}{f(s'|q^*)} \approx 0$

Negative disclosure: $\beta(q,s,T^*) \approx 0$

$n_{(q^*,s)}\frac{f(s|q)}{f(s|q^*)} \approx 0$ → either n(q*,s) ≈ 0 or f(s|q) ≈ 0

Drawbackes:

- l-diversity negatively impacts the utility of the data set → it changes the semantics of sensitive data.
- l-diversity is difficult to and could be unnecessary.
- l-diversity could be insufficient to prevent attribute disclosure → it can suffer from Skewness attack and Similarity attack.

Skewness attack:

	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

Table 1. Original Patients Table

→ Prob of cancer in the original table is low

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

Table 2. A 3-Anonymous Version of Table 1

→ Prob of cancer in the anonymized table is much higher than the global prob

Similary attack:

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

→ Salary is low

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

→ Has some kind of stomach diseas

T-CLOSENESS

The root of previous attacks to l-diversity is in the difference between the global distribution and local (in a q*-block) of sensitive values. Now we want to make e the global and the local distributions as close as possible.

Principle: An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this 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.

An observer has a prior belief B0 about an individual's sensitive attribute. A fully-generalized (with QI generalized to the most general value) table is released → the observer knows Q, the distribution of the sensitive attributes in the table ⇒ her belief evolves to B1. By knowing the quasi-identifier values of the individual, the observer is able to identify the equivalence class that the individual's record is in and learn the distribution P of sensitive attribute values in this class. The observer's belief changes to B2.

So:

- l-diversity aims at limiting the difference between B0 (priori knowledge) and B2 (posteriori knowledge) by forcing P (sensitive attribute values inside the class of equivalence where there is the individual) to have a level of diversity.
- t-closeness aims at limiting the difference B1 and B2 by keeping P as much as possible similar to Q. Intuitively, if P=Q ⇒ B1 = B2

	Utility Measure			Privacy	
Algorithms	**Query**	**Classification**	**Distribution**	**Provable Privacy**	**Robustness**
Randomization (additive)	Gaussian noise perturbs the data with a range. Hence, the impact on the data values is minimal. If the maximum and minimum of data values are within the query range, then it will support that kind of querying	Additive Gaussian noise will support classification	Gaussian noise maintains distributions	Poor	Poor
k-Anonymity	k-Anonymity supports queries, depending on the level of generalization	k-Anonymity supports classification if the equivalence classes are within the entropy	k-Anonymity maintains the distribution	1/k as the probability of identification is 1 in k records	k-Anonymity is robust but fails with homogeneous SD data
l-Diversity	*l*-Diversity does not support querying	*l*-Diversity does not support classification as it may introduce values that are not part of the classifiers	*l*-Diversity changes the distribution in search of better privacy	Supports	Not robust
t-Closeness	Querying is possible	Supports classification	Supports	Does not guarantee high privacy	Not robust

Assessment of Privacy Preserving Algorithms—Complexity

	Complexity		
Algorithms	**Computation**	**High Dimensionality**	**Vertical Scaling**
Randomization (additive)	Gaussian noise is not expensive	High dimensionality affects randomization.	Supports
k-Anonymity	Expensive	High dimensionality affects k-anonymization	Affects k-anonymization
i-Diversity	Expensive	High dimensionality affects *l*-diversity	Affects *l*-diversity
t-Closeness	Not expensive	High dimensionality does not affect t-closeness	Affects t-closeness

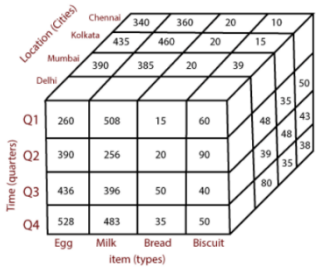
COMPLEX DATA

There exist 5 kinds of data:

- **Multidimensional Data**:
    - *Features*:
        - Correspond to Relational Data;
        - Its attributed are divided into 4 sets: Explicit Identifier (EI), Quasi-Identifier (QI), Sensitive Data (SD) and Non-Sensitive Data (NSD);
        - Records/rows are independent from each other, hence, anonymizing a few of the records will not affect the others;
        - There are two possible categories to privacy preservation techniques:
            - random perturbation methods;
            - group anonymization techniques.

- *Challenges*:
  - Difficulty in identifying the boundary between QI and SD (w.r.t. the adversary background knowledge);
  - High dimensionality of data (could make complex the privacy preservation);
  - Difficulty in achieving realistic balance between privacy and utility.

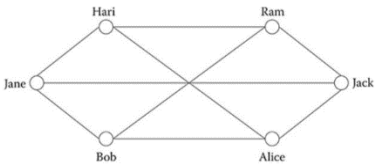
	Location="Chennai"				Location="Kolkata"				Location="Mumbai"				Location="Delhi"			
	item				item				item				item			
Time	Egg	Milk	Bread	Biscuit	Egg	Milk	Bread	Biscuit	Egg	Milk	Bread	Biscuit	Egg	Milk	Bread	Biscuit
Q1	340	360	20	10	435	460	20	15	390	385	20	39	260	508	15	60
Q2	490	490	16	50	389	385	45	35	463	366	25	48	390	256	20	90
Q3	680	583	46	43	684	490	39	48	568	594	36	39	436	396	50	40
Q4	535	694	39	38	335	365	83	35	338	484	48	80	528	483	35	50



- **Graph Data**: defined as G = (V, E), where V is a set of vertices and E a set of vertex pairs.

  - *Features*:
    - Contain many personal data and are complex data, so it could be easier to identify an element;
    - The useful information are both in the vertices and arcs.

  - *Challenges*:
    - Identity disclosure (divulgazione dell'identità): when it is possible to identify the users in the network;
    - Link disclosure: links between users are highly sensitive and can be used to identify relationships between users;
    - Content disclosure: content is associated with each node (entity). This sensitive content is classified into EI and QI.



- **Transaction Data**: these data grow horizontally (on columns).

  - *Features*:
    - Sparse high-dimensional data;
    - Sensitivity depends on the kind of product;
    - The sensitivity is in the transaction, not in single products.

  - *Challenges*:
    - High dimensionality;
    - Sparsity.

Name	P_1	P_2	P_3	P_4	P_5	P_6	P_n
Hari			1			1	
Nancy	1			1			
Jim		1					1

- **Longitudinal Data**: these data grow vertically (on rows).

  - *Features*:
    - The main goal of longitudinal data is to characterize the response of the individual to the treatment (healthcare domain);
    - Correspond to repeated measurements obtained from a single individual at different points in time;
    - Are clustered and within a cluster they are correlated and time ordered.

- *Challenges*:
  - Anonymization is not easy and aim to prevent identity and attribute disclosure.

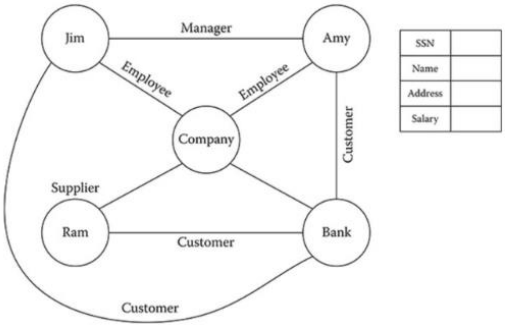
ID	Name	DOB	ZIP	Service Date	Diseases	Systolic (mmHg)	Diastolic (mmHg)
1	Bob	1976	56711	30/05/2012	Hypertension	180	95
2	Bob	1976	56711	31/05/2012	Hypertension	160	90
3	Bob	1976	56711	01/06/2012	Hypertension	140	85
4	Bob	1976	56711	02/06/2012	Hypertension	130	90
5	Bob	1976	56711	03/06/2012	Hypertension	125	85
6	Bob	1976	56711	04/06/2012	Hypertension	120	80
7	Alice	1969	56812	31/03/2012	Hypertension	160	90

- **Time Series Data**: result from taking measurements at regular intervals of time from a process sequence of observations indexed by the time of each observation). These data grow horizontally (on columns).
  - *Features*:
    - Less correlation between measurements than longitudinal data.
  - *Challenges*:
    - High dimensionality;
    - Maintaining the statistical properties, such as mean, variance, etc..

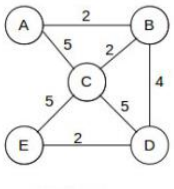
ID	Company Name	Address	Week 1	Week 2	Week 3	Week 4	Week 5
1	ABC	Park Street, 56001	10,000	12,000	17,000	8,000	11,000
2	ACME	Kings Street, 56003	15,000	17,000	18,000	20,000	21,000
3	XYZ	Main Street, 56022	20,000	23,000	25,000	26,000	30,000
4	PQR	Queen Street, 56021	14,000	18,000	19,000	19,500	21,000

GRAPH DATA

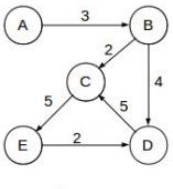
Definition: A graph G = (V, E) is a complex data structure made by a set vertices V and a set of edges E, where an edge is a pair of vertices. Each vertex is a multidimensional data.



Direction: Graphs can be directed or undirected.

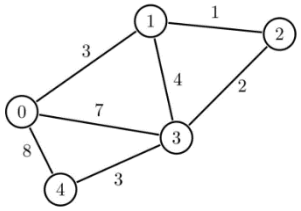


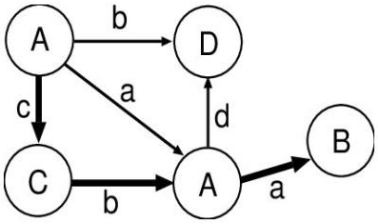
Undirected



Directed

Weights: Graphs can be weighted.





Label: Graphs can be labeld (etichettati) (in this case, edges become complex data).

ANONYMIZING GRAPHS

The data sources of a graph are: vertex properties, vertex labels and link relationships.

All these charateristics can be a problem for the graph anonymization because any change a node or a edges can change them and can also modify the utilty of the graph.

The privacy of a graph is related to its data sourses, in fact:

- Identity protection is useful for entities identification;
- Content protection is useful for entities information;
- Link protection is useful for relationships between entities.

In general, there exist three different anonymization methods:

- Naive anonymization: it consists in replacing (or removing) identifiers (which are equivalent to EI) with random values. Naive anonymization brings:
    - High utility
    - Low privacy
    - Weak against adversary external knowledge
- Random perturbation: it consists in a graph modification so it can bring an utility loss. Since the degree of a node is informative, it could allow to reveal identities, so a possible solution is the k-degree anonymity.

- Clustering: it consists in partition the nodes of the original graph transforming it in an anonymized super graph. The output of this process is a generalized graph, which consists of a set of super-nodes one for each partition and a set of super-edges which report the density of edges (in the original graph) between the partitions they connect. The advantage of the generalized graph can be used to study graph properties by randomly sampling a graph that is consistent with the generalized graph description and then performing complex analyses on it. These sampled graphs have the same properties and characteristics of the original graph.

IDENTITY PROTECTION

- K-degree anonymity: for each node *v* there exists other k − 1 nodes with the same degree as v.

  d_G = degree sequence := vector of size n (= |V|) such that $d_G(i)$ is the degree of the i-th node of G.

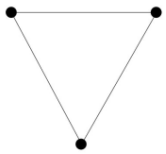
  Assumptions: elements in d_G are ordered in decreasing order

  $(d_G(1) \geq d_G(2) \geq \ldots \geq d_G(n))$.

  Definition 1: A vector of integers *v* is k-anonymous if every distinct value in v appears at least k times.

  Definition 2: A graph G (V, E) is k-degree anonymous if the degree sequence of G, d_G is k-anonymous.

  N.B. (degree = n° of edges connected to that node. In a directed graph the degree means the number of outgoing links)



3-degree anonymous graph

$$d_G = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}$$

- Graph anonymization problem: Given a graph G(V, E) and an integer k, find a k-degree anonymous graph Gˆ(V, Eˆ) with E ∩ Eˆ = E such that G_A(Gˆ, G) is minimized → it has at least one solution (same number of vertices, but the k-degree anonymous graph E^ can't be lower that the edges of the given graph)
- How to calculate the minimum number of edges to add to obtain k-degree anonymity?

$$L_1\left(\hat{\mathbf{d}} - \mathbf{d}\right) = \sum_i \left|\hat{\mathbf{d}}(i) - \mathbf{d}(i)\right|$$

- L_1 should be minimized (the number of edges we add for a node depends on the characteristics of the node)

General approach:

1. First, starting from $\mathbf{d}$, we construct a new degree sequence $\hat{\mathbf{d}}$ that is k-anonymous and such that the *degree-anonymization* cost

$$\text{DA}(\hat{\mathbf{d}}, \mathbf{d}) = L_1(\hat{\mathbf{d}} - \mathbf{d}),$$

is minimized.

2. Given the new degree sequence $\hat{\mathbf{d}}$, we then construct a graph $\hat{G}(V, \hat{E})$ such that $\mathbf{d}_{\hat{G}} = \hat{\mathbf{d}}$ and $\hat{E} \cap E = E$ (or $\hat{E} \cap E \approx E$ in the relaxed version).

CONTENT PROTECTION

Content of a node (e.g., address, zip codes, phone numbers, ...) are tuples of a relational database → you can use k-anonymity, l-diversity, t-closeness, etc..

LINK PROTECTION

Links between entities should be protected to avoid link prediction, which is possible with Machine Learning, because link prediction means to predict if two entities that are not connected now, they will be connected in the future, based on their current networks.

- Naive Anonymization: given G(V, E), $E_s \in$ E is the set of sensitive edges. Naive anonymization removes all sensitive edges. However, they can be reconstructed through the other edges. But a challenge is to identify which edges are sensitive.

- Random Perturbation: constructs an anonymized graph G' from the original graph G by:
  - Randomly selecting m existing edges;
  - Randomly adding m non-existent edges.

  The outcome obtained has a high privacy level, at the cost of utility.

GRAPHIC METRICS

- Centrality: how much a node is central in a graph. More connections a node has, respects to others, and more it is central in the graph;
- Betweenness: betweenness of a node *v* is the number of shortest paths from two other vertices *a* and *b* that pass through v;
- Closeness: closeness of a node *v* is the sum of the metric distances of *v* from all its neighboring nodes;
- Reachability: it is a property of undirected graphs. It is satisfied when there exists a sequence of nodes to reach from node *a* to node *b* which starts with *a* and ends with *b*.

Anonymization techniques affect in some way these graph metrics, in particular:

- Random perturbation changes the shape of the graph → impacts reachability, closeness and centrality;
- Clustering (on edges or vertices) changes neighboring → betweenness and closeness.

All these techniques lead to a loss of utility.

TIME SERIES DATA

ANONYMIZING TIME SERIES DATA

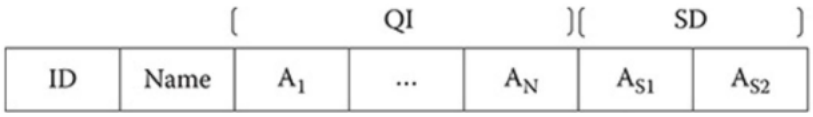
Additionally, with what we said before, Time Series Data:

- Are used for forecasting (prediction);
- Can be univariate (refers to a dataset that consists of a single variable observed over a period of time. It captures the values of a single attribute or feature at different time points) or multivariate (involves multiple variables observed over time. It captures the values of multiple attributes or features simultaneously at different time points)
- They are represented both in time domain and frequency domain.

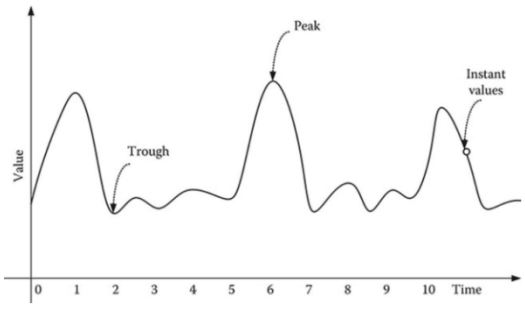
A Time Series data set contains: EI (SSN, names, …), QI (contain a series of time-related that SHOULD be anonymized), SD (series of time-related data that SHOULD NOT be anonymized).

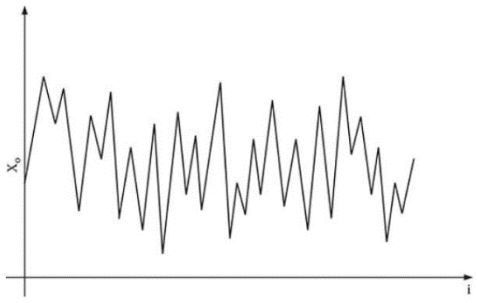
CHALLENGES IN PRIVACY PRESERVATION OF TIME SERIES DATA

- High Dimensionality (univariate time series data of 500 values has 500 dimension to choose from);
- Background Knowledge of the Adversary which is impossible to model. Its difficult to identify the boundary between QI and SD when the dimension is high.

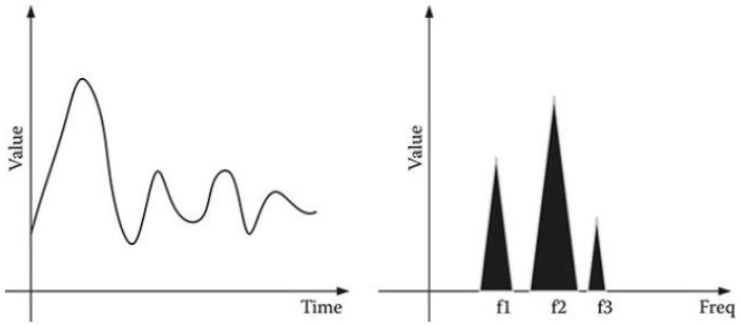


PATTERN AND PROPERTIES PRESERVATION





Pattern **Statistical properties**



FREQUENCY-DOMAIN PROPERTIES

Patterns, statistical properties (mean, variance, …) and correlation between time and frequency domains should be maintained/grated/preserved after anonymization.

To preserve the privacy of subjects:

- EI are removed.
- SD MUST be kept original.
- QI must be anonymized.

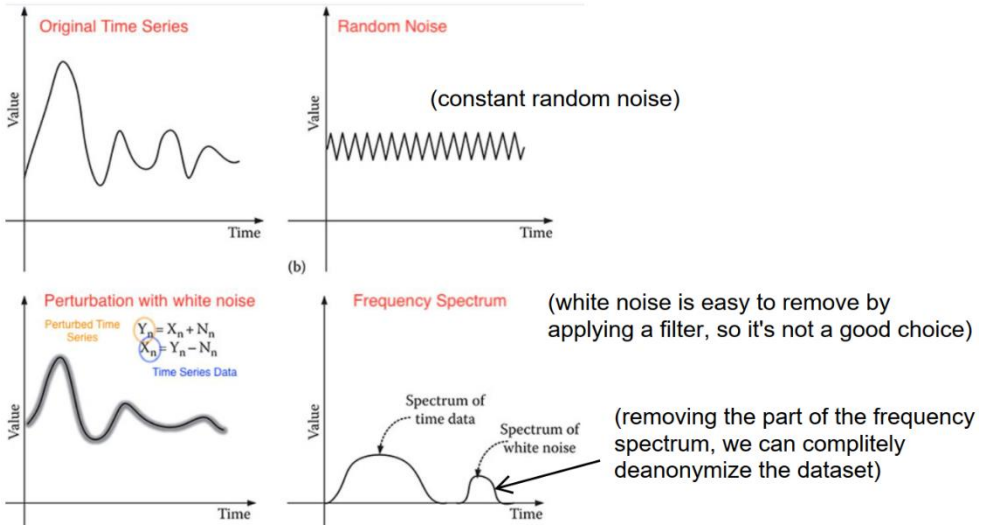
In Time Series Data, the data protection methods can be divided in two categories:

- Perturbative methods: possible introducing some additive random noise. Two ways:
    - White noise
    - Correlated noise
- Generalization: by k-anonymization.

WHITE NOISE

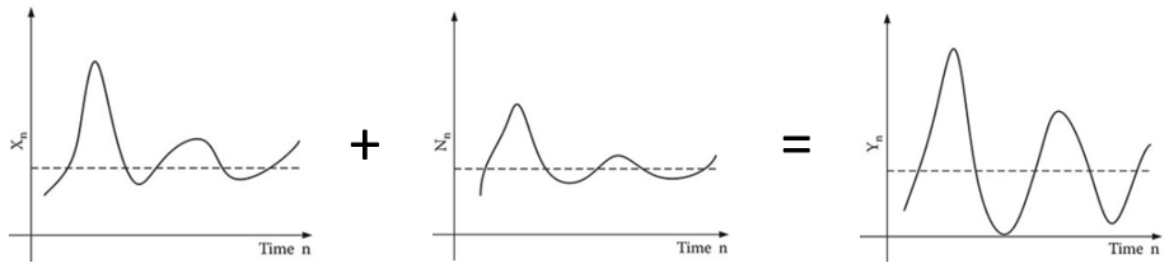
QI are perturbed with high-frequency white noise (i.e., by adding/removing random values). Additive white noise is the simplest way but the weakest in terms of privacy because re-identification can be obtained by filtering or regression.

As advantage, white noise is optimal in terms of utility, because the anonymized time series maintains statistical properties, the pattern and the frequency domain properties.



CORRELATED NOISE

It allows to avoid filtering attacks but it changes pattern and frequency of the time series data. Re-identification is possible through regression.



GENERALIZATION: (K, P)-ANONYMITY

(k, P)-Anonymity is an approach that aims at granting both k-anonymity and pattern preservation.

In (k, P)-Anonymity, the goal is to ensure that each time series in a dataset is indistinguishable from at least k-1 other time series, while also satisfying a privacy parameter P. The parameter P determines the maximum allowable difference between the original values and the generalized values.

The choice of k and P in (k, P)-Anonymity should be determined based on the specific privacy requirements and the sensitivity of the data being anonymized.

ANONYMIZING LONGITUDINAL DATA

As we said before, the characteristics of Longitudinal Data are:

- Data are clustered and comprise repeated measurements from a single individual;
- Records are still divided into EI, QI and SD but strong relation between the patient and the SD and a strong correlation among records in the cluster;
- Data in the cluster have a temporal order that implies the presence of a pattern (i.e., how the patient reacts to a treatment) in the data.

CHALLENGES IN ANONYMIZING LONGITUDINAL DATA

- Identity disclosure → prevent record linkage.
- Attribute disclosure → prevent sensitive data linkage.
- Correlation and dependency between records;
- Patterns;
- Unknown background knowledge of the adversary.

ANONYMIZING TRANSACTION DATA

The characteristics of Transaction Data are:

- EI are not part of the transaction data table (not interest on the particular record, but just to QI);
- The table is sparse: very few cells have entries in this high-dimensional space;
- There are few sensitive transactions that are classified as sensitive;
- There is no fixed length for QIs and SD;
- A large set of transactions considered non-sensitive data from the QI data set;
- QIs have very high dimensions;
- The sensitivity in the transaction needs to be protected.

CHALLENGES IN ANONYMIZING TRANSACTION DATA

K-anonymity and l-diversity are not suitable as they lead to high information loss, because columns grown over the time.

THREATS TO ANONYMIZED

Threat modeling helps in identifying possible threats to the system. Identifying threats is key to building an appropriate protection mechanism. With data privacy, threat models include a broad range of de-anonymization attacks.

There are 3 different threat levels to analyze: location and user complexity (adversary) → Background and external knowledge, data structure complexity dimensionality, sparsity, clusters... and anonymization algorithm.

Aim: I want to understand the sensitivity of data and disclosure risk for a given environment and setting.

We make a distinction between external knowledge and background knowledge:

- External knowledge is obtained from external sources (OSINT).
- Background knowledge is the information an adversary has about an individual or individuals in the data set (identify attributes (QI), distribution (statistical) of identify attributes, values of some sensitive data, distribution of sensitive attribute values, knowledge of the anonymization algorithm used for data protection, outliers in the data, associations in the data).

The background information an internal adversary possesses will be higher than an external adversary does. An internal adversary has more background information than an external adversary. Also, with today's social networks QI and SD are difficult to separate and depend on the individual. In general, internal adversary has application and organizational context, background knowledge and external knowledge, instead of external adversary will rely heavily on (farà molto affidamento su) external knowledge and he may or not may have background knowledge. If we talk about offshore internal (tipo specifico di minaccia interna che proviene da fonti esterne all'organizzazione, ma coinvolge individui o gruppi che agiscono come se fossero interni all'azienda stessa perchè hanno accesso privilegiato all'infrastruttura o ai sistemi dell'azienda, dando loro la capacità di agire internamente e accedere a informazioni riservate o sistemi sensibili) for sure he will have some or limited application context and will rely on external knowledge (no background knowledge). If we talk about outsourced (same geo) for sure he will have some or limited application context, he is from same geo so he will have geo context and will rely on external knowledge. If we talk about outsourced (other geo) for sure he will not have any kind of context and it's least harmful (meno dannoso) when data is anonymized.

Enterprise's data can be stored with different data structures (multidimensional, text, time series, graph data, ...). Privacy preservation of complex data structures is a challenge and an open problem as they provide an adversary more info for attack.

- Multidimensional data.

  Multidimensional data can be considered as an n×m matrix, where n is the number of records and m is the number of attributes or columns. Relational data represented as multidimensional data are the most widely used data structure. Multidimensional data have three disjoint sets of data: explicit identifiers (EI), quasi-identifiers (QI), and sensitive data (SD):

  - EIs by default are completely masked (perturbed).
  - QIs are anonymized.
  - SD are left as is to enable analysis.

  Most of the attacks are directed toward QI (identity disclosure) and SD (attribute disclosure).

Multidimensional Data and Attack Types

Target	Attack Type
Identity (quasi-identifiers)	Linkage attacks—links to external data sources. This happens when QI attributes can be linked to an external data source.
	Background knowledge attacks.
	Inference attacks—An adversary knows that the record owner is present in the data set.
	Data distribution attack—An adversary has knowledge about the statistical distribution of QI attributes.
	Outlier identification, for example, only Asian in the population.
	Probabilistic attacks.

Multidimensional Data and Attack Types

Target	Attack Type
Attribute (sensitive data)	Homogeneity attacks—Presence of clusters in sensitive data records.
	Background knowledge attack—An adversary has knowledge about a record owner's QIs and knows that he or she is in the data set and some aspect of SD and hence can infer. For example, the adversary knows that Alice (record owner) smokes heavily and also some of her QIs. With this background information, the adversary can infer that Alice suffers from lung cancer by referring to the released medical records.
	Association attack—An adversary is able to identify shopping patterns of a record owner with the help of background information of the record owner.
	Data distribution attacks.
	Outlier identification.

- Graph data.

  Graph is a very complex data structure. A graph G (V, E) has many vertices V, linked through a set of edges E. Many dimensions in a graph that can be exploited by an adversary and graph's structural information can be used to attack the graph (like, for example, vertices, sensitive vertex labels, edge labels).

Graph of Data and Attack Types

Target	Attack Type
Identity—vertex existence	An adversary can use the vertex degree or node degree to identify the existence of a particular individual in the graph network.
Identity—sensitive vertex labels	A vertex represents an individual in a social network. An individual has associated personally identifiable information and sensitive data. The individual can be reidentified using different attack techniques:
	Identity disclosure—linkage attacks.
	Identity disclosure—background knowledge attacks.
	Sensitive attribute disclosure—background knowledge attacks.
	Background knowledge consists of both attribute knowledge and structural information—attributes of vertices, specific link relationships between some target individuals, vertex degrees, neighborhoods of some target individuals, embedded subgraphs, and graph metrics.

Graph of Data and Attack Types

Target	Attack Type
Link relationship	Background knowledge attacks—an adversary who attacks a graph always has some background knowledge of the network and the individuals in the network and some properties of the network without which it is difficult to attack an anonymized network. Re-identifying link relationships is generally based on the knowledge of the individuals in the network.
Identity and link relationship identification	Cross-reference attacks—an adversary who wants to identify individuals and their link relationships in an anonymized network G_{main} can use an auxiliary network G_{aux}. The adversary has background knowledge that the individuals are also members of the auxiliary network. He uses this information to cross-refer with G_{main} to identify the individuals and their relationships.
Identity and link relationship identification	Neighborhoods—an adversary has background knowledge of the neighborhood of a target individual.
Identity and link relationship identification	Graph metrics—an adversary uses graph metrics such as closeness, betweenness, degree, centrality, and so on to identify individuals in the network.

They're used in domains as social networks, electronics, transportation, software, and telecom.
Problems:

- Identity disclosure. It occurs when it is possible to identify the users in the network.
- Link disclosure. Links between users are highly sensitive and can be used to identify relationships between users.
- Content disclosure. Just as in relational table, sensitive content is associated with each node (entity). This sensitive content is classified into explicit identifiers like name, SSN, and QI, such as demographics, gender, date of birth, and other sensitive data such as preferences and relationships.

It is important to maintain the statistical properties of the original time series data like mean, variance, and so on.

- Time series.

  Time series data are characterized by high dimensionality, pattern and frequency-domain characteristics. Anonymization techniques should ensure that all these characteristics are preserved in the anonymized data set. Generally, the attacks focus on identity, pattern, and time series values.

Time Series Data and Attack Types

Target	Attack Type
Identity	Even though EIs are masked, one could re-identify them using QI attributes. Background information about an entity can be used to re-identify. For example, a patient who has undergone an ECG would not want to reveal or publish his ECG values to others as he feels it is his personal data. But if an adversary knows that the patient has undergone an ECG test, then this is itself a loss of privacy even though the adversary has no knowledge of the ECG values. This is the fundamental difference between anonymity and privacy.
Time series patterns	A time series has a pattern. For example, a car rental company will have a maximum sale during the holiday season. An adversary having this kind of background knowledge will be able to re-identify.
Time series values	Filtering attacks—time series data perturbed with white noise can be subjected to filtering attacks. Specialized filters can be used to remove the noise and reveal the time series values.
Time series values	Regression attacks—time series data perturbed with correlated noise can be subjected to regression attacks. An adversary with some specific values of the time series can build a regression model to predict the values of the time series.

Differences between TS and LD:

- Time series data result from taking measurements at regular intervals of time from a process.
- Longitudinal data has stronger correlations between measurements w.r.t. time series.

- Longitudinal data have very small dimensions compared to time series data that have high dimensional and keep growing.

- Longitudinal data.

  Longitudinal data are extensively used in healthcare domain, especially in clinical trials because they are a series of measurements taken over a period of time from a patient in response to medication or treatment. They are sensitive data and it's so difficult to anonymize them. The threats to longitudinal data occur on both identifying attributes and sensitive data.

Longitudinal Data and Attack Types

Target	Attack Type
Identity	Record linkage attacks.
	The identity of a record owner can be reidentified using external data if QIs are not anonymized properly.
Sensitive data	Background knowledge.
	An adversary having background knowledge of a patient such as admission date, disease, and so on can re-identify the record owner.
Sensitive data	Probabilistic attack.

- Transaction data.

  Transaction data hold transactions of customers and they are characterized by high dimensionality and sparsity:

  - High dimensionality means that there are too many columns or attributes in the database.
  - Sparsity means that each individual record contains values only for a small percentage of the columns.

  Sensitivity depends on the kind of product and it is in the transaction, not in single products.

  Sparsity increases the probability that re-identification succeeds, reduces the amount of background knowledge required in the re-identification process and makes it difficult to effectively anonymize transaction data and balance privacy against utility.

Transaction Data and Attack Types

Target	Attack Type
Identity	Removing identification information is not sufficient for anonymity. An adversary having background knowledge of a target individual and her shopping preferences could still be able to re-identify.
Sensitive transaction	Background knowledge attacks—an adversary having some background knowledge of a target individual will be able to find the sensitive transaction.

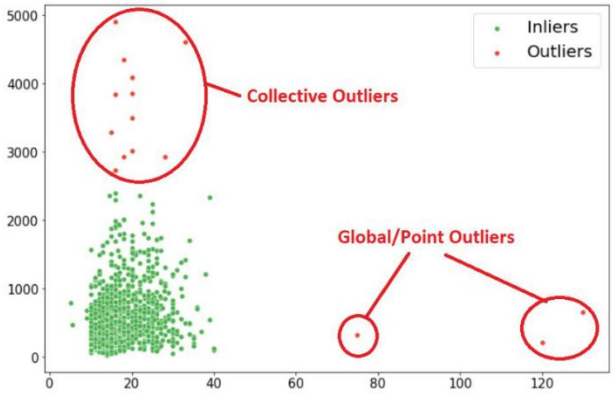
PRIVACY PRESERVING DATA MINING

Massive amounts of data are being collected by companies in different ways: online trackers, smart devices, ... These data are an asset for the companies, and they are mined (estratti/prelevati) to extract knowledge.

Data mining is a process where critical business data are analyzed to gain new insights (approfondimenti) about customers, businesses, and markets. This new knowledge gained can be used to:

- improve customer relationships,
- improve website navigation,
- define advertising plans,
- produce better-quality products and services.

Mined data are generally stored in a relational or multidimensional format and stored in companies' central data warehouses. Nowadays some enterprises started to use different data structures, such as: graphs data, time series data, longitudinal data, big data, etc…



SECURITY REQUIREMENTS

Different data repositories are needed to store all these diverse data, in fact analytics is carried out on the data in the repositories. The access to these data repositories is strictly controlled by access control rights and it is subject to strong security measures these data are very sensitive and contain customer-identifying information.

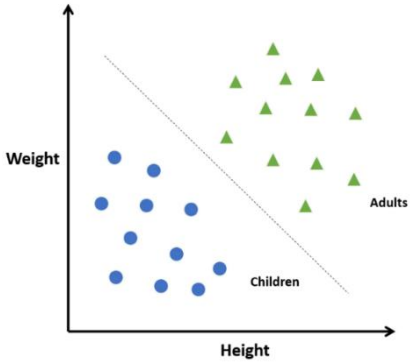
Companies need to ensure that the data are anonymized before being used for analytics or mining and also they have to be protected in the case they are shared with specialized analytics businesses.

KEY FEATURES OF DATA MINING

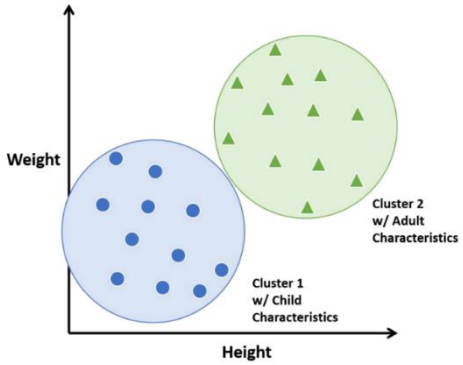
The goal of data mining is to extract knowledge from the data. Some of the key functions of data mining are:

- Clustering → partitioning a data set into clusters of similar data;
- Classification → used for prediction. In predictive modeling, a model is built to predict a value of a single variable based on the values of the other variables;
- Association rule → find associations between the transactions of a customer;
- Outliers → identifying outlying data, that is, the data whose value is way outside or away from other data values.

Classification



Clustering



Outliers

THREATS

Clustering, classification, and association rule mining, generate an output that does not contain any customer data but generalized models, it means that there are no threats (no risk) to de-identification.

However, they should be protected in any case because:

- they could be provided to third parties;
- it is impossible to make assumptions on the background knowledge of an attacker;
- regulatory compliance needs.

ASSOCIATION RULE MINING

Goal: find associations between the transactions of a customer.

Problem: find relationships among items in a database D.

A relation is defined as follows:

$$X \rightarrow Y \text{ where } X \subseteq I, Y \subseteq I \text{ and } X \cap Y = \emptyset$$

Where:

I = $\{i_1, i_2, \dots, i_m\}$ be a set of items

T = $\{t_1, t_2, \dots, t_n\}$ be a set of transactions on the database, where $t_i \subseteq$ I, t_i is a subsets of the available items

Support := the number of transactions containing X.

Low support implies that the transaction randomly occurs → a minimum support (minSup) should be defined to prune rare transactions.

Confidence := the percentage of transactions in T that contain X and that also contain Y.

Low confidence implies that it is impossible to predict Y from X → minConf should be defined to remove weak associations.

Additional note:

A privacy risk, for a transaction database, comes only when it is linked or joined with customer identity data. When a transaction table is associated with customer data, then the table becomes sensitive.

Example of association rule mining:

Sample Transaction Database

Transaction ID	Bread i_1	Bread i_2	Butter	Eggs	Milk	Chocolate	Cheese	Flour	Beer	Meat i_m
t_1	1		1	1		1	1			
t_2	1		1	1		1		1		
t_3			1	1						
t_4	1		1	1						
t_5	1		1				1			
t_6	1		1						1	1
t_7	1		1	1	1			1		
t_8				1		1				

Consider:

- I = {Bread, butter, eggs, cheese, ...}
- T = {t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8}
- X = {Bread, butter}
- Y = {Eggs}

We have that:

- Support = 6/8 (t_1, t_2, t_4, t_5, t_6, t_7)
- Confidence = 3/6 (t_1, t_4, t_7)

t_1	Bread, butter, eggs, cheese, chocolates
t_2	Chocolates, bread, butter, cheese
t_3	Eggs, flour, butter
t_4	Bread, butter, eggs
t_5	Bread, butter, cheese
t_6	Bread, butter, meat, beer
t_7	Bread, butter, eggs, milk
t_8	Eggs, flour, chocolates

CLUSTERING

Data clustering is a method of creating groups of objects in such a way that objects in one cluster are very similar and objects in different clusters are quite distinct (data clustering is also referred to as unsupervised learning in ML). Each cluster has a center point.

The goal of clustering is to find the intrinsic grouping of data for which a distance function is used; simpler by finding the optimal clustering solution, we aim to maximize the similarity within clusters and minimize the similarity between different clusters.

A cluster is made by all the data that has an Euclidean distance less than a given threshold. Consider that the mean of a group is denoted by m_i and the data in the group is denoted by x_i.

The Euclidean distance between x_i and m_i is: $dist(x_i, m_i) = ||x_i - m_i|| = \sqrt{\sum(x_i - m_i)^2}$.

Cluster quality: indicates that similar data points form a cluster and dissimilar points are in different groups of clusters.

Data points like one another and also close to the mean come together to form a cluster.

A cluster quality is controlled by the:

- Similarity measure: similar data points constitute a cluster, and dissimilar points are not in the same cluster;
- Center point;
- Distance measure (Euclidean distance);
- Structure.

These aspects of cluster quality are important when privacy preservation techniques are applied before clustering. When an organization wants to carry out data mining activities such as clustering, they generally outsource the task to specialized analytics firms. Outsourcing data has a major issue: data need to be protected before outsourcing.

PRIVACY PRESERVING TEST DATA

Testing is an important part of the systems development life cycle (SDLC). The quality of software application depends on the quality of testing.

Software testing is a process, or a series of processes, designed to ensure that computer code does what it was designed for and not anything else. High-quality testing requires high-quality data.

Problem of test outsourcing: guarantee that the entities of data that are used, are protected at a certain level while retaining testing efficacy.

To anonymize Multidimensional Data for testing perturbative data anonymization techniques are preferred, like transformation, rotation, and noise addition.

Testing Data can be subdivided in:

- Functional Testing: system testing is aimed at evaluating specific parts of the system w.r.t. specific use cases;
- Non-Functional testing: stress or load, scalability, responsiveness, reliability, security, etc., look at various nonfunctional characteristics of the software system.

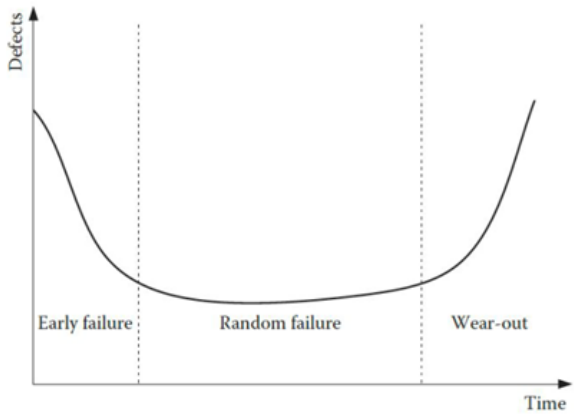
Good test case requires testing all possible inputs but that's impossible due to the fact that data are limited.

The best source of data is production data, but they must be anonymized as they carry personal information.

TEST DATA AND RELIABILITY

Any device or software has three phases of reliability:

- Early failure: the software exhibits many defects that are uncovered by initial rounds of testing. Defects are high at this time;
- Random failure: the software stabilizes and exhibits a steady state with random defects that get discovered intermittently, while minor updates are made to the software system;
- Wear-out: the software system is becoming old and needs updating to keep up with changes in policy/business/technology that require numerous amendments (modifiche). During this phase, again the number of defects begins to increase and continues to do so until the software system is unable to adapt any further and is abandoned.



As test environments do not get the same kinds of resources as production, there is the need to reduce the amount of data and to do it data should be sampled. These subsets of data are picked from original data in a way that they represent the entire data in syntax, semantics and statistics.

UTILITY OF TEST DATA: COVERAGE

Coverage is a quality assurance metric that determines how thoroughly a test suite exercises a given program. The loss of test coverage can be measured using the lines of code that were covered using original and anonymized data.

The utility loss of test coverage can be measured in terms of lines of codes that can be tested using the original and the anonymized data.

Example:

Let's suppose we have a software S to test with data D.

For each test phase t the test coverage (with N the total number of test cases) is given by the formula:

$TC(t) = \sum_{i=1}^{N} tc(t_i)$ where the test case tc_i is $\rightarrow$ $tc_i = f(S_i, d_i)$

S_i = system being tested | d_i = data provisioned

Utility metrics:

- Syntax
- Semantics
- Statistics

Test data anonymization starts from defining EI, QI and SD:

- EI should be anonymized;
- QI depends on the application context;
- SD represent the facts in a software system;
- Outliers: they have significance in test data as they are important since they tend to invoke parts of code that do not execute often.

(slide 14 ???)

PRIVACY PRESERVATION OF TEST DATA: EI

EI must be masked, however, they cannot be randomly removed when dealing with test data. Instead, the masking algorithms should grant, on the anonymized test data:

- Referential integrity: if there are primary keys that appear as foreign keys in other tables, we to care to propagate the same masked value to all respective rows of tables where this field appears, which is essential to preserve data integrity;
- Consistency: if there are semantic relationships between EI and QI, such relationship these must be maintained.

Example:

Sample Salary Data Table

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

Referential Integrity

ID	Name	ID	Designation
12345	John	12345	Project Manager
56789	Harry	56789	Architect
52131	Hari	52131	Developer II
85438	Mary	85438	Program Mgr
91281	Srini	91281	Tester I
11253	Chan	11253	Consultant

Consistency

ID	First Name	....	....	Full Name
12345	John	....	....	John Bailey
56789	Harry	....	....	Harry Wagner
52131	Hari	....	....	Hari Krishna
85438	Mary	....	....	Mary Allen
91281	Srini	....	....	Srini Iyengar
11253	Chan	....	....	Chan Nair

DATA GENERATION

There are areas where anonymization techniques are not secure enough to protect against threats. The alternative, in some case, is to generate synthetic data, information that's artificially generated rather than produced by real-world events. It is observed that although easy to create (sebbene siano facili da creare), artificial data can lead to (portare) results that are either significantly different or opposite in certain cases. The necessity of synthetic data arises due to high volume, high sensitivity of data, no historical data availability, and bridging incomplete data.

Steps:

- Classify the metadata.
- Create the model according to rules that are created using metadata information, referential integrity constraints, dependent and independent variables, correlated fields data, domain context and application scenarios. The model represents the application context in the form of instructions that help to produce useful data.

EI GENERATION

Privacy is more important to preserve in the case of EIs than utility. In data mining, specific names, social security numbers (SSNs), and phone numbers are not important. It is the collective profile of record owners that are important, which cannot be obtained from EIs.

SDG Technique	Brief Explanation
Substitution	Prepopulated sets of data are created. For example, first name, last name, and middle name that are directly substituted for original data. Substitution is difficult to implement when consistency is a requirement due to the randomness involved in picking the replacement.
Credit card, social security number, aadhar number	Format is preserved while replacing original digits and characters with randomly generated ones. The randomness needs to be carefully introduced wherever the meaning of data is not impacted. A PAN number in India has meaning for some characters in it. For example, the fourth character "P" in AAZPE3479P means the PAN belongs to an individual, whereas a 'C' in the same place means it belongs to a company.
E-mail address	Based on standards being followed, either an e-mail address is generated for entire record set uniquely or a common e-mail address is assigned to each row.
Mobile phone numbers	Most often, mobile numbers are 10-digit numbers that can be generated using a random number generator.
Flat value	A single value is assigned to all rows within a set.

QI GENERATION

Generate EIs requires only conforming to the syntax of the field instead of generating QIs requires also to preserve the semantics as well (mean, variances, boundary values).

We must remember that QIs are data that help identify a record owner when combined with background knowledge or external data sources and their utility is linked to the preservation of the correlation with SD fields.

SDG Technique	Brief Explanation
Randomization	A random value is generated to replace the original value. Based on the data format and range of original data, a number is randomly picked. This technique is applicable to date fields specifically.
Geographical area	Addresses and zip codes can be generated to be valid zip codes, or if the data are too clustered (like the voters list of a particular city), geographical area can be substituted against a zip code. Of course, the prerequisite is that altering the format of data is permissible.
Generalization	Generalizing a set of nominal categorical data values like making master's degree or PhD into postgraduate.

SD GENERATION

Sensitive data form the most important ingredient of a data set. In data mining, facts are always attributed to sensitive data while profiling is attributed to QI. Numerical SD attributes are easier to generate. It is always desirable to have synthetic data created as close to original values as possible but however, from a privacy preservation perspective, the closer the SD to original values, the greater the privacy concerns are.