Data Protection and Privacy

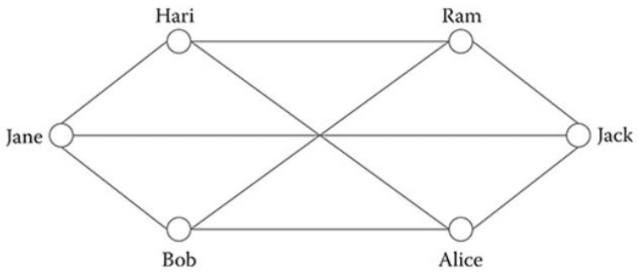
University of Genoa

Lesson 3: Complex data

Gaspare Ferraro <ferraro@gaspa.re>

Types of Data: Multidimensional Data

- Features
  - Multidimensional Data = Relational Data
  - Attributed are divided into 4 sets: EI, QI, SD and NSD
  - Each record or row is independent of others; therefore, anonymizing a few of the records will not affect the others
  - Anonymizing a tuple in a record will not affect other tuples in the record
  - Privacy preservation techniques are divided into two categories: random perturbation methods and group anonymization techniques
- Challenges
  - Difficulty in identifying the boundary between QI and SD (w.r.t. the adversary background knowledge)
  - High dimensionality of data poses a big challenge to privacy preservation
  - Clusters in sensitive data set
  - Difficulty in achieving realistic balance between privacy and utility

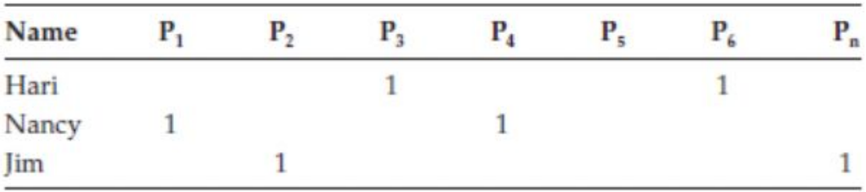


Types of Data: Graph Data

è piu difficile anonimizzare i dati perche ci sono dati che non sono dei singoli nodi ma sono info tra le relazioni dei nodi e quini le info non sono dei nodi ma nel mezzo

- Features
  - A graph: G = (V, E), where V is a set of vertices and E a set of vertex pairs
  - Used in domains as social networks, electronics, transportation, software, and telecom
  - Contain many personal data and are complex -> easier to identify
- Challenges
  - 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

Types of Data: Transaction Data



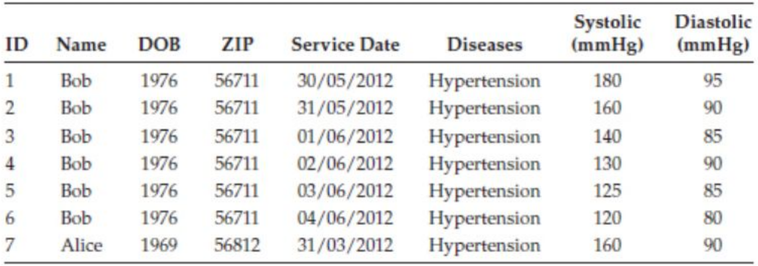
Name	P_1	P_2	P_3	P_4	P_5	P_6	P_n
Hari			1			1	
Nancy	1			1			
Jim		1					1

- Features
  - they hold transactions of customers
  - 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
  - Conventional privacy preservation techniques used for relational tables that have fixed schema are not applicable on transaction data

comportamento nel tempo

Types of Data: Longitudinal Data

crescono verticalmente e non orizzontalmente



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

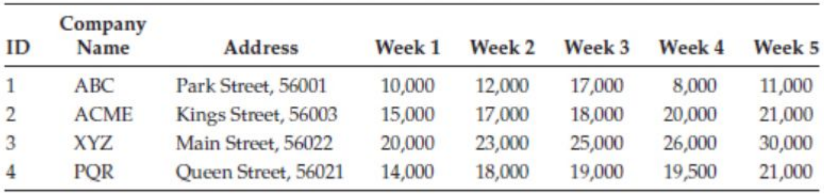
- Features

puo diventare facilmente molto grande

  - Typical of healthcare domain. The goal of longitudinal study is to characterize the response of the individual to the treatment
  - Data are clustered, composed of repeated measurements obtained from a single individual at different points in time
  - The data within the cluster are correlated and have a temporal order
- Challenges
  - The characteristics of longitudinal data in the anonymized data set D should be maintained
  - Anonymization designs aim to prevent identity and attribute disclosure

crescono orizzontalmente ma non verticalmente= le righe sono fisse le colonne indicano il passare del tempo

Types of Data: Time Series Data



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

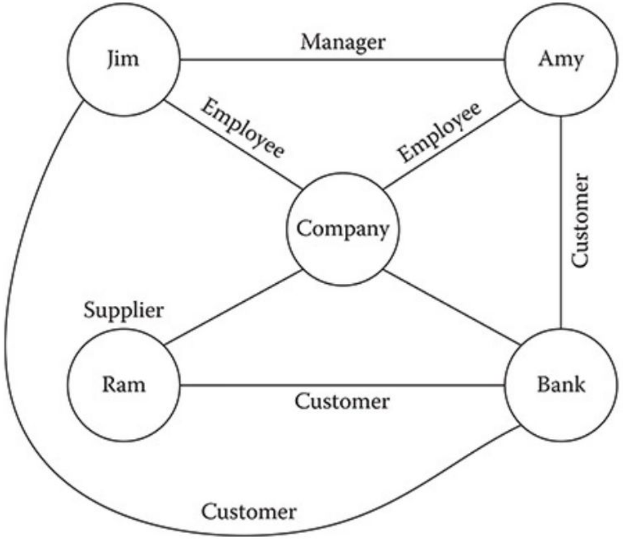
crescono con il tempo

- Features
  - 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
- Challenges
  - High dimensionality
  - Retaining the statistical properties of the original time series data like mean, variance, and so on
  - Supporting various types of queries like range query or pattern matching query
  - Preventing identity disclosure and linkage attacks

Graph data

V set vertix(nodi-rishe) edges(unioni tra i nodi)

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 (i.e. e ∈ E, s.t. e ∈ V×V).

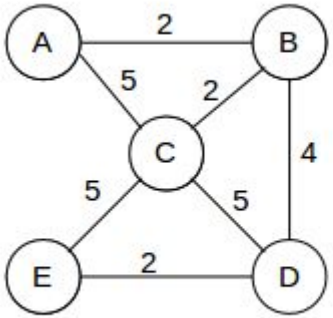


SSN	
Name	
Address	
Salary	

le righe non sono piu indipendenti ma ci sono connesioni tra diverse righe

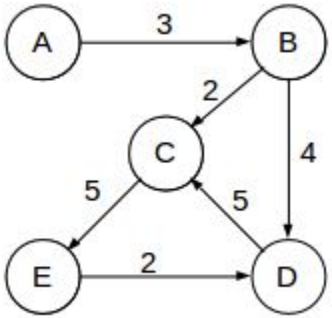
Graph data

Edges, and graphs, can be directed or undirected



Undirected

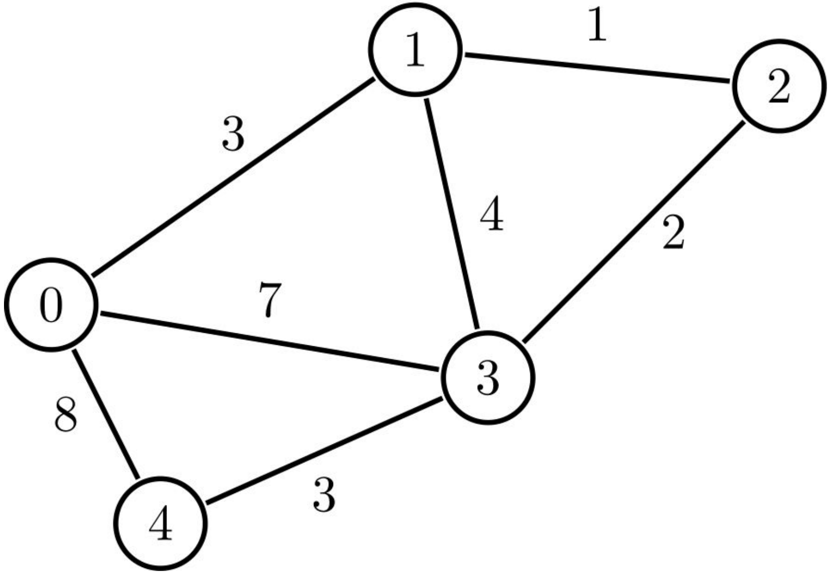
si possono trasformere in directed mettendo entrambi i versi in ogni segmento



Directed

Graph data

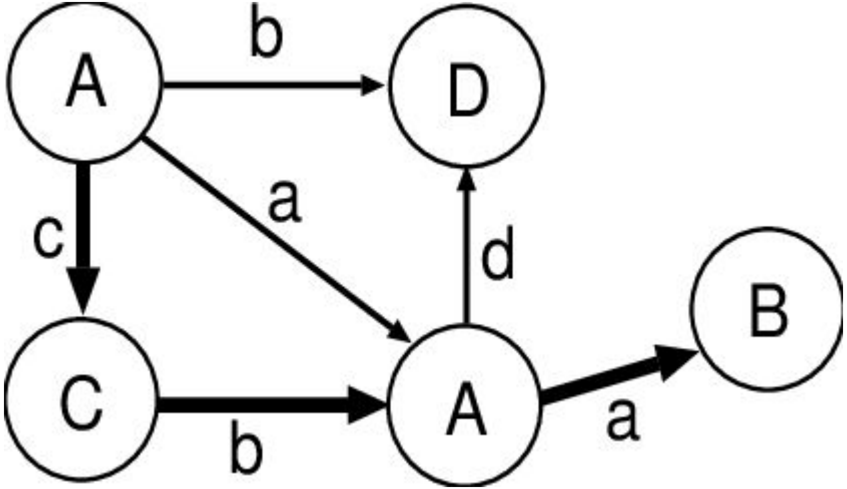
Edges, and graphs, can also be weighted



Graph data

Edges, and graphs, can also be labeled

nel senso che l'unione puo contenere dati, puo essere piu complicato rispetto ad una generica connesione



Anonymizing Graphs I

- Graph: data sources
  - Vertex properties, Vertex labels, Link relationships
  - Graph metrics (betweenness, closeness, centrality, path length, and reachability)
- Graphs vs. Multidimensional data
  - In multidimensional data each record or tuple can be transformed independent of each other.
  - Because of its simple structure, it is easier to prevent both identity and attribute disclosure;
  - Problem: with graph data, there are more dimensions to be taken care of → any change in the nodes or edges affects the characteristics of the graph and also the utility of the anonymized graph.

Anonymizing Graphs II

- Privacy of graph: challenges
  - Identity protection → entities identification (Multid. equivalent: EI)
  - Content protection → entities information (Multid. equivalent: QI, SD)
  - Link protection → Relationship between entities (Multid. equivalent: None)
- Anonymization methods
  - Naive anonymization → effective when adversary has no background knowledge
  - Random perturbation → graph modification → utility loss
  - Clustering → grouping similar objects (i.e., vertices, edges, vertices-edges)

clustering =anonimizzare raggruppando gli utenti sotto una generalizzazione

Identity Protection I

- Naive Anonymization
    - Identifiers (EI) are replaced by random values
        - high utility
        - low privacy
        - Weak against external knowledge
- Graph modification
    - Idea: the degree of a node is informative (i.e., social networks)
        - it could allow to reveal identities
    - k-degree Anonymity
        - for each node v there exists other k − 1 nodes with the same degree as v.

anonimizzare togliere info (dai nodi) in una struttura a dati non è sensato poiche la struttura stessa puo fornire molte info

degree=numero di nodi connessi

Identity Protection II

quando trattiamo grafi molto grandi come puo essere quello di facebook con 3 miliardi di nodi. per eseguire operazioni su grafi cosi grandi si lavora sempre con clustering quindi grafi semplificati raggruppando i nodi con una caratteristica in comune

- Clustering
  - Anonymized super-graphs:
    - Robust w.r.t. naive anonymization.
    - it anonymizes a graph by partitioning nodes and then describing the graph at the level of partitions.
    - The output 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.
  - Advantages:
    - 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 a standard analysis on this synthetic graph.
    - These sampled graphs retain key properties of the original degree distribution, path lengths, and transitivity allowing complex analyses.

applichiamo i concetti visti sui grafi.

K-degree Anonymity I

- Problem formulation: Let G(V, E) be a simple graph; V a set of nodes and E a set of edges in G. $\mathbf{d}_G$ denotes the degree sequence of G. That is, $\mathbf{d}_G$ is a vector of size n = |V| such that $\mathbf{d}_G(i)$ is the degree of the ith node of G.
- Assumptions: Entries in d are ordered in decreasing order of the degrees they correspond to, that is, $\mathbf{d}_G(1) \geq \mathbf{d}_G(2) \geq \dots \geq \mathbf{d}_G(n)$. Additionally for i < j, **d**[i,j] denotes the subsequences of dG that contains elements i, i+1, ..., j-1, k
- Definition 1: A vector of integers v in 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, $\mathbf{d}_G$ is k-anonymous.

Example of K-degree Anonymous graphs

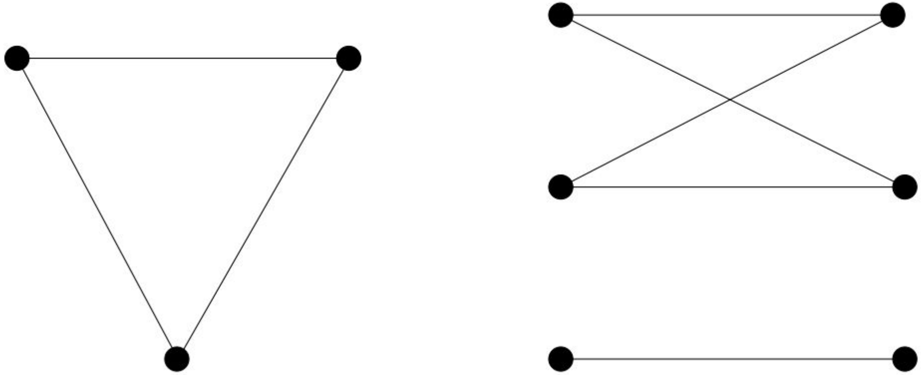


Figure 1: Examples of a 3-degree anonymous graph (left) and a 2-degree anonymous graph (right).

K-degree Anonymity II

- 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
- How to calculate the minimum number of edges to add to obtain k-degree anonymity?

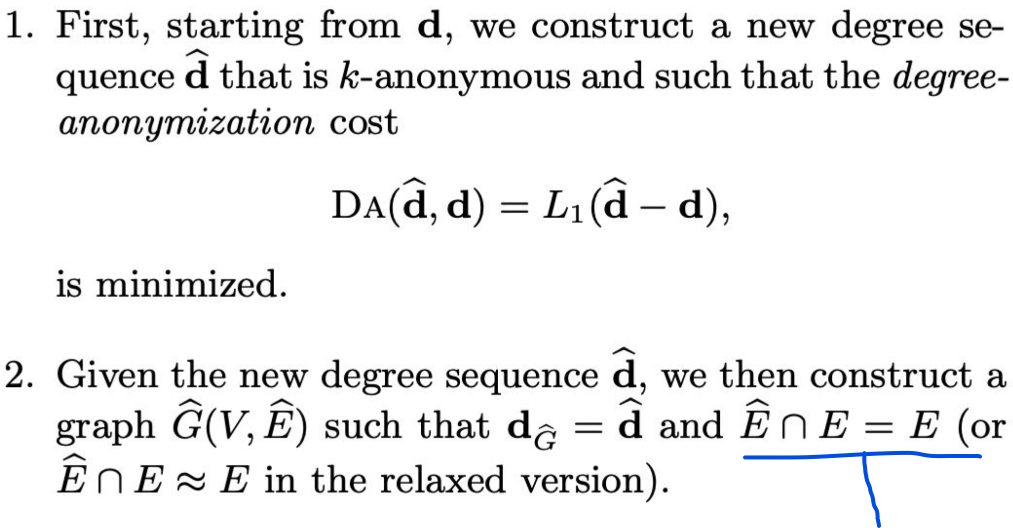
$$L_1\left(\widehat{\mathbf{d}}-\mathbf{d}\right)=\sum_i\left|\widehat{\mathbf{d}}(i)-\mathbf{d}(i)\right|$$

- L_1 should be minimized

$$\mathrm{GA}(\widehat{G},G)=\left|\widehat{E}\right|-|E| \quad = \quad \frac{1}{2}L_1\left(\widehat{\mathbf{d}}-\mathbf{d}\right)$$

K-degree Anonymity III

General approach:



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

$$\mathrm{DA}(\widehat{\mathbf{d}}, \mathbf{d}) = L_1(\widehat{\mathbf{d}} - \mathbf{d}),$$

is minimized.

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

se togliessi edge l'uguaglianza non sarebbe verificata intersezione tra nuovi edges e vecchi edges = vecchi edges

Link Protection I

Link prediction:predizione sulla probabilita di unione tra due nodi (saranno amici su facebook due utenti in futuro). quindi se tu elimini edges ma poi questi sono facili da indovinare

- Notes on Content Protection:
  - Content of a node (e.g., address, zip codes, phone numbers, favorite soccer team, ...) are tuples of a relational database → you can use k-anonymity, l-diversity, t-closeness, ...
- Links should be protected to avoid link prediction:
  - Is it possible to predict whether two entities that are not connected currently will connect in 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.
  - Q: Which edges are sensitive?

anonimizzando i nodi popolari puoi evitare la profilizzazione dell'utente. es utente instagram togli tutte le pagine famose (calciatori , giornali ,calcio ecc) in questo modo le info sui gusti di quell apersono sono ben protette

Link Protection II

- Random perturbation: Random perturbation constructs an anonymized graph G’ from the original graph G by:
  - Randomly selecting m existing edges
  - Randomly adding m non-existent edges
- Obtained results:
  - high privacy at the cost of utility
- How to define m?
- Does it work like k in k-anonymization?

Graph Metrics I

metriche importanti dei dati

centralita=come quel grafo di enkk
cose molto famose sono molto
cenness per cui sono piu centrali

- **Centrality**: the importance in a graph. It is related to the application.
- **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.

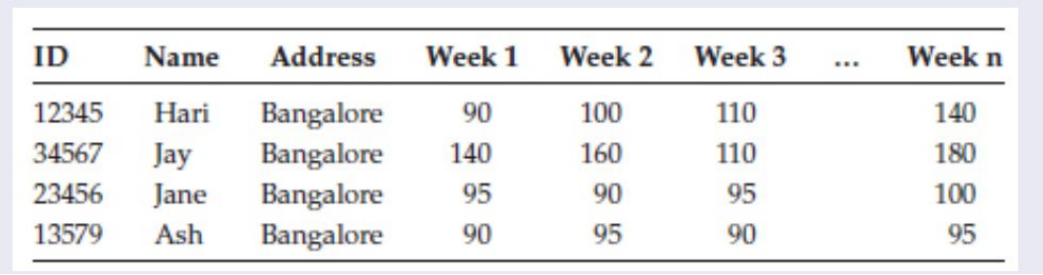
Effect of Anonymization

- Random perturbation changes the shape of the graph → impacts reachability, closeness and centrality.
- Clustering (on edges or vertices) changes neighboring → betweenness and closeness.
- All this techniques lead to a loss of utility

le varie tecniche cambiano la forma del grafo per cui portano ad una perdita di info ,in sostanza cambiano tutte le metriche viste nella slide sopra

Anonymizing Time Series Data I

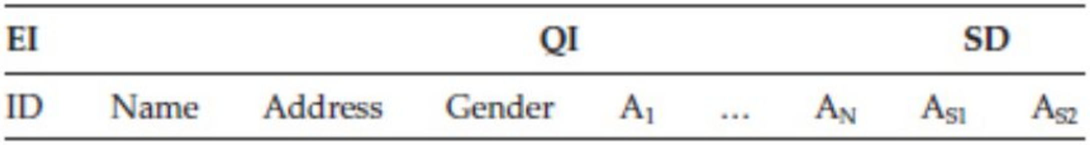
- Time Series: a sequence of observations indexed by the time of each observation (e.g., stock prices, bond, interest rate, blood pressure, ...).
- They are mined for forecasting
- They can be univariate or multivariate → large number of points in time → they are highly dimensional.
- They are represented both in time domain and frequency domain.



ID	Name	Address	Week 1	Week 2	Week 3	...	Week n
12345	Hari	Bangalore	90	100	110		140
34567	Jay	Bangalore	140	160	110		180
23456	Jane	Bangalore	95	90	95		100
13579	Ash	Bangalore	90	95	90		95

Anonymizing Time Series Data II

- As data increases, new data stream will be appended → how will this affect the table?

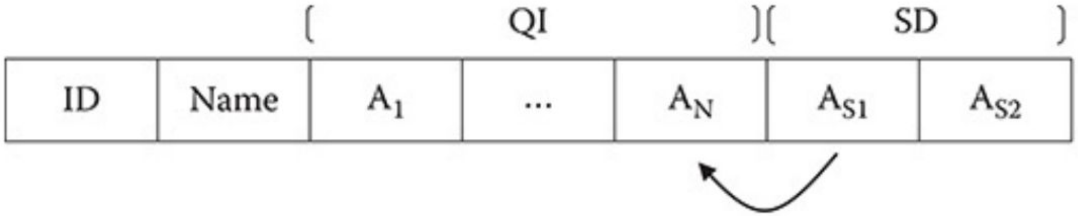


EI			QI				SD	
ID	Name	Address	Gender	A_1	...	A_N	A_{S1}	A_{S2}

- The data set contains:
  - EI: SSN, names, . . .
  - QI: contain a series of time-related data (i.e., A_1, ... , A_N) that SHOULD be anonymized.
  - SD: as series of time-related data that SHOULD NOT be anonymized.

Anonymizing Time Series Data III

- 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: it is simply impossible to model → high protection and poor utility.

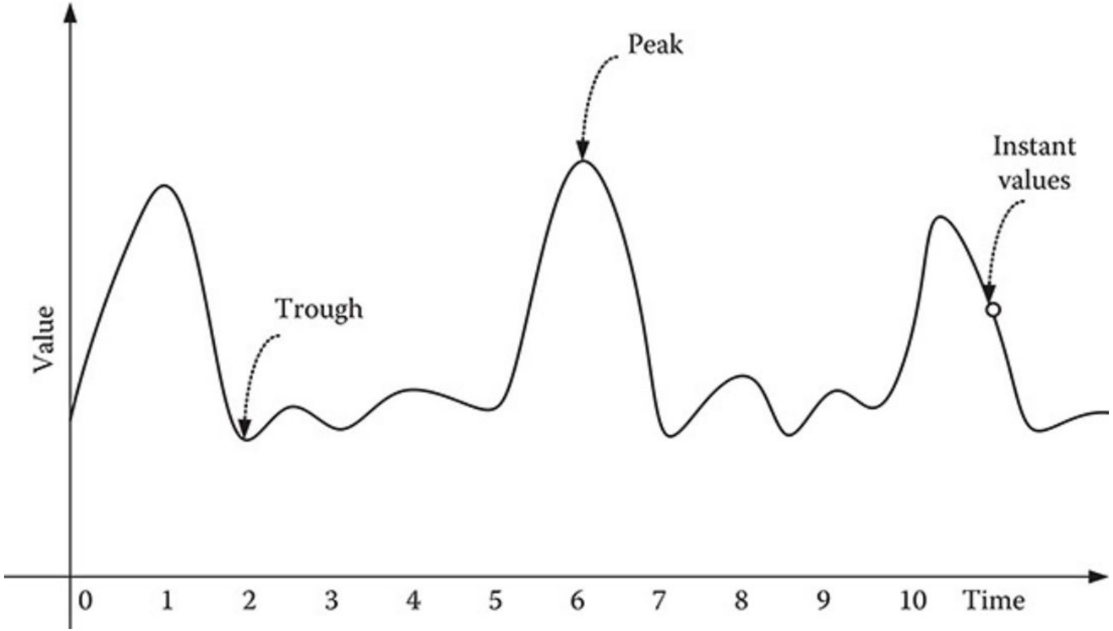


Classify as QI or SD? – Boundary between QI ans SD is “blurred” when dimensions are high. Because of the unknown background knowledge of the adversary, it is difficult to classify the attributes as QI or SD

Anonymizing Time Series Data IV

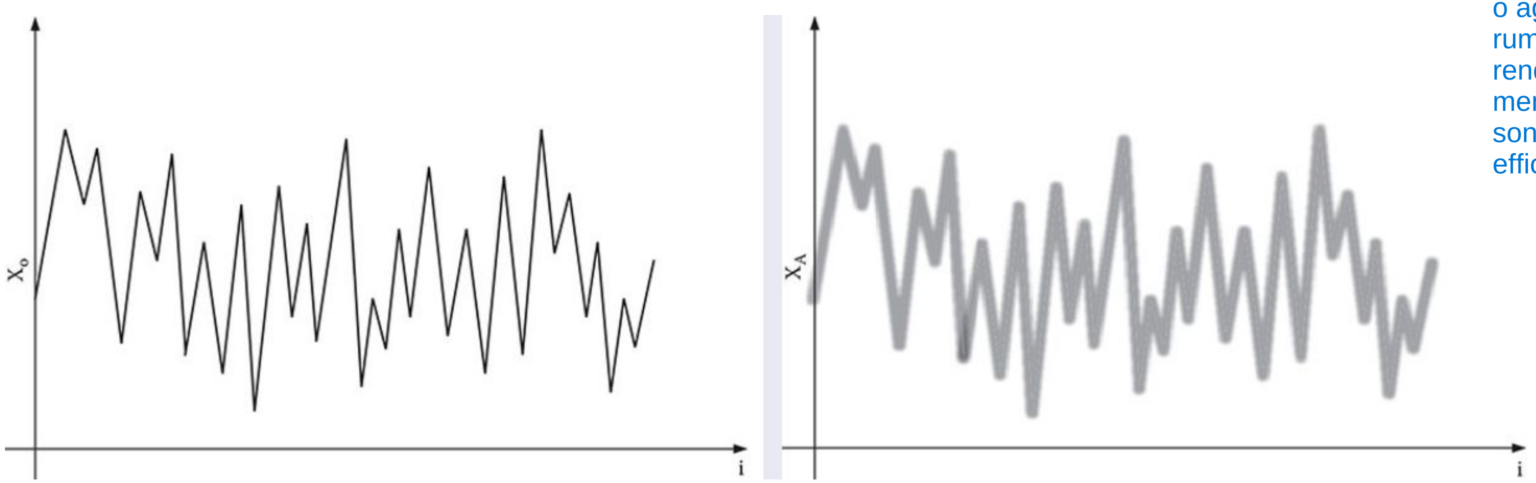
nelle time series è importante il grafico nel tempo

- **Pattern preservation**: Time series data have both instant values and a pattern.



Anonymizing Time Series Data V

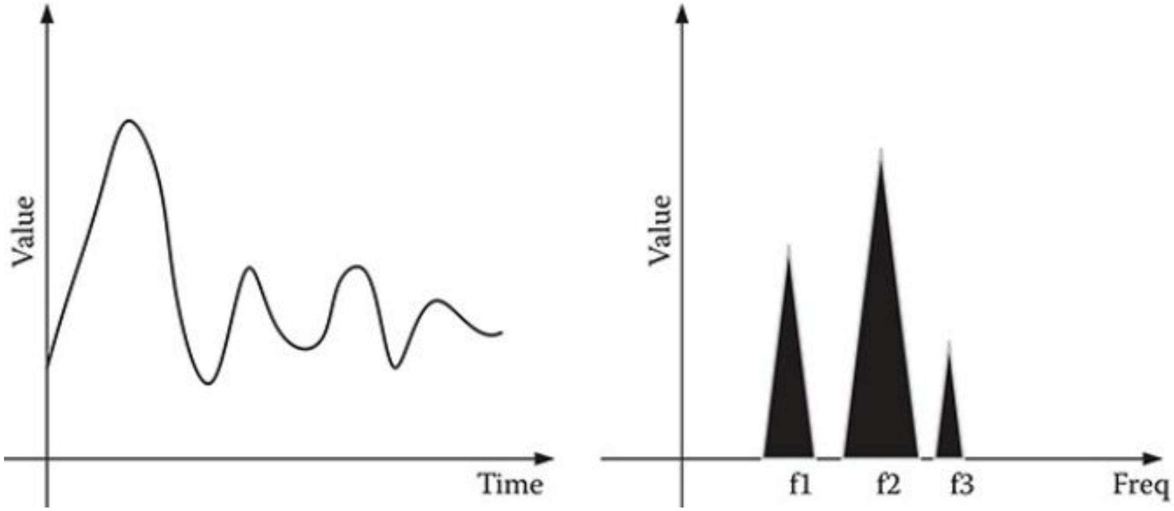
- **Preservation of Statistical Properties**: mean, variance and other statistical properties of the time series should be granted after anonymization.



alterare la forma o aggiungere del rumore per rendere i dati meno precisi non sono molto efficaci

Anonymizing Time Series Data VI

- **Preservation of Frequency-domain Properties**: correlation between time and frequency domains should be preserved in each anonymized time series.



Anonymizing Time Series Data VII

- Time Series Data Protection Methods
- Privacy issue same as Multidimensional data. Data:
  - EI are removed
  - SD MUST be kept original
  - QI must be anonymized
- Two categories:
  - Perturbative Methods → additive random noise
  - Generalization → k-anonymization! High level of generalization may impact the pattern
- Additive random noise:
  - white noise
  - correlated noise

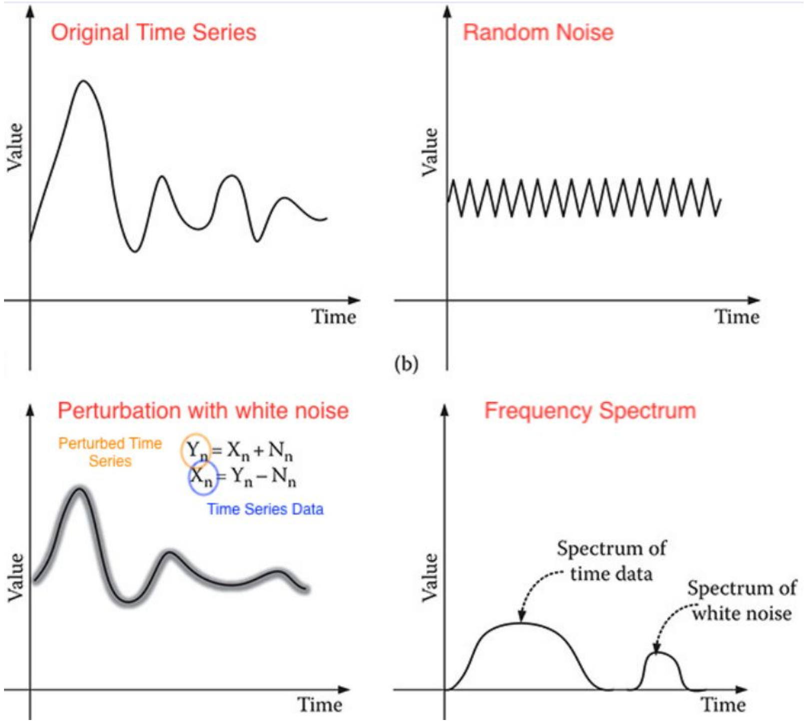
tipi di rumore che possiamo aggiungere

Anonymizing Time Series Data VIII

- White Noise
  - QI are perturbed with high-frequency white noise (i.e., by adding/removing random values).
  - The simplest but the weakest in terms of privacy, but optimal in terms of utility → the anonymized time series maintains statistical properties, the pattern and the frequency domain properties.
- Re-identification of perturbed time series data can be achieved through:
  - *Filtering* through low-pass filters.
  - *Regression*, but ineffective if independent noise is added to the time series data.

Anonymizing Time Series Data IX

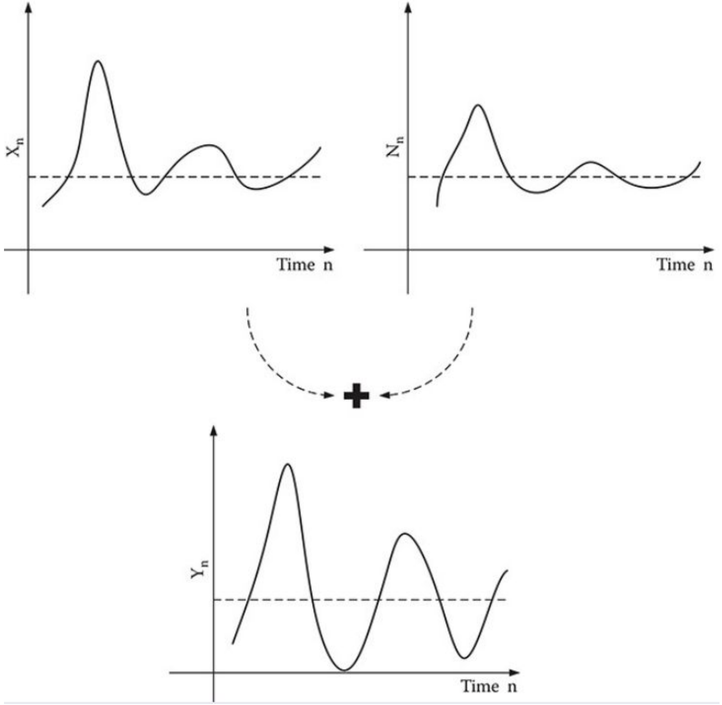
White Noise



Anonymizing Time Series Data IX

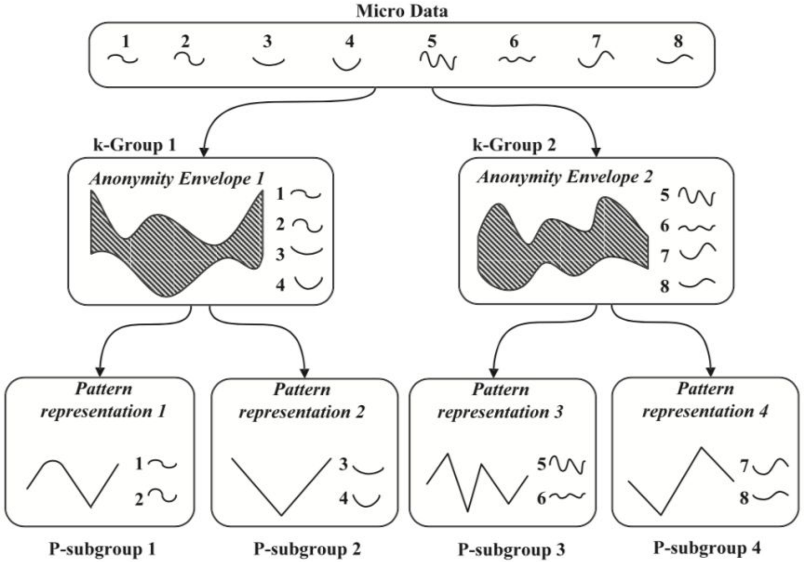
Correlated noise:

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



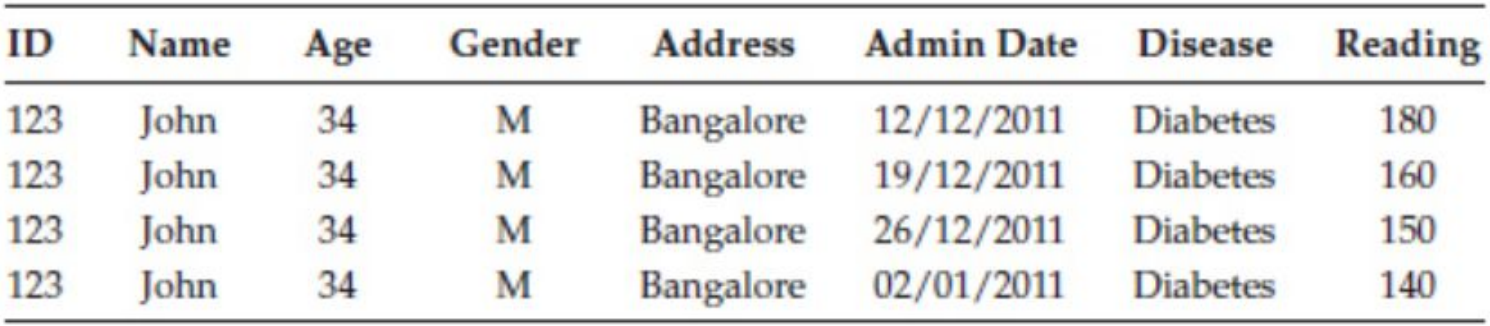
Generalization: (k, P)-Anonymity

(k, P)-Anonymity is an approach that aims at granting both k-anonymity and pattern preservation on time series data



Anonymizing Longitudinal Data I

- **Longitudinal Data**: They are commonly used in the healthcare domain and represent series of measurements.



ID	Name	Age	Gender	Address	Admin Date	Disease	Reading
123	John	34	M	Bangalore	12/12/2011	Diabetes	180
123	John	34	M	Bangalore	19/12/2011	Diabetes	160
123	John	34	M	Bangalore	26/12/2011	Diabetes	150
123	John	34	M	Bangalore	02/01/2011	Diabetes	140

Anonymizing Longitudinal Data II

Characteristics of Longitudinal Data

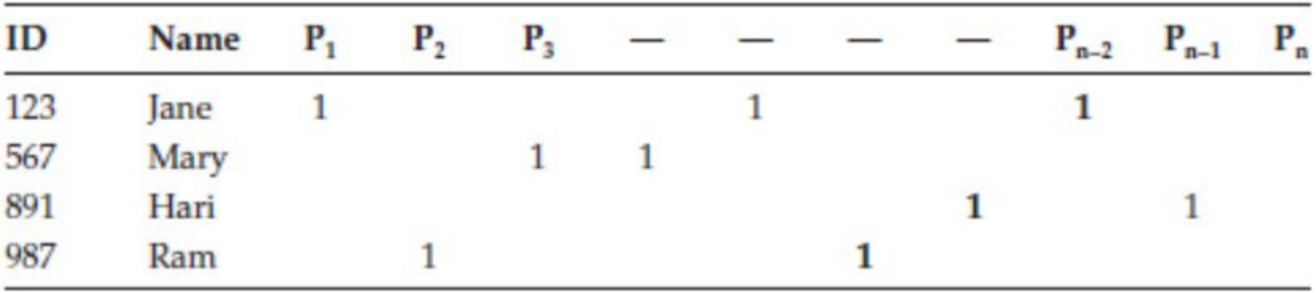
- Differences with Time Series Data:
  - Data are clustered and comprise repeated measurements from a single individual
  - Records are still divided into EI, QI and SD but strong relation between the patient and the SD and a strong correlation among records in the cluster.
  - Data in the cluster have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

Anonymizing Longitudinal Data III

- Challenges in Anonymizing Longitudinal Data
    - Identity disclosure → prevent record linkage.
    - Attribute disclosure → prevent sensitive data linkage.
    - Correlation in the cluster.
    - Pattern in the clustered data set.
    - Take into account that the records in the clustered data set cannot be treated independent of each other as in the case of multidimensional data (relational data).
    - Unknown background knowledge of the adversary.
- State of the Art in Anonymizing Longitudinal Data:
    - Anonymization of longitudinal data is an open issue → there is a lack of proposals right now.

Anonymizing Transaction Data I

- Transaction Data
  - They are complex, and suffers from high dimensionality and sparsity
  - Used in retail domain to log customers’ transactions
  - Mined to extract associations or correlations among transactions



ID	Name	P_1	P_2	P_3	—	—	—	—	P_{n-2}	P_{n-1}	P_n
123	Jane	1				1			1		
567	Mary			1	1						
891	Hari							1		1	
987	Ram		1				1				

Anonymizing Transaction Data II

- Characteristics of Transaction Data
  - EI such as ID and names are not part of the transaction data table.
  - The table is sparse; very few cells have entries in this high-dimensional space.
  - There are few sensitive transactions that are classified as sensitive data (shown in bold in the sample table). 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.
- State of the Art in Anonymizing Transaction Data:
  - Anonymization of transaction data is also an open issue → only a couple of approaches ATM