Priacy in Data Publication

5 novembre 2024

0.1 Privacy and Data Protection in Emerging Scenarios

La continua crescita riguarda:

- database governativi e aziendali;
- contenuti generati dagli utenti (Youtube, Flickr, ...);
- informazioni personali identificabili (creazione account, compilazione questionario, ..).

0.1.1 Data Sharing and Dissemination

La condivisione dei dati serve per:

- studiare tendenze e fare inferenze statistiche utili;
- condividere conoscenza;
- accedere a servizi online.

0.1.2 External Data Storage and Computation

L'archiviazione e il calcolo esterni (es. Cloud) offrono:

- risparmio sui costi e benefici di servizio;
- maggiore disponibilità e protezione più efficace dai disastri.

Outline È fondamentale garantire che la privacy e l'integrità dei dati siano adeguatamente protette.

- Privacy nella pubblicazione dei dati: rilascio e disseminazione dei dati.
- Privacy nel data outsourcing: terze parti memorizzano e gestiscono i dati.

0.2 Privacy in Data Publication

0.2.1 Statistical DBMS vs Statistical Data

DBMS statistici - Interazione interattiva tra Client e DBMS:

- Il sistema risponde solo a query statistiche.
- Necessario un controllo dinamico per proteggere la privacy e prevenire il rilascio indiretto di informazioni.

Dati statistici - Interazione non interattiva tra Client e DBMS:

- Pubblicano statistiche.
- Il controllo sul rilascio indiretto viene eseguito prima della pubblicazione.

Privacy Issues

- **Observation Window**: Si verifica quando più query successive rivelano informazioni sensibili.
- **Collusion**: Si verifica quando utenti collaborano per dedurre dati non deducibili singolarmente.

0.2.2 Macrodata vs Microdata

In passato, i dati erano principalmente rilasciati in forma tabellare (macrodata: dati aggregati) e attraverso database statistici. Oggi, molte situazioni richiedono il rilascio dei dati specifici memorizzati (microdata: dati disaggregati) che offrono maggiore flessibilità e disponibilità di informazioni per gli utenti. Tuttavia, i microdata sono soggetti a un maggiore rischio di violazioni della privacy, come i linking attacks: attacchi di collegamento.

Classification of Macrodata Tables Classifizione in due gruppi:

- **Count/Frequency**: Ogni cella della tabella contiene il numero di rispondenti (count) o la percentuale di rispondenti (frequency) che hanno lo stesso valore su tutti gli attributi analizzati.
- **Magnitude Data**: Ogni cella della tabella contiene un valore aggregato di una quantità di interesse su tutti gli attributi analizzati.

0.2.3 Information Disclosure

Disclosure: attribuzione di informazioni sensibili a un rispondente (individuo o organizzazione).
Si ha divulgazione quando:

- **Identity Disclosure**: un rispondente è identificato dai dati rilasciati;
- **Attribute Disclosure**: info sensibili su un rispondente vengono rivelate dai dati rilasciati;
- **Inferential Disclosure**: i dati rilasciati consentono di inferire il valore di alcune caratteristiche di un rispondente, anche se nessun record rilasciato fa riferimento a quel rispondente.

Identity Disclosure La divulgazione dell'identità si verifica quando una terza parte può identificare un rispondente dai dati rilasciati. La rivelazione di un individuo come rispondente in una raccolta di dati può o meno violare i requisiti di riservatezza:

- **Macrodata**: La rivelazione dell'identità generalmente non è un problema, a meno che non porti a divulgare informazioni riservate (divulgazione degli attributi).

- **Microdata**: L'identificazione è considerata un problema, poiché i record di microdata sono dettagliati; la divulgazione dell'identità implica solitamente anche la divulgazione degli attributi.

Attribute Disclosure La divulgazione degli attributi si verifica quando informazioni riservate su un rispondente vengono rivelate e possono essere attribuite a quest'ultimo.
Le informazioni possono essere rivelate esattamente o stimate con un certo grado di approssimazione (è un problema quando il grado di incertezza è inferiore a quello atteso).

Inferential Disclosure La divulgazione inferenziale si verifica quando informazioni riservate e/o sensibili possono essere dedotte con alta confidenza dalle proprietà statistiche dei dati rilasciati.
La divulgazione inferenziale non rappresenta sempre un rischio:

- I dati vengono rilasciati per consentire agli utenti di comprendere le relazioni tra le variabili.
- Le inferenze sono progettate per prevedere comportamenti aggregati, non attributi individuali.

0.2.4 Restricted Data and Restricted Access

La scelta dei metodi di limitazione della divulgazione statistica dipende dalla natura dei dati di cui si vuole garantire la riservatezza. Rimuovere identificatori espliciti presenti nei microdati (nome, CF, ...) è un primo passo per preparare il rilascio di tali dati. La riservatezza può essere protetta mediante:

- limitazione della quantità di informazioni presenti nelle tabelle rilasciate (dati riservati);
- imposizione di condizioni sull'accesso ai dati rilasciati (accesso limitato);
- una combinazione di queste due strategie.

0.2.5 Protection for Count/Frequencies Macrodata

I dati raccolti dalla maggior parte dei sondaggi sulle persone sono pubblicati in tabelle che mostrano conteggi (numero di persone per categoria) o frequenze (frazione o percentuale di persone per categoria). Le tecniche di protezione includono:

- **Sampling**: pubblicare un sondaggio anziché un censimento;
- **Special Rules**: definire restrizioni sul livello di dettaglio che può essere fornito in una tabella (ad esempio, non pubblicare o rendere deducibili i redditi all'interno di un intervallo di $1.000);

- **Threshold Rules**: definire una cella di una tabella come sensibile se il numero di rispondenti è inferiore a una certa soglia specificata.

0.2.6 Disclosure Protection Techniques for Microdata

Le tecniche di protezione classiche, spesso applicate prima del calcolo delle statistiche, possono essere classificate come segue:

- **Masking Techniques**: trasformano il dataset non rilasciando o perturbando i suoi valori.
  - ***Non-Perturbative***: i dati originali non vengono modificati, ma alcuni dati sono soppressi e/o alcuni dettagli vengono rimossi (es. campionamento, soppressione locale, generalizzazione);
  - ***Perturbative***: i dati originali vengono modificati (es. arrotondamento, scambio).
- **Synthetic Data Generation Techniques**: rilasciano valori plausibili ma sintetici.
  - ***Fully Synthetic***: il dataset rilasciato contiene solo dati sintetici;
  - ***Partially Synthetic***: il dataset rilasciato contiene un mix di dati originali e sintetici.

The Anonymity Problem La quantità di registri di proprietà privata che descrivono le finanze, gli interessi e i dati demografici di ciascun cittadino è in costante aumento. Questi dati vengono de-identificati prima del rilascio, ovvero qualsiasi identificatore esplicito viene rimosso. Tuttavia, la de-identificazione non è sufficiente. Molti comuni vendono registri della popolazione che includono le identità degli individui insieme a dati demografici di base. Questi dati possono quindi essere utilizzati per linkare identità con informazioni de-identificate, portando alla **re-identificazione**.

0.2.7 Classification of Attributes in a Microdata Table

Gli attributi nella tabella di microdata originale possono essere classificati come segue:

- **Identifiers**: attributi che identificano univocamente un rispondente (es. SSN).
- **Quasi-Identifiers**: attributi che, in combinazione, possono essere utilizzati con info esterne per reidentificare alcuni rispondenti o ridurre l'incertezza sulla loro identità (es. Race, DoB, ...).
- **Confidential**: attributi che contengono informazioni sensibili (es. Disease).
- **Non-Confidential**: attributi che non sono considerati sensibili dai rispondenti e il cui rilascio non causa divulgazione.

0.2.8 Factors Contributing to Increase the Disclosure Risk

I possibili fattori che contribuiscono al rischio di divulgazione dei microdata includono:

- **Esistenza di Registri ad Alta Visibilità**: alcuni registri possono rappresentare rispondenti con caratteristiche uniche, come lavori molto insoliti (es. star del cinema) o redditi molto elevati.
- **Possibilità di Matchare i Microdata con Informazioni Esterne**: esistono individui che possiedono una combinazione unica o peculiare delle variabili caratteristiche nei microdata.
  - Se alcuni di questi individui vengono scelti nel campione della popolazione, il rischio di divulgazione aumenta, è importante mantenere segreta l'identità degli individui scelti.

La possibilità di collegamento o la sua precisione aumentano con:

- l'esistenza di un elevato numero di attributi comuni tra la tabella di microdata e le fonti esterne;
- l'accuratezza o la risoluzione dei dati;
- il numero e la ricchezza delle fonti esterne, non tutte le fonti disponibili potrebbero essere conosciute dall'agenzia che rilascia i microdata.

0.2.9 Factors Contributing to Decrease the Disclosure Risk

I seguenti fattori possono contribuire a diminuire il rischio di divulgazione nei microdata:

- **Sottogruppo della Popolazione**: una tabella di microdata spesso contiene solo un subset dell'intera popolazione $\rightarrow$ le informazioni di un rispondente potrebbero non essere incluse.
- **Disallineamento Temporale e/o di Formato tra le Varie Fontie**: il rischio di divulgazione nei microdata è influenzato da un disallineamento temporale, poiché le informazioni nelle tabelle possono non essere aggiornate (spesso obsolete di uno o due anni). Inoltre, le fonti esterne possono esprimere i dati in formati diversi, il che riduce la capacità di collegare informazioni e aumenta il rischio di errore nell'identificazione dei rispondenti.
- **Presenza di Rumore**: sia le tabelle di microdata che le fonti esterne contengono naturalmente rumore, il che riduce la capacità di collegare le informazioni.

Measures of Risk La valutazione del rischio di divulgazione richiede di considerare diversi fattori:

- **Probabilità di Rappresentazione**: la probabilità che il rispondente di interesse sia presente sia nei microdata che in un file esterno.
- **Variabili di Collegamento**: la probabilità che le variabili di matching siano registrate in modo linkabile nei microdata e nel file esterno.
- **Unicità del Rispondente**: la probabilità che il rispondente sia unico o peculiare rispetto alla popolazione del file esterno (cruciale per il rischio di divulgazione).

Ogni rispondente unico nella popolazione è anche unico nel campione, ma non viceversa.

0.2.10 k-anonymity

La k-anonymity, insieme alla sua applicazione attraverso la generalizzazione e la soppressione, mira a proteggere le identità dei rispondenti durante la pubblicazione di informazioni veritiere. Essa cerca di garantire che i dati rilasciati siano indistinguibili rispetto ad un numero minimo di rispondenti. La scelta di k definisce il grado di sicurezza → **TRADE OFF: INFO LOSS** VS **PRIVACY GAIN**.
I **QI: Quasi-identificatori** sono insieme di attributi che possono essere sfruttati per il linking, di conseguenza la pubblicazione di quest'ultimi deve essere controllata.

Basic Idea Tradurre il requisito di k-anonymity sui dati rilasciati
Ogni combinazione di valori dei quasi-identificatori deve essere abbinata indistintamente ad almeno k rispondenti. Di conseguenza, i rispondenti devono essere indistinguibili (all'interno di un dato insieme) rispetto a un insieme di attributi. Inoltre, per soddisfare il requisito di k-anonymity, ogni valore di quasi-identificatore che appare nella tabella rilasciata deve avere almeno k occorrenze, il che rappresenta una condizione sufficiente per la soddisfazione del requisito.

Generalization and Suppression

Tecniche di mascheramento non-perturbative per raggiungere la k-anonymity:

- **Generalization** I valori di un dato attributo vengono sostituiti con valori più generali, basandosi sulla definizione di una gerarchia di generalizzazione.
- **Suppression** Per proteggere informazioni sensibili, si rimuovono tali informazioni.

L'introduzione della soppressione può ridurre la quantità di generalizzazione necessaria per soddisfare il vincolo di k-anonymity. Inoltre, tali tecniche non introducono rumore.

Domain Generalization Hierarchy

Una relazione di generalizzazione $\leq_D$ definisce un mapping tra il dominio D e le sue generalizzazioni. Dati due domini $D_i, D_j \in \text{Dom}$, $D_i \leq_D D_j$ indica che i valori nel dominio D_j sono generalizzazioni dei valori in D_i.
$\leq_D$ implica l'esistenza, per ogni dominio D, di una gerarchia di generalizzazione del dominio $DGH_D = (\text{Dom}, \leq_D)$:

- $\forall D_i, D_j, D_z \in \text{Dom} : D_i \leq_D D_j, D_i \leq_D D_z \Rightarrow D_j \leq_D D_z \vee D_z \leq_D D_j$.
- Tutti gli elementi massimali di Dom sono singleton.

Data una tupla di dominio $DT = \langle D_1, \ldots, D_n \rangle$ tale che $D_i \in \text{Dom}$, $i = 1, \ldots, n$, la gerarchia di generalizzazione del dominio di DT è $DGH_{DT} = DGH_{D_1} \times \ldots \times DGH_{D_n}$.

Value generalization hierarchy

La relazione di generalizzazione dei valori $\leq V$ associa a ogni valore nel dominio D_i un valore unico nel dominio D_j, rappresentando una generalizzazione diretta di D_i. Questa relazione implica l'esistenza di una gerarchia di generalizzazione dei valori (VGH_D) per ciascun dominio D.
La VGH_D ha una struttura ad albero:

- **Foglie**: Rappresentano i valori nel dominio D.
- **Radice**: È il valore più generale, situato nell'elemento massimo di DGH_D.

Generalized Table with Suppression

Una tabella T_j è detta una generalizzazione (mediante soppressione di tuple) della tabella T_i ($T_i \preceq T_j$), se soddisfa le seguenti condizioni:

- $|T_j| \leq |T_i|$ (la cardinalità di T_j è minore o uguale a quella di T_i).
- Il dominio $\text{dom}(A, T_j)$ di ogni attributo A in T_j è uguale o una generalizzazione del dominio $\text{dom}(A, T_i)$ dell'attributo A in T_i.
- È possibile definire una funzione iniettiva che associa ogni tupla t_j in T_j con una tupla t_i in T_i, tale che il valore di ogni attributo in t_j sia uguale o una generalizzazione del valore dell'attributo corrispondente in t_i. Non tutti gli elemnti del dominio hanno un immagine nel codominio.

k-minimal Generalization with Suppression

Siano $T_i(A_1, \ldots, A_n)$ e $T_j(A_1, \ldots, A_n)$ due tabelle tali che $T_i \preceq T_j$. Il **vettore di distanza** di T_j da T_i è definito come il vettore

$$DV_{i,j} = [d_1, \ldots, d_n],$$

dove ogni d_z per $z = 1, \ldots, n$ è la lunghezza del percorso unico tra $dom(A_z, T_i)$ e $dom(A_z, T_j)$ nella gerarchia di generalizzazione del dominio DGH_{D_z}.
Siano T_i e T_j due tabelle t.c. $T_i \preceq T_j$, e sia $MaxSup$ la soglia specificata di soppressione accettabile. La tabella T_j è detta una **generalizzazione k-minimale** della tabella T_i se e solo se:

1. T_j soddisfa la k-anonymity garantendo la soppressione minima richiesta se per ogni tabella T_z che soddisfa la k-anonymity e t.c. $T_i \preceq T_z$ e $DV_{i,z} = DV_{i,j}$, allora deve valere $|T_j| \geq |T_z|$.
2. $|T_i|$ - $|T_j| \leq MaxSup$
3. $\forall T_z$ t.c. $T_i \preceq T_z$ e T_z soddisfa le condizioni 1 e 2 $\Rightarrow \neg(DV_{i,z} < DV_{i,j}) \iff DV_{i,z} >= DV_{i,j}$

Computing a Preferred Generalization

Diversi criteri di preferenza possono essere applicati nella scelta di una generalizzazione minima:

- **minimum absolute distance:** preferisce le gen con la distanza assoluta minore (minor numero di passaggi di gen - indipendentemente dalle gerarchie da cui provengono).
- **minimum relative distance:** preferisce le gen con la distanza relativa minore (un passaggio è considerato relativo se viene diviso per l'altezza della gerarchia di dominio a cui si riferisce).
- **maximum distribution:** preferisce le gen che hanno il maggior numero di tuple distinte.
- **minimum suppression:** preferisce le gen che sopprimono meno tuple (massima cardinalità).

Questi criteri aiutano a scegliere tra le diverse generalizzazioni possibili, tenendo conto di fattori come la distanza dalla tabella originale, il numero di info distinte mantenute e la quantità di dati soppressi.

Classification of k-anonymity Techniques

La generalizzazione e la soppressione possono essere applicate a diversi livelli di granularità:

- **Generalization:** può essere applicata a livello di singola colonna (ad ogni passo generalizzazione di tutti i valori nella colonna) oppure a livello di singola cella.
- **Suppression:** può essere applicata a livello di riga (eliminazione/rimozione di un'intera tupla), a livello di attributo (eliminazione/rimozione di tutti i valori di una determinata colonna) e/o a livello di singole celle (eliminazione/rimozione di determinate celle di una data tupla/attributo).

Algorithms for Computing a k-anonymous Table

Computare tabelle k-anonime minime, con generalizzazione degli attributi e soppressione delle tuple (`AG_TS`), è computazionalmente difficile $\rightarrow$ *problema* $\in$ $NP\text{-}Hard$.

- La maggior parte degli algoritmi esatti proposti in letteratura ha un tempo computazionale esponenziale nel numero degli attributi che compongono il quasi-identificatore.
- Quando il numero di attributi nel quasi-identificatore ($|QI|$) è piccolo rispetto al numero di tuple nella tabella privata (PT), questi algoritmi esatti con `AG_TS` risultano praticabili.
- Sono stati proposti molti algoritmi esatti per compuatare tabelle k-anonime basati su `AG_TS`.

0.2.11 Algorithms for AG_TS and AG_

Computing a k-minimal Solution

- Ogni percorso in DGH_{DT} rappresenta una strategia di generalizzazione per PT
- Chiamiamo *locally minimal generalization* il nodo con indice minore in ogni percorso che soddisfa la k-anonymity
- Proprietà sfruttate dall'algoritmo:
  1. Ogni k-minimal gen è localmente minima rispetto a un percorso, ma il contrario non è vero
  2. Salendo nella gerarchia, il # di tuple da rimuovere per garantire la k-anonymity diminuisce
- Se non esiste una soluzione che garantisca la k-anonymity sopprimendo meno di MaxSup tuple all'altezza h, non può esistere una soluzione con altezza inferiore a h che lo garantisca.

Binary Search Algorithm on Distance Vectors

L'algoritmo adotta una ricerca binaria sul reticolo dei vettori distanza:

1. Valuta le soluzioni all'altezza $\lfloor \frac{h}{2} \rfloor$
2. Se esiste almeno una soluzione che soddisfa la k-anonymity:
   - Valuta le soluzioni all'altezza $\lfloor \frac{h}{4} \rfloor$
3. Altrimenti valuta le soluzioni all'altezza $\lfloor \frac{3h}{4} \rfloor$
4. Fino a quando l'algoritmo min(h) per la quale esiste un DV che soddisfa la k-anonymity

Per ridurre il costo computazionale, l'algoritmo utilizza una matrice di vettori distanza (DVs' Matrix).

k-Optimize algorithm

- Ordinare gli attributi nel quasi-identificatore (QI) e i valori nei rispettivi domini.
- Associare un indice intero a ciascun valore del dominio, seguendo l'ordine definito.

Ad esempio:

Race	**ZIP**
⟨[asian: 1] [black: 2] [white: 3]⟩	⟨[94138: 4] [94139: 5] [94141: 6] [94142: 7]⟩

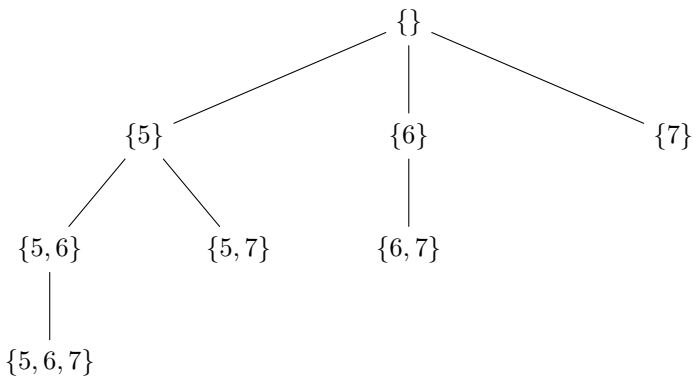
- Una generalizzazione è l'unione dei singoli valori di indice.
- Il valore più basso in un dominio di attributi viene omesso. Ad esempio, {6} corrisponde a:
  - **Race**: {1}, cioè: ⟨[asian or black or white]⟩
  - **ZIP**: {4, 6}, cioè: ⟨[94138 or 94139], [94141 or 94142]⟩
- L'ordine dei valori all'interno dei domini ha un impatto sulla generalizzazione.

L'algoritmo **k-Optimize** costruisce un **albero di enumerazione** per l'insieme degli indici I.
La radice dell'albero è l'insieme vuoto $\emptyset$, e i figli di ciascun nodo n sono ottenuti aggiungendo un singolo elemento i dell'insieme I, tale che $\forall i' \in n, i > i'$. Ogni nodo ha un **costo** che riflette la quantità di generalizzazione e soppressione associata all'anomizzazione rappresentata dal nodo.
L'algoritmo cerca l'anonimizzazione con il costo minimo attraverso una **visita dell'albero** tramite **ricerca in profondità**. Tuttavia, poiché l'albero ha $2^{|I|}$ nodi, la visita completa non è praticabile. Quindi viene adottata una strategia di **potatura (pruning)**:

- Un nodo n viene potato se nessuno dei suoi discendenti può fornire una soluzione ottimale.
- Questo si determina calcolando un **lb:limite inferiore** sul costo dei nodi nel sottoalbero radicato in n. Se il limite inferiore è maggiore del miglior costo corrente, il nodo n viene potato.

Set Enumeration Tree Example

Considerando il dominio ZIP rispetto all'associazione tra valori e indici definita precedentemente. L'albero di enumerazione ha la seguente struttura:



Incognito Algorithm

L'algoritmo **Incognito** verifica **k-anonimity** con riferimento a un adeguato sottoinsieme del QI.
Esso adotta un approccio **bottom-up** per visitare le gerarchie di generalizzazione dei domini (DGHs). La condizione di k-anonimity rispetto a un sottoinsieme di QI è necessaria, ma non sufficiente per garantire la k-anonimity rispetto a tutto il QI. Il processo iterativo dell'algoritmo procede come segue:

- **Iterazione 1**: si controlla la **k-anonimity** per ciascun attributo singolo in QI, scartando le generalizzazioni che non soddisfano la k-anonimity.
- **Iterazione 2**: si combinano le generalizzazioni rimanenti in coppie, verificando la **k-anonimity** per ciascuna coppia ottenuta. Scartando le coppie che non soddisfano la k-anonimity.
- **Iterazione n**: si considerano tutte le n-uple di attributi ottenuti dalle generalizzazioni che soddisfavano la k-anonimity nell'iterazione $i-1$, scartando le soluzioni che non la rispettano.
- ...

- **Iterazione** |**QI**|: restituisce il risultato finale, che rappresenta una generalizzazione che soddisfa la k-anonimity rispetto all'intero quasi-identificatore (QI).

L'algoritmo procede dunque costruendo progressivamente soluzioni, partendo da singoli attributi e combinandoli in gruppi via via più grandi fino a considerare tutti gli attributi del quasi-identificatore.

Heuristic Algorithms

Gli algoritmi esatti presentano una complessità esponenziale rispetto al numero di QI considerati. Per questo motivo, sono stati proposti algoritmi euristici:

- basato su algoritmi genetici, risolve il problema della k-anonymity utilizzando un metodo di ricerca stocastica incompleto.
- basato sul *simulated annealing* per trovare soluzioni localmente minime; richiede un elevato tempo computazionale e non garantisce la qualità della soluzione.
- approccio euristico top-down per rendere una tabella k-anonima; inizia dalla soluzione più generale e specializza iterativamente alcuni valori fino a violare il requisito di k-anonymity.

Non è possibile fornire limiti sull'efficienza e sulla bontà delle soluzioni ottenute ma è possibile utilizzare risultati sperimentali per valutare la qualità delle soluzioni recuperate.

0.2.12 Algorithms for _CS and CG_

Mondrian Multidimensional Algorithm

L'algoritmo **Mondrian Multidimensional** si basa su una rappresentazione spaziale delle tuple e dei quasi-identificatori:

- Ogni attributo nel quasi-identificatore (**QI**) rappresenta una dimensione.
- Ogni tupla nel set di dati privati (**PT**) rappresenta un punto nello spazio definito da **QI**.
- Le tuple con lo stesso valore di **QI** sono rappresentate assegnando una molteplicità ai punti.
- Lo spazio multidimensionale viene partizionato dividendo le dimensioni in modo tale che ogni area contenga almeno k occorrenze di valori dei punti.
- Tutti i punti in una regione vengono generalizzati a un valore unico.
- Le tuple corrispondenti sono sostituite dalla generalizzazione calcolata.

L'algoritmo Mondrian è flessibile e può operare:

- **Su un numero diverso di attributi:**
  - *Single or Multi-dimension.*
- **Con diverse strategie di generalizzazione:**
  - *Global or Local recoding*: colonna o cella.
- **Con diverse strategie di partizionamento:**
  - *Strict or Relaxed partitioning*: senza o con possibili sovrapposizioni.
- **Utilizzando metriche diverse per determinare come dividere ogni dimensione.**

k-anonymity Revisited

La **k-anonymity** cambia a seconda del livello di generalizzazione applicato:

- **AG:** Ogni n-upla di quasi-identificatori deve apparire almeno k volte.
- **CG:** La condizione di avere almeno k occorrenze è sufficiente ma non necessaria. È possibile utilizzare un requisito meno restrittivo:
  1. Per ogni sequenza di valori pt in $PT[QI]$, ci devono essere almeno k tuple in $GT[QI]$ che contengono una sequenza di valori che generalizzano pt.
  2. Per ogni sequenza di valori t in $GT[QI]$, ci devono essere almeno k tuple in $PT[QI]$ che contengono una sequenza di valori per cui t è una generalizzazione.

La gen a livello di cella permette una maggiore flessibilità rispetto alla gen a livello di attributo.

0.2.13 Attribute Disclosure

La **k-anonymity** è suscettibile a diversi attacchi:

- **Homogeneity of the Sensitive Attribute Values:** Tutte le tuple con lo stesso QI in una tabella k-anonima possono avere lo stesso valore per l'attributo sensibile.
  - Ad esempio, un avversario sa che Carol è una donna di colore e che i suoi dati sono inclusi nella tabella. Se tutte le tuple con quel quasi-identificatore condividono lo stesso valore dell'attributo sensibile, l'avversario può dedurre che Carol soffre di mancanza di respiro.

- **Background Knowledge:** Un avversario può utilizzare informazioni esterne già note per dedurre informazioni sensibili.
  - Ad esempio, un avversario sa che Hellen è una donna bianca presente nella tabella. Se le opzioni possibili per la sua malattia sono "dolore al petto" o "mancanza di respiro", e l'avversario sa che Hellen corre per 2 ore al giorno, può escludere la "mancanza di respiro" e dedurre che Hellen soffre di "dolore al petto".

ℓ-Diversity

Un *q-block* (cioè, un insieme di tuple con lo stesso valore per i *quasi-identifiers*) è detto ℓ-diverse se contiene almeno ℓ valori differenti e *ben rappresentati* per l'attributo sensibile.

- *ben rappresentati* può essere definito tramite entropia o ricorsione.
- la ℓ-diversity implica che un avversario deve eliminare almeno $\ell - 1$ valori possibili per inferire un valore sensibile di un rispondente.

Una tabella viene definita ℓ-diverse se tutti i suoi *q-blocks* sono ℓ-diverse, rendendo impossibili attacchi di omogeneità e più difficili quelli basati su *background knowledge*. La ℓ-diversity è monotona rispetto alle gerarchie di generalizzazione usate per la *k-anonymity*. Tuttavia, la ℓ-diversity può lasciare spazio a nuovi attacchi basati sulla distribuzione dei valori all'interno dei *q-blocks*, come:

Skewness Attack

Un *skewness attack* avviene quando la distribuzione dei valori in un *q-block* è diversa da quella della popolazione originale.

Similarity Attack

Un *similarity attack* si verifica quando un *q-block* contiene valori diversi ma semanticamente simili per l'attributo sensibile.

Group Closeness

Un *q-block* rispetta la *t-closeness* se la distanza tra la distribuzione dei valori dell'attributo sensibile nel *q-block* e nella popolazione è inferiore a una soglia t.

- Una tabella rispetta la *t-closeness* se tutti i suoi *q-blocks* rispettano tale condizione.
- La *t-closeness* è monotona rispetto alle gerarchie di generalizzazione per la *k-anonymity*.

Qualsiasi algoritmo per la *k-anonymity* può essere esteso per rispettare la *t-closeness*, ma tale proprietà può essere difficile da ottenere. Inoltre, un osservatore potrebbe usare conoscenze esterne o pregresse per inferire informazioni.

Types of Background Knowledge

Le conoscenze possono riguardare:

- l'individuo target
- altri individui, il che potrebbe comunque rivelare informazioni sensibili
- famiglie di valori uguali, come informazioni genomiche che collegano un gruppo di persone.

Multiple Releases

I dati possono essere soggetti a frequenti cambiamenti e necessitare di pubblicazioni regolari. Tuttavia, rilasci multipli di una tabella di microdati possono causare fughe di informazioni poiché un destinatario malevolo può correlare i dataset rilasciati. Pertanto, i rilasci multipli (o longitudinali) non possono essere indipendenti, e devono essere protetti contro attacchi di intersezione.

m-invariance

Per affrontare il problema dei rilasci longitudinali, una sequenza $T_1, ..., T_n$ di tabelle di microdati rilasciate soddisfa la proprietà di *m-invariance* se:

- ogni classe di equivalenza contiene almeno m tuple;
- nessun valore sensibile appare più di una volta in ciascuna classe di equivalenza;
- per ogni tupla t, le classi di equivalenza a cui appartiene t nella sequenza sono caratterizzate dallo stesso insieme di valori sensibili.

Ciò implica che la correlazione delle tuple in $T_1, ..., T_n$ non permette a un destinatario malevolo di associare meno di m valori sensibili differenti a ciascun rispondente.

Extended Scenarios

Le tecniche di *k-anonymity*, ℓ-diversity e *t-closeness* si basano su ipotesi che non sempre sono applicabili in scenari specifici. In particolare:

- **Multiple tuples per respondent**
- **Rilascio di più tabelle con dipendenze funzionali**
- **Più quasi-identificatori**
- **Quasi-identificatori non predefiniti**
- **Rilascio di stream di dati**
- **Preferenze di privacy fine-grained**

k-anonymity in Various Applications

Oltre al classico problema del rilascio di microdati, il concetto di *k-anonymity* e le sue estensioni possono essere applicati in diversi scenari, come ad esempio:

- **Social Networks**
- **Data Mining**
- **Location Data**

Neighborhood Attack in Social Networks

Un *neighborhood attack* si verifica quando, dato un grafo de-identificato G' di una rete sociale G, un avversario sfrutta la conoscenza sui vicini di un utente u per re-identificare il vertice che rappresenta u.

k-anonymity in Social Networks

L'idea è di adattare il requisito di *k-anonymity* alle reti sociali. Un vertice u è *k-anonymous* se esistono almeno $k-1$ altri vertici $v_1, ..., v_{k-1}$ tali che i sottografi indotti dal vicinato di u e dal vicinato di $v_1, ..., v_{k-1}$ sono isomorfi. Un grafo G' è *k-anonymous* se ogni vertice u in G' è *k-anonymous*.
Intuizione: aggiungere archi fittizi per soddisfare il requisito di k-anonimato.
Se G' è *k-anonymous*, con una neighborhood background knowledge, qualsiasi vertice in G non può essere re-identificato in G' con una confidenza maggiore di $1/k$.
Obiettivo: calcolare una versione *k-anonymous* di una grafo minimizzando il numero di archi aggiunti.

k-anonymous Data Mining

Le tecniche di *privacy-preserving data mining* dipendono dalla definizione di privacy, catturando quali informazioni sono sensibili nei dati originali e dovrebbero essere protette. Il *k-anonymous data mining* mira a garantire che i risultati del data mining non violino i requisiti di *k-anonymity* sui dati originali. Alcuni esempi di tecniche per compromettere la k-anonymity sfruttando il data mining ncludono:

- **Association Rule Mining**: tecniche per trovare regole di associazione possono compromettere la k-anonymity.
- **Classification Mining**: tecniche di classificazione possono portare a minacce per la privacy.

k-anonymity in Location-Based Services

Per proteggere l'identità degli utenti in base alla loro posizione geografica, è possibile adottare il concetto di *k-anonymity*, come segue:

- Considerare solo le aree che contengono almeno k individui
- Ingrandire l'area per includere almeno altri $k-1$ utenti (*k-anonymity*)
- Obfuscazione delle aree (*location privacy*) per ridurre la precisione o la confidenza dei dati
- Protezione del percorso degli utenti (*trajectory privacy*) tramite mix/modifica delle traiettorie

Re-identification with Any Information

Qualsiasi tipo di informazione può essere utilizzata per re-identificare dati anonimi. Questo rende la protezione della privacy particolarmente difficile a causa della crescente quantità e varietà di dati raccolti sugli individui. Di seguito, due esempi noti.

AOL Data Release

Nel 2006, AOL pubblicò 20 milioni di query di ricerca effettuate da 650,000 utenti per favorire la comunità di ricerca. Nonostante l'anonimizzazione di username e indirizzi IP, l'uso di numeri identificativi unici permise la re-identificazione di utenti attraverso query specifiche. Il caso più noto riguarda *Thelma Arnold*, re-identificata tramite ricerche locali e mediche. Questo evidenzia come dati apparentemente anonimi possano essere utilizzati per risalire all'identità delle persone.

Netflix Prize Data Study

Nel 2006, Netflix lanciò una sfida per migliorare il proprio algoritmo di raccomandazione fornendo un dataset di 100 milioni di record sui rating dei film di circa 500,000 utenti. Studi successivi mostrarono che pochissime informazioni ausiliarie sono necessarie per de-anonimizzare un utente:

- Con 6 rating e date ($\pm$ 2 settimane), il 99% degli utenti può essere identificato univocamente
- Con 2 rating e date ($\pm$ 3 giorni), il 68% degli utenti può essere identificato univocamente

Informazioni ausiliarie possono essere ottenute da fonti esterne, come IMDb. Questo sollevò preoccupazioni legate alla privacy, poiché le preferenze cinematografiche possono rivelare orientamenti politici, religiosi o sessuali.

Other Privacy Breaches

L'uso di app per il fitness che tracciano la posizione degli utenti ha mostrato come mappe dettagliate possano esporre informazioni sensibili sulla posizione e sull'identità delle persone.

Syntactic vs Semantic Privacy Definitions

- **Syntactic Privacy Definitions** Le definizioni di privacy sintattiche misurano il grado di protezione di una persona nei dati con un valore numerico. Ad esempio:
  - Ogni rilascio di dati deve essere indistinguibilmente associato ad almeno un certo numero di individui nella popolazione.
- **Semantic Privacy Definitions** Le definizioni di privacy semantiche soddisfano un requisito di privacy semantico. Ad esempio:
  - Il risultato di un'analisi eseguita su un dataset rilasciato non deve essere influenzato dalla presenza o assenza di una singola tupla nel dataset.

Differential Privacy

- **Informal Definition** La *differential privacy* mira a prevenire che un avversario sia in grado di rilevare la presenza o assenza di un individuo in un dataset. Ad esempio, il conteggio degli individui affetti da una malattia in un database medico, con un meccanismo che fornisce probabilmente lo stesso risultato su dataset che differiscono per un solo individuo.
- **Formal Definition** Una funzione randomizzata K fornisce ε-differential privacy se per tutti i dataset D e D' che differiscono per al massimo una riga, e per ogni insieme $S \subseteq \text{Range}(K)$:

$$\Pr[K(D) \in S] \leq e^{\varepsilon} \times \Pr[K(D') \in S]$$

Differential Privacy Scenarios

La differential privacy si applica a due scenari principali:

- **Scenario Interattivo**: valutazione di query in tempo reale (Statistical DBMS).
- **Scenario non Interattivo**: rilascio di tabelle di macro-dati pre-calcolate (Statistical Data).

Essa viene solitamente implementata tramite l'aggiunta di rumore casuale, ma ciò non preserva necessariamente la veridicità dei dati.

Differential Privacy Variations

Per ridurre la quantità di rumore aggiunto, sono state proposte diverse varianti:

- **(ϵ, δ)-differential privacy**: l'approssimazione ϵ può essere violata con bassa probabilità (controllata da δ).
- Metodi basati su trasformate wavelet per migliorare l'utilità dei dati.

Differential Privacy Applications

Meccanismi basati su differential privacy sono stati sviluppati per vari domini:

- **Social networks**
- **Data mining**
- **Location data**

Is Differential Privacy Enough?

La differential privacy limita l'inferenza sulla presenza di una tupla, ma non necessariamente l'inferenza sulla partecipazione di un individuo al processo di generazione dei dati. Ad esempio, la partecipazione di Bob in un social network potrebbe influenzare le relazioni tra i suoi amici, non solo la sua tupla.

k-anonymity vs Differential Privacy

- **k-anonymity**:
  - **Pro**: Cattura bene i requisiti del mondo reale.
  - **Contro**: Non offre una protezione completa.
- **Differential Privacy**:
  - **Pro**: Offre garanzie di protezione migliori.
  - **Contro**: Non facile da capire o applicare, e non garantisce una protezione completa.

0.3 Some Examples of Other Privacy Issues

0.3.1 Sensitive Value Distributions

- Le tuple individuali non sono intrinsecamente sensibili.
- Una raccolta di tuple potrebbe rivelare informazioni sensibili non esplicitamente riportate, in particolare a causa di distribuzioni di valori peculiari.

Example: Soldiers' Medical Records

- I record individuali non sono sensibili.
- La distribuzione dell'età dei soldati in una località può indicare il tipo di località:
  - Soldati giovani suggeriscono tipicamente un campo di addestramento.
  - Funzionari più anziani indicano un quartier generale.

Counteracting Inference Channels

- La valutazione dell'esposizione dei dati rilasciati può essere effettuata attraverso:
  - ll calcolo a priori del numero massimo di tuple rispetto alla baseline distribution, inclusi il numero di rilasci per diversi valori di attributi.
  - La valutazione delle metriche di esposizione sulle tuple richieste.

0.3.2 Privacy and Genomic Data

Le informazioni genomiche presentano opportunità in medicina ma sollevano anche diversi problemi di privacy:

- Il genoma umano può identificare il suo proprietario.
- Contiene info sensibili sulla provenienza etnica, predisposizione a malattie e altri tratti fenotipici.
- I dati genomici possono rivelare informazioni sui parenti e sui discendenti sulla base del genoma.

Individuals' Re-identification

Example: The 1000 Genomes Project:

Un'iniziativa internazionale avviata nel 2008 per definire un catalogo della variazione genetica umana. Durante il progetto, cinque uomini coinvolti sia in questo progetto sia in uno studio su famiglie mormoni nello Utah sono stati riidentificati attraverso un'analisi incrociata condotta dal Whitehead Institute for Biomedical Research. Il processo di identificazione includeva l'estrazione degli haplotipi del cromosoma Y, l'inserimento di questi dati in database genealogici per identificare possibili cognomi e l'uso di tali cognomi in database demografici per raccogliere ulteriori informazioni sui donatori e i loro familiari.

0.3.3 Sensitive Inference from Data Mining

The Target case:

Target, il secondo rivenditore di sconti più grande negli Stati Uniti, assegna a ogni cliente un numero di identificazione (Guest ID) collegato a informazioni personali come carta di credito, nome e indirizzo email. Questa identificazione consente a Target di raccogliere e analizzare la storia degli acquisti dei clienti per scopi pubblicitari mirati. Gli analisti di Target hanno identificato circa 25 prodotti che consentono di assegnare a ciascun acquirente un punteggio di previsione della gravidanza. Ad esempio, una donna di 23 anni che acquista lozione al burro di cacao, una borsa sufficientemente grande da fungere da borsa per pannolini, integratori di zinco e magnesio e un tappeto blu brillante potrebbe avere un'87% di probabilità di essere in attesa a fine agosto. Queste informazioni vengono utilizzate per inviare coupon in momenti specifici durante la gravidanza. L'analisi dei dati ha rivelato eventi significativi nella vita dei clienti, come lauree, nuovi lavori o traslochi, rendendo le abitudini di acquisto più flessibili e prevedibili, il che si traduce in un potenziale profitto significativo per i rivenditori. Tra il 2002 e il 2010, le entrate di Target sono cresciute da 44 miliardi a 67 miliardi di dollari grazie a queste campagne di marketing mirato.

0.3.4 Social Media

Profiling in Social Media

Le nostre attività sui social media e i "like" possono rivelare informazioni sensibili. È importante notare che i social media condividono frequentemente i nostri dati con terze parti, come inserzionisti e aziende di analisi, il che può portare a violazioni della privacy.

Cambridge Analytica Scandal

Un esempio eclatante è lo scandalo di Cambridge Analytica, dove i dati di milioni di utenti di Facebook sono stati raccolti senza il consenso degli stessi per influenzare campagne politiche. Questo caso ha messo in luce le vulnerabilità legate alla privacy degli utenti e all'uso improprio di tali informazioni.

User Profiling - OCEAN Model

Inoltre, la profilazione degli utenti avviene attraverso modelli come il modello OCEAN (apertura, coscienziosità, estroversione, amicalità e nevroticismo), che categorizza gli utenti in base a tratti della personalità, consentendo una pubblicità mirata e la manipolazione comportamentale.

Biometric Data Privacy

Infine, la privacy dei dati biometrici, come il riconoscimento facciale, solleva ulteriori preoccupazioni. Questi sistemi possono identificare gli individui senza

il loro consenso, sollevando interrogativi etici e legali sulla sorveglianza e sulla protezione dei dati personali.