**Micro vs Macro data**

I dati vengono rilasciati per scopi statistici o per statistiche in sé, può avvenire un’inferenza che non si voleva ottenere dalla release dei dati, essa può dereivare solamente dai dati in sé oppure dai dati in combinazione con una sorgente esterna di dati.

Statistical DBMS: rispondono solo a query statistiche.

Statistical data: rappresentano una statistica

Problema statistical dbms: l’informazione di un individuo può rivelarsi dalla correlazione di più statistiche o da una statistica e dell’external knowledge, esempio sommo in una query, sommo in un’altra query e poi faccio differenza e ottengo il valore di un individuo, questo non va bene per la privacy.(combinazione di query sensibile).

Macrodata count/freq table, magnitude table già viste capitolo prima.

**Macrodata Disclosure Protection Techniques: Tables of Counts or Frequencies**

Sampling, special rules, threshold rules

Sampling: prendo una parte della popolazione, i risultati sono combinati con un peso per rendere meno identificabili i singoli individui.

Special rules: provvedimento quando la popolazione usata nei macrodati è utilizzata per intera, limitano il dettaglio rispetto all’ottenimento dell’informazione esempi: an individual’s age within a five-year interval, earnings within a $1,000 interval, benefits within a $50 interval.

Esempio della tabella che non rispetta l’intervallo di 5k, combina le righe ma è possibile farlo anche su colonne, per togliere il problema.

Threshold rules: una cella è sensibile se l’informazione deriva da meno di tot individui.

Tecniche per evitare celle sensibili: table restructuring and category combination ◦ cell suppression ◦ random rounding ◦ controlled rounding ◦ confidentiality edit.

La tecnica più usata per le celle sensibili è la soppressione che però dev’essere acompagnata da una soppressione secondaria, fondamentale altrimenti il valore della prima soppressione viene ritrovato dalla differenza con i valori marginali, la soppressione secondaria può avvenire su colonna o su riga. (si usa linear programming)

Comunque la soppressione secondaria NON garantisce protezione.

Rounding (random o controlled) evita così tanta data loss causata dalla soppressione

1. Random: il valore viene aumentato o diminuito, il totale dei marginali può non corrispondere
2. Controlled: se aumento uno diminuisco l’altro il totale dei marginali corrisponde

(linear programming)

Confidentiality edit: creato per proteggere informazioni del census.(censimento)

In sostanza le informazioni che si reputano sensibili vengono scambiate per alcuni records con altri records in base a dei matching. Procedimento:

1. Prendo un sample dal microdata file
2. Trovo un match su certi attributi (matching on a specified set of important attributes)
3. Scambio il valore degli attributi di quei records

L’obiettivo è perturbare un po’ i dati ma non troppo, quindi se mi serve education level e company, il matching lo trovo sugli altri attributi da un’altra parte (race e salary),

Dopo cambio attributi, quindi a parità di sesso e razza avrò education level un pochino perturbato.

Il matching posso farlo io sul mio dataset ma anche altri lo possono fare sul mio, quindi oltre al mio 10% dei dati ne cambierà un'altra porzione di dati un altro 10% derivante dal matching con altri datasets i quali hanno scambiato tuple con il mio.

Limito lo statistical disclosure.

**Macrodata Disclosure Protection Techniques: Tables of Magnitude Data**

I valori outliers sono i più colpiti dalla disclosure e le tecniche di protection vanno a limitare la visibilità di questi elementi.

Le regole di soppressione cercano di capire se il valore da reperire è abbastanza difficile da trovare o da stimare, cioè è abbastanza difficile per un rispondente stimare il valore di un altro rispondente in modo abbastanza vicino.

Regole di soppressione:

P-percent rule: una cella è considerata sensibile se il valore stimato come upper bound e il valore stimato come lower bound sono più vicini al valore reale rispetto al valore di una percentuale p.

Formula su slides

Pq rule: In questo caso si assume che sia presente un knowledge precedente, quindi prima della pubblicazione dei dati era possibile inferire un valore pari a q. Il resto rimane uguale alla p-percent rule (p < q < 100)

Q = 100 non c’è nessun knowledge precedente alla release, in sostanza la q dice quanto lontana è la stima del valore prima della release.

Q = 80% alice = 100 inferisco prima della release alice prende da 20 a 180. Stima di lontananza non vicinanza.

Formula slides

(n,k) rule: Se in un gruppo di persone, alcune di esse (n o meno) contribuiscono ad una grossa percentuale (k o +) rispetto al valore totale allora la cella è da considerarsi sensibile.

In sostanza significa che un gruppo di poche persone contribuisce alla maggior parte del valore.

Quelle tecniche viste di soppressione di macrodati riguardavano l’identificazione di una cella sensibile, una volta identificata ho due opzioni:

O ricostruisco la tabella in modo che tutte le celle sensibili se ne vanno, oppure sopprime la cella sensibile e un’altra (soppressione secondaria).

Per la soppressione secondaria, sia che ricostruisco la tabella, sia che sopprimo altre celle, è importante verificare che i valori soppressi non possano essere inferiti dagli altri valori tramite equazioni di riga o colonna o altre celle.

Le celle da scegliere per la soppressione secondaria di solito vengono scelte a mano se la tabella è piccola.

Per vedere se la tabella garantisce abbastanza protezione si applica un audit che è una tecnica di programmazione lineare per stimare i bound delle celle soppresse, se la stima è troppo vicina al valore reale allora la cella è sensibile.

È importante che le celle soppresse secondariamente non apportino una grande information loss, anche se non c’è una definizione precisa di information loss essa si può intuire dal contesto.

**Microdata**

Utilizzati sempre di più in contesti recenti, esempi machine learning però più sensibili ad attacchi.

Le tecniche principali per proteggere questi dati sono o cambiare il significato dei dati aggiungendo noise oppure riducendo il contenuto dei dati

**Microdata disclosure protection technique**

Metodi principali: • including data from a sample of the whole population only • removal of identifiers • limiting geographic details • limiting the number of variables

**Classification of microdata protection techniques**

• masking techniques • synthetic data generation techniques

Masking = perturbare i dati oppure non rilasciarli (soppressione i guess)

Synthethic = generare dati verosimili ma che non siano uguali a quelli originali.(però fungano per lo stesso scopo)

**Microdata Disclosure Protection Techniques: Masking Techniques**

Consiste nel rilasciare dati utili per scopi statistici ma in modo da mantenere confidenzialità nella release (partendo dai dati originali ovviamente).

Queste tecniche possono essere perturbative o non perturbative.

Perturbative: modificano il valore dei dati

Non perturbative: eliminano alcuni dettagli o intere informazioni, NON modificano il valore dei dati

Sampling: Usare un campione della popolazione totale, la protezione è: il rispondente può esserci o non esserci (uncertainity about data)

Local suppression: elimina (sopprime) il valore di un attributo, tipicamente si elimina il valore dell’attributo più sensibile che può provocare maggiore disclosure per quel rispondente.

Quali attributi scegliere e considare più sensibili dipende dal contesto.

Global recording: al posto che tenere il valore preciso per ogni rispondente ad ognuno di essi ci associo l’etichetta relativa all’intevallo di cui quel valore fa parte.

Top-coding: definisce un valore oltre il quale i valori sono considerati sensibili e quindi non vengono riportati in tabella ma vengono sostituiti da una flag che indica che quel valore supera il limite indicato

Bottom-coding: definiscono una soglia sotto la quale i valori sono definiti sensibili, andare sotto questa soglia non è consentito infatti i valori più bassi della soglia non vengono rilasciati ma viene indicato che sono al di sotto di una certa soglia tramite una flag.

Generalization: rimpiazza i valori con valori più generici es: data di nascita.

È basato sul concetto di gerarchia di generalizzazione, in cui la root è il valore + generale possibile e le foglie sono i valori originali (+ specifici possibili).

Random noise: perturba i dati aggiungendo o moltiplicando ad essi il valore di una variabile data da una certa distribuzione, la pubblicazione della distribuzione è sconsigliato in quanto aumenta il disclosure risk

Swapping: introduce incertezza, trovo un matching su certi attributi su tuple nello stesso file e scambio il valore degli attributi non matchati.

Micro-aggregation (blurring): raggruppo le singole tuple in piccoli gruppi (es su certi attributi sex and marital status) e al posto di pubblicare i singoli valori dei singoli rispondenti pubblicherò la media di quei valori.

**Microdata Disclosure Protection Techniques: Synthetic Techniques**

I dati sintetizzati devono presentare la stessa qualità per lo scopo a cui servono, cioè devono fornire gli stessi risultati per quanto riguarda possibili statistiche.

Il vantaggio principale è che questi dati non sono correlati a nessun rispondente e quindi non possono portare alla re-identificazione dei rispondenti stessi