Protezione di Macrodata e Microdata

Parte II

Indice

Spesso il rilascio di dati statistici può inferire a dati non intesi per il rilascio. La rivelazione può avvenire:

- con i soli dati rilasciati
- dalla combinazione dei dati rilasciati con informazioni disponibili al pubblico
- dalla combinazione dei dati rilasciati con informazioni provenienti da altre fonti

Capitolo 1

Statistical DBMS

Un **DBMS statistico** è un DBMS che offre accesso a statistiche su gruppi di invidui. Non deve rivelare nessuna informazione su nessun individuo in particolare.

Le informazioni confidenziali possono essere dedotte:

- combinando i risultati di statistiche differenti
- combinando i risultati delle statistiche con conoscenza esterna

Name	Sex	Major	Class	Income
Allen	Female	CS	1980	68k
Baker	Female	EE	1980	50k
Cook	Male	EE	1978	70k
Davis	Female	CS	1978	80k
Evans	Male	EE	1981	60k
Frank	Male	CS	1978	76k
Good	Male	CS	1981	64k
Hall	Male	EE	1978	60k
Iles	Male	CS	1979	70k

Query 1: sum of the incomes of individuals with major in EE (240k) −
Query 2: sum of the incomes of males with major in EE (190k)
= sum of the incomes of females with major in EE (50k)
income of Baker

$\Longrightarrow$ The combination of queries is sensitive

Una *sensitive query* è una query che può provocare una *disclosure*, ovvero la rivelazione di informazioni sensibili su un individuo. Le query, prese singolarmente, potrebbero non essere sensibili, ovvero non rivelare informazioni confidenziali. Tuttavia, un insieme di query considerate nel loro complesso può

diventare sensibile. Questo fenomeno è noto come ***collusione***. Attraverso la collusione, le informazioni aggregate da query non sensibili possono portare alla deduzione di dati privati o confidenziali.

Questa è la ragione per cui mi serve il controllo basato sulla storia: devo tenere traccia di quello che mi chiedi e della conoscenza che hai, e quindi di cosa puoi inferire.

Capitolo 2

Protezione per Tabelle di Conteggio o Frequenze

La protezione di questo tipo di tabelle si divide in tre fasi:

1. Sampling
2. Identificazione delle celle sensibili
   - special rules
   - threshold rules
3. Protezione delle celle sensibili
   - table restructuring
   - soppressione
   - rounding
   - confidentiality edit

2.1 Sampling

Stabilisco un campione della popolazione totale (che sia rappresentativo, senza bias, ...) e faccio la statistica su tale campione.

Prima di aggregare i dati, i singoli valori vengono moltiplicati per un peso (*sampling weight*); in questo modo viene mantenuta la correlazione statistica dei dati ma introducendo del rumore (se i pesi non vengono pubblicati), rendendo più difficile identificare i dati dei singoli rispondenti dai valori pubblicati.

2.2 Identificazione delle celle sensibili

2.2.1 Regole Speciali

Le regole speciali definiscono il livello di dettaglio oltre il quale non è consentito rilasciare dati.

Vengono chiamate in questo modo perché dipendono dall'agenzia e dal tipo di tabella (dominio di applicazione).

Per soddisfare le regole speciali si può utilizzare:

- table restructuring
- category combination

2.2.2 Threshold rules

Una cella viene considerata sensibile se il numero di rispondenti è inferiore a un certo numero specificato.

2.3 Protezione delle celle sensibili

2.3.1 Table Restructuring

La tabella viene ristrutturata e righe o colonne vengono combinate (*rolling-up categories*).

2.3.2 Soppressione delle celle

È una delle tecniche di protezione più usata. La sola soppressione delle celle sensibili non è sufficiente (**soppressione primaria**): è necessaria una seconda soppressione (**soppressione complementare**) per ogni riga e colonna in cui viene soppressa una cella sensibile, altrimenti il valore della cella sensibile potrebbe essere calcolabile dal totale marginale.

La scelta delle celle per la soppressione complementare è un problema difficile; possono essere utilizzati modelli di programmazione lineare in cui l'obiettivo è massimizzare o minimizzare una funzione obiettivo, soggetta a dei vincoli:

- la funzione obiettivo potrebbe essere la minimizzazione delle celle soppresse o la massimizzazione della protezione dei dati
- i vincoli possono essere dei requisiti di riservatezza, come il numero minimo di celle da sopprimere o la necessità di mantenere la validità statistica dei dati

2.3.3 Rounding

Per ridurre la perdita di dati dovuta alla soppressione, si utilizza il *rounding* dei valori a un multiplo della soglia di sensibilità. Esistono due approcci possibili:

- **Random:** viene scelto in maniera casuale se arrotondare i valori per eccesso o per difetto
  - la conseguenza è che la somma dei valori in una riga o in una colonna potrebbe differire dai totali marginali
- **Controllato:** garantisce che la somma delle righe e colonne siano uguali ai totali marginali
  - **Vantaggi:** garantisce che i dati pubbliati siano coerenti con i totali marginali
  - **Svantaggi:** richiede l'uso di programmi informatici specializzati per il calcolo delle soluzioni di arrotondamento; non sempre potrebbero esistere soluzioni

Nota

Tutti i valori delle celle devono essere multiplo del valore di *sensitivity threshold*; è fondamentale per mantenere la riservatezza e l'integrità dei dati.

2.3.4 Confidentiality Edit

La *Confidentiality Edit* è stata sviluppata dal U.S. Census Bureau per fornire protezione alle tabelle preparate utilizzando il Censimento del 1990. Questa tecnica si basa su due approcci distinti:

- **Protezione dei dati del Censimento decennale regolare** (100% della popolazione).
- **Protezione Long-Form del Censimento** (campione della popolazione).

Entrambi gli approcci applicano tecniche di **statistical disclosure limitation** ai microdati sui quali sono calcolate le statistiche, protezione ottenuta mediante la modifica dei dati di input.

Procedura di Switching per il file di microdati al 100% → Per il file di microdati al 100% (1° caso), la *confidentiality edit* si applica attraverso un processo di switching:

1. Prendere un campione di record dal file di microdati.
2. Trovare una corrispondenza per tali record in un'altra regione geografica, effettuando il matching in base a un insieme specifico di attributi importanti.
3. Scambiare tutti gli attributi sui record corrispondenti.

Per piccoli blocchi, la frazione di campionamento può essere aumentata per fornire una protezione aggiuntiva. Se aumentiamo la frazione di campionamento e decidiamo di includere più soggetti nel campione, ci saranno più dati a disposizione e più variabilità, riducendo il rischio di identificare singoli residenti basandosi sui risultati. Inoltre, il file di microdati può essere utilizzato direttamente per preparare tabelle di macrodata, mantenendo le correlazioni statistiche ottenibili dai *raw data*.

2.4 Macrodata Disclosure Prot. Techniques: Tables of M

2.4.1 Data Protection in Magnitude Tables

I *Magnitudine Data* sono generalmente quantità non negative riportate in sondaggi o censimenti. È probabile che la distribuzione di questi valori sia asimmetrica (skewed). Le tecniche di limitazione della divulgazione si concentrano sulla prevenzione della stima precisa dei valori per gli outlier.

- Esempi: informazioni su redditi, spese o altre quantità aggregate.

- Le tecniche di protezione si concentrano principalmente su come gestire i valori estremi (*outliers*), maggiormente a rischio di disclosure di informazioni sensibili.
- Il campionamento è meno efficace nel fornire protezione (NO riservatezza *outliers*).

La protezione delle tabelle di dati di magnitudine implica diversi passaggi fondamentali:

1. **Identification of sensitive cells**
   - **p-percent**: soglia percentuale sotto la quale i valori possono essere considerati sensibili.
   - **pq**: combinazione di valori che possono rivelare informazioni riservate.
   - **(n,k)**: criterio basato su dimensioni per identificare celle sensibili.
2. **Protection of sensitive cells**
   - **Suppression**: tecnica comune utilizzata per prevenire la divulgazione di dati riservati.
3. **Verify result**
   - **Audit**: processo che verifica se le tecniche di protezione sono state applicate correttamente.
   - **Information loss**: analisi dell'*information loss* per garantire la significatività dei dati.
   - **Parameters are not disclosed**: importante per mantenere la riservatezza dei dati.

2.4.2 Suppression Rules - 1

Le *regole di soppressione* sono fondamentali per proteggere le informazioni sensibili nelle tabelle di dati. Esse si concentrano su due aspetti principali:

- **Regole di Soppressione Primarie**: Determinano se una cella potrebbe rivelare informazioni su un singolo rispondente. Se una cella è considerata sensibile, non può essere rilasciata.
- **Regole di Soppressione Comuni**: Le regole più comuni includono:
  - *p-percent rule*: Stabilisce una soglia percentuale al di sotto della quale i valori delle celle sono considerati sensibili.
  - *pq rule*: Si basa sulla combinazione di variabili per determinare la sensibilità delle celle.
  - *(n,k) rule*: Utilizza un criterio dimensionale per identificare le celle sensibili.

Verifica della Sensibilità Queste regole sono utilizzate per identificare celle sensibili verificando se è sufficientemente difficile per un rispondente stimare il valore riportato da un altro rispondente in modo troppo preciso.

2.4.3 Suppression Rules - 2

Le **Primary Suppression Rules** determinano se una cella potrebbe rivelare informazioni su un singolo rispondente → in tal caso tali celle sono considerate sensibili e non possono essere rilasciate. Le regole di soppressione più comuni sono:

- la *the p-percent rule*
- la *the pq rule*
- la *the (n,k) rule*

Queste regole vengono utilizzate per identificare celle sensibili verificando se è sufficientemente difficile per un rispondente stimare con accuratezza il valore riportato da un altro rispondente.

2.4.4 Primary Suppression Rule: p-percent

Tale regola stabilisce che la divulgazione di informazioni sensibili a partire dai *Magnitude Data* si verifica se l'utente può stimare troppo accuratamente il contributo di un rispondente.

- Una cella è considerata **sensibile** se le stime superiori e inferiori per il valore del rispondente sono più vicine al valore riportato di una percentuale predefinita p.
- Formalmente, una cella è considerata protetta se:

$$\sum_{i=c+2}^{N} x_i \geq \frac{p}{100} x_1$$

  dove:

  - $x_1, x_2, \ldots, x_N$: valori dei rispondenti in ordine decrescente,
  - c: dimensione di una coalizione di rispondenti interessati a stimare x_1 (*collusione*).
  - il valore più grande x_1 è il più esposto (*outlier*).

2.4.5 Primary Suppression Rule: pq

La **p-percent rule** assume che non ci sia alcuna conoscenza precedente sui valori dei rispondenti. Tuttavia, le agenzie non dovrebbero fare questa assunzione.

- Nella regola pq, le agenzie possono specificare quanto sia nota precedentemente l'informazione assegnando un valore q, che rappresenta quanto accuratamente i rispondenti possono stimare il valore di un altro rispondente prima che i dati vengano pubblicati ($p < q < 100$).
- Il parametro q rappresenta l'errore nell'estimazione prima che la cella venga pubblicata.
- Formalmente, una cella è considerata protetta se:

$$\frac{q}{100} \sum_{i=c+2}^{N} x_i \geq \frac{p}{100} x_1$$

  dove:

  - $x_1, x_2, \ldots, x_N$: valori dei rispondenti in ordine decrescente,
  - c: dimensione di una coalizione di rispondenti interessati a stimare x_1 (*collusione*).
  - il valore più grande x_1 è il più esposto (*outlier*).
  - la *pq rule* si riduce alla *p-percent rule* quando $q = 100$ (cioè, nessuna capacità di stima).

2.4.6 Primary Suppression Rule: (n,k)

Tale regola stabilisce che, indipendentemente dal numero di rispondenti in una cella, se un numero ridotto (n o meno) di questi rispondenti contribuisce a una grande percentuale (k% o più) del valore totale della cella, la cella è considerata sensibile.

- **Regola Intuitiva**: Se una determinata cella è dominata da un solo rispondente, il totale pubblicato rappresenta una stima superiore per il suo valore.
- Il valore n è selezionato per essere maggiore del numero di eventuali coalizioni sospettate.
- Molte agenzie utilizzano una regola (n, k) con $n = 1$ o $n = 2$.

Esempio: n = 2 & k = 70: cella sensibile se 2 tuple contengono f(info sensibile) > 70% tot
(f: sum or avg or ...)

2.4.7 Secondary Suppression

Una volta identificate le celle sensibili, ci sono due opzioni:

- **Ristrutturare la tabella** e combinare le celle fino a quando non rimangono più celle sensibili.

- **Soppressione delle celle**: non pubblicare celle sensibili (*Primary Suppression*) e rimuovere altre celle (*Secondary Suppression*). Un modo amministrativo per evitare la soppressione delle celle consiste nell'ottenere un permesso scritto dai rispondenti.

- **Secondary Sup:** è necessario selezionare altre celle non sensibili per la *suppression* per garantire che i dati a livello di rispondente nelle celle sensibili non possano essere stimati con troppa accuratezza → i dati di un rispondente non possono essere stimati troppo dettagliatamente.

- Le celle sensibili potrebbero essere divulgate a causa del fatto che:
  - le unioni delle celle soppresse possono essere sensibili secondo la regola di sensibilità adottata,
  - le equazioni delle righe e delle colonne rappresentate dalla tabella pubblicata possono essere risolte, e il valore per una cella soppressa stimato con troppa accuratezza.

Qualsiasi soppressione complementare è accettabile fintanto che le celle sensibili sono protette:

- Per tabelle piccole, la selezione delle celle complementari può essere fatta manualmente. Gli analisti dei dati sanno quali celle sono di maggiore interesse e non dovrebbero essere utilizzate per la soppressione complementare. La selezione manuale delle celle complementari è accettabile purché la tabella risultante fornisca una protezione sufficiente per le celle sensibili.

- Un audit automatizzato dovrebbe essere applicato per garantire che ciò sia vero.

2.4.8 Audit

Se i totali vengono pubblicati, la somma delle celle soppresse (primary/secondary) può essere derivata. È necessario applicare la regola di sensibilità a queste somme per garantire che non siano sensibili.

- Le righe e le colonne possono essere viste come un grande sistema di equazioni lineari.

- Stimare un *lower bound* e *upper bound* di ciascuna cella soppressa utilizzando *linear programming*.

- Se i limiti sono troppo vicini al valore originale, la cella è considerata sensibile.

Semplice per tabelle piccole, ma potrebbe risultare computazionalmente infattibile per tabelle grandi.

2.4.9 Information Loss

La selezione delle celle complementari dovrebbe comportare una minima perdita di informazioni. Non esiste una definizione unica di perdita di informazioni.

- Ad esempio, possiamo cercare di minimizzare:
  - la somma dei valori soppressi (alto numero di celle con valori piccoli può essere soppresso),
  - il numero totale di celle soppresse.

2.4.10 Information in Parameter Values

Mentre le *sup rules* possono essere pubblicate, i valori dei parametri dovrebbero rimanere riservati. Una volta che il valore di una cella soppressa è stato determinato in modo univoco, i valori delle altre celle sono facilmente derivabili.

2.5 Microdata

2.5.1 Macrodata vs Microdata

In passato, i dati venivano principalmente rilasciati in forma tabellare (macrodata) e DBMS. Oggi, molte situazioni richiedono che i dati specifici memorizzati, chiamati microdata, siano rilasciati.

- Aumento della flessibilità e disponibilità delle informazioni per i destinatari.
- I microdata sono soggetti a un maggiore rischio di violazioni della privacy (linking attacks).

2.5.2 Microdata

Per proteggere la privacy dei rispondenti, i data owner spesso rimuovono/crittografano identificatori espliciti come nomi, indirizzi e numeri di telefono. Tuttavia, la de-identificazione dei dati non offre alcuna garanzia di anonimato. Le informazioni rilasciate spesso contengono altri dati quasi identificativi (ad esempio, razza, data di nascita, sesso e codice postale) che possono essere collegati a informazioni disponibili pubblicamente per reidentificare i rispondenti o ridurre l'incertezza sulle loro identità. I destinatari dei dati possono determinare (o limitare l'incertezza) a quale rispondente si riferiscono alcuni dati rilasciati. Questo ha creato una crescente domanda nel dedicare risorse per una protezione adeguata dei dati sensibili. Le tecniche di protezione dei microdata seguono due strategie principali:

- ridurre il contenuto informativo (P),
- modificare i dati in modo che il contenuto informativo venga mantenuto il più possibile (NP).

2.5.3 Microdata Disclosure Protection Techniques

Per limitare il rischio di divulgazione, devono essere applicate le seguenti procedure:

- Inclusione di dati provenienti solo da un campione dell'intera popolazione.
- Rimozione degli identificatori.
- Limitazione dei dettagli geografici.
- Limitazione del numero di variabili.

2.5.4 Classification of Microdata Protection Techniques

Queste tecniche si basano sul principio che la reidentificazione può essere contrastata riducendo la quantità di informazioni rilasciate:

- Mascheramento dei dati (ad esempio, non rilasciando o perturbando i loro valori).
- Rilascio di valori plausibili al posto di quelli reali.

Secondo questo principio, le tecniche di protezione possono essere classificate in due categorie principali:

- **Masking techniques (perturbative or not perturbative)**
- **Synthetic data generation techniques**

2.5.5 Microdata Types

Le tecniche di protezione possono operare su diversi tipi di dati:

- **Continuous**. attr. definito continuo se numerico e su di esso sono definite operazioni aritmetiche. **Esempio**: data di nascita, temperatura, ecc.
- **Categorical**. attr. definito categorico se può assumere un insieme limitato e specificato di valori e le operazioni aritmetiche non hanno senso su di esso. **Esempio**: stato civile, razza, ecc.

2.6 Microdata Disclosure Prot. Techniques: Masking

2.6.1 Masking Techniques

Le tecniche di masking trasformano i dati originali per produrre nuovi dati che sono validi per l'analisi statistica e che preservano la riservatezza dei rispondenti. Esse sono classificate come segue:

- **Non-perturbative**: i dati originali non vengono modificati, ma alcuni dati vengono soppressi e/o alcuni dettagli vengono rimossi.
- **Perturbative**: i dati originali vengono modificati.

2.6.2 Sampling

La tabella di microdata protetta è ottenuta come un campione della tabella di microdata originale. Incertezza riguardo alla partecipazione di un rispondente → il rischio di reidentificazione diminuisce.

2.6.3 Local Suppression

La local suppression sopprime il valore di un attributo (cioè lo sostituisce con un valore mancante), limitando così le possibilità di analisi. Questa tecnica annulla alcuni valori degli attributi/celle sensibili che potrebbero contribuire in modo significativo al rischio di divulgazione della tupla coinvolta.

2.6.4 Global Recoding

Il global recoding comporta la suddivisione del dominio di un attributo in diversi intervalli disgiunti, tipicamente della stessa ampiezza, e ad ogni intervallo è associata *label*. La tabella di microdata protetta è ottenuta sostituendo i valori dell'attributo con la label associata all'intervallo corrispondente.

2.6.5 Top-coding e Bottom-coding

Il top-coding definisce un limite superiore (top-code) per ciascun attributo da proteggere. Qualsiasi valore maggiore di questo limite viene sostituito con un flag che informa l'utente del top-code e indica che il valore supera questo limite. Questa tecnica può essere applicata ad attributi categorici che possono essere ordinati linearmente, così come ad attributi continui. Analogo per bottom-coding rispetto al limite inferiore (bottom-code).

2.6.6 Generalization

La generalization sostituisce i valori con valori più generali. È tipicamente basata sulla definizione di una gerarchia di generalizzazione, dove il valore più generale è la radice e le foglie corrispondono ai valori più specifici. Possono essere costruite diverse tabelle di microdata generalizzate, a seconda del numero di passi di generalizzazione applicati.

2.6.7 Random Noise

Il random noise perturba un attributo sensibile aggiungendo o moltiplicando il valore di quest'ultimo con una variabile casuale di una distribuzione specificata. È necessario decidere se pubblicare o meno la distribuzione/i utilizza-

ta/e per aggiungere rumore ai dati. La pubblicazione della distribuzione/i potrebbe aumentare il rischio di divulgazione dei dati (Disclosure di Informazioni Sensibili).

2.6.8 Swapping

Una piccola percentuale di record viene abbinata con altri record nello stesso file, forse in diverse regioni geografiche, su un insieme di variabili predeterminate. I valori di tutte le altre variabili nel file vengono quindi scambiati tra i record selezionati. Questa tecnica riduce il rischio di reidentificazione poiché introduce incertezza riguardo al valore reale dei dati di un rispondente.

2.6.9 Micro-aggregation (Blurring)

La micro-aggregation consiste nel raggruppare tuple individuali in piccoli gruppi di dimensione fissa k. Viene pubblicata la media di ciascun gruppo invece dei valori individuali. I gruppi sono formati utilizzando criteri di massima similarità. Ci sono diverse variazioni della micro-aggregazione:

- la media può sostituire il valore originale solo per una tupla nel gruppo o per tutte;
- diversi attributi possono essere protetti attraverso la micro-aggregazione utilizzando lo stesso o diversi raggruppamenti;
- ...

2.7 Microdata Disclosure Prot. Techniques: Synthetic

2.7.1 Synthetic techniques

Poiché il contenuto statistico dei dati non è correlato alle informazioni fornite da ciascun rispondente, un modello che rappresenti bene i dati potrebbe, in linea di principio, sostituire i dati stessi. Un requisito importante per la generazione di dati sintetici è che tali dati e quelli originali devono presentare la stessa qualità nell'analisi statistica. Il principale vantaggio di questa classe di tecniche è che i dati sintetici rilasciati non sono riferiti a nessun rispondente e, pertanto, il loro rilascio non può portare a reidentificazione.