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 di Macrodata: 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* viene applicata attraverso un processo di switching:

1. si prende un campione di record dal file di microdati
2. si trova una corrispondenza per tali record in una popolazione contenente altri record (ad esempio un'altra regione geografica)
3. si scambiano tutti gli attributi sui record che corrispondono

Nota

Si opera non sulla statistica ma sui dati su cui viene prodotta tale statistica; la protezione è garantita dal fatto che sto introducendo del rumore nei dati.

Capitolo 3

Protezione di Macrodata: tabelle di magnitudo

È probabile che la distribuzione dei valori riportati nelle tabelle di magnitudo sia asimmetrica (*skewed*); le tecniche di limitazione di disclosure si concentrano sulla prevenzione della stima precisa dei valori per gli outlier.

3.1 Identificazione delle celle sensibili

Per identificare le celle sensibili si utilizzano le *regole di soppressione*, che cercano di verificare se è sufficientemente difficile per un rispondente stimare il valore di un altro rispondente in modo troppo preciso.

Queste regole vengono definite **regole di soppressione primaria**.

3.1.1 *p-percent* rule

Questa regola stabilisce una soglia percentuale al di sotto della quale i valori delle celle sono considerati sensibili.; verifica se è sufficientemente difficile per un rispondente 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 stabilita p. (è possibile inferire sotto a un certo intervallo di incertezza il dato)
- 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*).

Esempio

- Consider the respondents that contribute to the total income in a city, which is equal to 250K, to be (in decreasing order)
  - Alice: 100K
  - Bob: 80K
  - Carol: 30K
  - David: 20K
  - Eve: 10K
  - Frank: 3K
  - ...
- Which is the coalition of $c = 3$ respondents that can better estimate **Alice**'s income?
  Bob, Carol, David, whose total income is 130K
  can estimate that Alice's income is between **80K** and **120K**
  $\Longrightarrow$ sensitive for any p$\geq$20

3.1.2 *pq* rule

Possiamo definire questa regola come un affiniamento della regola precedente, che introduce la conoscenza pregressa con il valore q: rappresenta quanto accuratamente i rispondenti possono stimare il valore di un altro rispondente prima che i dati vengano pubblicati ($p < q < 100$).

- q indica l'errore nella stima prima della pubblicazione
- 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).

3.1.3 *(n,k)* rule

Questa regola stabilisce che, indipendentemente dal numero di rispondenti in una cella, se un numero ridotto ($\leq n$) di questi rispondenti contribuisce a una grande percentuale ($\geq k$) del valore totale della cella, allora la cella viene considerata sensibile (spesso si usa $n = 1$ o $n = 2$).

Regola intuitiva

Se una cella è dominata da un solo rispondente, il totale pubblicato rappresenta una stima superiore per il suo valore.

Esempio

$(2, 70) \rightarrow$ sensibile se ≤ 2 rispondenti fanno il $\geq 70\%$ del totale.

3.2 Protezione delle celle sensibili

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 sensibili

Soppressione secondaria

È necessario selezionare altre celle non sensibili da sopprimere, per garantire che i dati nelle celle sensibili non possano essere stimati con troppa accuratezza.
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.

3.3 Verifica dei risultati

3.3.1 Audit

L'audit è una fase di verifica in cui controllo che tutto sia protetto.
Se i totali vengono pubblicati, la somma delle celle soppresse 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
- viene stimato un *lower bound* e un *upper bound* di ciascuna cella soppressa utilizzando la programmazione lineare
- se i limiti sono troppo vicini al valore originale, la cella è considerata sensibile

Questa operazione è semplice per tabelle di piccole dimensioni, ma potrebbe essere computazionalmente infattibile per grandi tabelle.

3.3.2 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.

3.3.3 Informazioni dei valori dei parametri

Mentre le regole di soppressione possono essere pubblicate, i valori dei parametri dovrebbero rimanere riservati.

Capitolo 4

Protezione di Microdati

Oggi sempre più spesso vengono rilasciati microdati; questo tipo di dati sono soggetti ai *linking attack*.

Per proteggere la privacy dei rispondenti, si ricorre al rimozione o crittografia degli identificatori espliciti; tuttavia questo non offre la garanzia di anonimato: spesso le informazioni rilasciate contengono dei quasi identificatori, che collegati ad altri dati (pubblici o conoscenza pregressa) permettono di reidentificare i rispondenti o ridurre l'incertezza sulla loro identità.

Le tecniche di protezione di microdata seguono due strategie:

- **non perturbative:** ridurre il contenuto informativo, diminuiscono il livello di dettaglio senza introdurre rumore
- **perturbative:** modificare i dati in modo che il contenuto venga mantenuto il più possibile

Queste tecniche si basano sul principio che la reidentificazione può essere contrastata riducendo la quantità di informazioni rilasciate; possiamo classificare le tecniche di protezione come:

- **Masking techniques** (perturbative o non perturbative)
- **Synthetic data generation**

Le tecniche possono essere applicate su due diversi tipi di dati:

- **Continui:** dati numerici, su cui ha senso fare operazioni matematiche
- **Categorici:** dati che possono assumere un insieme limitato di valori, su cui non ha senso fare operazioni matematiche

4.1 Masking Techniques

In questa sezione esaminiamo diverse tecniche di *masking*, usando come esempio di riferimento la tabella seguente.

SSN	Name	Race	DoB	Sex	ZIP	MarStat	Holidays	Income
		Asian	64/09/27	F	94139	Divorced	13	260
		Asian	64/09/30	F	94139	Divorced	1	170
		Asian	64/04/18	M	94139	Married	40	200
		Asian	64/04/15	M	94139	Married	17	280
		Asian	64/03/09	M	94138	Married	10	190
		Black	63/03/13	M	94138	Married	2	190
		Black	63/03/18	M	94138	Married	13	185
		Black	64/03/18	M	94141	Married	60	290
		Black	64/09/13	F	94141	Married	15	200
		Black	64/09/07	F	94141	Married	60	290
		White	61/05/02	M	94138	Single	22	140
		White	61/05/14	M	94138	Single	17	170
		White	61/05/08	M	94138	Single	10	300
		White	61/09/15	F	94142	Widow	15	200

4.1.1 Sampling [NP]

La tabella protetta viene ottenuta facendo un campione della tabella di micro-data originale. Viene introdotta dell'incertezza riguardo alla partecipazione di un rispondente, dunque diminuisce il rischio di reidentificazione.

SSN	Name	Race	DoB	Sex	ZIP	MarStat	Holidays	Income
		Asian	64/09/27	F	94139	Divorced	13	260
		Black	64/09/13	F	94141	Married	15	200
		White	61/09/15	F	94142	Widow	15	200

4.1.2 Local Suppression [NP]

Sopprime il valore di un attributo che potrebbe contribuire in modo significativo al rischio di divulgazione, limitando le possibilità di analisi.

SSN	Name	Race	DoB	Sex	ZIP	MarStat	Holidays	Income
		Asian	64/09/27	F	94139	Divorced	13	260
		Asian	64/09/30	F	94139	Divorced	1	170
		Asian	64/04/18	M	94139	Married	40	200
		Asian	64/04/15	M	94139	Married	17	280
		Black	63/03/13	M	94138	Married	2	190
		Black	63/03/18	M	94138	Married	13	185
		Black	64/09/13	F	94141	Married	15	200
		Black	64/09/07	F	94141	Married	60	290
		White	61/05/14	M	94138	Single	17	170
		White	61/05/08	M	94138	Single	10	300
		White	61/09/15	F	**94142**	**Widow**	15	200

4.1.3 Global Recoding [NP]

Comporta la suddivisione del dominio di un attributo in diversi intervalli disgiunti, tipicamente della stessa ampiezza, dove per ogni intervallo viene associata una *label*.

La tabella protetta è ottenuta sostituendo i valori dell'attributo con la *label* associata all'intervallo corrispondente.

SSN	Name	Race	DoB	Sex	ZIP	MarStat	Holidays	Income
		Asian	64/09/27	F	94139	Divorced	13	med
		Asian	64/09/30	F	94139	Divorced	1	low
		Asian	64/04/18	M	94139	Married	40	med
		Asian	64/04/15	M	94139	Married	17	med
		Black	63/03/13	M	94138	Married	2	low
		Black	63/03/18	M	94138	Married	13	low
		Black	64/09/13	F	94141	Married	15	med
		Black	64/09/07	F	94141	Married	60	high
		White	61/05/14	M	94138	Single	17	low
		White	61/05/08	M	94138	Single	10	high
		White	61/09/15	F	94142	Widow	15	med

4.1.4 Top-coding e Bottom-coding [NP]

Il *top-coding* stabilisce un limite superiore per ciascun attributo da proteggere; qualsiasi valore maggiore di tale limite viene sostituito con una *flag* che indica all'utente che tale valore supera il limite.

La procedura è ananloga per il *bottom-coding* rispetto a un limite inferiore.

SSN	Name	Race	DoB	Sex	ZIP	MarStat	Holidays	Income
		Asian	64/09/27	F	94139	Divorced	13	260
		Asian	64/09/30	F	94139	Divorced	<10	170
		Asian	64/04/18	M	94139	Married	>30	200
		Asian	64/04/15	M	94139	Married	17	280
		Black	63/03/13	M	94138	Married	<10	190
		Black	63/03/18	M	94138	Married	13	185
		Black	64/09/13	F	94141	Married	15	200
		Black	64/09/07	F	94141	Married	>30	290
		White	61/05/14	M	94138	Single	17	170
		White	61/05/08	M	94138	Single	10	300
		White	61/09/15	F	94142	Widow	15	200

4.1.5 Generalizzazione [NP]

I valori vengono sostituiti con altri più generali. È basata sulla definizione di una gerarchia di generalizzazione.

SSN	Name	Race	DoB	Sex	ZIP	MarStat	Holidays	Income
		Asian	64/09	F	94139	Divorced	13	260
		Asian	64/09	F	94139	Divorced	1	170
		Asian	64/04	M	94139	Married	40	200
		Asian	64/04	M	94139	Married	17	280
		Black	63/03	M	94138	Married	2	190
		Black	63/03	M	94138	Married	13	185
		Black	64/09	F	94141	Married	15	200
		Black	64/09	F	94141	Married	60	290
		White	61/05	M	94138	Single	17	170
		White	61/05	M	94138	Single	10	300
		White	61/09	F	94142	Widow	15	200

4.1.6 Rumore casuale [P]

Il rumore casuale perturba un attributo sensibile aggiungendo o moltiplicando il suo valore con una variabile casuale di una distribuzione specificata.

È necessario decidere se pubblicare o meno la distribuzione usata per aggiugere rumore ai dati; la pubblicazione potrebbe aumentare il rischio di disclosure.

La somma del rumore introdotto deve essere pari a 0 per preservare la statistica.

Race	DoB	Sex	ZIP	MarStat	Holidays	Noise	Income
Asian	64/09/27	F	94139	Divorced	13	+2	260
Asian	64/09/30	F	94139	Divorced	1	+1	170
Asian	64/04/18	M	94139	Married	40	-10	200
Asian	64/04/15	M	94139	Married	17	+3	280
Black	63/03/13	M	94138	Married	2	+5	190
Black	63/03/18	M	94138	Married	13	+8	185
Black	64/09/13	F	94141	Married	15	+4	200
Black	64/09/07	F	94141	Married	60	-11	290
White	61/05/14	M	94138	Single	17	-2	170
White	61/05/08	M	94138	Single	10	-3	300
White	61/09/15	F	94142	Widow	15	+3	200

Race	DoB	Sex	ZIP	MarStat	Holidays	Income
Asian	64/09/27	F	94139	Divorced	15	260
Asian	64/09/30	F	94139	Divorced	2	170
Asian	64/04/18	M	94139	Married	30	200
Asian	64/04/15	M	94139	Married	20	280
Black	63/03/13	M	94138	Married	7	190
Black	63/03/18	M	94138	Married	21	185
Black	64/09/13	F	94141	Married	19	200
Black	64/09/07	F	94141	Married	49	290
White	61/05/14	M	94138	Single	15	170
White	61/05/08	M	94138	Single	7	300
White	61/09/15	F	94142	Widow	18	200

4.1.7 Swapping [P]

Spiegazione con esempio

Alcuni record potrebbero *matchare* con altri record dello stesso file su alcuni attributi prefissati, ma appartenere a diverse zone geografiche.

I valori di tutte le altre variabili vengono scambiati. Questa tecnica riduce il rischio di reidentificazione dato che introduce incertezza sulla veridicità del dato di un rispondente.

SSN	Name	Race	DoB	Sex	ZIP	MarStat	Holidays	Income
		Asian	64/09/27	F	94139	Divorced	13	260
		Asian	64/09/30	F	94139	Divorced	1	170
		Asian	64/04/18	M	94139	Married	40	200
		Asian	64/04/15	M	94139	Married	17	280
		Black	63/03/13	M	94138	Married	2	190
		Black	63/03/18	M	94138	Married	13	185
		Black	64/09/13	F	94141	Married	15	200
		Black	64/09/07	F	94141	Married	60	290
		White	61/05/14	M	94138	Single	17	170
		White	61/05/08	M	94138	Single	10	300
		White	61/09/15	F	94142	Widow	15	200

Identify 3 pairs of tuples with same **Sex** and **MarStat**

SSN	Name	Race	DoB	Sex	ZIP	MarStat	Holidays	Income
		Asian	64/09/27	F	94139	Divorced	13	260
		Asian	64/09/30	F	94139	Divorced	1	170
		Asian	64/04/18	M	94139	Married	2	190
		Asian	64/04/15	M	94139	Married	17	280
		Black	63/03/13	M	94138	Married	40	200
		Black	63/03/18	M	94138	Married	13	185
		Black	64/09/13	F	94141	Married	60	290
		Black	64/09/07	F	94141	Married	15	200
		White	61/05/14	M	94138	Single	10	300
		White	61/05/08	M	94138	Single	17	170
		White	61/09/15	F	94142	Widow	15	200

Swap **Holidays** and **Income**

4.1.8 Micro-aggregation [P]

La micro-aggregazione consiste nel raggruppare più tuple in dei gruppi di dimensione k. Viene poi pubblicata la media di tale gruppo al posto dei singoli valori.

I gruppi sono formati usando critieri di massima similarità.

Race	DoB	Sex	ZIP	MarStat	Holidays	Income
Asian	64/09/27	F	94139	Divorced	13	215
Asian	64/09/30	F	94139	Divorced	1	215
Asian	64/04/18	M	94139	Married	40	213
Asian	64/04/15	M	94139	Married	17	213
Black	63/03/13	M	94138	Married	2	213
Black	63/03/18	M	94138	Married	13	213
Black	64/09/13	F	94141	Married	15	245
Black	64/09/07	F	94141	Married	60	245
White	61/05/14	M	94138	Single	17	235
White	61/05/08	M	94138	Single	10	235
White	61/09/15	F	94142	Widow	15	200

4.2 Tecniche Sintetiche

Consistono nel generare dati sintetici per sostituire quelli dei rispondenti, in modo che venga mantenuta la correlazione statistica.

Il vantaggio nell'utilizzo di queste tecniche è che i dati sintetici rilasciati non sono riferiti a nessun rispondente e il loro rilascio non può portare a reidentificazione.