Problemi di inferenza

Controllare l'accesso ai singoli dati non è sufficiente per garantire protezione.

Il controllo dell'accesso deve essere complementato con controlli di flusso, di inferenza, e di aggregazione.

Controlli di flusso

Controllano i flussi di informazioni causati da esecuzione di programmi per assicurarsi che informazione non sia scritta in oggetti meno protetti di quelli da cui è stata letta.

Es., "**if** x=0 **then** z:=false **else** z:= true" fa fluire informazione da x a z.

- Classi statiche: classi fisse di accesso associate agli oggetti e istruzioni bloccate se provocano flusso improprio
- Classi dinamiche: la classe di accesso associata a un oggetto è il least upper bound della classe degli oggetti che fanno fluire informazione nell'oggetto

Controlli di flusso possono essere effettuati a

- Tempo di compilazione
- Run time

Controlli di inferenza

La correlazione esistente fra le informazioni può permettere agli utenti di inferire dati sensibili da dati non sensibili a loro rilasciati.

Esiste un canale di inferenza da un insieme di dati X a un insieme di dati Y ($X \rightarrow Y$) se conoscendo X posso sapere il valore di Y (inferenza certa) o ridurre la mia incertezza su questo.

Es., Livello di occupazione e specializzazione possono fare inferire lo stipendio di un impiegato

Sapendo se uno studente ha una borsa di studio posso inferire qualcosa sulla sua media o livello economico

Per proteggere da canali di inferenza ogni insieme di dati X deve essere protetto almeno quanto tutti i dati Y che possono essere inferiti da X.

Inferenza spesso modellata in basi di dati multilivello $\Rightarrow$ la classificazione di X deve dominare quella di Y

Determinazione delle classificazione di informazione

Deve tenere in considerazione

- inferenza: Se Y può essere dedotto da $x_1, \ldots, x_n$, allora $\mathsf{lub}\{\lambda(\mathsf{x}_1), \ldots, \lambda(\mathsf{x}_n)\} \geq \lambda(\mathsf{Y})$
- associazione: la classe di più attributi deve avere almeno una certa classificazione (più alta di quella degli attributi singolarmente presi). Esempio non voglio che l'associazione fra nome e salario sia nota a soggetti sotto il livello C. $\mathsf{lub}\{\mathsf{lambda}(\mathsf{nome}), \lambda(\mathsf{salario})\} \geq \mathsf{C}$
- aggregazione: la aggregazione di certe informazioni ha una classe più alta di quella dei singoli attributi singolarmente presi. Es., il numero di telefono di un impiegato (qualsiasi) è pubblico ma la lista dei telefoni è protetta.

Inferenza, associazione, e aggregazione

Non facili da gestire.

L'aggregazione deve essere applicata necessariamente in modo dinamico (difficile però tenere traccia delle richieste).

Inferenza e associazione possono applicati staticamente (con classificazione dei dati).

Nella risoluzione di problemi di inferenza e associazione voglio

- garantire protezione
- massimizzando la visibilità dell'informazione.

La massimizzazione della visiilità è un problema complesso (può essere NP completo).

Rilascio di dati statistici

Spesso vengono rilasciati dati statistici o dati per uso statistico.

Il rilascio di questi dati inevitabilmente rivela informazione circa i dati dei singoli soggetti.

Il rilascio provoca divulgazione (disclosure) quando dati che dovrebbero essere trattati come confidenziali sono resi noti.

Divulgazione può:

- essere basata solamente sui dati rilasciati
- risultare dalle combinazione di dati rilasciati con informazione esternamente disponibile (record linkage) disponibili pubblicamente o a un insieme ristretto di persone.

Quando un ente rilascia dati (agenzie statistiche in particolare) dovrebbe assicurarsi che il rischio di divulgazione dai dati rilasciati sia minimo.

Metodi di rilascio

La maggior parte dei dati statistici sono rilasciati in forma di macrodati (tabelle) o microdati.

- Macrodati Ogni entrata in una tabella statistica rappresenta il valore aggregato di una quantità su tutte le unità di analisi che appartengono a una unica cella statistica.
- Microdati Un file di microdati è composto da record individuali ognuno contenente valori di variabili per una singola persona, entità commerciale, o altra unità.

Tabelle di macrodati

La tabelle di macrodati possono essere distite in due categorie

- Tabelle di frequenza (count) Mostrano il numero, o la percentuale, di elementi della popolazione che hanno certe caratteristiche.
- Tabelle di grandezza (magnitude data) Riportano l'aggregazione di una "quantità di interesse" su tutte le unità di analisi nella cella. I dati possono essere presentati come una media dividendo l'aggregato per il numero di unità nella cella.

Per distinguere fra tabelle di frequenza e di grandezza, la "quantità di interesse" deve misurare qualcosa di più che non la semplice appartenenza ad una cella. Ad esempio, tabelle che mostrano il numero di aziende in un certo settore manifatturiero raggruppate per codice industriale e contea sono tabelle di frequenza. Tabelle che presentano il fatturato totale di tali aziende in ogni cella sono tabelle di grandezza.

Tabelle di frequenza – Esempio

Numero di beneficiari di un certo benefit per importo e contea

	Importo						
Contea	**$0-19**	**$20-39**	**$40-59**	**$60-79**	**$80-99**	**$100+**	**Totale**
A	2	4	18	20	7	1	52
B	-	-	7	9	-	-	16
C	-	6	30	15	4	-	55
D	-	-	2	-	-	-	2

File di microdati – Esempio

Tutti i record della contea Alfa relativi a bambini delinquenti

N	Bambino	Contea	Educ. CF	Stipendio CF	Razza CF
1	John	Alfa	molto alto	201	nera
2	Jim	Alfa	alto	103	bianca
3	Sue	Alfa	alto	77	nera
4	Pete	Alfa	alto	61	bianca
5	Ramesh	Alfa	medio	72	bianca
6	Dante	Alfa	basso	103	bianca
7	Virgil	Alfa	basso	91	nera
8	Wanda	Alfa	basso	84	bianca
9	Stan	Alfa	basso	75	bianca
10	Irmi	Alfa	basso	62	nera
11	Renee	Alfa	basso	58	bianca
12	Virginia	Alfa	basso	56	nera
13	Mary	Alfa	basso	54	nera
14	Kim	Alfa	basso	52	bianca
15	Tom	Alfa	basso	55	nera
16	Ken	Alfa	basso	48	bianca
17	Mike	Alfa	basso	48	bianca
18	Joe	Alfa	basso	41	nera
19	Jeff	Alfa	basso	44	nera
20	Nancy	Alfa	basso	37	bianca

Divulgazione di informazione

Esistono definizioni differenti di divulgazione e tipi differenti di divulgazione sono stati proposti. Una possibile definizione di divulgazione è la seguente.

“Divulgazione è l’attribuzione impropria di informazione a un soggetto, sia esso individuo o organizzazione.

Divulgazione esiste quando:

- un soggetto è identificato all’interno di un file rilasciato (divulgazione di identità)
- informazione sensibile circa un soggetto è rivelata attraverso il file rilasciato (divulgazione di attributo)
- i dati rilasciati rendono possibile la determinazione di alcune caratteristiche di un individuo più accuratamente che altrimenti possibile (divulgazione inferenziale)”

Divulgazione identità

Esiste divulgazione di identità quando una terza parte può identificare un soggetto dai dati rilasciati. Rivelare che un individuo è un rispondente dei dati rilasciati può o meno violare i requisiti di confidenzialità.

Per dati tabellari, rilvelare identità non è generalmente considerato divulgazione, a meno che l’identificazione porti a divulgare informazioni confidenziali (divulgazione di attributo).

Per i microdati l’identificazione è generalmente considerata come divulgazione, poichè i record sono così dettagliati che la probabilità di identificazione senza rivelazione di informazione addizionale è minima. Per questa ragione i metodi per la limitazione della divulgazione applicati ai microdati limitano o modificano l’informazione che potrebbe essere utilizzata per identificare i rispondenti dell’informazione.

Divulgazione attributo

Esiste divulgazione di attributo quando informazione confidenziale circa un soggetto è rivelata e può essere attribuita a un soggetto.

Occorre sia nel caso in cui l'informazione sia rivelata esattamente sia che possa essere stimata.

Comprende sia l'identificazione del soggetto sia la divulgazione della informazione confidenziale che gli pertiene.

Divulgazione inferenziale

Esiste divulgazione inferenziale quando l'informazione può essere inferita con alta confidenza da proprietà statistiche dei dati.

Ad esempio i dati statistici pubblicati possono mostrare una elevata correlazione fra prezzo della casa e stipendio del proprietario. Il prezzo della casa, pubblicamente disponibile, può permettere di inferire informazione sul salario del proprietario.

In generale agenzie statistiche non trattano il problema dell'integrità inferenziale per due ragioni principali:

- equiparare divulgazione e inferenza, implicherebbe che nessun dato potrebbe più essere rilasciato
- inferenze predicono comportamento di aggregati, non di singoli attributi, quindi non sono buoni predittori di singoli dati.

Dati ristretti e accesso ristretto

La scelta dei metodi per limitare divulgazione da dati statistici dipende dalla natura dei dati la cui confidenzialità deve essere garantita.

Alcuni file di microdata includono identificatori espliciti quali nome, indirizzo, o codice fiscale. La rimozione di tutti questi identificatori è il primo passo per la confidenzialità della informazione che deve essere protetta.

Le tecniche di limitazione di divulgazione utilizzate per tabelle di grandezza possono essere utilizzate per proteggere tabelle di frequenze.

Per tabelle di frequenza sono disponibili anche altre tecniche.

Protezione di informazione

In generale la confidenzialità può essere protetta

- Restringendo la quantità di informazione nelle tabelle o file di microdati rilasciati (dati ristretti)
- Imponendo condizioni sull'accesso ai dati prodotti (accesso ristretto)
- Combinazione dei due metodi precedenti.

Dati ristretti

Dati di uso pubblico sono generalmente rilasciati da agenzie statistiche al pubblico senza restrizioni o condizioni, ad eccezione del pagamento di una modica somma. È quindi necessario che il rischio di divulgazione sia basso.

L'applicazione di metodi per il controllo della divulgazione molto spesso richiede di restringere molto il contenuto dei file rilasciati, a volte fino al punto in cui i dati stessi perdono valore.

L'utilizzo di metodi di restrizione all'accesso permette di limitare restrizioni sul contenuto, permettendo a particolari utenti di accedere a dati più dettagliati in dipendenza da restrizioni su chi può accedervi, da quale locazione, a quale *scopo*, etc.... L'utilizzo di restrizioni all'accesso è molto spesso accompagnato da contratti scritti che vincolano penalmente gli utenti che acquisiscono informazioni rispetto a divulgazioni o uso improprio.

Pubblicazione di dati statistici

Agenzie statistiche raccolgono dati relativi a singoli individui e producono dati statistici per rilascio alla comunità.

Alcuni dati gestiti da agenzie statistiche sono considerati di proprietà del respondente al quale si riferiscono.

Quando i dati dai quali le statistiche sono calcolate sono confidenziali è necessario limitare il rischio di possibili divulgazioni dei dati che possono essere causate dalle statistiche.

Tutti i metodi che limitano la divulgazione di dati causano perdita di informazione. Tale perdita può arrivare al punto in cui i dati possono non essere più adatti a certi usi statistici.

Lo scopo è rilasciare la maggior quantità di informazione possibile senza rivelare singoli dati.

Tabelle di frequenze

I dati raccolti da molti survey sulla popolazione sono pubblicati in tabelle che mostrano somme (numero di persone per categoria) o frequenze (frazioni o percentuali di persone per categoria).

Techiche di protezione includono:

- Campionamento
- Regole speciali
- Regole di soglia

Campionamento

Condurre un survey campione anzichè un censimento.

Le stime sono date moltiplicando le risposte dei signoli individui per un peso di campionamento di aggregare i dati.

Se i pesi di campionamento non sono resi noti, il loro utilizzo aiuta a rendere il risultato meno vulnerabile a divulgazioni.

In aggiunta, alcune agenzie chiedono che le stime raggiungano una certa accuratezza prima di essere pubblicate. Dati che non raggiungono accuratezza non sono pubblicati perchè poco significativi (poco corretti).

Campionamento – 2

Quando le tabelle di frequenza sono basate su dati di tutta la popolazione (es., censimento), procedure di limitazione della divulgazione devono essere applicate. Possiamo individuare due classi di tecniche:

- Regole speciali specifiche di ogni tabella. Differiscono per differenti agenzie.
- Regole generali. Una cella è considerata sensibile se il numero di rispondenti è minore di una specifica soglia.

Regole speciali

Impongono restrizioni sul livello di dettaglio che può essere riportato in una tabella.

Esempio
Social Security Amministration (SSA) proibisce tabelle nelle quali i dettagli di una cella sono

- uguali al totale parziale o
- permetterebbero all'utente di determinare
  - l'età di un individuo all'interno di un intervallo di cinque anni
  - stipendi all'interno di un intervallo di $1000
  - benefits all'interno di un intervallo di $50

Regole speciali – Esempio

Esempio: Numero di beneficiari di un certo benefit per importo e contea

	Importo						
Contea	**$0-19**	**$20-39**	**$40-59**	**$60-79**	**$80-99**	**$100+**	**Totale**
A	2	4	18	20	7	1	52
B	-	-	7	9	-	-	16
C	-	6	30	15	4	-	55
D	-	-	2	-	-	-	2

Non può essere pubblicata. I dati per B e D violano le regole.

- D c'è una sola cella non vuota. Due regole violate:
  - La cella dettagliata è uguale al totale
  - Si inferisce che tutti i beneficiari della contea ricevono tra $40 e i $59 di benefit. Meno dell'intervallo consentito.
- B ci sono due celle non vuote ma l'intervallo dei possibili benefit è tra i $40 e i $79. Intervallo più piccolo dei $50 consentiti.

Regole speciali – Esempio- 2

Per evitare divulgazione non autorizzata i dati delle contee possono essere combinati.

La combinazione di categoria ("rolling-up categories") può essere effettuata per righe o per colonne.

Supponiamo di combinare per righe. Combinando B con A e D con C otteniamo.

	Importo						
Contea	**$0-19**	**$20-39**	**$40-59**	**$60-79**	**$80-99**	**$100+**	**Totale**
A& B	2	4	25	29	7	1	68
C & D	-	6	32	15	4	-	57

Nota: combinare B con D non avrebbe risolto il problema.

Regole di soglia

Una cella di una tabella è definita sensibile se il numero dei rispondenti è minore di un certo numero specificato (alcune agenzie richiedono 5, altre 3).

Celle sensibili non possono essere rilasciate.

Per soddisfare la regola una agenzia può utilizzare le seguenti tecniche:

- Ristrutturazione di tabelle e combinazione categorie
- Soppressione di celle
- Arrotondamento casuale
- Arrotondamento controllato
- Inserimento confidenziale

Tabella con divulgazione

Esempio

Numero di bambini delinquenti per contea e livello di educazione del capo famiglia

	Livello di educazione				
Contea	**Basso**	**Medio**	**Alto**	**Molto Alto**	**Totale**
Alfa	15	1	3	1	20
Beta	20	10	10	15	55
Gamma	3	10	10	2	25
Delta	12	14	7	2	35
Totale	50	35	30	20	135

Celle con meno di 5 rispondenti sono considerate sensibili.

Soppressione di celle

Una delle tecniche maggiormente diffuse è quella della soppressione di celle.

Sopprimere le celle sensibili (soppressione primaria) non è sufficiente a garantire protezione.

Almeno una altra cella deve essere soppressa (soppressione complementare) per ogni riga e colonna nella quale è stata soppressa una cella sensibile. Altrimenti il valore della cella soppressa potrebbe essere calcolato come differenza fra il totale e le altre entrate.

Anche in presenza di soppressione complementare è difficile garantire che il risultato fornisca protezione adeguata.

Tabella con divulgazione - non protetta da soppressione

Esempio: Numero di bambini delinquenti per contea e livello di educazione del capo famiglia

	Livello di educazione				
Contea	**Basso**	**Medio**	**Alto**	**Molto Alto**	**Totale**
Alfa	15	D_1	D_2	D_3	20
Beta	20	D_4	D_5	15	55
Gamma	D_6	10	10	D_7	25
Delta	D_8	14	7	D_9	35
Totale	50	35	30	20	135

Sembra garantire protezione, ma.....

(15 + D_1 + D_2 + D_3) + (20 + D_4 + D_5 + 15) - (D_1 + D_4 + 10 + 14) - (D_2 + D_5 + 10 + 7) = 20 + 55 - 35 - 30.

$\Longrightarrow$ $D_3 = 1$.

Soppressione di celle

La scelta delle celle da eliminare con soppressione complementare è quindi più complicata di quanto appare.

Tecniche di programmazione lineare sono utilizzate per la scelta automatica di celle da eliminare con soppressione complementare.

Tecniche di audit possono essere utilizzate per valutare la bontà della soppressione per vedere se fornisce protezione adeguata.

Tabella con divulgazione - protetta da soppressione

Esempio: Numero di bambini delinquenti per contea e livello di educazione del capo famiglia

	Livello di educazione				
Contea	**Basso**	**Medio**	**Alto**	**Molto Alto**	**Totale**
Alfa	15	D	D	D	20
Beta	20	10	10	15	55
Gamma	D	D	10	D	25
Delta	D	14	D	D	35
Totale	50	35	30	20	135

La tabella sopra garantisce protezione, ma .....

delle 16 celle componenti la tabella solo 7 sono pubblicate, mentre 9 sono soppresse.

Arrotondamento casuale

Per cercare di limitare la quantità di dati persi con la soppressione, possono essere utilizzati metodi alternativi, quali l' arrotondamento di valori, che può essere

- casuale Nell'arrotondamento casuale i valori sono arrotondati ma, invece di utilizzare convenzioni standard, viene fatta una scelta casuale per decidere se arrotondare per difetto o per eccesso.

  Poichè la scelta sull'arrotondamento è fatta cella per cella la somma per righe e per colonne può non corrispondere più al totale dato.

  Come conseguenza gli utenti potrebbero perdere "fiducia" nei dati rilasciati.

- controllato vengono cambiati anche i totali.

Arrotondamento casuale - esempio

Livello di educazione (originale)

Contea	Basso	Medio	Alto	Molto Alto	Totale
Alfa	15	1	3	1	20
Beta	20	10	10	15	55
Gamma	3	10	10	2	25
Delta	12	14	7	2	35
Totale	50	35	30	20	135

Livello di educazione (arroitondamento casuale)

Contea	Basso	Medio	Alto	Molto Alto	Totale
Alfa	15	*0	*0	*0	20
Beta	20	10	10	15	55
Gamma	*5	10	10	*0	25
Delta	*15	*15	*10	*0	35
Totale	50	35	30	20	135

Arrotondamento controllato

Per risolvere il problema delle addizioni è stata sviluppata una procedura di arrotondamento controllato.

È una forma di arrotondamento ma richiede che ogni riga e colonna ammonti esattamente ai totali pubblicati.

Techniche di programmazione lineare possono essere utilizzate per identificare gli arrotondamenti da apportare.

Svantaggi:

- richiede l'uso di programmi specializzati (non largamente disponibili).
- per tabelle complesse potrebbero non esistere soluzioni per l'arrotondamento controllato.

Arrotondamento controllato - esempio

Livello di educazione (originale)

Contea	Basso	Medio	Alto	Molto Alto	Totale
Alfa	15	1	3	1	20
Beta	20	10	10	15	55
Gamma	3	10	10	2	25
Delta	12	14	7	2	35
Totale	50	35	30	20	135

Livello di educazione (arrotondamento controllato)

Contea	Basso	Medio	Alto	Molto Alto	Totale
Alfa	15	*0	*5	*0	20
Beta	20	10	10	15	55
Gamma	*5	10	10	*0	25
Delta	*10	*15	*5	*5	35
Totale	50	35	30	20	135

Inserimento confidenziale

L'inserimento confidenziale è una procedura relativamente recente sviluppata per proteggere confidenzialità di dati di censimento pubblicati da U.S. Census Bureau nel 1990.

In quella occasione sono stati utilizzati due diversi approcci:

- Il primo utilizzato per proteggere dati regolari decennali (100 % della popolazione).
- Il secondo utilizzato per proteggere dati relativi soltanto a una parte della popolazione.

Entrambi gli approcci sono applicati direttamente sui file di microdati sui quali le statistiche sono calcolate $\Longrightarrow$ Le statistiche vengono protette modificando i dati di input.

Inserimento confidenziale

Per il 100 percento file di microdati, l'inserimento confidenziale include switching.

Modifica direttamente il file di microdati utilizzato per calcolare le statistiche che saranno rilasciate come segue.

1. Prendi un campione di record dai file di microdati
2. Trova un match per questi record in un'altra regione geografica che abbia lo stesso valore per un insieme di attributi definiti come importanti
3. Scambia (switch) tutti gli sui record trovati come match.

Per blocchi piccoli la frazione del campione viene aumentata per fornire una maggiore protezione.

Il file di microdati così modificato è quindi utilizzato per produrre statistiche.

Inserimento confidenziale – esempio

Esempio Tutti i record della contea Alfa relativi a bambini delinquenti

N	Bambino	Contea	Educ. CF	Stipendio CF	Razza CF
1	John	Alfa	molto alto	201	nera
2	Jim	Alfa	alto	103	bianca
3	Sue	Alfa	alto	77	nera
4	Pete	Alfa	alto	61	bianca
5	Ramesh	Alfa	medio	72	bianca
6	Dante	Alfa	basso	103	bianca
7	Virgil	Alfa	basso	91	nera
8	Wanda	Alfa	basso	84	bianca
9	Stan	Alfa	basso	75	bianca
10	Irmi	Alfa	basso	62	nera
11	Renee	Alfa	basso	58	bianca
12	Virginia	Alfa	basso	56	nera
13	Mary	Alfa	basso	54	nera
14	Kim	Alfa	basso	52	bianca
15	Tom	Alfa	basso	55	nera
16	Ken	Alfa	basso	48	bianca
17	Mike	Alfa	basso	48	bianca
18	Joe	Alfa	basso	41	nera
19	Jeff	Alfa	basso	44	nera
20	Nancy	Alfa	basso	37	bianca

Inserimento confidenziale – esempio 2

Modifica dei microdati

1. Prendi un campione dei record dai microdati (ad esempio, 10%). Assumiamo che 4 e 17 fanno parte del campione.
2. Trova un match in qualche altra contea sulle altre variabili (razza, salario). (Come risultato di questo matching, i totali della contea per queste variabili rimangono invariati dopo lo swapping).
   - Un match per il record 4 (Pete) è trovato nella contea Beta. Il match è con Alonso che appartiene a una famiglia il cui capofamiglia ha un livello di educazione molto alto.
   - Un match per il record 17 (Mike) è trovato nella contea Delta. Il match è con George che appartiene a una famiglia il cui capofamiglia ha un livello di educazione medio.

Inserimento confidenziale – esempio 2

2. In aggiunta supponiamo che parte dei record del campione dalle altre contee ha un match con la contea Alfa.
   - Un record della contea Delta (June con livello di educazione alto) ha un match con Virginia (12).
   - Un record dalla contea Gamma (Heather con livello di educazione bassa) ha un match con Nancy (record 10).
3. Dopo che tutti i match sono stati calcolati scambia i dati nei record che hanno match.
4. Utilizza il file prodotto per calcolare le statistiche.

Inserimento confidenziale nel file di microdati - esempio

Esempio Record per la contea Alfa dopo lo swapping

N	Bambino	Contea	Educ. CF	Stipendio CF	Razza CF
1	John	Alfa	molto alto	201	nera
2	Jim	Alfa	alto	103	bianca
3	Sue	Alfa	alto	77	nera
4*	Alonso	Alfa	molto alto	61	bianca
5	Ramesh	Alfa	medio	72	bianca
6	Dante	Alfa	basso	103	bianca
7	Virgil	Alfa	basso	91	nera
8	Wanda	Alfa	basso	84	bianca
9	Stan	Alfa	basso	75	bianca
10	Irmi	Alfa	basso	62	nera
11	Renee	Alfa	basso	58	bianca
12*	June	Alfa	alto	56	nera
13	Mary	Alfa	basso	54	nera
14	Kim	Alfa	basso	52	bianca
15	Tom	Alfa	basso	55	nera
16	Ken	Alfa	basso	48	bianca
17*	George	Alfa	medio	48	bianca
18	Joe	Alfa	basso	41	nera
19	Jeff	Alfa	basso	44	nera
20*	Heather	Alfa	basso	37	bianca

*: nome e livello di educazione presi da entrate in altra contea.

Inserimento confidenziale - esempio

Livello di educazione (originale)

Contea	Basso	Medio	Alto	Molto Alto	Totale
Alfa	15	1	3	1	20
Beta	20	10	10	15	55
Gamma	3	10	10	2	25
Delta	12	14	7	2	35
Totale	50	35	30	20	135

Livello di educazione (con inserimento confidenziale)

Contea	Basso	Medio	Alto	Molto Alto	Totale
Alfa	13	2	3	2	20
Beta	20	10	10	15	55
Gamma	5	9	11	0	25
Delta	14	12	8	1	35
Totale	50	35	30	20	135

Inserimento confidenziale

Per quanto riguarda il file campione c'è già una protezione data dalla campionatura. Studi hanno mostrato che la protezione era sufficiente ad eccezione di piccole aree geografiche.

Per fornire protezione addizionale a piccole aree geografiche, fu scelta casualmente una famiglia per ogni area e i suoi dati cancellati, sostituendoli con valori esplicitamente inseriti (blank and impute).

Il file di microdati trattato in quel modo può quindi essere utilizzato per calcolare statistiche senza applicazione di ulteriori tecniche.

Protezione di tabelle di grandezza

Valori di grandezza sono generalmente quantità non negative riportate in survey o censimenti di aziende.

In tali tabelle è molto probabile che la distribuzione sia distorta (skewed), con poche delle entità che hanno valori molto grandi.

Le tecniche di limitazione della divulgazione cercano di assicurare che i dati pubblicati non possano essere utilizzati per stimare i valori alti. Notare che proteggendo i valori più elevati proteggiamo in effetti tutti i valori.

Per tabelle di grandezza è poco probabile che il campionamento da solo fornisca protezione poichè molti survey includono molte entità che sono incluse con certezza.

Di conseguenza le unità che sono più visibili a causa della loro dimensioni non ricevono alcuna protezione dal campionamento.

Regole di soppressione

Regole chiamate regole di soppressione primaria sono applicate per determinare se una data cella potrebbe rivelare informazione relativa a un singolo rispondente. Tali celle sono considerate sensibili e non possono essere rilasciate.

Le regole di soppressione primaria più comunemente utilizzate per identificare celle sensibili sono:

- regola (n,k)
- la regola del p-percento
- la regola pq

Tutte queste regole sono basate sull'intenzione di rendere difficile per un rispondente la stima dei valori riportati dagli altri rispondenti. Il valore più grande è più facilmente stimabile.

Regole di soppressione primaria - regola del p-percento

Nella regola del p-percento o del “livello stimato di equivocazione del p-percento”, esiste divulgazione di informazione se un utente può stimare il valore di un attributo per un rispondente all’interno di un intervallo ristretto.

Esiste divulgazione se limite superiore e limite inferiore stimati per il valore di un rispondente sono più vicini al valore riportato di una percentuale p specificata.

Regole di soppressione primaria - regola pq

Nella regola del p-percento non c’è alcuna assunzione sulla conoscenza sui rispondenti che gli utenti hanno a priori.

Molto spesso, agenzie che pubblicano i dati non possono basarsi su questa ipotesi.

Nella regola pq le agenzie specificano quanta conoscenza a priori c’è specificando una valore q ($p < q < 100$) che rappresenta quanto accuratamente i rispondenti possono stimare i valori di un altro rispondente prima che i dati siano pubblicati.

Regole di soppressione primaria - regola (n,k)

Nella regola (n,k), a prescindere dal numero di rispondenti in una cella, se un piccolo numero di rispondenti (n o meno) contribuiscono a una grande percentuale ($k\%$ o più) del valore totale della cella, la cella è considerata sensibile.

Da molti questa è considerata una regola intuitiva perchè, per esempio, se una cella è dominata da un rispondente il valore della cella (totale) da solo è una stima naturale di upper bound per il valore del più grande rispondente.

Sebbene coalizioni non siano specificatamente discusse nella regola (n,k), agenzie scelgono il valore di n maggiore del numero di coalizioni sospettate. Molte agenzie utilizzano n=1 o 2.

Soppressione secondaria

Dopo aver identificato le celle sensitive, ci sono due opzioni:

- Ristrutturare la tabella e raggruppare celle fino a che non rimangono celle sensibili
- Sopprimere celle. La soppressione delle celle sensibili è chiamata soppressione primaria. Altre celle possono essere soppresse suppressione secondaria per evitare che le celle sensibili possano essere derivate sottraendo al totale le altre celle pubblicate.

  Tecniche e problemi sono simili a quelli visti per tabelle di frequenza.

Un approccio a volte seguito per evitare soppressione di celle utilizzato da certe agenzie è quello di ottenere un permesso scritto per pubblicare una cella sensibile da parte dei rispondenti che contribuiscono alla cella.

Soppressione secondaria

Le scelta delle cella da cancellare con soppressione secondaria deve assicurare che i valori delle celle sensibili non possano essere stimati troppo accuratamente.

Tale requisito è generalmente interpretato ad indicare che i dati di un rispondente non possono essere stimati all'interno di un intervallo dato dal valore $+$ e $-$ una certa percentuale.

Ci sono due modi in cui i valori possono essere compromessi.

- Unioni di celle soppresse possono essere sensibili (rispetto alla regola di sensitività adottata).
- Equazioni di riga e colonna possono essere risolte e il valore di una cella soppressa stimato con una certa accuratezza.

Soppressione secondaria

Ogni insieme di celle proposte per la soppressione secondaria è accettabile fintanto che le celle sensibili sono protette.

Questo significa che la scelta potrebbe anche essere fatta a mano.

Tipicamente un analista di dati conosce quali celle sono di maggior interesse all'utente (e non dovrebbero essere utilizzate per soppressione secondaria, se possibile) e quali celle non sono di interesse (e possono quindi essere utilizzate per soppressione secondaria).

Per assicurare che la scelta fatta manualmente fornisce protezione viene effettuata una procedura automatica di audit.

Audit

Se celle sensibili sono protette cancellando anche altre celle (secondarie) nella tabella ma pubblicando i totali, implicitamente sono anche pubblicati i totali delle celle soppresse.

Un modo di applicare audit consiste quindi nell'applicare tecniche di programmazione lineare a tali informazioni per vedere cosa può essere inferito sui valori soppressi.

Intuitivamente righe e colonne della tabella possono essere viste come un grande sistema di equazioni lineari. Le celle soppresse rappresentano valori non noti. Utilizzando tecniche di programmazione lineare si può stimare il minimo e il massimo valore che ogni cella soppressa può avere. La cella è considerata protetta se tali valori non sono vicini al valore vero più di una certa percentuale.

Per tabelle piccole il problema dell'audit è semplice. Per tabelle di grandi dimensioni il problema può diventare computazionalmente intrattabile.

Soppressione secondaria

Per grandi sistemi la scelta di celle secondarie è abbastanza complessa. Può essere fatta con tecniche

- manuali. Rischiano di lasciare non protette celle sensibili o introdurre inconsistenze fra tabelle (che possono esserte utilizzate per inferire informazione).
- automatiche con utilizzo di tecniche di programmazione lineare.

L'obiettivo di minimizzare la perdita di informazione rende il problema più complesso e, in alcuni casi intrattabile.

La maggior parte delle tecniche oggi disponibili applica un approccio di tipo greedy (ad ogni passo l'ottima scelta di soppressione è scelta). Tale approccio non garantisce però soluzione ottima e può portare a più perdita di informazione di quanto sia necessario.

Perdita di informazione

La scelta delle celle da cancellare deve minimizzare la perdita di informazione che può verificarsi.

Non esiste una unica definizione perdita di informazione e quindi di soluzione ottima.

Ad esempio di può voler minimizzare

- la somma dei valori soppressi (però tante celle con piccoli valori possono essere soppressa).
- il numero di celle soppresse.

Informazione sui parametri

Le regole di soppressione possono essere rese pubbliche.

I parametri utilizzati dovrebbero essere invece considerati confidenziali. Essi infatti potrebber causare divulgazione.

Esempio Supponiamo di applicare la regola del p-percento con $p = 20$ e che lo stesso valore è utilizzato per la soppressione complementare. Supponiamo che una cella abbia valore 100 e sia stata soppressa. Supponiamo che risolvendo equazioni lineari sui dati pubblicati un utente restringe determina per una cella x un limite inferiore di 80 e superiore di 120, $80 \leq x \leq 120$.
$\Longrightarrow x = 100$.
Avendo determinato precisamente x l'utente può anche effettuare altre inferenze.

Protezione di tabelle di grandezza

Numero di studenti per genere e anno

Genere	1978	1979	1980	1991	Totale
Femmine	1	2	2	1	6
Maschi	3	2	0	2	7
Totale	4	4	2	3	13

Totale esame SAT per genere e anno

Genere	1978	1979	1980	1991	Totale
Femmine	800	1330	1120	500	3750
Maschi	1930	1150	0	1180	4260
Totale	2730	2480	1120	1680	8010

Regola (n,k)

Totale esame SAT per genere e anno

Genere	1978	1979	1980	1991	Totale
Femmine	800	1330	1120	500	3750
Maschi	1930	1150	0	1180	4260
Totale	2730	2480	1120	1680	8010

Regola (n,k) con n=1, k=90 $\Rightarrow$ non pubblicare celle in cui un rispondente contribuisce a più del 90% del totale.

Togliamo le celle in cui c'era uno studente solo.

Totale esame SAT per genere e anno

Genere	1978	1979	1980	1991	Totale
Femmine	D	1330	1120	D	3750
Maschi	1930	1150	0	1180	4260
Totale	2730	2480	1120	1680	8010

Regola (n,k)

Totale esame SAT per genere e anno

Genere	1978	1979	1980	1991	Totale
Femmine	D	1330	1120	D	3750
Maschi	1930	1150	0	1180	4260
Totale	2730	2480	1120	1680	8010

Soppressione secondaria

Totale esame SAT per genere e anno

Genere	1978	1979	1980	1991	Totale
Femmine	D	1330	1120	D	3750
Maschi	D	1150	0	D	4260
Totale	2730	2480	1120	1680	8010

Microdati

Un metodo alternativo alle tabelle per pubblicare risultati raccolti da censimenti o survey è quello di rilasciare file di microdati di uso pubblico.

Un file di microdati può essere visto come una tabella in cui ogni tupla della relazione corrisponde ad un singolo rispondente. I record riportati possono includere

- età, razza, e genere della persona (per file demografici).
- codice di classificazione, numero di dipendenti, fatturato (per file commerciali).

Il rischio di divulgazione per file di microdati relativi a aziende è sempre stato più elevato di quello di dati demografici.

Microdati – 2

Informazioni raccolte su aziende commerciali da agenzie statistiche sono generalmente rilasciate in forma di tabelle di grandezza.

I microdati contengono spesso record di visibilità di rispondenti, che potrebbero quindi essere identificati attraverso altra informazione disponibile.

Storicamente microdati relativi a aziende non sono stati generalmente rilasciati (salvo poche eccezioni). Anche file demografici di microdati sono sempre stati rilasciati in forma limitata.

Sempre più spesso oggi i dati sono rilasciati in forma di microdati.

Microdati risultano più convenienti per i riceventi dell'informazione che possono calcolare statistiche in base ai loro bisogni.

Microdati – 3

È riconosciuto che la protezione dei microdati è più complessa della protezione di tabelle.

Una delle principali difficoltà nella protezione dei microdati è la possibilità di combinare l'informazione rilasciata con altra informazione disponibile.

Non esistono misure largamente accettate per il rischio di divulgazione per i microdati, quindi non esistono "standard" che potrebbero essere applicati per valutare se una protezione è adeguata.

Fattori che aumentano il rischio di divulgazione

Ci sono due maggiori sorgenti di rischio

- Esistenza di record ad alta visibilità. Alcuni record possono rappresentare rispondenti con caratteristiche uniche, come lavori non comuni (es., ministro, attore) o stipendi molto elevati.
- Possibilità di combinare l'informazione con altre informazioni pubblicamente disponibili (record linkage).

  Individui o aziende possono possedere una unica combinazione di valori per certe variabili caratteristiche (quasi identificatori). Esiste rischio di divulgazione se alcuni di questi individui fanno parte del campione della popolazione rappresentato dal file.

Nota: l'identità dei rispondenti nel file campione non dovrebbe essere nota perchè potrebbe essere utilizzata per ri-identificare l'individuo nel file di microdati campione.

Rischio di divulgazione da microdati

Medical Data Released as Anonymous

SSN	Name	Ethn	DOB	Sex	ZIP	Mar. Status	Problem
		asian	09/27/64	female	94139	divorced	hypertension
		asian	09/30/64	female	94139	divorced	obesity
		asian	04/18/64	male	94139	married	chest pain
		asian	04/15/64	male	94139	married	obesity
		black	03/13/63	male	94138	married	hypertension
		black	03/18/63	male	94138	married	shortness of breath
		black	09/13/64	female	94141	married	shortness of breath
		black	09/07/64	female	94141	married	obesity
		white	05/14/61	male	94138	single	chest pain
		white	05/08/61	male	94138	single	obesity
		white	09/15/61	female	94142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	.....
.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....
Sue J. Carlson	900 Market St.	San Fran.	94142	9/15/61	female	democrat	.....
.....	.....	.....	.....	.....	.....	.....	.....

Re-identificazione dei rispondenti

Nel 1997 uno studente dell'MIT analizzò il registro comunale degli abitanti (54,805) di Cambridge, MA trovando che le seguenti percentuali di popolazione potevano essere identificate sulla base di certi dati.

- data di nascita: 12%
- data di nascita & genere: 29%
- data di nascita & CAP di 5 cifre: 69%
- data di nascita & CAP di 9 cifre: 97% (53,033)

Caratteristiche uniche possono identificare i pazienti (es., un paziente cinquantenne in un ospedale pediatrico)

Nello stesso anno lo studente analizzò informazioni anonimizzate rilasciate da un ospedale del Massachussets e riuscì a scoprire (sulla base di dati caratterizzanti) i dati medici relativi a un noto personaggio politico.

Fattori che aumentano il rischio di divulgazione – 2

Il rischio di divulgazione da microdati aumenta notevolmente se il file contiene informazioni amministrative o altri tipi di dati da sorgenti esterne, che potrebbero quindi essere utilizzate per collegare l'informazione.

La possibilità di combinare informazioni aumenta all'aumentare di

- numero di variabili in comune fra file di microdati e sorgenti esterne
- accuratezza dei dati
- numero di sorgenti esterne

La grande quantità di dati oggi più o meno pubblicamente disponibile unita all'aumentata potenza di calcolo ha reso critico il problema della protezione dell'informazione da divulgazione.

Fattori che aumentano il rischio di divulgazione – 3

Il rischio di divulgazione aumenta inoltre quando:

- al crescere della complessità della struttura del file.

  Ad esempio se è noto che due record appartengono allo stesso gruppo (esempio gruppo familiare) c'è rischio maggiore che qusti di questi possa essere identificato anche se nessuna informazione circa il gruppo in quanto tale rilasciata.

- alcuni record del file sono rilasciati anche in un altro file in forma più dettagliata o con codifiche con intersezione delle stesse variabili.

- alcuni record del file sono rilasciati anche in un altro file contenente alcune variabili in comune e altre variabili addizionali.

Fattori che aumentano il rischio di divulgazione – 4

- agenzie statistiche collegano esplicitamente l'informazione nel file di microdati rilasciati con altri file di microdati.

  Questo succede spesso per survey "longitudinali", riguardanti cioè survey dello stesso campione in differenti periodi di tempo.

  Il rischio di divulgazione aumenta quando dati registrati in tempi differenti possono essere collegati allo stesso rispondente.

  Modifiche di valori che un esterno può osservare nei record di un rispondente (esempio, cambiamenti di occupazione, stato civile, o stato economico) nel tempo possono portare all'identificazione del rispondente.

Fattori che diminuiscono il rischio di divulgazione

I fattori che hanno un ruolo maggiore nel diminuire il rischio di divulgazione sono

- campionamento: la maggior parte dei file di microdati rappresentano solo un campione non noto della popolazione.
- "età" dei dati. I dati rilasciati si riferiscono a periodi passati.

Fattori che diminuiscono il rischio di divulgazione – 2

Il campionamento diminuisce attacchi volti a trovare record di individui specifici (probabilità che l'individuo che l'intrusore sta cercando non sia rappresentato).

È sempre però possibile per l'intrusore cercare di re-identificare i rispondenti nel campione.

Grazie al campionamento, anche se le caratteristiche di un rispondente sono uniche all'interno del campione, potrebbero esserci molte entità nel mondo esterno con le stesse caratteristiche del rispondente.

Il rischio di divulgazione può inoltre diminuire se solo i dati relativi ad a un sottoinsieme del campione della popolazione sono pubblicati. (In questo modo anche se un intrusore sa che un individuo o azienda ha partecipato ad un survey non necessariamente lo ritroverà nei dati pubblicati.)

Fattori che diminuiscono il rischio di divulgazione – 3

Un altro fattore che diminuisce il rischio di divulgazione è l'"età" dei dati. Spesso agenzie pubblicano dati che sono almeno vecchi di due anni.

- Le caratteristiche dei rispondenti possono essere cambiate nel frattempo.
- La differenza nel tempo dei dati del file di microdati e i file esterni disponibili informazione diminuisce la possibilità di collegare informazione.

Anche quando ci sono variabili in comune fra il file di microdati e sorgenti esterne, le variabili possono essere registrate in modo diverso (domini diversi di valori).

Un fattore molto importante che limita il rischio di divulgazione è la complessità del problema e quindi la quantità di risorse (in termini di tempo, spazio, e soldi) necessaria all'intrusione per inferire informazione.

Misure di rischio di divulgazione

Valutano il possibile rischio di divulgazione al quale i dati sono esposti. Includono considerazioni su:

- probabilità che un rispondente per il quale un intrusore può cercare informazione sia rappresentato sia nel file di microdati sia in un file esterno che può essere utilizzato per collegare informazione
- probabilità che le variabili in comune siano registrate allo stesso modo nei due file
- probabilità che un rispondente per i quali l'intrusore può cercare informazione sia unico nella popolazione del file esterno

La percentuale di record relativi a rispondenti che sono unici nella popolazione (campioni unici) gioca un ruolo principale nella divulgazione.

Notare che ogni record unico nella popolazione rimane unico nel campione. Il vice-versa non è vero.

Tecniche di protezione per microdati

Per ridurre la possibilità di divulgazione, la maggior parte dei file di microdati rilasciati:

- Include dati solo di un campione della popolazione
- Non include identificatori espliciti
- Limita i dettagli geografici
- Limita il numero di variabili del file

Limitazione dei dettagli geografici

La locazione geografica è una caratteristica che

- appare molto spesso nei file di microdati
- è spesso sfruttata per re-identificazione di rispondenti ($\Longrightarrow$ riduce la popolazione alla quale ci si riferisce)

È quindi necessario limitare i dettagli geografici.

Ad esempio, il Census Bureau non fornisce il codice di alcuna locazione geografica con meno di 100000 persone nel campione (negli anni '80 il limite era 250000). Il limite di 250000 è ancora utilizzato per dati ad alto rischio.

..... però .....

molti file di microdati contengono "variabili contestuali". Sono variabili che identificano l'area nella quale un rispondente risiede senza identificare l'area. Un esempio di variabile contestuali è la temperatura di una zona.

Tecniche di protezione per microdati - 2

Metodi addizionali includono

- Codifica al massimo e minimo (top e bottom-coding). Valori troppo alti o troppo bassi (sopra o sotti una certa soglia) sono riportati solo indicando la soglia.
- Registrazione per intervalli o arrotondamenti, include generalizzazione di valori
- Aggiunta o moltiplicazione per numeri casuali (noise)
- Scambio (swapping)
- Scelta casuale di record, cancellando determinate variabili e definendo valori per esse (blank and impute)
- Aggregazione per piccoli gruppi di rispondenti e sostituzione dei valori individuali con la media (blurring)

File di microdati – Esempio

Esempio Tutti i record della contea Alfa relativi a bambini delinquenti

N	Bambino	Contea	Educ. CF	Stipendio CF	Razza CF
1	John	Alfa	molto alto	201	nera
2	Jim	Alfa	alto	103	bianca
3	Sue	Alfa	alto	77	nera
4	Pete	Alfa	alto	61	bianca
5	Ramesh	Alfa	medio	72	bianca
6	Dante	Alfa	basso	103	bianca
7	Virgil	Alfa	basso	91	nera
8	Wanda	Alfa	basso	84	bianca
9	Stan	Alfa	basso	75	bianca
10	Irmi	Alfa	basso	62	nera
11	Renee	Alfa	basso	58	bianca
12	Virginia	Alfa	basso	56	nera
13	Mary	Alfa	basso	54	nera
14	Kim	Alfa	basso	52	bianca
15	Tom	Alfa	basso	55	nera
16	Ken	Alfa	basso	48	bianca
17	Mike	Alfa	basso	48	bianca
18	Joe	Alfa	basso	41	nera
19	Jeff	Alfa	basso	44	nera
20	Nancy	Alfa	basso	37	bianca

Campionamento, Rimozione di identificatori, e limitazione di dettagli geografici

1. Considera solo un campione della popolazione (es. 10%)

   Nel nostro esempio solo due record (tuple) relative alla popolazione della contea Alfa.

2. Rimuovi identificatori espliciti.

   Nel nostro esempio i nomi dei bambini.

3. Limita i dettagli geografici.

   Ad esempio non mostrare i dati di una contea se ha meno di 30 bambini delinquenti nella popolazione.

   $\Longrightarrow$ Non possiamo mostrare i dati per la contea Alfa e Gamma. Possiamo però combinare i dettagli geografici di Alfa e Gamma.

File di microdati - esempio

Esempio Campionamento, rimozione di identificatori, limitazione di dettagli geografici.

N	Contea	Educ. CF	Stipendio CF	Razza CF
1	AlfaGamma	alto	61	bianca
2	AlfaGamma	basso	48	bianca
3	AlfaGamma	medio	30	nera
4	AlfaGamma	medio	52	bianca
5	AlfaGamma	molto alto	117	bianca
6	Beta	molto alto	138	nera
7	Beta	molto alto	103	bianca
8	Beta	basso	45	bianca
9	Beta	medio	62	bianca
10	Beta	alto	85	bianca
11	Delta	basso	33	nera
12	Delta	medio	51	nera
13	Delta	basso	59	bianca
14	Delta	basso	72	nera

Protezione di microdati

Nel nostro esempio ci sono solo 5 variabili per ogni bambino.

Possiamo pensare che tali variabili siano state scelte da un insieme più completo che include: nome dei genitori, nome e numero dei fratelli, età del bambino, età dei fratelli, indirizzo, scuola, etc.

Più variabili sono incluse più è probabile che un bambino possa essere identificato.

È possibile che informazione disponibile ad altri nella popolazione sia utilizzata con i valori di stipendio pubblicati per identificare univocamente la famiglia del bambino delinquente.

Ad esempio, il datore di lavoro di una persona conosce il suo stipendio con precisione.
Tali variabili sono chiamate variabili ad alta visibilità e necessitano di protezione addizionale.

Codifica per intervalli o arrotondamenti

La ricodifica dei valori in categorie è uno dei metodi di protezione di informazione maggiormente utilizzati.

Ricodifica i valori riportando l'intervallo o la classe di appartenza.

Es.

- stipendi potrebbero essere raggruppati in intervalli di 10 mil. Invece di riportare un valore specifico viene riportato l'intervallo in cui cade.
- Anzichè riportare la data di nascita in modo completo, si può riportare soltanto l'anno e il mese.... soltanto l'anno..... o intervalli di anni

Rende l'informazione meno precisa (ma sempre corretta). La riduzione di precisione diminuisce la possibilità di correlazione dell' informazione poichè diminuisce l'unicità dei valori.

Codifica al massimo e al minimo

- Codifica al massimo (top coding) Un top-code per una certa variabile è un limite superiore su tutti i valori pubblicati per quella variabile. Ogni valor superiore a quel limite non è pubblicato.
  Al suo posto qualche tipo di flag indica che il valore è superiore al top-code. Ad esempio invece di pubblicare un valore di stipendio di 700 mil. possiamo solo dire che è superiore a 500 mil.

  Top-coding è generalmente applicato a variabili quali età, stipendio per individui, o fatturato per aziende.

- Codifica al minimo (bottom coding) Analogamente, rappresenta un limite inferiore per i valori pubblicati per una certa variabile. Valori piccoli (inferiori ad una certa soglia) sono riportati come minori della soglia

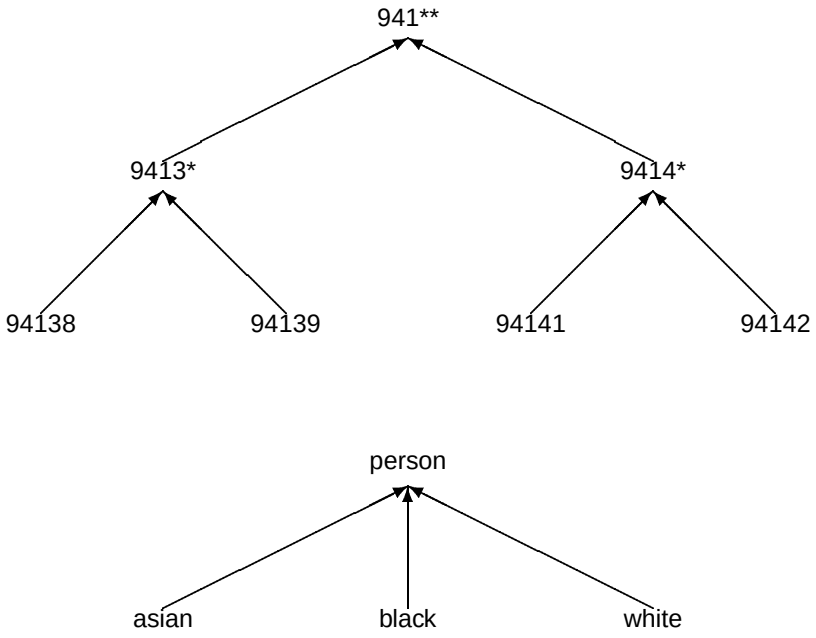
  Esempi di variabili che possono essere codificate al minimo sono anno di nascita per individui o aziende, anno di costruzione.

File di microdati - esempio

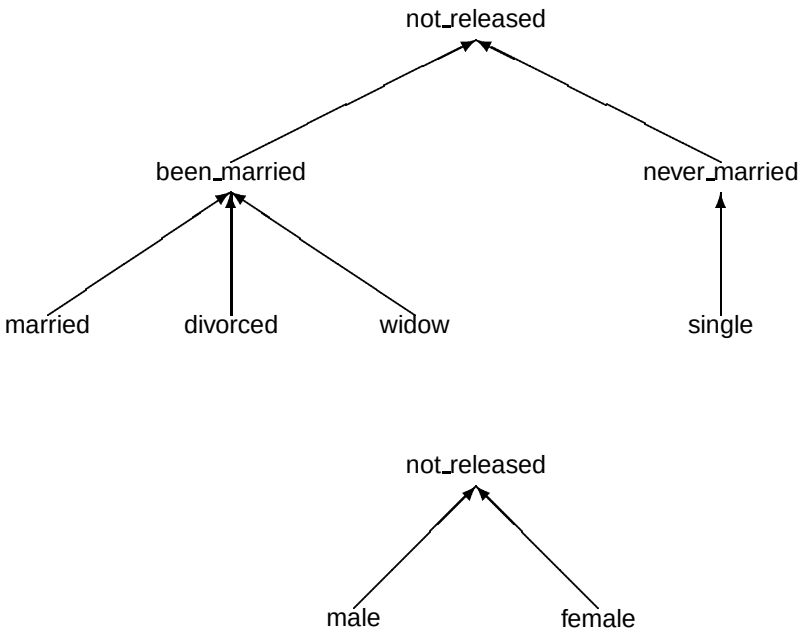
Esempio Campionamento, rimozione di identificatori, limitazione di dettagli geografici, stipendio codificato al massimo e minimo e ricodificato a intervalli.

N	Contea	Educ. CF	Stipendio CF	Razza CF
1	AlfaGamma	alto	60-69	bianca
2	AlfaGamma	basso	40-49	bianca
3	AlfaGamma	medio	<40	nera
4	AlfaGamma	medio	50-59	bianca
5	AlfaGamma	molto alto	>100	bianca
6	Beta	molto alto	>100	nera
7	Beta	molto alto	>100	bianca
8	Beta	basso	40-49	bianca
9	Beta	medio	60-69	bianca
10	Beta	alto	80-89	bianca
11	Delta	basso	<40	nera
12	Delta	medio	50-59	nera
13	Delta	basso	50-59	bianca
14	Delta	basso	70-79	nera

Generalizzazione di valori



Generalizzazione di valori



Generalizzazione di valori – esempio

Quasi identificatori nella tabella originale

Race	DOB	Sex	ZIP	MaritalStatus
asian	09/27/64	female	94139	divorced
asian	09/30/64	female	94139	divorced
asian	04/18/64	male	94139	married
asian	04/15/64	male	94139	married
black	03/13/63	male	94138	married
black	03/18/63	male	94138	married
black	09/13/64	female	94141	married
black	09/07/64	female	94141	married
white	05/14/61	male	94138	single
white	05/08/61	male	94138	single
white	09/15/61	female	94142	widow

PT

Generalizzazione di valori – esempio

Generalizzazione sui diversi attributi [0,2,1,2,2]

Race	DOB	Sex	ZIP	Mar.Status
asian	64	not_rel.	941**	not_released
asian	64	not_rel.	941**	not_released
asian	64	not_rel.	941**	not_released
asian	64	not_rel.	941**	not_released
black	63	not_rel.	941**	not_released
black	63	not_rel.	941**	not_released
black	64	not_rel.	941**	not_released
black	64	not_rel.	941**	not_released
white	61	not_rel.	941**	not_released
white	61	not_rel.	941**	not_released
white	61	not_rel.	941**	not_released

$GT_{[0,2,1,2,2]}$

Ogni tupla ha almeno due rispondenti.

Generalizzazione di valori – esempio

Generalizzazione sui diversi attributi [1,3,0,1,1]

Race	DOB	Sex	ZIP	MaritalStatus
person	[60-64]	female	9413*	been_married
person	[60-64]	female	9413*	been_married
person	[60-64]	male	9413*	been_married
person	[60-64]	male	9413*	been_married
person	[60-64]	male	9413*	been_married
person	[60-64]	male	9413*	been_married
person	[60-64]	female	9414*	been_married
person	[60-64]	female	9414*	been_married
person	[60-64]	male	9413*	never_married
person	[60-64]	male	9413*	never_married
person	[60-64]	female	9414*	been_married

$GT_{[1,3,0,1,1]}$

Ogni tupla ha almeno due rispondenti.

Generalizzazione e soppressione

Sopprimendo alcune tuple (fino a una certa percentuale accettata) posso diminuire la quantità di generalizzazione necessaria.

Race	DOB	Sex	ZIP	MaritalStatus
asian	09/64	female	94139	divorced
asian	09/64	female	94139	divorced
asian	04/64	male	94139	married
asian	04/64	male	94139	married
black	03/63	male	94138	married
black	03/63	male	94138	married
black	09/64	female	94141	married
black	09/64	female	94141	married
white	05/61	male	94138	single
white	05/61	male	94138	single

$GT_{[0,1,0,0,0]}$

Ogni tupla ha almeno due rispondenti.

Ho soppresso una tupla ma posso pubblicare valori più precisi.

Soppressione di celle

La complicazione maggiore sta nel fornire protezione minimizzando la perdita di informazione dovuta a generalizzazione e soppressione.

La soppressione potrebbe essere applicata a livello di cella. Più complessa.

- devo determinare le caratteristiche uniche e cercare di eliminare il numero minimo di celle
- richiede soppressione complementare (secondaria). Oltre alle celle che elimino potrei doverne eliminare altre per impedire linking.

Altre tecniche

Le tecniche viste finora (top and bottom-coding, generalizzazione, arrotondamento) mantengono la veridicità dei dati (i dati pubblicati sono corretti.... anche se meno precisi).

Anche la soppressione mantiene la veridicità dei dati (se comunichiamo che certi valori/tuple sono stati soppressi).

Esistono altri metodi di protezione, che però hanno lo svantaggio di modificare i dati (introducendo incertezza sul loro valore)

- aggiunta di disturbo casuale
- scambio
- blank and impute
- offuscamento (blurring)

Aggiunta di disturbo casuale

Un metodo alternativo per mascherare i valori delle variabili ad alta visibilità è quello di aggiungere o moltiplicarli a numeri casuali.

Ad esempio, possiamo aggiungere una variabile con media 0 e deviazione standard 5 allo stipendio.

Ad esempio applicando aggiunta di disturbo casuale ai microdati campionati scegliamo 14 valori casuali e li aggiungiamo ai valori degli stipendi.

La probabilità di distribuzione utilizzata per il calcolo del disturbo da applicare può essere o meno pubblicata. Renderla nota può comunque aiutare possibili intrusori nell'inferire informazione.

File di microdati - esempio

Esempio Campionamento, rimozione di identificatori, limitazione di dettagli geografici, aggiunta di disturbo casuale.

N	Contea	Educ. CF	Stipendio CF	Razza CF
1	AlfaGamma	alto	61	bianca
2	AlfaGamma	basso	42	bianca
3	AlfaGamma	medio	32	nera
4	AlfaGamma	medio	52	bianca
5	AlfaGamma	molto alto	123	bianca
6	Beta	molto alto	138	nera
7	Beta	molto alto	94	bianca
8	Beta	basso	46	bianca
9	Beta	medio	61	bianca
10	Beta	alto	82	bianca
11	Delta	basso	31	nera
12	Delta	medio	52	nera
13	Delta	basso	55	bianca
14	Delta	basso	61	nera

Aggiunta di disturbo casuale – 2

L' aggiunta di disturbo deve essere fatta con molta cura per garantire la correttezza di statistiche che possono dover essere calcolate sui dati.

Se una agenzia conosce come i dati verranno usati (cioè che statistiche saranno calcolate) può fare sì che il disturbo non comprometta le proprietà statistiche.

I livelli di disturbo necessario per garantire protezione può rendere il risultato finale inutilizzabile per certe applicazioni.

Per questo motivo il disturbo non è molto utilizzato.

Scambio

Con lo scambio scegliamo un campione dei record, troviamo un altro record nel file che ha un match su certe variabili specificate e scambiamo i due record.

Nella applicazione dello swapping per il calcolo di tabelle di frequenza abbiamo scelto record con match su razza e stipendio.
In questo esempio protezione possiamo cercare di trovare un match su livello di educazione e razza e scambiare gli stipendi.

Blank and impute

Vengono scelti alcuni (pochi) record dal file di microdati. Tali record sono sostituiti con valori esplicitamente introdotti.

Con riferimento al nostro esempio.

1. Viene scelto un record a caso da ogni contea pubblicabile (AlfaGamma, Beta, e Delta)
2. Il record scelto è sostituito con un valore fittizio introdotto.

Ad esempio potremmo scegliere il record 2 in AlfaGamma, 6 in Beta, e 13 in Delta e sostituire i valori degli stipendi con 63, 52, 49.

Nota che i numeri inseriti sono fittizi. Possiamo però immaginare che siano calcolati come media su tutte le famiglie della contea con stessa razza e educazione.

Offuscamento (Blurring)

Sostituisce un valore con una media.

Ci sono molti modi di implementare il blurring.

- Gruppi di record per calcolare la media possono essere formati mediante matching su altre variabili o ordinamento di variabili di interesse.
- Il numero di record in un gruppo (per il quale sarà calcolata la media) può essere fisso o variabile.
- La media associata a un gruppo può essere assegnata a tutti i membri o a una parte.
- La media può essere calcolata su più di una variabile utilizzando gruppi differenti per le diverse variabili.

Blurring – esempio

Con riferimento al nostro esempio.

Nel file completo possiamo fare un match su variabili importanti quali: contea, razza e due gruppi di educazione (molto alto, alto) e (medio, basso).

Può poi essere calcolata la media all'interno di ogni gruppo, ad esempio di due record per volta.
Nella contea Alfa il salario per il gruppo $\{$John,Sue$\}$ sarebbe sostituito dalla media dei loro stipendi (139), quello di per $\{$Jim,Pete$\}$ da 82, e così via.

Dopo il blurring il file di dati può essere sottoposto a campionamento, rimozione di identificatori, e limitazione di dettagli geografici.

Basi di dati statistiche

Nelle basi di dati statistiche tutti i dati sono mantenuti on-line in un DBMS (generalmente relazionale) che effettua controlli di inferenza dinamicamente.

Le relazioni nella base di dati statistica sono simili ai file di microdati ma mantengono specifiche informazioni sensibili, quali stipendio, che possono essere rilasciate solo in forma statistica (ad esempio solo la somma o la media degli stipendi può essere rilasciata).

Problema: rispondere alle query statistiche senza rilasciare i singoli dati sensitivi (o diminuire l'incertezza su di essi).

La protezione delle basi di dati statistiche è più matura della protezione di macro e microdati. Ma la tendenza oggi è rilasciare direttamente macro e microdati.