

Privatezza e Protezione dei Dati - Papers

Francesco Fontana, Alessandro Marchetti, Alfredo Santarcangelo

2024

Contents

0.1	Introduction	2
1	<i>k</i>-anonymity	3
1.1	<i>k</i> -anonymity e Table <i>k</i> -anonime	3
1.1.1	<i>Generalization</i>	4
1.1.2	<i>Suppression</i>	5
1.2	Generalizzazione <i>k</i> -Minima	5
1.3	Classificazione tecniche di <i>k</i> -anonymity	7
1.4	Algoritmo Samarati (AG_TS)	7
1.4.1	Evitare il calcolo delle table generalizzate	8
1.5	Bayardo-Agrawal: <i>k-Optimize</i> (AG_TS)	9
1.6	LeFevre-DeWitt et al.: Incognito (AG_TS)	10
1.7	Algoritmi Euristici	10
1.8	Algoritmi per _CS e CG_	11
1.9	Ulteriori studi riguardo <i>k</i> -anonymity	11
1.9.1	<i>k</i> -anonymity multidimensionale	11
1.9.2	<i>l</i> -diversity	12
1.9.3	Valutazione della <i>k</i> -anonymity	12
1.9.4	Algoritmi distribuiti	12
1.9.5	<i>k</i> -anonymity con viste multiple	12
1.9.6	<i>k</i> -anonymity con Micro-Aggregazione	13
1.9.7	<i>k</i> -anonymity per la Location Privacy	13
1.9.8	<i>k</i> -anonymity per i Protocolli di Comunicazione	13
2	Protezione Microdati	14
2.1	Introduzione	14
2.2	Macrodata vs Microdata	14
2.3	Classificazione delle tecniche di protezione riguardo il Disclosure dei microdati	14
2.4	Tecniche di Mascheramento	17
2.4.1	Tecniche non perturbative	17
2.4.2	Tecniche perturbative	18
2.5	Tecniche per la Generazione di Dati Sintetici	20
2.6	Misure per valutare la confidenzialità e l'utilità dei Microdati	20
2.6.1	Rischio di Disclosure (Disclosure Risk)	21

2.6.2	Perdita di Informazione (Information Loss)	24
2.6.3	Combnazione di Rischio di Disclosure e Perdita di Infor- mazione	26
2.7	Conclusioni	27
2.8	Introduzione	28
2.9	CIA nel Cloud	28
2.10	Problemi e sfide	29
2.10.1	Protezione dei dati al rilascio	29
2.10.2	Accesso a grana fine ai dati nel cloud	29
2.10.3	Accesso selettivo ai dati nel cloud	30
2.10.4	Privacy dell'utente	30
2.10.5	Privacy delle query	30
2.10.6	Integrità della computazione e delle query	31
2.10.7	Esecuzione collaborativa delle query con provider multipli	31
2.10.8	SLA e Auditing	31
2.10.9	Multi-locazione e virtualizzazione	31
2.11	Conclusioni	32
3	Supporting User Privacy Preferences in Digital Interactions	33
3.1	Introduzione	33
3.2	Concetti Base e <i>Desiderata</i>	34
3.2.1	Client Portfolio	34
3.2.2	Atomicità Credenziali	35
3.2.3	Policy di Disclosure	35
3.2.4	Trust Negotiation	36
3.2.5	Client Privacy Preferences	36
3.2.6	Preferenze Privacy del Server	37
3.3	Cost-Sensitive Trust Negotiation	37
3.3.1	Minimizzare il Costo	38

0.1 Introduction

Chapter 1

k -anonymity

Contesto è il rilascio di *microdata*. De-identificazione non garantisce anonimità.

1.1 k -anonymity e Table k -anonime

Il concetto di k -anonymity cerca di catturare sulla Private Table (PT) il vincolo che i dati rilasciati dovrebbero essere associabili in maniera indistinguibile a non meno di un certo numero di respondent.

Il set di attributi disponibili esternamente e quindi sfruttabili per fare linking è chiamato *quasi-identifier*.

Definizione 1 (k -anonymity requirement)

Ogni rilascio di data deve essere tale che ogni combinazione di valori del Quasi-Identifier può essere matchata in maniera indistinguibile con almeno k respondent.

k -anonymity richiede che ogni valore del *Quasi-Identifier* abbia almeno k occorrenze nella table rilasciata, come da def 1.1 che segue:

Definizione 2 (k -anonymity)

Date una table $T(A_1, A_2, \dots, A_m)$ e un insieme di attributi QI , Quasi-Identifier sulla table T :

T soddisfa k -anonymity rispetto a QI se e solo se ogni sequenza di valori in $T[QI]$ appare almeno con k occorrenze in $T[QI]$ ¹

Definizione 1.1 è sufficiente per k -anonymity. Applicazione di k -anonymity richiede una preliminare identificazione del *Quasi-Identifier*.

Il *Quasi-Identifier* dipende dalle informazioni esterne disponibili al recipiente poichè determina le capacità di linking dello stesso. Diversi *Quasi-Identifier* possono potenzialmente esistere per una data table.

Per semplicità a seguire nel paper si assume che:

¹ $T[QI]$ denota la proiezione con tuple duplicate degli attributi QI in T

- PT ha unico *Quasi-Identifier*.
- *Quasi-Identifier* è composto da tutti gli attributi nella PT disponibili esternamente.
- PT contiene al massimo una sola tupla per ogni respondent.

k -anonymity si concentra su due tecniche di protezione: *Generalization* e *Suppression*, le quali preservano la veridicità dei dati (diversamente da swapping e scrambling).

1.1.1 *Generalization*

Sostituzione dei valori di un attributo con valori più generali. Consideriamo:

- *Domain*: set di valori che un attributo può assumere.
- *Generalized domains*: contiene valori generalizzati e relativo mapping tra ogni domain e ogni sua generalizzazione.
- *Dom*: set di domini originali con le loro generalizzazioni.
- *Generalization relationship* \leq_D : dati $D_i, D_j \in \text{Dom}$, $D_i \leq D_j$ significa che i valori in D_j sono generalizzazioni dei valori in D_i .

\leq_D definisce ordinamento parziale su *Dom* ed è richiesto nelle seguenti condizioni:

Condizione 1 (C1 - Determinismo nel processo di generalizzazione)

$\forall D_i, D_j, D_z \in \text{Dom}$:

$$D_i \leq_D D_j, D_i \leq_D D_z \implies D_j \leq_D D_z \vee D_z \leq_D D_j^2.$$

Condizione 2 (C2 -)

Tutti gli elementi massimali di *Dom* sono singoletti (*singleton*)³.

- **DGH_D** - *Domain Generalization Hierarchy*: gerarchia di ordinamento totale per ogni dominio $D \in \text{Dom}$.

Per quanto riguarda i valori nei domini consideriamo:

- *Value generalization relationship* \leq_V : associa ogni valore in D_i ad un unico valore in D_j , sua generalizzazione.
- **VGH_D** - *Value Generalization Hierarchy*: albero dove
 - Foglie sono valori in D .
 - Radice è il valore, singolo, nell'elemento massimale di DGH_D

²Questo comporta che ogni dominio D_i ha al massimo un solo dominio di generalizzazione diretta D_j

³La condizione assicura che tutti i valori in ogni dominio possano essere generalizzati ad un singolo valore

1.1.2 Suppression

Consideriamo Soppressione di Tupla. Questa "modera" la *Generalization* quando un numero limitato di *outlier*⁴ forzerebbe una generalizzazione elevata.

1.2 Generalizzazione *k*-Minima

Definizione 3 (Table Generalizzata con Soppressione)

Consideriamo T_i e T_j due table sugli stessi attributi.

T_j è generalizzazione (con soppressione di tupla) di T_i , riportata come $T_i \preceq T_j$, se:

1. $|T_j| \leq |T_i|$
2. Dominio $\text{dom}(A, T_j)$ è uguale o una generalizzazione di $\text{dom}(A, T_i)$, dove A indica ogni attributo in $T_{i,j}$
3. E' possibile definire funzione iniettiva che associa ogni tupla $t_j \in T_j$ con una tupla $t_i \in T_i$, per la quale ogni attributo in t_j è uguale o generalizzazione del corrispondente in t_i .

Definizione 4 (Distance Vector)

Siano $T_i(A_1, \dots, A_n)$ e $T_j(A_1, \dots, A_n)$ tali che $T_i \preceq T_j$.

il distance vector di T_j da T_i è il vettore

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

dove ogni $d_z, z = 1, \dots, n$ è la lunghezza dell'unico percorso tra $\text{dom}(A_z, T_i)$ e $\text{dom}(A_z, T_j)$ nella DGH_{D_z}

Corollario 1 (Ordine Parziale tra DV)

$DV = [d_1, \dots, d_n] \leq DV' = [d'_1, \dots, d'_n]$ se e solo se $d_i \leq d'_i$ per $i = 1, \dots, n$.

Si costruisce una gerarchia di distance vectors come lattice (diagramma) corrispondente alla DGH_D come in fig. 1.1

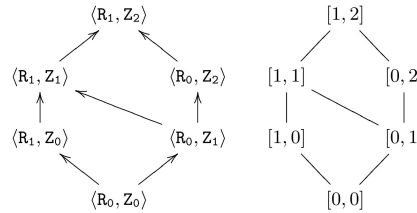


Figure 1.1

⁴Secondo paper riportato, per *Outlier* si intende respondent che sono "molto diversi" o "molto distanti" dalla maggior parte degli altri individui. DOI: https://doi.org/10.1007/978-3-662-46497-7_11.

Per bilanciare tra perdita di precisione dovuta a *Generalization* e perdita di completezza dovuta a *Suppression* si suppone che data holder determini la soglia **MaxSup**, che indica il numero di tuple che possono essere sopprese.

Definizione 5 (Generalizzazione *k*-minima con Soppressione)

Siano T_i e T_j due table tali che $T_i \preceq T_j$, e sia **MaxSup** la soglia di soppressione accettabile scelt. T_j è una generalizzazione *k*-minima di T_i se e solo se:

1. T_j soddisfa *k*-anonymity applicando soppressione minima, ossia T_j soddisfa *k*-anonymity e: $\forall T_z : T_i \preceq T_z, DV_{i,z} = DV_{i,j}, T_z$ soddisfa *k*-anonymity $\implies |T_j| \geq |T_z|$.
2. $|T_i| - |T_j| \leq \text{MaxSup}$.
3. $\forall T_z : T_i \preceq T_z$ e T_z soddisfa le condizioni 1 e 2 $\implies \neg(DV_{i,z} < DV_{i,j})$.

Ultima espressione rende meglio come $DV_{i,z} \geq DV_{i,j}$. Il concetto che esprime è che "non esiste un'altra *Generalization* T_z che soddisfi 1 e 2 con un *DV* minore di quello di T_j "

Diversi **preference criteria** possono essere applicati nella scelta della generalizzazione minimale preferita:

- **Distanza assoluta minima:** minor numero totale di passi di generalizzazione (indipendentemente dalle gerarchie di *Generalization* considerate).
- **Distanza relativa minima:** minimizza il numero relativo di passi di generalizzazione (passo relativo ottenuto dividendo per l'altezza del dominio della gerarchia a cui si riferisce).
- **Massima distribuzione:** maggior numero di tuple distinte.
- **Minima soppressione:** minor tuple sopprese (maggior cardinalità).

1.3 Classificazione tecniche di k -anonymity

Classificazione delle tipologie di composizione *Generalization* e *Suppression* sono mostrati in fig.1.2.

Generalization	Suppression			
	<i>Tuple</i>	<i>Attribute</i>	<i>Cell</i>	<i>None</i>
<i>Attribute</i>	AG_TS	AG_AS \equiv AG_	AG_CS	AG_- \equiv AG_AS
<i>Cell</i>	CG_TS not applicable	CG_AS not applicable	CG_CS \equiv CG_	CG_- \equiv CG_CS
<i>None</i>	_TS	_AS	_CS	- not interesting

Fig. 8. Classification of k -anonymity techniques

Figure 1.2

Casi *not applicable* (**CG_TS** e **CG_AS**): supportare *Generalization* a grana fine (cella) implica poter applicare soppressione allo stesso livello.

Vengono espote qui di seguito alcune soluzioni algoritmiche in letteratura, riassunte in fig.1.3

Algorithm	Model	Algorithm's type	Time complexity
Samarati [26]	AG_TS	Exact	exponential in $ QI $
Sweeney [29]	AG_TS	Exact	exponential in $ QI $
Bayardo-Agrawal [5]	AG_TS	Exact	exponential in $ QI $
LeFevre-et-al. [20]	AG_TS	Exact	exponential in $ QI $
Aggarwal-et-al. [2]	_CS	$O(k)$ -Approximation	$O(kn^2)$
Meyerson-Williams [24] ²	_CS	$O(k \log k)$ -Approximation	$O(n^{2k})$
Aggarwal-et-al. [3]	CG_-	$O(k)$ -Approximation	$O(kn^2)$
Iyengar [18]	AG_TS	Heuristic	limited number of iterations
Winkler [33]	AG_TS	Heuristic	limited number of iterations
Fung-Wang-Yu [12]	AG_-	Heuristic	limited number of iterations

Figure 1.3: Alcuni approcci a k -anonymity(n è numero di tuple in PT).

1.4 Algoritmo Samarati (AG_TS)

Il primo algoritmo per garantire k -anonymity è stato proposto insieme alla definizione di k -anonymity. La definizione di k -anonymity è basata sul QI quindi l'algoritmo lavora solo su questo set di attributi e su table con più di k tuple.

Data una *DGH* ci sono diversi percorsi dall'elemento in fondo alla gerarchia alla radice. Ogni percorso è una differente *strategia* di generalizzazione. Su ogni percorso c'è esattamente una *Generalization* minima localmente (nodo più basso che garantisce k -anonymity).

In maniera naif si può cercare su ogni percorso il minimo locale per poi trovare il minimo globale tra questi ma non è praticabile per l'elevato numero di percorsi.

Per ottimizzare la ricerca si sfrutta la proprietà che salendo nella gerarchia la soppressione richiesta per avere k -anonymity diminuisce:

- Ogni nodo in DGH viene associato ad un numero, **height**, corrispondente alla somma degli elementi nel Distance Vector associato.
- Altezza di ogni DV nel diagramma (*distance vector lattice VL*) si scrive come $height(DV, VL)$.

Se non c'è soluzione che soddisfi k -anonymity sopprimendo meno di $MaxSup$ ad altezza h non può esistere soluzione che soddisfi ad una altezza minore.

L'algoritmo usa binary search cercando la minore altezza in cui esiste un DV che soddisfa k -anonymity rispettando $MaxSup$ e ha come primo passo:

$$\text{Cerco ad altezza } \lfloor \frac{h}{2} \rfloor : \begin{cases} \text{trovo vettore che soddisfa } k\text{-anonymity} \implies \lfloor \frac{h}{4} \rfloor \\ \text{altrimenti} \implies \text{cerco in } \lfloor \frac{3h}{4} \rfloor \end{cases} \quad (1.1)$$

La ricerca prosegue fino a trovare l'altezza minore in cui esiste vettore che soddisfa k -anonymity con $MaxSup$.

1.4.1 Evitare il calcolo delle table generalizzate

Algoritmo richiederebbe il calcolo di tutte le table generalizzate. Per evitarlo introduciamo il concetto di DV tra tuple.

Definizione 6 (Distance Vector tra tuple - Antenato Comune)

Sia T una table.

Siano $x, y \in T$ due tuple tali che $x = \langle v'_1, \dots, v'_n \rangle$ e $y = \langle v''_1, \dots, v''_n \rangle$ con v'_i e v''_i valori in D_i con $i = 1, \dots, n$.

Il **distance vector** tra x e y è $V_{x,y} = [d_1, \dots, d_n]$. dove d_i è la lunghezza (uguale) dei due percorsi da v'_i e v''_i al loro comune antenato comune più prossimo v_i sulla VGH_{D_i} .

In altri termini ogni distanza in $V_{x,y}$ è una distanza uguale dal dominio di v'_i e v''_i al dominio in cui sono generalizzati allo stesso valore v_i .

Allo stesso modo $V_{x,y}$ per $x, y \in T_i$ equivale a $DV_{i,j}$ per $T_i \preceq T_j$ per cui x e y vengono generalizzate alla stessa tupla t .

Per il momento il resto è delirio TODO

1.5 Bayardo-Agrawal: *k-Optimize* (AG_TS)

Approccio considera che la generalizzazione di attributo A su dominio **ordinato** D corrisponde ad un partizionamento del dominio dell'attributo in intervalli. Ogni valore del dominio deve comparire in un intervallo e ogni valore in un intervallo precede ogni valore degli intervalli che lo seguono.

Si assume quindi un ordine tra gli attributi del *Quasi-Identifier*. Inoltre associa un valore intero chiamato *index* ad ogni intervallo di ogni dominio degli attributi del *Quasi-Identifier*.

Una *Generalization* è quindi rappresentata come l'unione dei valori di indice per ogni attributo (il valore più piccolo in un dominio viene omesso perché comparirà sicuramente nella generalizzazione per quel dominio).

k-Optimize costruisce un *set enumeration tree* sul set I di valori di indice, con radice vuota. Ogni nodo figlio è costruito dal padre inserendo in coda un index di I maggiore degli altri index già presenti nel padre (per ordinamento totale).

La visita dell'albero permette di valutare con Depth First Search ogni nodo, prunando ogni nodo e tutti i suoi figli se questi non possono corrispondere a soluzioni ottimali. Nello specifico *k-Optimize* pruna un nodo n quando determina che nessuno dei suoi discendenti può essere ottimale. Data una funzione di costo l'algoritmo calcola un limite inferiore sul costo che può essere ottenuto sul subtree del nodo n , prunando se nodo a costo minore è già stato trovato.

Se un nodo viene prunato allora anche altri nodi, non per forza del sottoalbero, possono essere prunati: supponendo di prunare il nodo $\{1,3\}$ in figura 1.4 allora posso prunare qualunque altro nodo contenente 1 E 3, come ad esempio $\{1,2,3\}$.

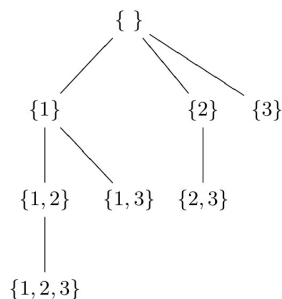


Figure 1.4: Set enumeration tree sull'insieme di indici $I=\{1,2,3\}$

1.6 LeFevre-DeWitt et al.: Incognito (AG_TS)

Idea di base è riassunta nella seguente definizione:

Definizione 7 (*k*-anonymity dei subset di *QI*)

Se una table *T* con Quasi-Identifier *QI* composto da $m = |QI|$ attributi soddisfa *k*-anonymity, *T* soddisfa *k*-anonymity anch per qualunque Quasi-Identifier *QI'* talce che $QI' \subset QI$.

Pertanto *k*-anonymity su un subset di *QI* è condizione necessaria (e non sufficiente) per *k*-anonymity di *T* sull'intero *QI*.

Algoritmo esclude in anticipo alcune generalizzazioni della gerarchia con un calcolo a priori.

Strategia: bottom-up BFS sulla *DGH*. Incognito genera tutte le possibili table *k*-anonime minime secondo i seguenti passaggi:

1. (Iterazione 1): verifica *k*-anonymity per ogni singolo attributo nel *Quasi-Identifier*, scartando quelle che non soddisfano.
2. (Iterazione 2): combina a coppie le *Generalization* non scartate al passo 1 verificando *k*-anonymity.
3. (...)
4. (Iterazione *m*): arriva a considerare l'intero set di attributi di *QI*

Utilizzando approccio bottom-up, per la condizione citata prima, se una *Generalization* soddisfa *k*-anonymity allora anche sue ulteriori dirette generalizzazioni soddisfano *k*-anonymity e pertanto non sono considerate ulteriormente.

Un esempio per $QI = \{Race, Sex, Marital Status\}$ è riportato nella figura 1.5 che segue:

1.7 Algoritmi Euristici

k-anonymity è un problema NP-difficile e pertanto ha complessità esponenziale nella dimensione del *QI*.

Alcuni approcci euristici alternativi proposti:

- Iyengar - algoritmi genetici e ricerca stocastica incompleta.
- Winkler - "simulated annealing" che non garantisce qualità ed è costoso.
- Fung, Wang, Yu - Top-down che, partendo dalla soluzione più generale e specializza iterativamente alcuni valori della soluzione corrente finchè non viola *k*-anonymity⁵.

Essendo metodi euristici non esiste un limite di efficienza per queste soluzioni.

⁵ Algoritmo determina ad ogni passo una "buona" specializzazione guidata da una metrica che compara sia *information gain* che *anonymity loss*

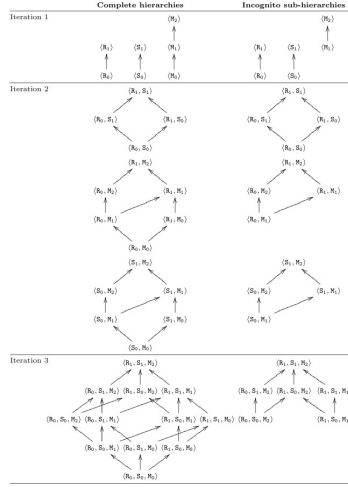


Figure 1.5

1.8 Algoritmi per $_CS$ e $CG_$

Algoritmi esatti per $_CS$ e $CG_$, operando a livello di cella, possono avere costo computazionale esponenziale nel numero di tuple della table. Per questo motivo si sfruttano algoritmi di approssimazione.

La descrizione degli algoritmi è qui omessa in quanto non presentati a lezione, ma indicazione sul nome e sul costo computazionale sono riportati in fig.1.3.

1.9 Ulteriori studi riguardo k -anonimity

Breve resoconto di alcuni studi riguardanti k -anonimity.

1.9.1 k -anonimity multidimensionale

Nella proposta originale di k -anonimity si assume che ogni valore abbia una sola generalizzazione. Tuttavia un *Generalization* step potrebbe portare a diversi valori generalizzati (e.g. $91412 \rightarrow 9141^*$ oppure 914^*2) e quindi avremmo, invece che un albero, un grafo per la DGH.

LeFevre, DeWitt e Ramakrishnan hanno proposto un modello multidimensionale NP-difficile con approssimazione greedy, sia per domini numerici che categorici. Complessità è $O(n \log n)$ con n numero di tuple, e la qualità della GT è migliore rispetto alla monodimensionale.

1.9.2 l -diversity

Qui riportata l -diversity come presentata a lezione, in considerazione dell'*homogeneity attack* e del *background knowledge attack*. Autori del concetto di l -diversity sono Machanavajjhala, Gehrke, and Kifer.

1.9.3 Valutazione della k -anonymity

Analizzando il risultato della k -anonymity con mining, Aggarwal mostra che all'aumentare della dimensione $|QI|$ la perdita di informazione potrebbe diventare elevata (perchè la probabilità che k tuple siano simili è molto bassa).

1.9.4 Algoritmi distribuiti

Anonimizzare dati distribuiti tra diverse parti interconnesse. Diverse proposte:

- Jiang e Clifton: partizione verticale della table e storage in siti diversi. Ricostruzione con join tramite una chiave comune. Necessaria strategia di k -anonymity comune.
- Wang, Fung e Dong: come Jiang ma strategia negoziata interattivamente tra le parti.
- Zhong, Yang e Wright: partizione orizzontale e due soluzioni di cui una prevede crypto valori sensibili e l'altra prevede .CS.

1.9.5 k -anonymity con viste multiple

Definizione 8 (Data Association)

Due o più attributi sono da considerare più sensibili quando i loro valori sono associati rispetto a quando questi valori appaiono separatamente.

Un problema di questo tipo risulta evidente nell'esempio in fig.1.6 in cui è facile ricostruire le associazioni John_Obesity e David_Obesity dalle viste v_1 e v_2 .

Name	Sex	Disease	Name	Sex	Sex	Disease
Sue	F	hypertension	Sue	F	F	hypertension
Claire	F	obesity	Claire	F	F	obesity
Ann	F	chest pain	Ann	F	F	chest pain
John	M	obesity	John	M	F	short breath
David	M	obesity	David	M	M	obesity
Mary	F	short breath	Mary	F		
Alice	F	short breath	Alice	F		
Carol	F	chest pain	Carol	F		
Kate	F	short breath	Kate	F		
PT			view v_1		view v_2	

Figure 1.6: Una PT e due sue possibili viste v_1 e v_2

Yao, Wang e Jajodia mostrano che il problema è NP^{NP} -difficile, mentre se non ci sono dipendenze funzionali tra le viste esiste soluzione polinomiale.

1.9.6 k -anonymity con Micro-Aggregazione

Domingo-Ferrer e Mateo-Sanz propongono di usare Micro-Aggregation al posto di *Generalization Suppression*.

Definizione 9 (Micro-Aggregation)

Micro-Aggregation consiste nel:

1. *dividere microdata in cluster di almeno k tuple ciascuno.*
2. *per ogni attributo, il valore medio sul cluster sostituisce il valore della tupla nel cluster.*

La partizione ottimale è quella che massimizza l'homogeneity nel cluster, ossia clusterizza valori simili. Questo riduce l'information loss.

La somma dei quadrati è il metodo tradizionale per determinare il cluster.

Essendo NP-difficile, si riduce complessità riducendo lo spazio delle soluzioni a cluster di dimensione tra k e $2k$ ⁶.

1.9.7 k -anonymity per la Location Privacy

Soluzioni che garantiscono k -anonymity non tra tuple in un database ma in un set di individui che inviano un messaggio nello stesso contesto spazio-temporale.

1.9.8 k -anonymity per i Protocolli di Comunicazione

Un protocollo di comunicazione è *sender k -anonymous* (*receiver k -anonymous*) se l'attaccante può rilevare solo un set di k possibili *sender* (*receiver*) er un messaggio.

⁶ k per garantire k -anonymity, $2k$ per ridurre information loss.

Chapter 2

Protezione Microdati

2.1 Introduzione

2.2 Macrodata vs Microdata

2.3 Classificazione delle tecniche di protezione riguardo il Disclosure dei microdati

Il controllo del disclosure dei microdati è un'importante problema pratico sia nel privato che nel pubblico.

La protezione dei microdati ha apparentemente due obiettivi tra di loro contrastanti. Da una parte si vuole evitare la re-identificazione, la quale accade ogni qualvolta l'informazione del respondent che appare nella table di microdati è identificata; il quale, è associata con le identità dei respondent corrispondenti.

Dall'altra invece le applicazioni di tecniche dovrebbero preservare le *chiavi delle proprietà statistiche* dei dati originali i quali sono state segnati come importanti dai destinatari dei dati (ossia chi riceverà i dati finali)

Precisamente data una table di microdati T , una tecnica di protezione dei dati dovrebbe trasformare questa table T in una table T_1 in modo che:

1. il rischio che un utente malevolo possa usare T_1 per determinare informazioni confidenziali o identificare un respondent, sia basso
2. l'analisi statistica su T e T_1 possa produrre risultati simili

Generalmente i seguenti fattori contribuiscono al rischio di disclosure:

- L'esistenza di tuple altamente visibili (ossia tutte quelle tuple con caratteristiche che hanno caratteristiche uniche)

- Possibilità di matching tra le table di microdati e le informazioni esterne.
- L'esistenza di un alto numero degli attributi comuni che possono aumentare la possibilità di linking.

Mentre i fattori che minimizzano il rischio di disclosure sono:

- Una table di microdati contenenti un subset della popolazione intera. Questo implica che le informazioni di un respondent specifico, il quale potrebbe essere un utente malevolo potrebbe voler sapere, potrebbe non essere incluso nella table di microdati.
- Le informazioni specificate nelle table rilasciate, quindi pubbliche non sono aggiornate. Potrebbero cambiare nel tempo, indi per cui il linking con le informazioni esterne potrebbero non essere accurati
- Una table di microdati e le informazioni esterne contengono rumore, riducendo la probabilità di linking delle informazioni
- Una table di microdati e le risorse esterne possono contenere dati espressi in forme diverse, riducendo l'abilità di linking

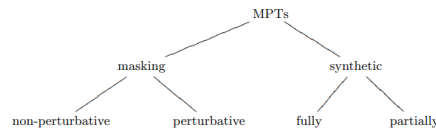


Fig. 3. Classification of microdata protection techniques (MPTs)

Figure 2.1: Classificazione delle tecniche di protezione dei microdati

Generalmente, per limitare il rischio di disclosure di una table di microdati è necessario sopprimere gli identifiers impliciti ed espliciti come step iniziale (e.g. SSN, Name)

Questo processo è noto come de-identificazione, quest'ultima non necessariamente rende le tuple anonime, siccome è possibile fare re-identificazione usando informazioni esterne.

Tipicamente le tecniche sono basate sul principio che la re-identificazione possa essere contrastata riducendo la quantità delle informazioni rilasciate, mascherando i dati (e.g., non rilasciando o perturbando i valori), o rilasciando valori plausibili modificati invece di quelli reali. Seguendo questo principio le tecniche si suddividono in due macro categorie:

- Tecniche di mascheramento: I dati originali sono trasformati producendo nuovi dati che non sono validi per le analisi statistiche in modo da preservare la confidenzialità dei respondent. Le tecniche di mascheramento si suddividono in:

- non perturbative: il dato originale non viene modificato, ma alcuni dati vengono soppressi o vengono rimossi dettagli.
- perturbativi: i dati originali sono modificati.
- Generazione dei dati sintetici: I set di tuple originale nelle table di microdati sono sostituiti da un nuovo set di tuple generati in modo di preservare le proprietà statistiche chiave del dato originale. Dal momento che le table dei microdati rilasciati contengono dati sintetici, il rischio di re-identificazione viene ridotto. Questi possono essere interamente sintetici (fully synthetic) o misti con i dati originali (partially synthetic).

Un'altra caratteristica importante delle tecniche di protezione è che possono essere usati su tipo di dati differenti. In particolare i tipi di dati possono essere categorizzati come

- *Continui*: Un attributo è detto continuo se le operazioni numeriche e aritmetiche sono definite sull'attributo. E.g. Data di nascita e temperatura sono attributi continui.
- *Categorici*: Un attributo è definito categorico se le operazioni aritmetiche non hanno alcun senso se fatto sull'attributo. E.g. Razza, su cui non si possono fare operazioni aritmetiche

Di seguito, sono descritti le tecniche di protezione dei microdati principali, e se sono applicabili a dati continui, categorici o entrambi.

Technique	Continuous	Categorical
Sampling	yes	yes
Local suppression	yes	yes
Global recoding	yes	yes
Top-coding	yes	yes
Bottom-coding	yes	yes
Generalization	yes	yes

Figure 2.2: Applicazione delle tecniche non perturbative su dati differenti

Technique	Continuous	Categorical
Resampling	yes	no
Lossy compression	yes	no
Rounding	yes	no
PRAM	no	yes
MASSC	no	yes
Random noise	yes	yes
Swapping	yes	yes
Rank swapping	yes	yes
Micro-aggregation	yes	yes

Figure 2.3: Applicazione delle tecniche perturbative su dati differenti

2.4 Tecniche di Mascheramento

Alcune tecniche di mascheramento

2.4.1 Tecniche non perturbative

Le tecniche non perturbative producono microdati eliminando dettagli dai dati originali. Di seguito alcune di esse:

- *Sampling*: I table dei microdati protetti includono solo tuple formate da sample (campioni) della popolazione totale. Siccome c'è dell'incertezza se uno specifico respondent sia presente nei sample, il rischio di re-identificazione viene ridotto. Per esempio possiamo decidere se pubblicare solo le tuple pari della table originale. Questa tecnica può operare solo su dati categorici
- *Local suppression*: Sopprime i valori degli attributi in modo da limitare la possibilità di analisi. Questa tecnica lascia uno spazio vuoto su alcuni attributi (sensitive cells) che contribuiscono in modo significativo al rischio di disclosure delle tuple coinvolte.
- *Global Recoding*: Il dominio degli attributi, ossia tutti i possibili valori, sono divisi intervalli disgiunti di grandezza uguale, e ogni intervallo viene associato a una label. La table dei microdati protetti sono ottenuti sostituendo i valori degli attributi con le label associate agli intervalli corrispondenti. Questa tecnica riduce i dettagli della table e quindi riduce il rischio di re-identificazione. Sono presenti due tecniche particolari di recoding, ossia *Top-Coding* e *Bottom-Coding*, qui sotto descritte:
 - *Top-Coding*: E' basata sulla definizione di limite superiore (upper limit), chiamato top-code, per ogni attributo che deve essere protetto. Ogni valore maggiore di questo valore è sostituito dal top-code. Questa tecnica può essere applicata agli attributi categorici che possono essere ordinati linearmente come gli attributi continui.
 - *Bottom-Coding*: Simile al top-coding ma definisce il limite inferiore (lower limit). Quindi, ogni valore più basso di questo non verrà ripubblicato e verrà sostituito con il bottom-code.
- *Generalization*: Consiste nel rappresentare i valori di un dato attributo usando valori più generali. Questa tecnica è basata sulla definizione di generalizzazione gerarchica, dove i valori più generali sono alla radice della gerarchia e le foglie corrispondono ai valori specifici. Un processo di generalizzazione perciò procede con la sostituzione dei valori rappresentati dai nodi foglia con dei valori al nodo superiore.

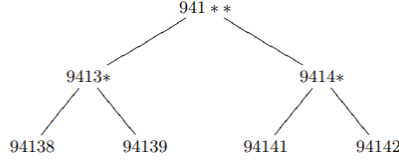


Fig. 6. Generalization hierarchy for attribute ZIP

Figure 2.4: Generalizzazione gerarchica

2.4.2 Tecniche perturbative

Con le tecniche perturbative le table dei microdati vengono modificate per la pubblicazione. Le modifiche possono portare a combinazioni dei valori originali uniche le quali svaniscono non appena vengono introdotte nuove combinazioni.

- *Resampling*: Questa tecnica sostituisce i valori degli attributi sensibili continui con un valore medio computato sul numero di samples della popolazione originaria. Precisamente, assumiamo che N sia il numero di tuple presenti in una table di microdati, e S_1, \dots, S_t con t i sample di dimensione N . Ogni sample viene rankato indipendentemente e successivamente viene calcolata la media del j_{esimo} valore ranked. I valori medi ottenuti sono riordinati prendendo in considerazione l'ordine dei valori originali, seguendo quindi l'ordinamento della tabella iniziale.

					Microdata Protection				
S ₁	S ₂	S ₃	S ₄		S ₁	S ₂	S ₃	S ₄	Average
260	220	170	210		170	150	170	170	165
170	280	290	190		170	180	185	185	180
200	210	220	230		185	190	190	190	188.75
280	310	270	200		190	210	200	200	200
190	290	185	185		200	220	220	210	212.5
185	180	300	260		200	265	250	220	233.75
200	285	250	220		200	270	260	230	240
290	265	260	290		260	280	270	230	260
170	150	190	230		280	285	270	260	273.75
300	270	270	310		290	290	290	290	290
200	298	200	170		300	310	300	310	305
(a) Initial samples					(b) Ordered samples				
Original value (S ₁)					Released value				
260					260				
170					165				
200					212.5				
280					273.75				
190					200				
185					188.75				
200					233.75				
290					290				
170					180				
300					305				
200					240				
(c) Released data									

Fig. 8. An example of resampling over attribute Chol

Figure 2.5: Esempio di resampling

- *Rounding*: Simile alla tecnica usata anche per la protezione dei macrodati ed è applicabile solo agli attributi continui. Sostituisce i valori originali con dei valori approssimati, quest'ultimi sono scelti da un set di *rounding points* p_i ognuno il quale definisce il *rounding set*. Per esempio:

- ***Rounding Points*** possono essere scelti come multipli di un valore base b , con $b = p_{i+1} - p_i$.
- ***Rounding Sets*** definiti come

$$* \begin{cases} [p_i - \frac{b}{2}, p_i + \frac{b}{2}) \text{ con } i = 2, \dots, r-1 \\ [0, p_1 + \frac{b}{2}) \\ [p_r - \frac{b}{2}, X_{max}], \text{ con } X_{max} \text{ il valore più grande possibile} \end{cases}$$

* rispettivamente per p_1 e p_r

* Un valore v di X viene sostituito dal round point corrispondente al rounding set dove è contenuto v .

- *Random Noise*: Perturba gli attributi sensibili aggiungendo o moltiplicando una variabile casuale con una distribuzione. Il rumore additivo (additive noise) è preferito al rumore moltiplicativo (multiplicative noise) e viene formalmente definito così:

Poniamo X_j come la j -esima colonna della table dei microdati originaria corrispondente a un attributo sensibile e supponiamo che ci siano N tuple. Ogni valore X_{ij} , con $i = 1, \dots, N$, viene sostituito con $x_{ij} + \epsilon_{ij}$, dove ϵ_j è il vettore degli errori normalmente distribuiti ricavati da una variabile casuale con la media uguale a 0, in generale, con una varianza che sia proporzionale agli attributi originari

Forse da approfondire

1

- *Swapping*: Consiste nel modificare il subset di una table di microdati facendo lo swapping dei valori di un set formato degli attributi sensibili, chiamato *swapped attributes*, tra coppie di tuple selezionate (le coppie sono selezionate secondo un criterio ben definito). Intuitivamente, questa tecnica riduce il rischio di re-identificazione perché introduce dell'incertezza sul valore reale legato al dato del respondent. Anche se questa tecnica è facile da applicare, generalmente ha lo svantaggio di non preservare le proprietà dei sottodomini. Inizialmente la tecnica originale fu presentata solamente per gli attributi categorici.

¹N.B ϵ viene usato per perturbare i dati ma lasciando intatta la correlazione con la tabella iniziale

- *Micro-Aggregation (o Blurring)*: Consiste nel raggruppare le tuple individuali in piccole aggregazioni di una dimensione fissata k : verrà pubblicata la media di ogni aggregato invece dei valori individuali. Anche se esistono diverse funzioni per calcolare la similarità, può essere difficile trovare una soluzione di raggruppamento ottimale. Ci sono diverse varianti di *micro-aggregation*, per esempio la media può sostituire il valore originale anche solo per una singola tupla nel gruppo o per tutte. Molti attributi possono essere protetti grazie a *micro-aggregation* usando lo stesso o un diverso raggruppamento e, la dimensione del gruppo può essere variabile o fissata ad un minimo.

2.5 Tecniche per la Generazione di Dati Sintetici

La generazione dei dati sintetici è una valida alternativa per la protezione dei microdati. Il principio base su cui queste tecniche sono basate riguarda la non correlazione tra i contenuti statistici dei dati e l'informazione provvista da ogni respondent, quindi un modello ben rappresentante il dato potrebbe sostituire il dato stesso. Un requisito molto importante per la generazione dei dati sintetici, che rende il processo di generazione un problema importante, è che i dati sintetici e il dato originale potrebbero presentare la stessa qualità delle analisi statistiche. Il vantaggio maggiore di questi tipi di tecniche è che i dati sintetici rilasciati non sono collegati a nessun respondent, indi per cui il rilascio non può essere soggetto al fenomeno di re-identificazione; questo, inoltre, lascia ai proprietari dei dati di porre maggiore attenzione sulla qualità del dato rilasciato, invece che al problema di re-identificazione.

2.6 Misure per valutare la confidenzialità e l'utilità dei Microdati

Come discusso precedentemente, ci sono svariate tecniche per la protezione dei microdati. Le performance di queste tecniche di protezione sono valutate in termini di *information loss* e *disclosure risk*.

- Information loss è la quantità di informazioni che esistono nei microdati originali e che, a causa delle tecniche di protezione, questo non si verifica nei microdati protetti.
- Disclosure risk è il rischio di incorrere nel disclosure quando i microdati vengono rilasciati.

Esistono due soluzioni estreme per il rilascio di microdati, le quali sono:

- Crittare i dati originali (rimuove il rischio di disclosure e il massimo di information loss)

- Rilasciare i dati originali (massimo rischio di disclosure e nessun information loss)

In questa sezione andremo a descrivere alcuni dei pmetodi più importanti usati per quantificare il rischio di disclosure e di perdita di informazione (information loss).

2.6.1 Rischio di Disclosure (Disclosure Risk)

Generalmente ci sono due tipi di disclosure: *disclosure di identità ed disclosure di attributi*. Il disclosure riguardo l'identità si occupa di una specifica identità la quale può essere linkata a tuple presenti nelle table di microdati. La disclosure riguardo gli attributi invece l'informazione che può essere spifferata di un attributo legato ad un individuo. In generale, due fattori che potrebbero avere un impatto sul disclosure di identità sono:

- *univocità della popolazione*, ossia la probabilità di identificare un respondent, il quale è unico, con una specifica combinazione di attributi è elevata se quei attributi sono presenti nella table di microdati.
- *re-identificazione*, ossia, che i mcirodati rilasciati sono linkabili ad un'altra table pubblicata, dove gli *identifiers* non sono stati rimossi.

Sono stati proposti diversi metodi per misurare il rischio di disclosure dei microdati rilasciati. Per esempio, *la combinazione minima non sicura degli attributi* ritorna il numero degli attributi con una combinazione unica presenti in una specifica tupla di microdati. Questo metodo può essere adottato solo con le tecniche di mascheramento non perturbative, e, più è alto questo valore tanto più basso sarà il rischio di disclosure.

Univocità

Ogni qual volta un sample unico è anche una popolazione unica, il disclosure d'identità diventa molto più probabile. Esistono metodi differenti per valutare il rischio di univocità, e tutti i metodi si affidano alla valutazione della probabilità. Il primo metodo misura la probabilità dell'univocità della popolazione (*population uniqueness (PU)*), ossia la probabilità che nella popolazione esista solo un individuo con una certa combinazione di valori su un determinato insieme di attributi. Questa probabilità è misurata come:

$Pr(PU) = \sum_j I_j$, con $\frac{F_j-1}{N}$, dove N è la dimensione della popolazione e F_j è il numero degli individui nella popolazione con la j_{esima} combinazione sugli attributi considerati, e $I_j()$ è una funzione e:

$$\begin{cases} I(A) = 1, & \text{se } A = true \\ 0, & \text{altrimenti.} \end{cases}$$

Il secondo metodo misura la probabilità che un sample unico *sample unique (SU)* sia anche una popolazione unica (*PU*). questa probabilità è misurata come:

$PR(PU|SU) = \frac{\sum_j I(f_j=1, F_j=1)}{\sum_j I(f_j=1)}$, dove f_j è il numero degli individui presenti nei sample con la j_{esima} combinazione sugli attributi considerati.

Questi metodi sono chiamati misure *file-level* perché assegnano lo stesso rischio a tutte le tuple. Il rischio di disclosure *tuple-level* è la probabilità di disclosure dell'identità di un individuo specifico.

Questi tipi di misure sono state introdotte a causa della non-omogeneità del rischio di re-identificazione su un'intera tabella di microdati.

Supponendo che ci siano K combinazioni differenti dei valori legati ai quasi-identifiers. Queste combinazioni producono una suddivisione sia nella popolazione che nei sample. Assumendo che F_k sia la frequenza della k_{esima} suddivisione, il rischio di disclosure per una tupla presente in un sample con la k_{esima} combinazione è $\frac{1}{F_k}$.

Il problema di questo metodo è che F_k non è generalmente noto per la popolazione. Poiché le frequenze della distribuzione campionaria f_k sono note, si considera la distribuzione delle frequenze F_k , data f_k ($F_k|f_k$ può essere modellata come una binomiale negativa).

Si noti che l'univocità può essere utilizzata come misura del rischio di divulgazione solo se i microdati sono stati protetti attraverso una tecnica di mascheramento non perturbativa.

Le tecniche perturbative modificano i valori dei dati e quindi non è possibile stabilire correttamente la frequenza di un valore nel campione rilasciato, perché possono essere introdotte nuove combinazioni uniche e possono scomparire le combinazioni uniche originali.

Record Linkage

Consiste nel trovare un match tra le tuple presenti nelle table di microdati protetti e le tuple presenti nelle fonti esterne di informazione pubbliche e non anonime. Poiché non è possibile conoscere a priori tutte le fonti esterne di informazione che possono essere utilizzate da un eventuale utente malintenzionato, viene eseguito un controllo probabilistico sui microdati protetti.

È necessario adottare diversi metodi di record linkage a seconda che la tabella di microdati e le informazioni esterne abbiano o meno attributi comuni. Se ci sono attributi comuni, è necessario innanzitutto adottare una rappresentazione unica per gli attributi comuni. Ad esempio, abbreviazioni diverse nel nome di una persona porterebbero a concludere che due tuple non sono correlate, mentre in realtà si riferiscono allo stesso respondent. È quindi possibile adottare una strategia per il record linkage. I metodi di record linkage possono essere suddivisi in tre grandi categorie:

- *deterministici*: Cerca una corrispondenza esatta su uno o più attributi tra tuple di diversi set di dati. Il principale svantaggio di questo metodo è che non prende in considerazione la rilevanza dell'attributo nel trovare un collegamento.

- *probabilistici*: Dati due dataset, D_1 e D_2 , viene calcolato l'insieme di tutte le possibili coppie di tuple (d_{1i}, d_{2j}) , dove

$$\begin{cases} d_{1i} \in D_1 \\ d_{2j} \in D_2. \end{cases}$$

A ogni coppia è associata una probabilità che rappresenta se la coppia è una corrispondenza reale.

- Se la probabilità è inferiore a una soglia fissa T_1 , la coppia viene scartata perché le tuple sono considerate non linked
 - Se la probabilità è superiore a una seconda soglia fissa T_2 , la coppia è considerata una corrispondenza reale
 - Se la probabilità è compresa tra T_1 e T_2 , è necessaria una valutazione umana per verificare se rappresenta una corrispondenza o meno. Tale probabilità viene calcolata considerando diversi pesi per i vari attributi e l'accordo o l'accordo parziale sui valori degli attributi. I pesi associati agli attributi e le due soglie T_1 e T_2 sono stabiliti dal proprietario dei dati.
- *basati sulla distanza*: Dati due insiemi di dati, D_1 e D_2 , ogni tupla $d_{1i} \in D_1$ viene abbinata alla tupla più vicina $d_{2j} \in D_2$.

Questo metodo richiede la definizione di una funzione di distanza f tra coppie di tuple. Per esempio, la definizione di f può sfruttare funzioni di distanza definite sugli attributi e può assegnare pesi diversi a ciascun attributo, a seconda della sua importanza nel processo di link. Un esempio di funzione di distanza è la *Distanza Euclidea*, che considera ogni tupla come un vettore e assegna lo stesso peso a ogni attributo. Questo metodo di record linkage non è adatto agli attributi categorici, perché è difficile definire la distanza tra due categorie, in particolare se il loro dominio non è ordinato.

Vengono usati anche altri metodi quando ci sono dataset senza attributi comuni. In questi casi la re-identificazione è più difficile. Un metodo proposto di recente si basa sul clustering. Fondamentalmente, un metodo di clustering viene applicato ai set di dati considerati. Il risultato è un insieme di cluster di tuple e ogni cluster all'interno di un set di dati viene mappato su un cluster all'interno dell'altro set di dati. Tale mappatura viene eseguita utilizzando una funzione di similarità. Sebbene il record linkage sia considerato una minaccia, ci sono molte situazioni in cui può essere utile. Il record linkage può essere utilizzato nella gestione di grandi database per estrarre informazioni importanti sullo stesso soggetto. Ciò è particolarmente utile quando i dati sono distribuiti su server diversi (ad esempio, le informazioni mediche della popolazione sono solitamente distribuite su sistemi diversi e una tecnica di record linkage può essere sfruttata per ricostruire le informazioni associate a un determinato individuo).

Interval Disclosure

La misura di disclosure dell'intervallo viene calcolata in modi diversi, a seconda del tipo di dato dell'attributo (continuo o categorico).

Nel caso di attributo categorico, per ogni tupla presente nella table di microdati, gli intervalli classificati (*ranked intervals*) sono costruiti come segue.

Ogni attributo è classificato indipendentemente e un intervallo classificato è definito attorno il valore assunto dall'attributo in ogni tupla t . I ranghi dei valori all'interno dell'intervallo costruito intorno alla tupla t devono differire meno del $p\%$ del numero totale di tuple.

Inoltre, il rango al centro dell'intervallo deve corrispondere al valore assunto dall'attributo considerato nella tupla t . Il rischio di divulgazione è quindi la proporzione dei valori originali che cadono nell'intervallo centrato sul valore protetto corrispondente. Se tale proporzione è pari al 100%, un potenziale aggressore(attackers) è sicuro che il valore originale si trovi nell'intervallo attorno al valore protetto.

Nel caso di dati continui, il metodo è simile al precedente. La differenza principale è il modo in cui vengono costruiti gli intervalli classificati: non è possibile sfruttare il ranking e la costruzione si basa sulla deviazione standard dell'attributo.

2.6.2 Perdita di Informazione (Information Loss)

La misura per la perdita di informazione è strettamente correlata allo *scopo* per cui l'informazione verrà usata. Poiché gli scopi potrebbero essere differenti a non conosciuti a priori, non è possibile stabilire una misura generale per la perdita di informazione basata sul solo scopo. I metodi usati sono quindi, basati sui concetti di *analiticamente validi* e *analiticamente interessanti*, cui vengono definiti come seguono:

- una tabella di microdati protetta è analiticamente valida se conserva approssimativamente le analisi statistiche (ad esempio, media e covarianza) che possono essere prodotte con i microdati originali;
- una tabella di microdati protetta è analiticamente interessante se contiene un numero sufficiente di attributi che possono essere validamente analizzati.

Generalmente, ci sono due strategie per calcolare la perdita di informazione:

- i) confrontando direttamente le tuple dei microdati protetti con le tuple dei microdati originali.
- ii) confrontando le statistiche calcolate sui microdati protetti con le stesse statistiche valutate sui microdati originali.

Descriviamo ora l'idea di base di alcune delle più comuni misure di perdita di informazioni, suddivise in due categorie in base al tipo di dati degli attributi. Sono stati proposti altri metodi, sia per tecniche specifiche di protezione dei

microdati sia per casi generici.

Dati Continui

Per misurare la perdita di informazioni, la statistica di interesse (ad esempio, matrici di co-varianza, matrici di correlazione o loro varianti) viene valutata sia sui dati originali che su quelli protetti e la differenza tra i due valori viene calcolata. Le discrepanze tra le due statistiche possono essere valutate in tre modi diversi:

- *errore quadratico medio*
- *errore medio assoluto*
- *variazione della media*

Oltre alle misure statistiche, i dati possono essere confrontati prima e dopo l'applicazione di una tecnica di protezione dei microdati, calcolando nuovamente la differenza con uno dei tre metodi sopracitati. È importante notare che il valore della perdita di informazioni deve avere un valore massimo (ad esempio, 100 se si utilizza una notazione percentuale) per confrontare metodi diversi che hanno la stessa scala di calcolo della perdita di informazioni.

Dati Categorici

Le misure di perdita di informazione introdotte brevemente per gli attributi continui non sono direttamente applicabili agli attributi categorici. In questo caso, le misure principali sono tre:

- *confronto diretto*
- *confronto tra tabelle di contingenza*
- *misura di entropia*

Il confronto diretto dei valori degli attributi categoriali richiede la definizione di una funzione di distanza tra le categorie. Nel caso di categorie non ordinate, la distanza tra la categoria c_1 nei microdati originali e la corrispondente categoria c_2 nei microdati protetti è:

$$\begin{cases} \text{uguale a } 0, & \text{se le due categorie sono le stesse} \\ 1, & \text{altrimenti} \end{cases}$$

Al contrario, se esiste un ordinamento tra le categorie, la distanza tra le categorie c_1 e c_2 è pari al numero di categorie tra c_1 e c_2 diviso per il numero totale di categorie. La misura di confronto delle tabelle di contingenza consiste nel confrontare le tabelle di contingenza corrispondenti. Una misura basata sull'entropia può essere utilizzata quando una tabella di microdati è stata protetta applicando le tecniche di soppressione locale, global recoding o PRAM. L'idea è che la perdita di informazione possa essere misurata utilizzando l'*entropia di Shannon*, poiché il processo di mascheramento viene modellato come il rumore aggiunto ai microdati originali quando vengono trasmessi attraverso un

canale rumoroso. La misura della perdita di informazione utilizza la probabilità condizionale (la probabilità di un valore nei microdati originali, una volta che il valore nei microdati protetti è dato).

2.6.3 Combinazione di Rischio di Disclosure e Perdita di Informazione

Le tecniche di protezione dei microdati descritte in questo capitolo hanno un impatto diverso sull'utilità dei dati e sul rischio di divulgazione. Per poter valutare tecniche alternative di protezione dei microdati, abbiamo bisogno di un quadro di riferimento per valutare la bontà di una tecnica di protezione.

Il rischio di divulgazione e la perdita di informazioni devono quindi essere combinati. Un metodo semplice consiste nel calcolare la media dei due valori e scegliere la tecnica (e l'impostazione dei parametri) con il punteggio più alto. Un altro metodo è quello delle *mappe di confidenzialità R-U*, che è un grafico in cui la misura dell'utilità dei dati (l'inverso della perdita di informazioni) è riportata sull'asse x e il rischio di divulgazione è riportato sull'asse y . Per ogni tecnica di protezione dei microdati, viene tracciata una linea sul piano cartesiano con un punto per ogni parametro impostato. Sulla base del grafico ottenuto, è possibile confrontare le varie tecniche di protezione e scegliere la più adatta. Una volta scelta la tecnica di protezione, le mappe di riservatezza R-U possono essere utilizzate anche per la selezione dei parametri. È importante notare che una mappa R-U è solo un metodo per correlare il rischio di divulgazione e la perdita di informazioni; tali misure devono essere calcolate utilizzando uno dei metodi sopra menzionati.

Un altro approccio per bilanciare l'utilità dei dati e il rischio di divulgazione è rappresentato dal concetto di tabella k -minima con k -anonymity (guarda il capitolo " k -anonymity"). Il k -anonymity stabilisce una soglia limite inferiore di rischio di divulgazione per una tabella, assicurando che ogni tupla della tabella non possa essere correlata a meno di k rispondenti. L'approccio alla k -anonimità mira a trovare (applicando tecniche di generalizzazione e soppressione) una tabella k -minima, cioè una tabella che non generalizza più di quanto sia necessario per raggiungere la soglia k . In altre parole, una tabella k -minima è una tabella che minimizza la perdita di informazioni.

Le misure descritte dovrebbero essere utilizzate prima che i dati vengano rilasciati per verificare se la protezione è adeguata alle "richieste di riservatezza dei respondent e alle esigenze di informazione dei destinatari dei dati".

Dopo l'applicazione delle tecniche di protezione, i microdati protetti possono essere controllati e rilasciati solo se sono conformi ad un certo grado di protezione. Queste misure possono essere usate anche dal destinatario dei dati per valutare la "protezione dei respondent e l'utilità dei dati".

2.7 Conclusioni

Oggi le società globali interconnesse hanno una grande richiesta sulla disseminazione e la condivisione di informazioni. Mentre nel passato le informazioni rilasciate erano principalmente in forma tabulare e statistica, molte situazioni oggi chiamano per il rilascio dei microdati specifici. Per affrontare questo problema sono state proposte numerose tecniche di protezione e, in questo capitolo, sono state descritte le tecniche di base per la protezione contro il disclosure dei microdati, classificando queste in tecniche di mascheramento e tecniche di generazione di dati sintentici. Per riassumere, le tecniche di generazione di dati sintentici proteggono i dati trasformando i loro valori. D'altro canto le tecniche di generazione di dati sintentici proteggono i dati sostituendoli con nuovi dati i quali conservano le proprietà statistiche originali. Inoltre sono state illustrate anche le principali misure adottate, solitamente, per valutare la riservatezza dei dati e l'utilità dei microdati protetti.

Garantire la sicurezza significa garantire Confidenzialità, Integrità e Accesso (=Disponibilità) ai dati. Nello scenario cloud tre sono i momenti chiave:

- memorizzazione dei dati
- gestione dei dati
- processazione dei dati

2.8 Introduzione

Usare il cloud è conveniente per benefici come ad esempio la scalabilità e l'elasticità, però comporta anche dei risvolti problematici perchè il proprietario dei dati non ne ha più il pieno controllo, con conseguente minaccia per la sicurezza che può minare la fiducia nell'adozione del cloud computing. Il NIST distingue 4 modelli di sviluppo del Cloud:

1. *privato*: mantenuto su una rete privata
2. *pubblico*: una organizzazione offre servizi cloud a terzi
3. *community*: infrastruttura condivisa da diverse compagnie ma con interessi comuni
4. *ibrido*: un cloud composto da una combinazione di cloud che siano 1,2 oppure 3

Sempre lo stesso NIST identifica 3 modelli di servizio:

1. *IaaS*: **Infrastructure** as a Service
2. *PaaS*: **Platform** as a Service
3. *SaaS*: **Software** as a Service

2.9 CIA nel Cloud

I problemi di sicurezza possono essere classificati con il paradigma CIA. **C** è garantire riservatezza delle informazioni salvate esternamente, **I** è garantire l'autenticità dei dati, **A** richiede che il provider soddisfi dei livelli di servizio attesi.

In uno scenario complesso si richiede l'esecuzione di query sui dati e comunque operazioni in cui entra in gioco anche un rapporto di fiducia con il cloud provider (o fornitore) che può essere **completamente affidabile** (assunto in caso di cloud privati), **curioso** (archiviazione ed elaborazione coinvolgono informazioni sensibili), **pigro** (ovvero potenzialmente non affidabile nel caso in cui debba gestire info sensibili), **dannoso** o bizantino (fornitore si comporta in modo improprio nella gestione, archiviazione ed elaborazione dei dati compromettendo CIA)

2.10 Problemi e sfide

Non vi è una ricetta come soluzione per preservare CIA, però nonostante i differenti aspetti si possono considerare diversi scenari:

2.10.1 Protezione dei dati al rilascio

Un problema dei dati rilasciati nel cloud è garantire la loro protezione (ovvero rispettare la CIA); solitamente gli utenti devono fidarsi completamente dei cloud provider. Questi garantiscono riservatezza da esterni, ma non vi è nessuno che garantisca che il proprietario del cloud possa accedere ai miei dati personali salvati.

- Una possibile soluzione potrebbe essere allora la **cifratura dei dati prima** di rilasciarli ai cloud provider. La cifratura quindi garantisce C e I. Si preferisce la critto simmetrica. Ma usare la critto presenta problemi quando si vuole fare un *recupero fine dei dati*.
- Si adotta allora la frammentazione, al posto della critto, che protegge da associazioni di dati:
 - dividendo le parti di informazioni
 - salvandole in frammenti separati non collegabili
- Si può optare anche, nella versione *two can keep a secret*, per fare sì che i frammenti separati siano proprio due provider non comunicanti tra loro, ciascuno dei quali detiene memorizzata una porzione di dati.
- Talvolta vengono cifrati solamente gli attributi sensibili oppure non si cifra nulla se vi è fiducia nei confronti del provider.

A è invece garantita dalla replicazione dei dati su più cloud provider.

2.10.2 Accesso a grana fine ai dati nel cloud

Mantenere i dati cifrati nel cloud impedisce la decifrazione da parte dei proprietari per eseguire query. Per realizzare questo approccio è necessario però operare query su dati cifrati, come ad esempio

- la **crittografia omomorfica**: questa però ha svantaggi come ad es:
 - realizzabilità solo su operazioni basilari
 - elevato costo computazionale
- **CryptDB**: supporta maggiori operazioni e migliora l'efficienza rispetto alla omomorfica. In questo modello ogni relazione è cifrata a livello di colonna con diversi livelli di cifratura, ognuno dei quali supporta l'esecuzione di una specifica operazione SQL. Funzionamento: quando server riceve la query, determina il livello di profondità di cifratura e il proxy invia

al provider la chiave di quel livello per rimuovere i livelli soprastanti di cifratura.

- Un altro approccio consiste nell'**allegare indici** ai dati cifrati per recuperare informazioni a grana fine ed eseguire query. L'indicizzazione può essere *diretta* (1:1) oppure *hash* (N:1) oppure *piatta* (N:1). Queste tre tecniche forniscono diverse garanzie di protezione. Quella hash e piatta garantiscono migliore riservatezza.

2.10.3 Accesso selettivo ai dati nel cloud

Idea è che differenti utenti possano avere differenti viste dei dati. Si parla allora della creazione di vere e proprie ACL, spesso però sono delegate al provider (che medierà ogni richiesta di accesso) e questo richiede fiducia da parte del cliente. Se non vi è totale fiducia si può combinare ACL con crittografia con chiavi diverse, a seconda delle autorizzazioni che si detengono all'atto della richiesta di accesso. Una possibile soluzione al problema del cambio di autorizzazioni nelle ACL è dato dalla *sovracrittografia*.

Inoltre, per alleggerire il carico computazionale, si può operare crittografia basata su attributi (ABE), critto a chiave pubblica che regola l'accesso ai dati in base agli attributi descrittivi associati ai dati stessi.

2.10.4 Privacy dell'utente

Idea è di consentire l'accesso ai dati a utenti non registrati nel sistema, cioè senza che dichiarino la propria identità. L'accesso deve essere quindi basato sulle proprietà degli utenti e non sull'identità. Tale accesso può essere basato su:

- attributi
- credenziali
- certificati

2.10.5 Privacy delle query

In alcuni scenari non sono privati solo i dati o gli utenti, ma anche gli accessi che questi ultimi fanno per accedere ai dati. Una soluzione è Private Information Retrieval, che però ha un alto costo computazionale. Idea è prelevare l'*i*-esimo bit di una stringa senza rivelare al server quale bit si sia prelevato. Una soluzione meno onerosa è **RAM Oblivious**: *ORAM* e *PathORAM*, basati su una struttura dati *gerarchica* e *dinamica* basata su chiave (detta *indice Shuffle*) modifica dinamicamente (shuffling) ad ogni accesso la posizione fisica dei dati, distruggendo la corrispondenza statica dati-blocchi fisici di memorizzazione. Viene creato un B+ albero contiene nodi con blocchi effettivi e blocchi fittizi per creare confusione al provider affinché non rintracci il blocco mirato.

L'indice di Shuffle predispone n ricerche false per non fare capire quale elemento si sia cercato veramente.

2.10.6 Integrità della computazione e delle query

In caso di mancata fiducia con il provider, il cliente deve poter verificare che il risultato di una query sia *corretto* (= risultato calcolato sui dati originari), *completo* (= non mancano dati dal risultato) e *fresco* (= query è stata eseguita sulla versione più recente dei dati).

Esistono soluzioni:

- **deterministiche:** strutture dati autentiche come ad es. schemi di concatenamento delle firme (consentono la verifica di ordinamento tra le tuple → verifica dell'integrità delle query) e Merkle hash tree (organizza i dati in una struttura ad albero basandosi su un attributo con cui si verifica il risultato della query).
- **probabilistiche:** come ad esempio l'inserimento di *tuple false* nei dati, *replica* di dati, *pre-calcolo dei token* associati ai risultati di una query. La probabilità di trovare una compromissione dell'integrità dipende dalla quantità di controlli applicati. Una possibile soluzione per la valutazione dell'integrità dei join calcolati da un provider non attendibile è usare server di archiviazione per inserire informazioni di controllo

2.10.7 Esecuzione collaborativa delle query con provider multipli

Scenario in cui ci sono più provider e anche più proprietari dei dati. Importante è garantire che le informazioni non vengano rese accessibili, rilasciate o trapelate: occorre un rilascio selettivo dei dati. [...]

2.10.8 SLA e Auditing

SLA è un accordo contrattuale che specifica le performance e disponibilità che un provider deve mantenere. In passato SLA era sinonimo di disponibilità, tempo di risposta,...) oggi indica anche criteri di sicurezza (crittografia, protezione perimetrale)

2.10.9 Multi-locazione e virtualizzazione

Multi-locazione significa capacità di fornire servizi informatici a diversi utenti usando infrastruttura cloud comune. Ogni utente dell'infrastruttura cloud condivide risorse di calcolo, come memoria e archiviazione... attività che vengono ben coordinate grazie alla virtualizzazione, che offre grande flessibilità ma introduce problemi di sicurezza come ad esempio **colpire** l'hypervisor (= sw che crea e coordina le macchine virtuali), riguardare l'**allocazione/deallocazione**

delle risorse. Il rischio è quello di perdere informazioni in modo improprio se le risorse virtuali non sono isolate nel loro assegnamento al singolo utente

2.11 Conclusioni

La rapida crescita di piattaforme cloud sollecita la considerazione di problemi e preoccupazioni riguardanti la sicurezza in questo scenario, soprattutto per quanto riguarda CIA.

Chapter 3

Supporting User Privacy Preferences in Digital Interactions

3.1 Introduzione

Per la disponibilità dei servizi online si verifica la situazione in cui il server che fornisce il servizio e l'utente che lo richiede sono ignoti l'uno all'altro. Di conseguenza sistemi di access control tradizionali non possono essere adottati in quanto non adatti a scenari open.

Soluzioni proposte sono basate su ABAC (Attribute-Based Access Control). Policy che regolamentano l'accesso ai servizi definiscono condizioni che il client richiedente deve soddisfare per avere accesso.

Il server, a seguito di una richiesta di accesso risponde con le condizioni che il client deve soddisfare per essere autorizzato ad accedere al servizio.

Per accedere il client rilascia **certificati digitali** come **credenziali** firmate da una **CA**. L'adozione delle credenziali nell'Access Control ha diversi vantaggi:

- Non serve $\langle username, password \rangle$ per ogni servizio.
- Offre migliore protezione dall'acquisizione impropria dei privilegi di accesso.

L'uso di credenziali è stata approfonditamente studiata dal lato server con definizione di diversi *policy languages*, *policy engines* e strategie per la comunicazione delle consizioni di accesso.

Le soluzioni trovate, sebbene potenti ed espressive, non supportano completamente gli specifici criteri di protezione dei client. Infatti i client necessitano di un sistema di preferenze che permetta di supportare una propria definizione di

sensitivity/privacy per i propri dati. Queste preferenze verranno poi usate nella situazione in cui più di un subset di credenziali soddisfano la Access Control Policy definita dal server.

Questo capitolo mostra una overview delle problematiche di privacy legate ai sistemi open, oltre che alcune soluzioni. Le sezioni riportano:

- Concetti Base e *Desiderata*.
- Soluzione Cost-Sensitive Trust Negotiation.
- V
- c
- d

3.2 Concetti Base e *Desiderata*

3.2.1 Client Portfolio

Definizione 10 (*Client Portfolio*)

*Insieme delle informazioni che un cliente può fornire al server per avere accesso ad un servizio è racchiuso in un **Portfolio** contenente*

- *Credenziali* firmate da *third party*.
- *Dichiarazioni* che comprendono proprietà non certificate dichiarate dall'utente.

Struttura Credenziale

Ogni credenziale c è composta da:

- Identificatore univoco $id(c)$.
- Emittente $issuer(c)$.
- Un set di attributi $attributes(c)$.
- La tipologia della credenziale $type(c)$.

Gerarchia Tipologie

I tipi di credenziale sono organizzati tipicamente in una gerarchia radicata, dove i nodi intermedi sono astrazione delle specifiche credenziali presenti sulle foglie. Un esempio è in fig. 3.1.

La gerarchia è nota sia al client che al server. Il server tuttavia crea richieste di credential types in quanto non può conoscere le effettive credenziali possedute dal client¹.

¹Inoltre client può avere più credenziali dello stesso tipo

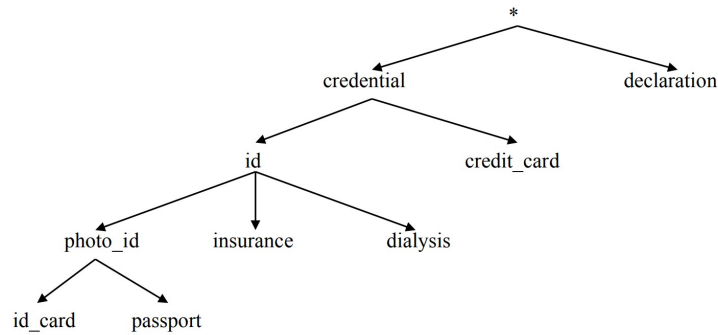


Figure 3.1

3.2.2 Atomicità Credenziali

Le credenziali si distinguono in **atomiche** e **non-atomiche**.

Le credenziali **atomiche** sono la tipologia più comune (vedi X.509) e possono essere rilasciato solo interamente. Per questo anche attributi non richiesti vengono rivelati al server.

Le credenziali **non-atomiche** permettono di limitare il rilascio di attributi rivelando selettivamente un subset degli attributi certificati dalla credenziale.

Ovviamente le *declaration* sono credenziali non-atomiche.

Attributi

Attributi nelle credenziali sono caratterizzati da:

- Tipo
- Nome
- Valore

Essi possono essere *credential-independent* e dipendere dal client ma non dalla credenziale, o *credential-dependent* e dipendere da client e credenziale che la certifica.

3.2.3 Policy di Disclosure

Le policy lato server che regolano l'accesso sono definite sugli attributi e le credenziali fornite dall'utente, questo perchè il server considera solo la gerarchia di credenziali (conosciuta a priori anche dal server).

Una Access Control Policy (policy di controllo di accesso) è quindi una espressione booleana composta da condizioni $term_1 op term_2$, dove op è un operatore del tipo $<, >, =$ etc.

Un esempio di policy: $(type(c) = insurance) \wedge (c.Company \neq 'A')$

3.2.4 Trust Negotiation

L'interazione tra client e server mira a stabilire una relazione di fiducia (trust) tra le parti per permettere l'accesso al servizio. Questa interazione si basa tipicamente sul rilascio di credenziali, spesso subordinato al rilascio di altre credenziali da parte della controparte.

Questa sequenza di scambi di credenziali è definita **strategia** e soddisfa le policy di server e client.

Alcune strategie adottate permettono il rilascio di credenziali solo se esiste una *strategia di successo* (successful strategy) che garantisce l'accesso al servizio, mentre altre rilasciano credenziali appena sono richieste.

Il fatto che esistano preferenze sul rilascio di credenziali, sia da parte del server che del client, comporta che le successful strategies non siano tutte equivalenti.

3.2.5 Client Privacy Preferences

Per quanto riportato, è necessario fornire i client di un modello flessibile ed espressivo per rappresentare le preferenze di privacy. Questo modello dovrebbe avere le seguenti caratteristiche:

- **Specific di Preferenze a Grana Fine:** il modello deve supportare la definizione di preferenze per ogni attributo per ogni **istanza** di credenziale.
- **Ereditarietà delle Preferenze di Privacy:** se le istanze di credenziali sono numerose, il modello deve poter applicare sensibilità ai tipi astratti e così a tutte le istanze per cui non era stato specificato un livello di sensibilità.
- **Rel. d'Ordine Parziale e Operatore Composizione:** ci deve essere ordinamento parziale per determinare se una informazione è più o meno sensibile di un'altra. Deve esistere un operatore di composizione \oplus che permette di calcolare il valore di sensibilità di un set di attributi/credenziali.
- **Associazioni Sensibili:** rilascio combinato di un set di attr./cred. risulta più o meno sensibile della somma dei rilasci singoli. Il modello deve comprendere anche questa definizione di preferenza.
- **Vincoli di Rilascio:** modello deve permettere di definire vincoli sul rilascio combinato di parti del portfolio, per evitare o limitare il rilascio dell'associazione tra questi.

- **Context-Based Preferences:** preferenze dipendono dal contesto (rilascio certificato medico a una farmacia, non a un hotel).
- **History-based Preferences:** rilascio dipende da interazioni passate (se a server X ho rilasciato un set, uso sempre quello e non fornisco nuove cred.).
- **Prova di Possesso:** nuove tecnologie permettono di rilasciare prove di possesso di certificati e di soddisfazione requisiti di accesso. Risulta necessario poter specificare preferenze anche per queste prove (proof).
- **User-Friendly Preference Specification:** necessario fornire interfacce ai client, che possono non essere abituati a sistemi di controllo dell'accesso, per poter specificare facilmente le preferenze.

3.2.6 Preferenze Privacy del Server

Mentre la comunicazione della policy server completa favorisce la privacy del client, la comunicazione dei soli attributi coinvolti favorisce la privacy del server.

Ricordare che anche server ha un portfolio di credenziali e attributi oltre alle policy.

Il sistema lato server deve considerare:

- **Disclosure Policy:** server deve poter decidere, a grana fine, come il rilascio policy dovrebbe essere regolato.
- **Comunicazione Policy:** necessario un meccanismo che trasforma la A.C.Policy in modo da proteggere privacy della policy e permettere al client di sapere cosa è richiesto.
- **Integrazione con Meccanismo Client:** Ci deve essere integrazione tra i sistemi lato server e lato client.

3.3 Cost-Sensitive Trust Negotiation

Sensibilità espressa con costo $w(c)$ associata ad ogni credenziale c nel portfolio e con ogni policy di access control p . p viene definita come espressione booleana sulle credenziali della controparte, mentre i valori booleani delle credenziali c sono espressi come *vero* se già rivelati, *false* altrimenti.

Il costo di sensibilità, definito da utente, modella quanto l'owner ritiene sensibile il rilascio della credenziale e delle informazioni sensibili che certifica.

In fig. 3.2 un esempio di portfolio per client e server.

L'obiettivo per client e server è quindi quello di minimizzare il costo di credenziali e policy scambiate **in una strategia di negoziazione che ha successo.**

$id(c)$	$w(c)$	Policy p regulating c	$w(p)$
MyIdCard	2	TRUE	0
MyPassport	4	TRUE	0
MyCreditCard	10	POS_register	5
MyDialysis	20	pharmacy_register	10
MyInsurance	15	pharmacy_register	10

Table 2 An example of client portfolio and policies regulating its disclosure

$id(c)$	$w(c)$	Policy p regulating c	$w(p)$
MyPOSRegister	2	TRUE	0
MyPharmacyRegister	5	passport v id_card	4

Table 3 An example of server portfolio and policies regulating its disclosure

Figure 3.2: Due esempi di portfolio con policy di rilascio

3.3.1 Minimizzare il Costo

Il problema della minimizzazione del costo viene formulata come segue:

Definizione 11 (Problema del Minimo Costo di Sensibilità)

Dati:

- C_s set di credenziali e servizi server;
- P_s set di policy server;
- C_c set di credenziali client;
- P_c set policy client;
- $w : C_s \cup P_s \cup C_c \cup P_c \rightarrow \mathbb{R}$ funzione costo;
- $s \in C_s$ il servizio richiesto.

Per minimizzare il costo in una successful strategy occorre trovare una sequenza di scambio tale che:

1. s è rilasciato al client;
2. ogni policy è soddisfatta prima del rilascio della corrispettiva credenziale;
3. la somma dei costi è **minima**.

Il problema è NP-difficile e con costo esponenziale nel numero di input (credenziali e policy). Tuttavia due soluzioni euristiche (non ottimali) ottengono una buona soluzione con costo polinomiale. Queste considerano o meno il costo delle policy e sono brevemente² descritte qui di seguito:

²La completa descrizione degli algoritmi penso esuli dallo scope di questo riassunto

Senza Costo delle Policy

Le policy non sono associate ad un costo, quindi possono essere pubblicate liberamente.

Viene definito il DAG **policy graph** $G(V, A, w)$ che modella le policy di rilascio delle credenziali sia lato server che lato client.

Un esempio del grafo è riportato in fig.3.3 che segue:

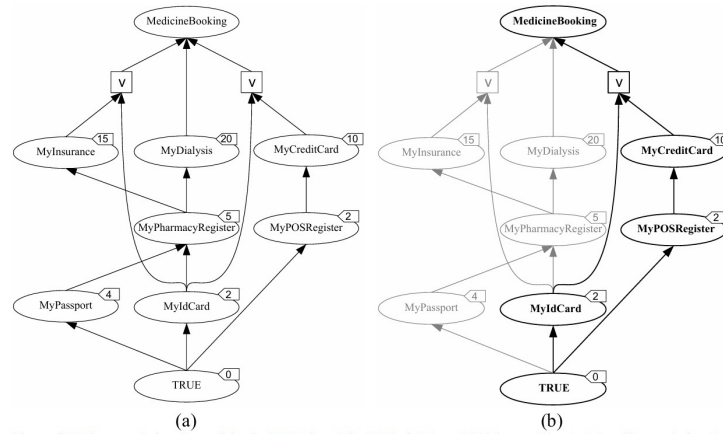


Figure 2 Policy graph for the policies in Table 2 and in Table 3 (a), and Minimum Directed Acyclic graph for the MedicineBooking service (b)

Figure 3.3: Policy graph e suo DAG minimo.

Il primo step riguarda sempre la disclosure delle policy, e ciò permette di costruire il grafo.

FFormalmente il problema si riduce quindi al calcolo del DAG minimo.