Tecniche di Anonimizzazione e Pseudonimizzazione

Anna Monreale

Ogni persona giornalmente lascia miriadi di tracce digitali svolgendo semplicemente delle attività comuni come parlare al telefono, fare la spesa al supermercato, scrivere un post su Facebook, usare il navigatore della propria macchina o del proprio smartphone. Tutti questi dati sono preziosi per sviluppare servizi sia personalizzati che per la collettività. Ad esempio, avendo a disposizione le tracce GPS, che descrivono i movimenti in una specifica città, si potrebbero sviluppare delle analisi che porterebbero alla determinazione di un profilo della mobilità urbana, utile all'amministrazione comunale per identificare particolari criticità da risolvere con un intervento mirato. Un esempio di uso dei dati per fornire servizi personalizzati, è quello di una banca che, avendo accesso a informazioni di diversa natura sul suo cliente (attività sui social, attività sul web, acquisti nei vari negozi, ecc.), potrebbe offrire servizi finanziari più adeguati alle abitudini del proprio cliente.

Per poter sfruttare i dati personali allo scopo di sviluppare servizi per il singolo o per la collettività sono necessari strumenti di *Big Data Analytics & Social Mining*, che permettono di capire, misurare e possibilmente predire il comportamento umano. Il problema principale è che tutto questo ha delle conseguenze etico-legali. In particolare, durante l'uso di questi strumenti non si può ignorare il diritto alla protezione dei dati personali. La figura professionale del DPO può supportare lo sviluppo di questi servizi ma per farlo deve poter innanzitutto conoscere la differenza tra *dato anonimo* e *dato pseudonimo*.

I **dati anonimi** sono privati di tutti gli elementi identificativi della persona. Essi non sono soggetti alle norme sulla privacy poiché non sono considerati dati personali. I **dati pseudonimi** sono dati in cui gli elementi identificativi sono stati sostituiti da elementi diversi, in modo tale che la re-identificazione della persona a cui il dato si riferisce non sia possibile, se non con informazione aggiuntiva (esterna). I dati pseudonimi, a differenza di quelli anonimizzati, sono comunque dati personali.

Tecniche di Pseudonimizzazione

Una tecnica di pseudonimizzazione ha lo scopo di sostituire un dato identificativo (es. nomi, codice fiscale, ecc.) con **un valore surrogato** che spesso è chiamato **token,** il quale deve essere irreversibile senza informazione aggiuntiva e distinguibile dal valore originale**.** Nella Figura 1 mostriamo il risultato di un processo di pseudonimizzazione che ha sostituito nella prima tabella il nome con un token (codice).

Nome	Sesso	Anno Nascita	CAP	DIAGNOSI
Verdi	F	1962	300122	Cancro
Rossi	F	1961	300133	Gastrite
Gialli	M	1950	300111	Infarto
Neri	M	1954	300112	Emicrania
Bianchi	F	1965	300200	Lussazione
Rosa	M	1953	300115	Frattura

Token	Sesso	Anno Nascita	CAP	DIAGNOSI
11779	F	1962	300122	Cancro
12121	F	1961	300133	Gastrite
21177	M	1950	300111	Infarto
41898	M	1954	300112	Emicrania
56789	F	1965	300200	Lussazione
65656	M	1953	300115	Frattura

Figura 1: Processo di pseudonimizzazione

L'obiettivo della pseudonimizzazione è quello di ridurre il rischio di rendere pubblici dei dati che permettono la re-identificazione diretta, cioè possibile senza informazione aggiuntiva. Per offrire questa garanzia, un processo di pseudonimizzazione deve mantenere la corrispondenza tra il valore originale e il relativo token in una locazione differente dalla tabella pseudonima. Quindi, nell'esempio in Figura 1 la corrispondenza tra la colonna "nome" della prima tabella e la colonna "token" della seconda tabella deve

essere mantenuta ad esempio in un computer diverso da quello in cui viene memorizzata la seconda tabella.

Esistono diverse tecniche di pseudonimizzazione che possiamo dividere in:

- tecniche di crittografia basate su chiave segreta: queste tecniche criptano i dati identificativi usando una chiave segreta. La decifratura può essere fatta solo usando questa chiave che è conosciuta solo dal controllore dei dati.
- tecniche basate su funzioni hash: queste tecniche usano una funzione che, dato un identificatore (composto da uno o più attributi), restituisce un valore di dimensione fissa. La funzione non deve essere invertibile.
- tecniche basate su funzioni di tipo keyed-hash: queste tecniche sono molto simili alle precedenti, solo che la funzione hash per calcolare il valore surrogato a partire da un identificatore ha bisogno anche di una chiave segreta, rendendo in questo modo il lavoro di un possibile attaccante più difficile.
- tecniche basate su funzioni di tipo keyed-hash con cancellazione della chiave: queste tecniche sono uguali alle precedenti, solo che dopo la generazione del valore surrogato, che sostituirà l'identificatore originale, la tabella di corrispondenza tra il valore originale e quello generato verrà cancellata.
- tecniche di tokenizzazione: queste tecniche sostituiscono l'identificatore con un token generato con un meccanismo di crittografia, con una funzione che genera un numero sequenziale o con un numero casuale.

Una volta che i valori surrogati sono stati generati devono essere mantenuti e gestiti. A questo fine, un'azienda o un'istituzione può decidere di gestire sia i valori surrogati che la base di dati internamente oppure può rivolgersi a una terza parte che offre il servizio di generazione dei valori surrogati.

Tecniche di Anonimizzazione

Rendere il dato pseudonimo non è sufficiente a garantire la protezione dei dati personali in presenza di informazione aggiuntiva, che può essere usata per la re-identificazione di un individuo in un dato reso pubblico. A questo fine è necessario trattare i dati originali con tecniche di anonimizzazione che proteggono contro attacchi basati sul collegamento dei dati da protegge con informazione esterna. Un attacco famoso di questo tipo è quello che ha permesso la re-identificazione del governatore del Massachussetts in una tabella contente informazioni mediche su alcuni pazienti. In un esperimento sono state collegate le informazioni presenti in due tabelle (Figura 2): una contenente le informazioni mediche sui pazienti dell'ospedale del Massachussetts, dove il governatore era stato ricoverato; e una contenente i dati sugli elettori del Massachussetts. Le due tabelle avevano in comune alcuni attributi come il sesso, la data di nascita e il CAP. Mettendo assieme le informazioni delle due tabelle è stato scoperto che 6 persone avevano lo stesso anno di nascita del governatore, di cui 3 erano uomini e 1 sola aveva il suo stesso CAP. Di conseguenza, conoscendo l'informazione su questi 3 attributi del governatore, è stato possibile venire a conoscenza della sua malattia. Ad esempio, supponendo che le due tabelle siano quelle raffigurate in Figura 2, nel caso in cui l'anno di nascita del governatore fosse 1950 e il suo CAP fosse uguale a 300111, allora il collegamento sulla base degli attributi "anno di nascita", "sesso" e "CAP" porterebbe all'inferenza che il governatore ha avuto un infarto.

ID	Sesso	Anno Nascita	CAP	DIAGNOSI
1	F	1962	300122	Cancro
2	F	1961	300133	Gastrite
3	M	1950	300111	Infarto
4	M	1954	300112	Emicrania
5	F	1965	300200	Lussazione
6	M	1953	300115	Frattura

Nome	Sesso	Anno Nascita	CAP	Indirizzo	Data Voto
Verdi	F	1962	300122	Via Lumi	1/1998
Rossi	F	1961	300133	Via Ali	1/1998
Gialli	M	1950	300111	Via Alba	3/1998
Neri	M	1954	300112	Via Alto	1/1998
Bianchi	F	1965	300200	Via Deli	3/1998
Rosa	M	1953	300115	Via Blu	3/1998

Figura 2: La prima tabella contiene dati medici mentre la seconda i dati dei votanti

Le tecniche di anonimizzazione che contrastano il rischio di un attacco del genere possono essere raggruppate in tecniche di anonimizzazione basate su generalizzazione e soppressione di dati, tecniche basate su randomizzazione dei dati, tecniche per sistemi distribuiti e tecniche per l'outsourcing di dati.

Le tecniche basate su generalizzazione e soppressione di dati prevedono una suddivisione degli attributi di una tabella in: identificatori, quasi-identificatori e attributi sensibili. Gli identificatori sono attributi che identificano univocamente le persone all'interno della tabella come ad esempio il codice fiscale. I quasi-identificatori sono attributi che, per natura non dovrebbero permettere una re-identificazione univoca (es. il sesso), ma combinando diversi quasi-identificatori questi possono identificare univocamente una persona all'interno della tabella. Questo è il caso dell'esempio del governatore che è stato identificato combinando sesso, anno di nascita e CAP. Infine, gli attributi sensibili contengono le informazioni da proteggere come la diagnosi di una malattia, il credo religioso, il pensiero politico, ecc.

Tra le tecniche appartenenti a questa categoria abbiamo quelle basate sul modello della k-anonymity, la quale richiede che ogni record della tabella anonimizzata sia indistinguibile, rispetto ai quasi-identificatori, da almeno altri k-1 record della stessa tabella. Una tabella può essere resa k-anonima usando un approccio basato sulla generalizzazione dei valori dei quasi-identificatori, oppure un approccio basato sulla soppressione di alcuni record. Comunque, tra i due approcci è preferibile sempre la generalizzazione perché evita la perdita completa delle informazioni su un individuo. In Figura 3 mostriamo l'effetto dell'applicazione di un processo di generalizzazione al fine di ottenere una tabella k-anonima. A partire dalla tabella originale, abbiamo generalizzato il valore dell'anno di nascita, sostituendo l'anno con un intervallo che include il valore dell'anno originale, e il CAP mantenendo solo le prime cifre del valore originale. In questo modo siamo riusciti a rendere la tabella originale 3-anonima. In altre parole, presa qualsiasi combinazione di valori per anno di nascita, sesso e CAP ci sono sempre almeno 3 persone con quelle caratteristiche. Da tutto ciò ne consegue che la probabilità di poter re-identificare uno dei pazienti e di inferire la sua malattia è al più 1/3.

ID	Sesso	Anno Nascita	CAP	DIAGNOSI
1	F	1962	300122	Cancro
2	F	1961	300133	Gastrite
3	M	1950	300111	Infarto
4	M	1954	300112	Emicrania
5	F	1965	300200	Lussazione
6	M	1953	300115	Frattura

ID	Sesso	Anno Nascita	CAP	DIAGNOSI
1	F	[1961-1965]	300***	Cancro
2	F	[1961-1965]	300***	Gastrite
3	M	[1951-1955]	30011*	Infarto
4	M	[1951-1955]	30011*	Emicrania
5	F	[1961-1965]	300***	Lussazione
6	M	[1951-1955]	30011*	Frattura

Figura 3: Processo di k-anonimizzazione

Il modello della k-anonymity ha delle debolezze come quella dovuta alla possibile omogeneità del valore dell'attributo sensibile in un gruppo. Supponiamo che con una tecnica di k-anonimizzazione troviamo un gruppo di record che hanno lo stesso valore per quanto riguarda gli attributi quasi-identificatori (sesso, anno di nascita e CAP nel nostro esempio). Cosa succede se tutti i record di questo gruppo hanno la stessa

malattia? Questo potrebbe essere un problema, poiché un possibile attaccante in questo caso non riesce a re-identificare il record dell'obiettivo (il governatore nel nostro esempio), ma sicuramente può inferire senza problemi la malattia del governatore dato che questa è uguale per tutti gli appartenenti al gruppo. Negli ultimi anni sono stati sviluppate diverse varianti della k-anonymity che hanno l'obiettivo di risolvere debolezze come quelle appena descritta.

Altre tecniche che permettono di garantire la privacy sono basate sulla **randomizzazione di dati.** Il metodo può essere descritto nel seguente modo: Data una tabella con attributi numerici come "salario" ed "età" si aggiunge ad ogni valore dell'attributo una quantità di rumore, estratto da una certa distribuzione di probabilità. Quindi, per ogni attributo il nuovo valore sarà dato dalla somma del valore originale e il valore di rumore aggiunto. Per esempio, se un record ha un valore di età uguale a 25 anni, dopo la randomizzazione potrebbe avere il valore di 30 anni dovuto al fatto che a 25 è stato sommata una quantità di rumore pari a 5. Negli approcci basati su randomizzazione si assume che la varianza del rumore aggiunto sia grande abbastanza da non permettere il recupero dei valori originali a partire da quello randomizzato. Tra le tecniche basate sulla randomizzazione dei dati merita una menzione la Differential Privacy che ha come obiettivo perturbare il risultato di un'interrogazione analitica sottoposta a una base di dati. L'idea della Differential Privacy è non permettere ad un attaccante l'inferenza di informazione sensibile su un individuo presente nella base di dati tramite interrogazioni successive. A tal fine, al risultato dell'interrogazione viene aggiunta una quantità di rumore in modo da rendere il risultato simile a quello ottenuto nel caso in cui l'interrogazione fosse sottoposta a una base di dati che non contiene le informazione sulla persona sotto attacco.

Alcune tecniche di anonimizzazione permettono di garantire la **privacy in sistemi distribuiti** in cui è necessario analizzare ed estrarre conoscenza a partire da basi di dati distribuite in locazioni differenti. In questo scenario difficilmente le varie parti hanno la possibilità o vogliono condividere i propri dati ai fini delle analisi. Quindi si ha la necessità di sviluppare dei processi analitici capaci allo stesso tempo di garantire la privacy e di permettere l'estrazione di conoscenza utile allo sviluppo di applicazioni specifiche. L'approccio tipico usato in questo contesto si basa sull'applicazione di tecniche di crittografia e in particolare tecniche basate sulla *secure multiparty computation.* I metodi di secure multiparty computation permettono di calcolare delle funzioni su dati forniti dalle varie parti senza condividere tali dati. Esistono metodi che permettono il calcolo sicuro della somma, dell'unione tra insiemi, della dimensione dell'intersezione tra insiemi e il prodotto scalare. Tutte queste funzioni possono essere usate come primitive di algoritmi utili per analizzare ed estrarre conoscenza intrinseca nei dati. Alcuni di questi metodi considerano che le basi di dati sono *distribuite verticalmente*, ovvero i record relativi a un insieme di individui sono distribuiti su siti diversi, in modo che in ogni locazione si possono trovare attribuiti che descrivono aspetti diversi dello stesso individuo. Ad esempio, un sito contiene i dati medici di alcuni pazienti, mentre un altro sito contiene le loro informazioni bancarie. Altri metodi invece sono adatti ad agire in un contesto in cui le basi di dati sono *distribuite orizzontalmente*, ovvero diversi siti hanno lo stesso tipo di informazioni ma su diversi individui. Ad esempio, in due filiali della banca abbiamo due basi di dati con la stessa struttura ma riguardanti clientele diverse.

Recentemente è stato evidenziato un grande interesse per il paradigma di analisi di dati basato sul cloud computing. L'idea è permettere a organizzazioni, con un potere computazionale limitato, di dare in outsourcing il task di analisi di dati da eseguire. L'outsourcing richiede però il trasferimento di dati che spesso contengono informazioni sensibili da proteggere. Le tecniche di anonimizzazione adeguate a questo scenario hanno un duplice obiettivo: garantire la protezione dei dati trasferiti alla terza parte e permettere l'analisi dei dati, garantendo che anche la conoscenza estratta dai dati anonimi non permetta inferenze sensibili. Inoltre, la conoscenza estratta dai dati può avere dei vantaggi competitivi quindi è ancor più

necessario che questa sia incomprensibile anche per il server cloud, in modo da garantire all'organizzazione la possibilità di usare i servizi offerti dal cloud senza perdere potere competitivo. Per ottenere questi obiettivi, gli approcci appartenenti a questa categoria prevedono che l'organizzazione prima del trasferimento dei dati debba applicare una trasformazione che aggiunge del rumore ai dati stessi e debba mantenere un'informazione compatta (che occupi poco spazio) sul rumore aggiunto. Questa informazione dovrà essere utilizzata per poter pulire i risultati dell'analisi che riceverà successivamente dal cloud. Infatti, il cloud applicherà l'analisi dei dati richiesta sui dati affetti di rumore, e di conseguenza anche i risultati saranno contaminati da questa distorsione. L'organizzazione quando riceverà il risultato dell'analisi potrà eliminare il rumore grazie all'informazione conservata all'atto del trasferimento iniziale dei dati.

Anonimizzazione su dati complessi

Il problema della privacy inizialmente è stato studiato per dati memorizzati in formato tabellare, come quelli presenti nell'esempio in Figura 1. L'avvento di tecnologie sempre più avanzate e sofisticate ha permesso di collezionare sugli individui anche forme di dati più complesse come liste di acquisti, attività nei social media come Facebook, ricerche sui motori di ricerca e la mobilità urbana.

Lo sviluppo di tecniche di anonimizzazione per queste forme di dati ha richiesto studi specifici. Le domande a cui rispondere in questo contesto sono: accedere alla lista dei prodotti acquistati da una persona oppure accedere agli spostamenti di una persona può violare la sua privacy? La risposta è certamente sì. Supponiamo di avere una base di dati che memorizza gli spostamenti di una persona, ovvero per ogni individuo memorizza la sequenza delle locazioni visitate. Un attaccante accedendo a questa base di dati potrebbe per ogni individuo estrarre le locazioni che quella persona visita più frequentemente nell'arco di un mese. Analizzando il tempo trascorso in quelle locazioni, potrebbe capire ad esempio che la locazione più frequente corrisponde alla sua abitazione, poiché tipicamente la persona sta lì durante la notte per tante ore. Mentre la seconda locazione più frequente è il suo posto di lavoro, poiché questa persona trascorre lì diverse ore nella fascia oraria 9-17. Andando a controllare chi abita nella prima locazione e chi lavora nella seconda, si potrebbe arrivare a identificare la persona a cui gli spostamenti fanno riferimento.

Per evitare problemi di questo genere è necessario rendere anonime anche queste forme di dati che apparentemente non sembrano nocivi per la privacy delle persone. Questi dati sono molto utili anche dal punto di vista analitico, infatti potrebbero essere utili per capire la mobilità urbana e identificare interventi specifici che poterebbero a un miglioramento in termini ad esempio di intensità del traffico. Di conseguenza è necessario che le tecniche di anonimizzazione garantiscano un buon bilanciamento tra il livello di privacy e la qualità dei dati ottenuta dopo l'anonimizzazione, in modo da permettere l'uso dei dati per analisi ancora significative.

Al fine di ottenere questi obiettivi è necessario definire un processo di anonimizzazione seguendo una metodologia orientata alla destinazione d'uso del dato. L'idea è quella di definire un processo di trasformazione dei dati che consideri quali rischi di privacy dobbiamo contrastare e quale tipo di analisi dobbiamo applicare sui dati dopo l'anonimizzazione. Conoscendo tutto questo, è possibile capire quali proprietà dei dati è necessario mantenere validi affinché i risultati di un'analisi non siano troppo affetti dalla trasformazione del processo di anonimizzazione. Un esempio di applicazione di questa metodologia ha permesso l'anonimizzazione di dati di mobilità garantendo, non solo un buon livello di protezione, ma anche la possibilità di usare i dati per un'analisi della mobilità tramite algoritmi di mining, che estraggono gruppi di movimenti frequenti in particolari zone della città. La tecnica di anonimizzazione in questione è basata su una forma di generalizzazione spaziale. L'idea è che invece di utilizzare le *locazioni* visitate dai singoli individui (informazione molto dettagliata), consideriamo le *aree* visitate dalle persone (informazione

più grossolana) in modo da avere in un dato istante delle aree popolose. In questo modo gli individui localizzati all'interno della stessa area sono considerati indistinguibili. Di conseguenza la tecnica generalizza di più, e quindi crea aree più grandi, in corrispondenza di zone poco popolose, e aree più piccole in corrispondenza di zone molto popolose. Dopo questa fase di generalizzazione, il metodo garantisce la proprietà di k-anonymity trasformando il dato in modo da avere per ogni sequenza di aree almeno k persone che seguono un percorso che coinvolge quelle aree.

Una metodologia per la gestione del rischio di privacy

L'applicazione delle tecniche di anonimizzazione richiede la valutazione del rischio di privacy per ogni individuo a cui i dati da anonimizzare fanno riferimento. In particolare, è necessario misurare il rischio di privacy, identificare la porzione di dati non anonima (con un alto rischio di re-identificazione) e applicare una tecnica di anonimizzazione che permette di mitigare i rischi. A tal fine una metodologia che, per la gestione del rischi di privacy si basa su questo approccio, è composta da due componenti principali: *Valutazione del rischio* e *Mitigazione del rischio.*

Il primo componente per la valutazione del rischio permette di simulare i possibili attacchi, che si possono condurre sui dati da anonimizzare, e per ogni attacco valuta il rischio di privacy. Questo componente, a partire dal dato originale da anonimizzare, deriva diverse aggregazioni del dato che potrebbero avere un impatto positivo sul rischio di privacy.

La valutazione del rischio di privacy permette di assegnare ad ogni individuo la sua probabilità di poter essere re-identificato. Di conseguenza, è possibile individuare gli individui ad alto rischio su cui la componente per la mitigazione del rischio di privacy può applicare una delle tecniche di anonimizzazione. L'idea fondamentale è di agire in modo mirato al fine di risolvere i problemi riscontrati durante la valutazione.