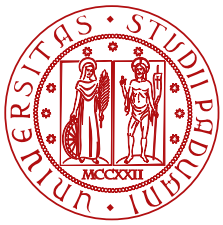


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA BIOMEDICA

Tecniche e Algoritmi per l'Anonimizzazione dei Dati Biomedici Sensibili

Relatore
Enrico Longato

Laureando
Andrea Pluchino

ANNO ACCADEMICO 2023-2024

Data di laurea 19/11/2024

Sommario

La pubblicazione delle informazioni personali delle persone, come ad esempio i dati sanitari, risulta essere significativamente vantaggiosa per istituzioni ospedaliere e organizzazioni governative di vario genere, supportando la ricerca scientifica specialmente nel contesto medico. Queste informazioni sensibili sono tuttora gestite all'interno di cartelle cliniche elettroniche (Electronic Health Record, EHR), le quali sono continuamente soggette ad opere di condivisioni e raccolta, essenziali per il beneficio individuale e collettivo. Tuttavia, i dati sanitari contengono numerosi dettagli sensibili dei pazienti, la cui pubblicazione potrebbe essere causa di violazioni non intenzionali della privacy. In questo elaborato, vengono presentate le tecniche di Privacy Preserving Data Publishing (PPDP) più conosciute ed efficaci per la creazione di un ambiente sicuro per la condivisione dei dati, inquadrando, al contempo, lo scenario su cui esse operano. La General Data Protection Regulation (GDPR) e l'Health Insurance Portability and Accountability Act (HIPAA) sono due tra le normative che maggiormente aiutano nel trattamento di questi dati sensibili. Dopo l'introduzione delle basi legislative che regolano questo ambito, si esplora sia il concetto di de-identificazione dei dati attraverso due approcci distinti, ovvero la pseudonimizzazione e l'anonimizzazione, sia un metodo più avanzato e innovativo, come la differential privacy. Ci si concentra, in particolare, sulle tecniche generali di anonimizzazione dei dati, ponendo l'accento su un confronto tra gli algoritmi conosciuti per questo scopo: k-anonymous, l-diversity e t-closeness. Il lavoro si conclude con alcune implementazioni pratiche degli algoritmi capaci di sottolineare punti di forza e limitazioni di quanto discusso, specialmente in relazione ai rischi di divulgazione della privacy che maggiormente si corrono in seguito alla pubblicazione dei dati anonimizzati.

Indice

1 Introduzione

L'espansione continua delle tecnologie digitali ha trasformato profondamente ogni settore, orientando la direzione del loro asse sul piano informatico, con il quale si è instaurata una forte connessione: su di esso, si riversa incessantemente un flusso di informazioni e dati importanti, che hanno richiesto numerose considerazioni riguardanti la loro gestione ed elaborazione. Uno dei campi maggiormente influenzati da questo cambiamento è il settore sanitario, dove la delicatezza dei dati trattati rende cruciale affrontare le problematiche legate alla loro protezione. Molte delle informazioni personali e sensibili delle persone, sono state trasportate e racchiuse in archivi in formato digitale, come le cartelle cliniche elettroniche (Electronic Health Record, EHR), e la loro condivisione e raccolta è fondamentale sia per il beneficio del singolo (come per la cura individuale), sia per il beneficio collettivo, poiché la ricerca medica permette di condurre studi che portano a nuove scoperte o approcci terapeutici.

Tuttavia, queste azioni potrebbero rischiare di violare i diritti fondamentali degli individui, sollevando interrogativi su come poter trattare questi dati senza ledere la libertà personale e i diritti alla privacy di ciascuno. Per garantirne la sicurezza, è diventato indispensabile adottare nuove misure di protezione avanzate dei dati capaci di difendere le informazioni dai nuovi tipi di attacchi informatici a cui potrebbero andare incontro. In tutti i casi in cui la raccolta dei dati o la loro condivisione è essenziale, si è sviluppata la necessità di tecniche di Privacy Preserving Data Publishing (PPDP), che mirano a proteggere la privacy individuale consentendo una condivisione sicura dei dati. Questo obiettivo richiede la creazione di sistemi efficaci e standardizzati, come algoritmi e schemi di de-identificazione, che permettano l'utilizzo dei dati per fini scientifici senza compromettere la riservatezza.

L'anonimizzazione dei dati, pertanto, non è solo una necessità legale in termine di tutela dei diritti individuali, ma anche una tecnica da studiare e affinare, che può essere implementata in più forme e tipologie, ciascuna più o meno efficace a seconda dei contesti, sebbene nessuna possa essere considerata una soluzione perfetta per tutti i casi. Nonostante privare i dati delle informazioni identificative degli individui a cui si riferiscono sia una soluzione fondamentale per la protezione della privacy, nei capitoli successivi di questa tesi verrà mostrato come il processo di anonimizzazione sia molto più complesso di quanto possa sembrare. Gli aggressori informatici, infatti, possono sfruttare numerose strategie e informazioni ausiliarie che, pur non permettendo di risalire direttamente all'identità degli individui, se intrecciate e combinate correttamente possono rivelare dettagli sensibili sui dati anonimizzati, dimostrando che la semplice anonimizzazione non è sempre sufficiente a garantire una protezione totale.

1.1 Scopo del lavoro

Lo scopo di questo elaborato è quello di presentare una panoramica sui meccanismi per la protezione dei dati sensibili che nel corso degli ultimi anni si sono maggiormente affermati. Nel Capitolo 2, dopo aver presentato le basi legislative su cui si inserisce l'anonimizzazione dei dati, si procede ad illustrare il concetto di de-identificazione, essenziale per comprendere e distinguere due approcci differenti: la pseudonimizzazione e l'anonimizzazione. Nello stesso capitolo viene argomentato un ulteriore approccio più attuale, che, al contrario delle due tecniche precedenti, mira ad ottenere risultati rilevanti dai dati senza comprometterne la struttura interna. Alla fine del capitolo, vengono, invece, evidenziati i rischi di re-identificazione dei dati che, anche nel caso di anonimizzazioni accurate, non possono essere esclusi.
Il Capitolo 3 descrive alcune delle tecniche generali più conosciute e documentate in letteratura, insieme ai tre algoritmi più utilizzati: k-anonymous, l-diversity e t-closeness, capaci di operare anonimizzazioni efficaci dei dati preservando l'utilità delle informazioni. L'elaborato si conclude con alcuni esempi di implementazioni degli algoritmi, consentendo di effettuare un confronto diretto tra le diverse soluzioni discusse in precedenza.

2 Strategie per la Protezione dei Dati

La protezione della privacy e dei dati personali si presenta come una delle maggiori sfide nella società digitale contemporanea. A causa del progressivo aumento della raccolta e condivisione di informazioni, è fondamentale tutelare quei dati che, se divulgati, potrebbero mettere a rischio la dignità o la sicurezza dell'individuo. I cosiddetti "dati sensibili" sono il fulcro di questa protezione, poiché riguardano aspetti profondamente personali. L'origine razziale o etnica di una persona, le sue convinzioni religiose e i dati relativi alla salute, sono alcuni tra i dati sensibili che la commissione europea ha definito e il cui diritto alla protezione è sancito già dalla Carta dei Diritti Fondamentali dell'Unione Europea [1].
Data la cruciale importanza di questi dati, si è reso necessario regolamentare e controllare il loro flusso attraverso normative specifiche, in grado di definire i margini del loro trattamento.

2.1 Aspetti normativi

Questa sezione esamina le principali normative che regolano la gestione e la protezione dei dati sensibili, che costituiscono sia l'origine dello sviluppo di strategie efficienti per la salvaguardia di queste informazioni, sia la base per una corretta comprensione del loro funzionamento, evidenziando l'impatto che la legislazione ha avuto sulle tecniche che verranno discusse nelle sezioni seguenti.

2.1.1 General Data Protection Regulation (GDPR)

L'entrata in vigore del General Data Protection Regulation (GDPR) il 24 maggio 2016, e la sua applicazione dal 25 maggio 2018, hanno costituito un passo essenziale nel rafforzamento e nella tutela della privacy, in un'era digitale in cui la raccolta, l'elaborazione e la condivisione dei dati personali sono diventate pratiche pervasive e potenzialmente rischiose per i diritti fondamentali degli individui.

L'introduzione del GDPR ha sostituito la precedente Data Protection Directive (DPD) introdotta nel 1995. L'uso crescente di dati personali nel settore pubblico e privato, facilitato dall'avvento delle tecnologie dell'informazione e della comunicazione (Information and Communication Technology, ICT), ha notevolmente aumentato il rischio di abuso dei dati sensibili, promuovendo il bisogno di creare sistemi di regolazione di questi dati che è culminato dapprima con l'entrata in vigore della DPD [2]. La presente direttiva è stata in grado di unificare alcuni principi chiave e regolamentare la protezione dei dati; tuttavia, ha lasciato

molta flessibilità agli stati membri, rendendo disomogenea l'applicazione delle norme. Non era inoltre sufficiente a proteggere correttamente i dati più vulnerabili, come quelli sensibili e personali, e inadeguata di fronte all'evoluzione rapida delle tecnologie digitali.
Per tali ragioni, l'adozione di una normativa più articolata, quale il GDPR, si è rivelata fondamentale per garantire una protezione più efficace di questi dati, rispondendo alle sfide poste dall'imminente sviluppo tecnologico. L'attuazione di quest'ultima normativa, non solo ha stabilito i criteri per la corretta regolamentazione dei dati, ma ha anche disposto le conseguenze in caso di violazione delle norme indicate, stravolgendo l'approccio alla gestione dei dati da parte delle organizzazioni.
Questo nuovo quadro normativo ha quindi dato avvio a una fase di controllo più stringente, in cui vengono emesse sanzioni significative verso chi non rispetta le disposizioni previste dal regolamento. Tale è il caso avvenuto nell'agosto del 2019, in cui il consiglio scolastico delle scuole superiori del comune di Skellefteå in Svezia è stato multato di 18.630 euro per aver violato diversi articoli del GDPR [3]. L'istituto aveva collaborato con una società privata al fine di sviluppare un sistema di riconoscimento facciale, e sebbene abbia ricevuto il consenso sia dagli studenti che dai tutor, data la sensibilità della tecnologia, era richiesta la consultazione dell'autorità per la protezione dei dati.
Un ulteriore caso è avvenuto il 23 ottobre 2019 in Austria, dove l'Österreichische Post (ÖPAG) è stata multata di 18 milioni di euro per aver elaborato illegalmente i dati di 2.2 milioni di persone al fine di calcolarne l'affinità politica e venderne le informazioni senza consenso. [4]

Per quanto la tutela dei dati personali degli individui sia di vitale importanza, rimangono indiscutibili i benefici che essi sono capaci di apportare nell'ambito della ricerca, della statistica e del progresso scientifico. L'introduzione del GDPR ha dunque spinto istituzioni e organizzazioni, sia pubbliche che private, a sviluppare meccanismi e algoritmi sempre più efficaci per il trattamento dei dati, legando al contempo la necessità del mantenimento della privacy con la loro utilità. Evolve quindi il concetto di de-identificazione, ponendosi come una vera e propria tecnica da studiare e affinare, e di cui verrà esposta e presentata un'analisi più dettagliata nella sezione seguente, insieme alle tecniche più utilizzate e conosciute per questo scopo.

2.1.2 Health Insurance Portability and Accountability Act (HIPAA)

L'elaborazione e archiviazione dei dati assume particolare rilevanza nel contesto medico, non solo a causa della mole di dati che giornalmente vengono generati durante le pratiche cliniche, ma anche a causa della natura sensibile di questi ultimi. Una gestione appropriata è fondamentale sia per garantire la privacy degli individui, sia per agevolarne l'accesso quando

richiesto, creando un sistema sicuro e immediato. Uno dei metodi sviluppati per questi scopi è l'Health Insurance Portability and Accountability Act (HIPAA), che è stato in grado di aiutare considerevolmente nella ricerca sanitaria. Sviluppato nel 1996, è un metodo molto sicuro per la gestione dei dati dei pazienti e della privacy. Essendo progettato specificamente per il settore sanitario negli Stati Uniti, fornisce regole dettagliate e specifiche per la protezione dei dati dei pazienti, specialmente per quanto riguarda l'adozione delle EHR [5]. Nonostante sia una normativa extraeuropea, L'HIPAA viene spesso presa come punto di riferimento anche da organizzazioni europee in contesti sanitari specifici che richiedono collaborazioni con enti statunitensi o adozioni di pratiche standardizzate a livello globale. Questa normativa assume quindi un'importanza maggiore per le organizzazioni sanitarie rispetto al GDPR, che è invece una legislazione più generale che si applica a tutti i settori in Europa.

2.2 Principi di de-identificazione

La delicatezza dei dati sensibili ha da sempre richiesto una protezione adeguata degli stessi, ma l'introduzione di ulteriori normative rispecchia l'importanza che questi assumono per la ricerca scientifica, soprattutto in ambito medico. Disporre di una mole vasta ed eterogenea di informazioni, dopo una sua analisi accurata, può supportare la previsione di esiti clinici dei pazienti, stimare la risposta a specifici trattamenti nonché individuare pattern e tendenze comuni, altrimenti non evidenziabili su scala ridotta.
Per esempio, i ricercatori infermieristici in oncologia utilizzano attualmente i *big data* per numerosi scopi utili [6]. I *big data* si riferiscono a un insieme di dati generati giornalmente, con caratteristiche di volume, velocità e varietà tali da rendere complessa la loro elaborazione, condivisione e il loro immagazzinamento, richiedendo tecniche e strumenti specializzati per una gestione precisa. Grazie ad essi, i ricercatori sono stati in grado di prevedere gli esiti dei pazienti a partire dalle note cliniche, identificare fenotipi distinti di sintomi e individuare i predittori della tossicità da chemioterapia.
Sebbene sia vero che i progressi nei metodi computazionali sono tali da offrire nuove e interessanti opportunità per avanzare nella scienza infermieristica oncologica attraverso l'uso dei *big data*, vi sono ancora diverse sfide legate all'accesso e al loro utilizzo. La sicurezza dei dati e la privacy dei partecipanti alla ricerca sono tuttora preoccupazioni rilevanti. Conseguentemente, l'adozione di sistemi di sicurezza dei dati sempre più avanzati è di interesse notevole per chi ha la necessità di usufruirne, costituendo la chiave per un progresso scientifico più veloce e consistente.

Come si è già anticipato, le normative in merito non negano la possibilità di utilizzare

questi dati, ma introducono la necessità di garantire che il loro uso non metta a repentaglio l'identità o i diritti fondamentali degli individui a cui appartengono le informazioni. Ciò è particolarmente rilevante nell'ambito medico, sia a causa della natura sensibile della maggior parte dei dati che vengono costantemente immagazzinati e utilizzati ogni giorno, sia per i numerosi vantaggi che, come abbiamo appena visto, possono derivarne.
La soluzione al problema, nonché strada da percorrere da parte di qualsiasi istituzione che voglia sfruttarli, converge nel processo di de-identificazione, che consiste essenzialmente nella rimozione della componente identificativa dei dati di cui si dispone. A seconda della metodologia utilizzata e del grado di de-identificazione applicato a un determinato dataset, il rischio di re-identificazione varia. In genere, questo rischio diminuisce con l'aumentare della rigorosità della de-identificazione. Quando il rischio di re-identificazione è assente, il dataset è considerato completamente anonimizzato. In questo modo, qualunque dato viene privato del "possessore" e, non essendo più riconducibile a nessuna persona, può essere interamente sfruttato per elaborazioni, analisi e addirittura condiviso con enti terzi, senza richiesta di consenso da parte di coloro di cui si trattano le informazioni.
"De-identificazione" e "anonimizzazione" sono spesso erroneamente interscambiate e, a causa dei molteplici significati, in quest'elaborato si farà riferimento alla definizione riportata in [7], secondo cui de-identificazione indica "l'insieme di tecniche che modificano i record individuali nel dataset originale per ottenere un dataset presumibilmente anonimo". Essendo un concetto più generale, l'introduzione di tecniche di de-identificazione ha lo scopo di rimuovere le componenti identificative dai dati, senza garantire l'effettiva irreversibilità di questo processo. L'anonimizzazione, al contrario, mira a rendere le informazioni definitivamente non riconducibili a individui specifici, anche se in combinazione con altre fonti di informazione. Una spiegazione più chiara del meccanismo di anonimizzazione verrà presentata nella sezione 2.2.2, che permetterà di evidenziare anche le complicazioni che emergono a seguito della sua attuazione.

Nel documento [8], Latanya Sweeney dimostra che l'87% della popolazione dell'U.S. è facilmente identificabile utilizzando solamente 3 dati personali: codice zip a 5 cifre, sesso e data di nascita. Risulta chiaro che per identificare correttamente una persona bastano pochi dati personali, e che la semplice rimozione di nome, cognome o indirizzo, non sia sufficiente a prevenire il riconoscimento di ciascun individuo da tutelare.
Apportare un efficace de-identificazione dei dati, riducendo il più possibile il rischio di re-identificazione da parte di terze parti può essere molto più difficile di quanto si possa pensare. Ciò è in buona parte connesso al bisogno di mantenere coeso il binomio privacy-utilità, che si riflette nell'esigenza di un'adeguata comprensione dei dati e della quantità di informazioni

personali che ciascun indicazione è in grado di trasmettere.
Presentare una classificazione degli attributi dei dati che si possono incontrare durante lo studio di eventuali dataset è dunque indispensabile per comprendere correttamente come procedere nel processo di de-identificazione, distinguendo e differenziando i vari parametri. Questo permette di valutare la loro utilità in relazione allo studio che si vuole condurre e al contempo la loro "sensibilità", ovvero quanto siano capaci di lasciar trasparire informazioni circa l'identità della persona a cui fanno riferimento.

Classificazione e struttura dei dati

Secondo quanto riportato nello studio [9], l'articolo 9 del GDPR definisce i dati personali come qualsiasi informazione diretta o indiretta relativa a una persona fisica identificata o identificabile. Inoltre, lo stesso studio distingue cinque categorie di dati, basandosi su [10]:

- Dati relazionali
- Dati transazionali
- Dati sequenziali
- Dati di traiettoria
- Dati grafici

I dati relazionali sono il tipo di dati più comune, specialmente in ambito sanitario, come ad esempio dati clinici in un registro di malattie o di popolazione. Di solito sono costituiti da un numero fisso di colonne e righe, che corrispondono rispettivamente agli attributi e ai record. Con questi ultimi si intende una singola unità di dati contenuta in un dataset e, nel contesto medico, si riferiscono ai dati di un singolo paziente che appare al più una volta per dataset. I dati transazionali hanno invece un numero variabile di colonne per ogni record, ad esempio pazienti diversi possono avere un numero diverso di transazioni. I dati sequenziali sono simili ai dati transazionali, ma vi è un ordine negli elementi in ogni record. I dati di traiettoria combinano dati sequenziali con informazioni di localizzazione. Ad esempio, dati sul movimento dei pazienti avrebbero informazioni di localizzazione sequenziate temporalmente. Infine, i dati grafici, comunemente usati nei social media, incapsulano le relazioni tra oggetti utilizzando tecniche della teoria dei grafi[1].

[1] La teoria dei grafi è una branca della matematica che studia le proprietà e le strutture dei grafi, costituiti da nodi e collegamenti, utilizzata per risolvere problemi di rete e molte altre applicazioni.

Come già anticipato, tipicamente, i dati sanitari sono relazionali e in forma tabulare. La gestione degli attributi nei processi di de-identificazione è fortemente correlata alla categoria a cui ciascun attributo appartiene. Continuando a seguire lo schema logico dello studio [9], posta una tabella, i suoi attributi sono racchiusi generalmente in 4 gruppi:

- Identificatori diretti **I**
- Identificatori indiretti o quasi identificatori **Q**
- Attributi sensibili **S**
- Altri attributi non sensibili **O**

Gli identificatori diretti **I**, come suggerisce il nome, possono essere direttamente usati per ricondursi all'identità del paziente, in quanto forniscono collegamenti specifici con i soggetti a cui si riferiscono. Un singolo identificatore diretto **I** è quindi sufficiente a risalire all'identità della persona. Esempi tipici del primo caso includono numeri di identificazione dei pazienti, numeri di telefono o indirizzo email. Supposto che in uno stesso centro di cura ci siano pazienti con uno stesso nome e cognome, in questo caso l'attributo "nome e cognome", non può essere considerato identificatore diretto. Tuttavia, una combinazione di nome completo e indirizzo di residenza, in quanto unico, è in grado di costituire un identificatore diretto.
Gli identificatori indiretti **Q**, anche denominati comunemente quasi identificatori, sono attributi che, se analizzati disponendo di conoscenze pregresse sul dataset di cui fanno parte, con alta probabilità, possono portare alla re-identificazione della persona. Si tenga a mente che un attributo non può essere considerato un quasi-identificatore **Q** se l'avversario che mira alla re-identificazione non ha a disposizione queste conoscenze. La distinzione tra i **Q** e gli **I** tiene conto dell'utilità analitica dell'attributo. Nello specifico, i **Q** sono utili per l'analisi dei dati, mentre gli **I** non lo sono. Alcuni esempi di **Q** includono data di nascita, genere o origine etnica. Gli attributi sensibili **S** contengono informazioni sanitarie sensibili sui pazienti, come ad esempio una particolare terapia; tuttavia non sono utili per determinare l'identità del paziente. Gli altri attributi **O** rappresentano invece il resto delle variabili che non sono considerate sensibili e sarebbero difficili da utilizzare per un avversario nella re-identificazione. Per un'interpretazione più chiare di quanto detto, si rimanda alla Tabella 2.1.
È importante specificare che nel contesto dell'anonimizzazione dei dati, con il termine avversario o aggressore, si fa riferimento ad un soggetto, umano o automatizzato, che tenta di re-identificare le informazioni de-identificate all'interno di un dataset anonimizzato, con l'obiettivo di risalire alle identità degli individui o anche solo di accedere ad alcuni dei loro dati sensibili. L'avversario sfrutta generalmente informazioni ausiliarie, come dati disponibili pubblicamente o conoscenze pregresse, per instaurare collegamenti tra i record anonimizzati e le persone corrispondenti.

I	Q	Q	Q	S	S	O
Nome	**Età**	**Genere**	**Cap**	**Stipendio**	**Malattia**	**Codice malattia**
Anna Verdi	46	F	96617	2k	Diabete	E11
Mario Rossi	33	M	96613	4k	Diabete	E10
Sara Gialli	41	F	96584	4k	Asma	J45
Carla Moretti	33	F	96586	3k	Anoressia	F50
Luca Bianchi	38	M	96491	2k	Asma	J45
Roberto Blu	52	M	96584	3k	Diabete	E11
Marta Neri	57	F	96492	2k	Cancro al seno	C50

Tabella 2.1: esempio di dati medici contenenti identificatori diretti **I**, quasi identificatori **Q**, attributi sensibili **S** e altri attributi non sensibili **O**. Le colonne rappresentano gli attributi, ovvero i diversi tipi di informazioni contenuti nel dataset e in rosso si evidenziano le categorie a cui ciascuno di questi attributi appartiene. Le righe sono i record di ciascun individuo presente nel dataset.

La distinzione più difficile da effettuare è sicuramente quella tra indicatori diretti e indiretti. A tal proposito, si adottano tre regole di determinazione utili a questo scopo e il loro flusso logico è schematizzato dalla Figura 2.1:

1. Un attributo può essere **I** o **Q** qualora risulti noto da un avversario come conoscenza pregressa;
2. Un attributo deve essere trattato come **Q** se è utile per l'analisi dei dati e come **I** altrimenti;
3. Un attributo dovrebbe essere specificato come **I** se può identificare un individuo in modo univoco.

Comprendere in maniera appropriata le informazioni che si posseggono e riuscire a classificare dati e attributi, è la base per selezionare la strategia più adatta. Qualsiasi tecnica presenta delle limitazioni, ma per essere valida, deve riuscire a garantire un alto grado di protezione dei dati minimizzando il rischio di re-identificazione, senza però snaturare del tutto l'utilità delle informazioni, altrimenti il beneficio che se ne può trarre è irrilevante. Riuscire a coniugare queste due esigenze è la ragione per cui esistono approcci differenti, ciascuno dei quali offre vantaggi e svantaggi, la cui scelta dipende dai risultati che da questi dati si vogliono ottenere.

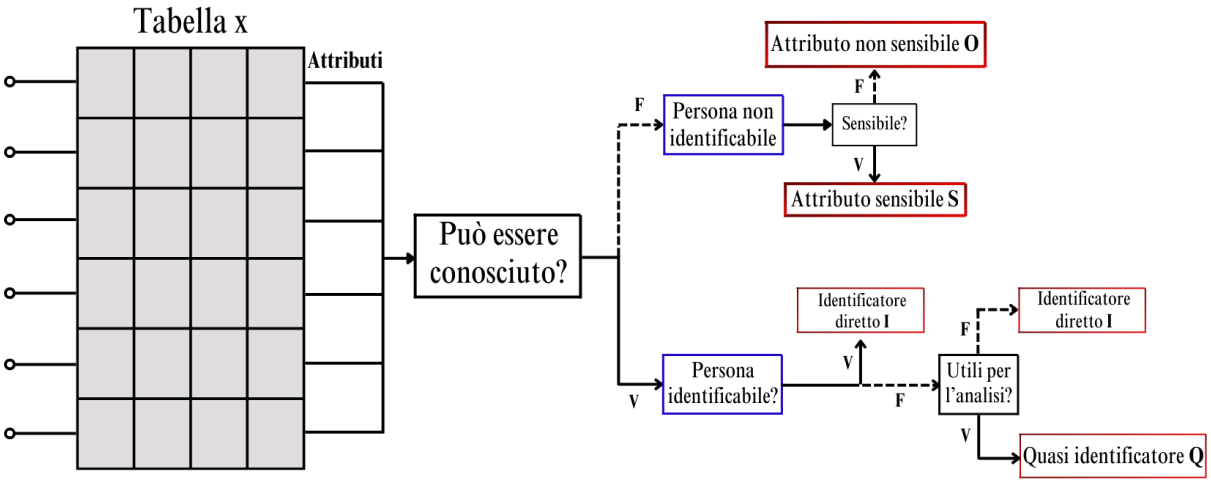


Figura 2.1: schema logico per classificare gli attributi di una tabella, ciascun attributo deve seguire la struttura logica raffigurata.

Si esploreranno più nel dettaglio due metodologie di protezione della privacy che includono il concetto di de-identificazione vera e propria: pseudonimizzazione e anonimizzazione. Nella sezione successiva si proporrà invece un approccio differente, la differential privacy, che non modifica strettamente i dati, bensì le informazioni che si possono ricavare da essi.
Tutte le tecniche esaminate sono però capaci di offrire prospettive diverse e chiarire la varietà di approcci e opzioni disponibili, specialmente in relazione al contesto sanitario.

2.2.1 Pseudonimizzazione

Presentare il concetto di pseudonimizzazione è importante al fine di interpretare correttamente l'evoluzione a cui le strategie di de-identificazione sono andate incontro a seguito dell'applicazione delle nuove normative. L'introduzione del GDPR ha infatti aperto nuove prospettive e messo in discussione meccanismi già consolidati, presentando le nuove strade che organizzazioni e istituzioni dovevano percorrere.
Riprendendo quanto detto in [7], la pseudonimizzazione è una tecnica che consiste nella rimozione di attributi identificatori diretti **I** da un dato, rimpiazzandoli con uno pseudonimo. Nome, email, numero di telefono o codice fiscale, sono esempi di questi identificatori diretti, che da soli o combinati, permettono di risalire facilmente all'identità dei pazienti. L'implementazione di uno pseudonimo che sostituisce i dati rimossi può avvenire sotto forma di un codice o di una falsa identità. Se non si ha accesso alla chiave che ne consente la decifrazione, non si possono associare le informazioni visibili con l'identità originaria del paziente.
Gli pseudonimi si distinguono in due tipi: pseudonimo unidirezionale e pseudonimo reversibile. Il primo non è invertibile e non permette di riassociare le informazioni con gli attributi oscurati, mentre il secondo consente la re-identificazione del paziente pseudonimizzato [11].

Gli algoritmi per calcolare lo pseudonimo possono essere basati su tecniche di crittografia o hashing[2].

La pubblicazione appena menzionata, illustra in maniera più dettagliata i procedimenti per eseguire correttamente la pseudonimizzazione di un dataset. Un passaggio chiave risiede nel dover scindere il dataset in due tabelle, una in cui sono mantenute le informazioni personali (Tabella 2.2) e un'altra in cui sono conservati i dati pseudonimizzati con i rispettivi pseudonimi (Tabella 2.3). La Tabella 2.2 è un esempio di ciò che potrebbe essere visualizzato dall'avversario, mentre la Tabella 2.3 rappresenta quella custodita, utilizzata per risalire alle associazioni. Il database pseudonimizzato è così in grado di provvedere correttamente alla privacy degli individui, rendendo impossibile un'associazione diretta tra persone specifiche e i loro dati. Il processo di identificazione e separazione dei dati personali da quelli correlati è denominato "depersonalizzazione".

Codice	**Età**	**Cap**	**Patologia**
CH382	26	80013	asma
CH361	52	80016	artrite
CH415	48	80121	ipertensione

Tabella 2.2: contiene quasi identificatori **Q** e attributi sensibili **S** protetti dallo pseudonimo, che viene sostituito al posto degli identificatori diretti **I**.

Codice	**Nome**	**Indirizzo**
CH382	Elena	121, via Venezia
CH361	Sara	49, via Firenze
CH415	Antonio	67, via Bologna

Tabella 2.3: contiene gli identificatori diretti **I** personali associati al corrispettivo pseudonimo sotto forma di codice. Utile per conservare le identità originarie.

Questa strategia sembrerebbe capace di annullare le possibilità di re-identificazione degli individui nel dataset pseudonimizzato, poiché l'unico elemento che lega le informazioni ai pazienti è lo pseudonimo, il quale offre protezione tramite una struttura robusta e sicura. In realtà, si evince facilmente dalla letteratura che questa tecnica presenta due debolezze che la rendono particolarmente vulnerabile:

1. Il rischio di decriptazione del sistema di sicurezza utilizzato per creare lo pseudonimo.
2. Il rischio di attacchi da parte di avversari tramite l'uso di informazioni ausiliari o database aggiuntivi in combinazione con le informazioni già in possesso.

La seconda problematica, in verità, costituisce un rischio per qualsiasi tipologia di de-identificazione adottata. Tuttavia, nel caso di pseudonimizzazione di un dataset, rappresenta un pericolo maggiore in conseguenza al fatto che gli attributi non oscurati non sono modificati, ma mantenuti al loro stato originario. Gli attacchi di collegamento (linkage attack), ossia attacchi combinati attraverso dati aggiuntivi, potrebbero quindi essere più semplici da condurre

[2] L'hashing è una tecnica di trasformazione dei dati che converte un input in una stringa di lunghezza fissa, solitamente utilizzata per proteggere l'integrità dei dati o per anonimizzarli.

rispetto ad altre forme di de-identificazione. Il suo meccanismo di funzionamento, insieme alle altre tipologie di attacchi di re-identificazione possibili, verrà illustrato in modo più approfondito dopo aver concluso la spiegazione delle tecniche di de-identificazione, in modo da comprendere meglio il campo di applicazione.
Poiché gli attributi non vengono modificati, la protezione vera e propria della pseudonimizzazione deriva dalla complessità del sistema di sicurezza usato per creare lo pseudonimo, introducendo, rispetto ad altre tecniche, un ulteriore fattore che aumenta il rischio di re-identificazione. Inoltre, qualora si verifichi il tracciamento della chiave utilizzata, si riesce a risalire alla totalità delle informazioni personali dei pazienti di un dataset, rappresentando un danno maggiore nell'ipotesi di attacchi avvenuti con successo.

Sebbene queste caratteristiche rendano più vulnerabili le informazioni de-identificate attraverso la pseudonimizzazione, sotto alcuni aspetti, rappresentano anche il punto di forza del sistema.
Esistono infatti contesti in cui il mantenimento delle informazioni al loro stato originale è indispensabile. La pseudonimizzazione si presta bene dunque in scenari di raccolta dei dati e archiviazioni in cui vengono raggruppati grandi volumi di informazioni provenienti da diverse fonti per l'elaborazione statistica e il data mining [12]. Quest'ultimo, consiste nel processo di analisi di grandi volumi di dati per scoprire modelli, tendenze e informazioni utili attraverso tecniche statistiche, algoritmi e machine learning. In pratica, si tratta di trasformare dati grezzi in conoscenza utilizzabile. In alcuni casi, è necessaria anche una via di accesso all'identità, perché aiuta a comprendere e valutare terapie e a condurre studi significativi. La pseudonimizzazione richiede però una prudenza maggiore in situazioni in cui è richiesta la condivisione e lo scambio di dati, poiché il loro flusso risulta pregnante di informazioni estremamente sensibili.
Prima dell'introduzione del nuovo regolamento, l'Ufficio del Commissario per l'Informazione del Regno Unito trattava i dati pseudonimizzati come anonimi per terze parti che non dispongono della chiave d'accesso [13]. Tuttavia, poiché le informazioni pseudonimizzate lasciano comunque la possibilità di risalire alle identità delle persone, risulta evidente che la pseudonimizzazione non è in grado di adoperare un'anonimizzazione effettiva dei dati, e che il suo utilizzo va adeguatamente contestualizzato e ponderato. Per tali ragioni, a seguito dell'introduzione del GDPR, i dati pseudonimizzati vanno trattati come dati personali, richiedendo l'impiego di tecniche più esaustive per il raggiungimento di un'anonimizzazione vera e propria che rispetti i requisiti della nuova normativa.
È anche importante sottolineare che realizzare la pseudonimizzazione di un dataset non impedisce la possibilità di attuare ulteriori tecniche di de-identificazione. Pertanto, non si deve considerare sufficiente per una corretta anonimizzazione se eseguita singolarmente.

2.2.2 Anonimizzazione

Esplorare e comprendere accuratamente il significato di anonimizzazione può essere semplice, ma realizzarla in una forma concreta rappresenta una sfida molto più impegnativa.
L'anonimizzazione è un processo costituito dall'insieme di algoritmi, tecniche e forme di de-identificazione che, se applicate ai dati, consentono di giungere alla creazione di una loro versione anonima. Il fulcro di questa tecnica risiede nella corretta comprensione del concetto di anonimato e di cosa stabilisce l'anonimità di un dato.
Riportando quanto affermato da Pfitzmann e Koehntopp, l'anonimato è "lo stato di essere non identificabile all'interno di un insieme di soggetti, il cosiddetto insieme di anonimato (anonymity set)" [14]. Nell'articolo 26 del GDPR, si definiscono i dati anonimi come "informazioni che non si possono ricollegare ad una identificata o identificabile persona naturale". Tali dati, sono quindi quelli che cadono effettivamente lontani dalla sfera di applicazione del GDPR.

In base a queste definizioni, lo scopo dell'anonimizzazione è quello di nascondere le relazioni che sussistono tra persone e informazioni. Affinché ciò avvenga correttamente, è essenziale rimuovere gli attributi identificativi diretti di un dato e modificare gli attributi rimanenti attraverso la metodologia più efficace. Si ricorda, infatti, che esistono numerose tipologie di attacchi e che, anche nei casi di dati ampiamente de-identificati, il rischio di una corrispettiva re-identificazione persiste.
L'ostacolo odierno più sostanziale riguarda la necessità di bilanciare la tutela della privacy degli individui con la preservazione dell'utilità delle informazioni in possesso. Un'anonimizzazione troppo efficace potrebbe ridurre notevolmente i benefici ricavabili dallo studio di un dataset, ma una troppo blanda non garantirebbe un'adeguata protezione dei dati, i quali sarebbero esposti a un pericolo costante.
Ottenere un dato anonimizzato, non costituisce quindi una difficoltà se considerato come problema a sé stante, ma risulta una sfida decisamente significativa quando si deve conciliare con il requisito dell'utilità.
A seguito del processo di anonimizzazione, i dati anonimi, a differenza di quelli pseudonimizzati, non dovrebbero permettere in alcun modo, almeno nella teoria, la re-identificazione del soggetto. Poiché ciò non può mai del tutto avvenire, l'obiettivo di questo metodo diventa quello di salvaguardare al meglio la privacy evitando di superare la soglia critica in cui l'utilità delle informazioni viene completamente compromessa (vedi Figura 2.2).

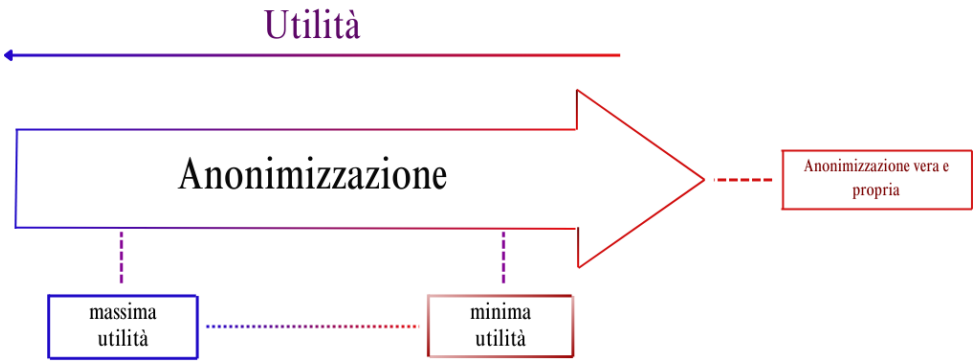


Figura 2.2: Relazione inversa tra anonimizzazione e utilità dei dati: massimo livello di anonimizzazione corrisponde a minima utilità dei dati.

L'anonimizzazione comprende una vasta gamma di meccanismi e tecniche, più o meno utilizzabili a seconda dello scopo dello studio condotto e dei risultati che si vogliono ottenere. Esistono inoltre algoritmi specifici che adoperano questo processo in maniera tale da trasformare e modificare dataset nei loro corrispettivi anonimizzati in modo soddisfacente. Nel Capitolo 3, si presenterà un'analisi di alcuni di questi algoritmi, così come delle tecniche generali che sono state maggiormente consolidate ed esplorate nella letteratura.
La versatilità di queste tecniche e la loro attuazione pratica, sono alcuni dei fattori che hanno permesso ad aziende, istituzioni e organizzazioni, di promuovere e investire su questa strada, che grazie alla forte protezione e libertà che garantisce, è in grado di adempiere ai criteri definiti dal GDPR.

2.3 Differential privacy

La differential privacy comprende una serie di meccanismi che hanno suscitato notevole interesse negli ultimi tempi. I modelli che adottano differential privacy sfruttano un approccio molto più dinamico per la protezione delle informazioni e sono stati proposti come metodologie qualificate anche nel contesto sanitario. Le tecniche presentate fino ad ora affondano le radici del loro funzionamento nell'assunzione iniziale che qualsiasi avversario che miri a rubare informazioni sensibili dispone, o può entrare in possesso, di ulteriori informazioni ausiliarie, inducendo il rischio di re-identificazione tramite attacchi che sfruttano dati combinati. Al contrario, il punto di forza di questa forma di protezione della privacy risiede nel non effettuare quasi nessuna assunzione circa informazioni già note da parte degli avversari, e che quindi la protezione degli individui è garantita senza tener conto delle conoscenze pregresse degli aggressori. Inoltre, un ulteriore scopo consiste nel far sì che l'aggiunta o rimozione di ulteriori record non modifichi l'uscita dovuta all'analisi del dataset in alcun modo [15]. Per chiarire

meglio la sfera di applicazione di questa metodologia, occorre un'analisi più specifica del suo funzionamento interno, di cui viene fornita una schematizzazione prendendo in esame l'articolo di revisione appena menzionato.

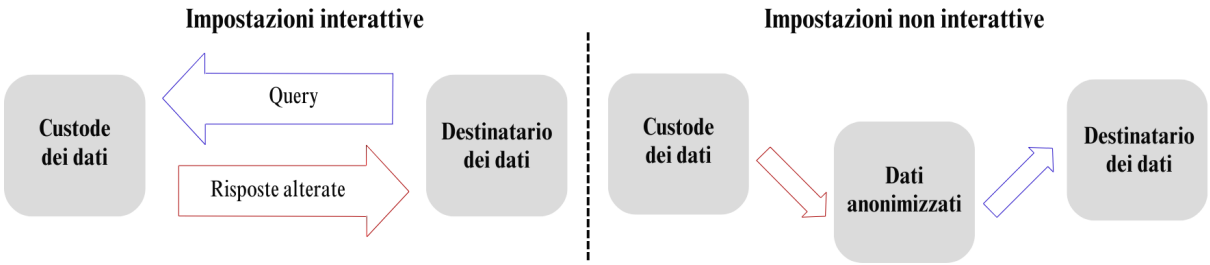


Figura 2.3: confronto tra impostazioni interattive e non interattive. Nell'impostazione interattiva si ha uno scambio diretto tra custode e destinatario di dati, mentre in quella non interattiva non sussiste nessun interazione diretta tra le due parti.

Prima di procedere con la spiegazione effettiva, è importante evidenziare che l'impostazione di questa tecnica differisce dall'impostazione delle altre tecniche analizzate precedentemente. L'impostazione rappresenta l'insieme delle regole e dei meccanismi attraverso cui vengono effettuate le interazioni con un dataset, determinando il modo in cui i dati vengono forniti, elaborati e protetti. Si distinguono due diversi tipi di impostazione e si riporta di seguito una spiegazione intuitiva di questi, in modo da collegare la tecnica in discussione con quelle esplorate antecedentemente e comprenderne le differenze (Figura 2.3):

- Impostazione non interattiva
- Impostazione interattiva

Come si deduce dai nomi, nell'impostazione non interattiva, per proteggere la privacy non vi è alcuna interazione tra il database e gli utenti, che ricevono solamente versioni già elaborate dei dati da parte del custode delle informazioni.
L'impostazione interattiva implica, invece, un'interazione diretta e continua tra l'utente e il server che detiene e custodisce le informazioni. L'utente pone delle query, ovvero "domande", al database che restituisce le risposte con l'informazione richiesta. Sotto quest'impostazione, la tutela della privacy avviene tramite la modificazione della risposta, che non corrisponde alla pura verità, ma che trasmette un'informazione ugualmente valida.
I processi di de-identificazione discussi in precedenza, appartengono alla categoria non interattiva, poiché sia la pseudonimizzazione che l'anonimizzazione mirano a modificare il dataset nel tentativo di rendere quanto più difficile possibile l'identificazione delle persone. La versione integrale è gestita solamente dal custode originario dei dati, mentre la loro archiviazione e condivisione utilizzerà versioni de-identificate della medesima.
Al contrario, la differential privacy si riconduce a una forma pressoché interattiva di protezione

dei dati, in cui l'utente stabilisce una comunicazione diretta col database attraverso le query sopracitate.

Entrando più nello specifico, la differential privacy richiede che, a prescindere dalla presenza o meno di una determinata riga nel database, la risposta a qualsiasi query sia "probabilisticamente indistinguibile" tra i due scenari possibili. In altre parole, posta l'esistenza di due database pressoché simili, che differiscono ad esempio a meno di una sola riga, un algoritmo coerente con gli scopi della differential privacy fornirà uscite (output) randomizzate che seguono distribuzioni di probabilità indistinguibili, ovvero quasi identiche, in entrambi i database (Figura 2.4), rendendo impossibile stabilire in quali dei due database è presente la riga specifica. Per uscite randomizzate si intendono risultati che includono un certo livello di rumore statistico, introdotto intenzionalmente per nascondere le differenze nei dati contenuti nel database.
Ciò vuol dire che un potenziale avversario non sarà in grado di discernere la presenza o meno di un particolare record relativo ad un individuo in un dataset, pur essendo a conoscenza di informazioni su tutti gli individui contenuti nel dataset in questione ad eccezione del record specificato.

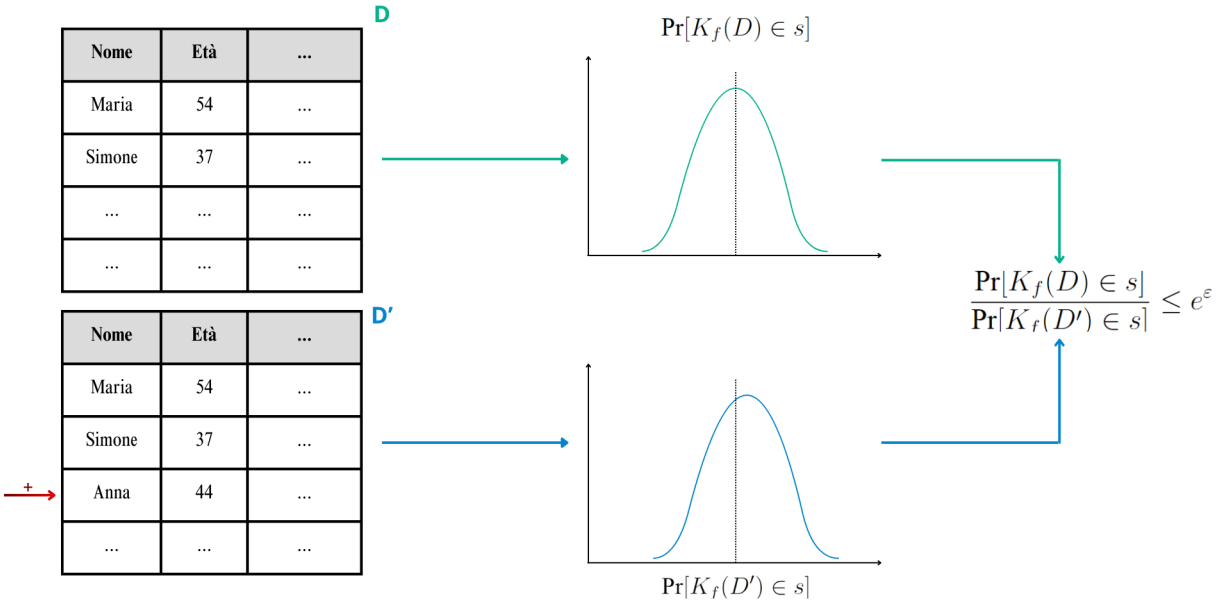


Figura 2.4: confronto delle distribuzioni di probabilità di due dataset *D* e *D'* con una sola riga di differenza. Le due uscite sono quasi identiche e per essere probabilisticamente indistinguibili tra di loro, il rapporto corrispondente deve essere limitato dal valore e^{ε}. Il parametro ε viene stabilito da chi ha il controllo del database, e valori più bassi implicano garanzie di privacy più elevate.

Precisamente, sia data una query arbitraria f con dominio $\mathcal{D}$ e range P (cioè una funzione $f : \mathcal{D} \rightarrow P$) e due dataset D e D' che differiscono in un solo record. Sia K_f una funzione randomizzata usata per produrre la risposta alterata alla query f, ovvero una funzione che prende in ingresso i risultati della query f (ossia il codominio P di f) e ne restituisce i valori randomizzati finali con Range(K_f) che dipende dal tipo di funzione randomizzata specifica adoperata (cioè $K_f : P \rightarrow$ Range(K_f)). Sia infine $\varepsilon \in [0, +\infty)$ un parametro stabilito da chi ha il controllo dei dataset, allora la funzione K_f fornisce ε-differential privacy se, per qualsiasi sottoinsieme $s \subseteq$ Range(K_f):

$$\Pr[K_f(D) \in s] \leq e^{\varepsilon} \Pr[K_f(D') \in s] \,, \tag{2.1}$$

cioè la probabilità che l'uscita della funzione randomizzata K_f applicata al dataset D appartenga al sottoinsieme s deve essere minore della probabilità che l'uscita dovuta alla stessa funzione randomizzata K_f applicata al dataset D' appartenga ad s modulata dal valore e^{ε}.
Il parametro ε è pubblico e permette di determinare quanto sia vicina la somiglianza tra le due uscite dei dataset randomizzati dopo l'applicazione della funzione randomizzata K_f. Ci si può riferire ad esso con "perdita di informazioni", e maggiore è il suo valore, minore è il livello di protezione della privacy garantito. Generalmente, oscilla tra 0.1 e 0.01, talvolta anche valori più grandi come $ln(2)$ o $ln(3)$. Idealmente, per $\varepsilon = 0$ le distribuzioni di probabilità dei risultati dei due dataset D e D' sono identiche, anche se, nella pratica, questo valore è irrealistico poiché restituirebbe risultati totalmente casuali e del tutto privi di informazioni utili. La disuguaglianza soprastante, stabilisce dunque il "rapporto di guadagno di conoscenza di un dataset rispetto all'altro" :

$$\frac{\Pr[K_f(D) \in s]}{\Pr[K_f(D') \in s]} \leq e^{\varepsilon}, \tag{2.2}$$

il quale sarà limitato dal fattore e^{ε}. Grazie a questa strategia, la rimozione di contenuti dal dataset non comporta cambiamenti significativi della probabilità di ciascuna uscita. Il valore di e^{ε} è sempre maggiore o uguale a 1 ma, sulla base dei valori che ε generalmente assume, si avranno risultati significativi solamente nell'intervallo tra poco più di 1 e 3.

Dalla letteratura, si riscontrano diversi metodi per modificare l'uscita dovuta alla funzione randomizzata applicata, ma il più ampiamente diffuso consiste nell'applicazione di rumore alla reale uscita dovuta ad una determinata query.
Se f è una query su un dataset D e r è la reale risposta a f non perturbata, allora la risposta alla query sarà $r + y$, dove y corrisponde alla quantità di rumore. Un'applicazione molto conosciuta è quella suggerita dagli autori di [16], che comporta l'adozione di una distribuzione Laplaciana del rumore. Di conseguenza, nell'enunciazione precedente, y è il rumore generato

casualmente da una distribuzione di Laplace con media 0 e scala $\Delta f/\varepsilon$, dove Δf rappresenta il valore massimo di $|f(D') - f(D'')|$ per tutti i $D', D'' \in \mathcal{D}$ che differiscono per una riga, altresì definibile come sensibilità locale della funzione f.

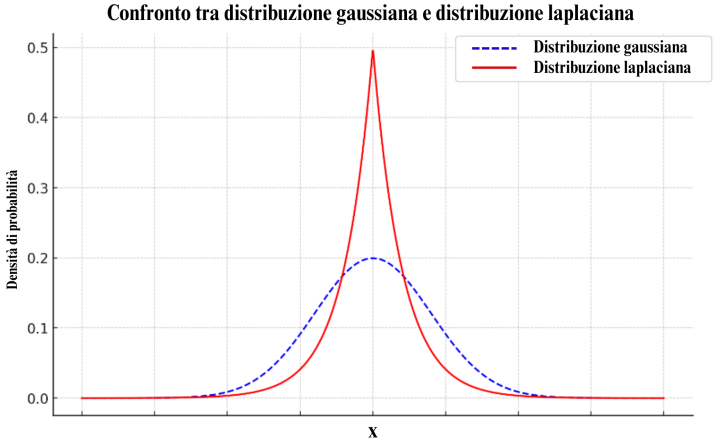


Figura 2.5: confronto tra la distribuzione normale gaussiana (in blu), e la distribuzione di Laplace (in rosso). Nell'asse delle ascisse, **x** rappresenta i possibili risultati di una query perturbata, mentre l'asse delle ordinate indica la densità di probabilità per ogni valore di **x**. La distribuzione di Laplace presenta una forma con code più marcate rispetto alla distribuzione gaussiana, il che implica una maggiore quantità di rumore applicata anche ai valori estremi. Ciò contribuisce a proteggere questi valori nei dati, riducendo il rischio di sbilanciamenti.

L'applicazione di un rumore di tipo Laplaciano permette di evidenziare un aspetto fondamentale della differential privacy, che ha lo scopo di gestire i dati in maniera efficace.
La distribuzione di rumore di Laplace presenta una forma che mantiene una significativa quantità di rumore anche intorno ai valori estremi, riducendo il rischio che vengano identificati. Ciò non avverrebbe nel caso in cui si applicasse rumore normale di tipo gaussiano, in quanto le code della distribuzione sono meno accentuate, il che potrebbe causare problemi in presenza di dati sbilanciati (vedi Figura 2.5). Il rumore normale o gaussiano è, inoltre, proporzionale alla varianza dei dati, e risulterebbe quindi elevato rispetto a quest'ultima; pertanto, l'alta sensibilità contribuirebbe a compromettere l'utilità dei dati.
Questo è anche il caso di alcune situazioni cliniche, in cui la varianza delle distribuzioni numeriche è spesso molto accentuata, comportando una sensibilità locale Δf marcata. Ciò rende più difficile ottenere risultati significativi dai dati, poiché sarebbe necessario applicare una distribuzione di rumore proporzionalmente elevata quando si adopera la differential privacy, il che riduce l'utilità dei dati [17]. Per ovviare a questa problematica, una strategia generalmente diffusa consiste nell'applicazione di una normalizzazione che permette di uniformare la sensibilità locale con quella complessiva, riducendo la possibilità che singoli valori possano influenzare in modo sproporzionato il risultato.

Un'altra considerazione importante da fare riguarda la differenza di importanza dei vari attributi. Indipendentemente dal contesto, ogni dataset contiene attributi con differente impatto sull'analisi complessiva, e l'applicazione di un eccessivo rumore su questi attributi può minare notevolmente la loro utilità. Sarebbe appropriato, quindi, sfruttare un sistema capace di identificare e priorizzare gli elementi dei dataset, così da applicare un rumore proporzionale alla loro importanza.

In una tecnica interattiva come la differential privacy, è altrettanto importante considerare come diversi tipi di query possano influenzare approcci differenti per mantenere la privacy. Nello studio [18], in cui vengono proposti meccanismi di differential privacy per analizzare dati medici e per migliorare le risposte alle query attraverso una partizione dei dati dei pazienti, viene fornita una distinzione tra due tipi di query, che offre spunti di riflessione interessanti:

1. **Query di conteggio:** Sono importanti per costituire modelli di apprendimento statistico attraverso operazioni del medesimo ambito e si concentrano sul numero di istanze che soddisfano determinate condizioni. Le risposte a queste query devono essere modificate per le ragioni già evidenziate, anche se l'aggiunta di rumore potrebbe influenzare l'utilità dei risultati, richiedendo che sia fatta con cautela.

2. **Query di carico:** Negli ambienti interattivi, i dati rimangono sotto il controllo del custode. Molte query sequenziali potrebbero però divulgare informazioni sensibili se correlate tra di loro. Si può ridurre questo rischio diminuendo il parametro ε o aumentando la sensibilità delle query. Ciò potrebbe tradursi in un'aggiunta di rumore graduale, anche se comporterebbe una probabile perdità di utilità al crescere del rumore. Una scelta più sensata potrebbe essere quella di raggruppare le query come un carico di lavoro, eseguendole ed elaborandole come parte di un'unica operazione o gruppo.

Si deduce quindi che la strategia più efficiente deve tener conto di numerose variabili, e dipende dall'obbiettivo che si intende raggiungere. La letteratura offre ulteriori soluzioni in aggiunta all'uso di rumore di Laplace, come la già discussa integrazione di rumore di tipo gaussiano e meccanismi esponenziali.
Nonostante la differential privacy non rientri nel significato specifico di de-identificazione in quanto non apporta delle vere e proprie modifiche ai dati, essa è comunque in grado di porsi come una delle migliori strategie attualmente studiate e in sviluppo, grazie alla sua versatilità e dinamicità, che la rendono pratica in una molteplicità di contesti diversi. Questa tecnica avanzata è di conseguenza ampiamente riconosciuta e, se adoperata nel modo corretto o in combinazione con altre tecniche, rispecchia i requisiti imposti dal GDPR ed è applicabile nei diversi settori, tra cui quello medico, che è di maggiore interesse per questo lavoro.

2.4 Problematiche sulla re-identificazione

Dopo aver esplorato le tecniche più comuni per la de-identificazione dei dati sensibili, come la pseudonimizzazione e l'anonimizzazione, e una strategia più avanzata e innovativa come la differential privacy, mettendone in risalto punti di forza e debolezze, è fondamentale procedere ad un confronto diretto, soprattutto in relazione ai rischi di re-identificazione che ciascuna di esse corre. Ciò è utile al fine di comprendere efficacemente le aree di applicazione di ognuna e come poterle combinare insieme assemblandone i corrispettivi vantaggi.

2.4.1 Rischi di re-identificazione

La re-identificazione è il processo mediante al quale si può risalire all'identità e alle informazioni sensibili di un individuo anche a seguito di avvenuta de-identificazione o anonimizzazione dei suoi dati in un determinato dataset. Come già evidenziato nelle sezioni precedenti, nonostante l'applicazione di tecniche più o meno avanzate di de-identificazione, l'anonimizzazione totale è un obiettivo non perseguibile, poiché è necessario mantenere un bilancio che permetta di ricavare informazioni utili dal materiale in possesso. L'oscurazione di attributi identificativi o modificazioni consistenti del dataset, possono comunque lasciare tracce sufficienti per ricollegarsi a una persona specifica attraverso attacchi mirati che sfruttano tecniche di analisi o incrocio con altre fonti di dati. Per queste ragioni, le tecniche di anonimizzazione e gli algoritmi sviluppati sono in continuo aggiornamento, in quanto, più alta è la loro efficienza, più facile è per le istituzioni poter usufruire di informazioni rilevanti, con particolare interesse per la ricerca scientifica e medica

Nello specifico, vi sono tre principali tipologie di rischi di re-identificazione che richiedono una gestione accurata [19]:

- **Prosecutor risk**: rischio di re-identificazione di un record sapendo che l'individuo esiste nel dataset;
- **Journalist risk**: rischio di re-identificazione di un record senza essere sicuri che l'individuo esista nel dataset;
- **Marketer risk**: rischio di re-identificare grandi volumi di dati.

A seguito di avvenuta quantificazione del rischio di re-identificazione di un dataset anonimizzato, espressa tramite indicatori che mostrano la percentuale di record a rischio, si può stabilire la condizione che mette in relazione le tre tipologie sopracitate, riassunata come:

$$\text{Prosecutor risk} \geq \text{Journalist risk} \geq \text{Marketer risk}$$

Come già specificato, la pseudonimizzazione da sola non è in grado di anonimizzare i dati, e le tecniche di anonimizzazione mirano a rendere i dati solo idealmente anonimi. Pertanto, da questo punto in avanti, ogni riferimento a dataset anonimizzati includerà implicitamente i rischi di re-identificazione associati.

2.4.2 Privacy disclosure

In relazione ai rischi di re-identificazione esistenti, è fondamentale prendere in considerazione i fattori che possono influenzare e complicare la protezione dei dati sensibili, tra cui spiccano i rischi di divulgazione della privacy.
Attraverso il termine inglese *privacy disclosure*, si definisce la divulgazione di informazioni personali che gli utenti cercano di proteggere da entità terze il cui accesso alle informazioni è vietato. Si conoscono tre tipi di privacy disclosure [20]:

- **Divulgazione dell'identità (identity disclosure):** corrisponde alla re-identificazione. Si verifica quando viene rivelata l'identità di un individuo da dati pubblici, ovvero quando un ipotetico avversario riesce a tracciare una corrispondenza tra un record e un paziente con alta probabilità di certezza.

- **Divulgazione degli attributi (attribute disclosure):** Si verifica quando un avversario riesce a collegare un paziente con il corrispettivo attributo sensibile **S**, il quale può essere uno specifico valore oppure un range di valori che contiene l'**S** in questione.

- **Divulgazione dell'appartenenza (membership disclosure):** Si verifica quando un avversario riesce a determinare l'appartenenza di una vittima target a un dataset pubblico con alta probabilità.

La divulgazione di informazioni importanti, è spesso la conseguenza di attacchi mirati da parte di avversari che puntano ad ottenere dati e informazioni sensibili, avvalendosi di conoscenze preliminari. Si individuano tre requisiti che potrebbero supportare la riuscita di attacchi pericolosi da parte degli avversari. Primo fra tutti è l'accesso al dataset anonimizzato, che può essere pubblico in quanto dovrebbe contenere solo informazioni protette. L'avversario deve inoltre disporre di tutti o alcuni tra i quasi identificatori **Q** di uno specifico bersaglio, i quali sono facili da acquisire da diverse fonti, come dati demografici esterni. Infine, sono spesso necessarie anche conoscenze sulla distribuzione degli **S** nel dataset considerato che potrebbero essere utilizzate per derivare ulteriori informazioni dal dataset.

2.4.3 Modelli di attacchi alla privacy

La realizzabilità degli attacchi alla privacy che mirano a divulgare informazioni sensibili è spesso connessa alla presenza dei quasi identificatori e alla loro capacità di stabilire correlazioni tra dati. Tenendo a mente che con **Q** ci si riferisce ai quasi identificatori, e con **S** agli attributi sensibili, si descrivono di seguito le tipologie di attacchi principali che rappresentano un pericolo per la divulgazione di identità o di attributi:

- **Attacco di collegamento (linkage attack):** la re-identificazione dell'identità dei proprietari dei record e dei suoi **S** da parte di un avversario, è causata dall'intreccio tra i **Q** ausiliari di cui l'avversario è in possesso, e il dataset pubblicato.

- **Attacco di omogeneità (homogeneity attack):** questo attacco permette di rivelare i valori di **S** di un obiettivo, grazie alla mancanza di omogeneità nei medesimi attributi. In altre parole, combinando i valori di **Q** ausiliari con quelli della tabella, ci si riconduce ad un solo valore di **S** plausibile.

- **Attacco basato su conoscenze pregresse (background knowledge attack):** questo tipo di attacco sfrutta il ragionamento logico e le informazioni aggiuntive che un avversario conosce di un determinato obiettivo al fine di violarne gli **S**. Va considerato che le inferenze tratte da queste informazioni possono essere imprecise o speculative e che deduzioni fuorvianti o non scontate possono talvolta condurre l'avversario a conclusioni errate.

- **Attacco di asimmetria (skewness attack):** I valori di **S** hanno diversi gradi di sensibilità. Una persona potrebbe non essere preoccupata della divulgazione di una sua condizione se questa è abbastanza comune o diffusa. Condizioni più peculiari potrebbero invece preoccupare altre persone, che desiderano mantenere nascosta quest'informazione. Quando però la distribuzione complessiva degli attributi **S** nei dati originali è sbilanciata, eventuali aggressori potrebbero dedurre i valori di questi **S**. Nel caso di **S** poco comuni, ma presenti in quantità maggiore in un dataset, la probabilità di essere associato a quella condizione sensibile che si vuole mantenere nascosta è maggiore rispetto alle normali aspettative.

- **Attacco di somiglianza (similarity attack):** Nei casi in cui la relazione semantica tra i distinti valori di **S** in una classe di equivalenza è molto stretta, si rischia di rivelare i valori di **S**. In molti casi, attributi semanticamente simili vengono raggruppati in classi di equivalenza o fasce e, nel caso in cui un avversario non conosca l'esatto attributo dell'avversario, può comunque risalire a **S** individuando la fascia corrispettiva.

Di seguito, per facilitare la comprensione di ciascuna tipologia di attacco, si supporta quanto detto attraverso considerazioni pratiche su un dataset fittizio.

I	Q	Q	Q	S	S	O
Id paziente	Età	Genere	Cap	Stipendio	Malattia	Codice malattia
001	40-50	F	966**	[2k-4k]	Condizione cronica	E**
002	30-40	M	966**	[2k-4k]	Condizione cronica	E**
003	40-50	F	965**	[2k-4k]	Respiratoria	J**
004	30-40	F	965**	[2k-4k]	Disturbo dell'alimentazione	F**
005	30-40	M	964**	[2k-4k]	Respiratoria	J**
006	50-60	M	965**	[2k-4k]	Condizione cronica	E**
007	50-60	F	964**	[2k-4k]	Neoplasia	C**

Tabella 2.4: Versione anonimizzata della Tabella 2.1

La Tabella 2.4, rappresenta un dataset anonimizzato attraverso la combinazione delle due tecniche di de-identificazione esaminate: il nome e il cognome sono stati pseudonimizzati attraverso l'introduzione di un identificativo del paziente, gli altri attributi sono stati generalizzati o sostituiti così da ridurre la specificità delle informazioni.
Nonostante l'unione di entrambe le tecniche, il dataset presentato non è esente da eventuali attacchi che, se adoperati correttamente, potrebbero condurre alla re-identificazione dei pazienti.

Ad esempio, supposto che un avversario A sia sicuro della presenza dell'individuo B nel dataset, ed essendo a conoscenza del fatto che B sia un uomo che vive nel cap 96584, A può facilmente dedurre che il record corrispettivo di B è il numero 006 **(linkage attack)**. Grazie a questo collegamento, A può entrare in possesso anche delle informazioni sensibili, che nonostante siano state generalizzate rispetto al dataset originale, possono comunque arrecare un danno personale alla privacy dell'utente.
Si supponga adesso che l'aggressore A sia a conoscenza del cap di un individuo B presente nel dataset, che è 96617. Ciò colloca B nella classe di equivalenza generata dai cap simili. L'aggressore, sebbene non possa capire con certezza quale sia il record corrispondente a B, può comunque giungere alla conclusione che B soffre di una condizione cronica a causa di mancanza di omogeneità negli attributi sensibili **S** corrispondenti a questa classe di equivalenza

(homogeneity attack).
Si ponga ora il caso che l'avversario A disponga di queste informazioni quasi identificatrici: B è una donna che vive nel cap 96584. Questi dati bastano a restringere il campo di possibilità ai record 003 e 004, ma non permettono di proseguire oltre con l'identificazione. Basandosi però sulle conoscenze pregresse di A, secondo cui B ha più difficoltà negli sport aerobici, A potrebbe giungere alla conclusione che B sia il paziente 003 poiché affetto da una malattia respiratoria, più affine con la problematica di B evidenziata **(background knowledge attack)**. Quest'attacco non garantisce la totale certezza del collegamento effettuato ma, se usato attentamente, può condurre con alta probabilità a deduzioni veritiere.
A causa del diverso grado di sensibilità di ciascun **S**, alcuni pazienti potrebbero non essere troppo preoccupati se la condizione sensibile che viene diffusa è abbastanza comune, come il diabete. Tuttavia, patologie meno comuni e delicate come l'anoressia, potrebbero danneggiare maggiormente la privacy dell'individuo. Se le classi di equivalenza della tabella contenessero un numero elevato di pazienti con queste malattie, allora la probabilità che venga identificato un paziente con questo disturbo sarebbe maggiore rispetto alla normalità, e ciò implicherebbe rischi più elevati per la privacy degli utenti **(skewness attack)**.
Si supponga infine che un avversario deduca che il possibile stipendio di una vittima obiettivo sia relativamente basso. Sebbene gli stipendi dei pazienti siano distinti, sono tutti categorizzati nella fascia [2K, 4K]. Pertanto, un avversario potrebbe dedurre che l'obiettivo appartenga ad una determinata fascia quando i valori **S** sono semantici simili come in questo caso. Anche se non si conoscono gli stipendi esatti, si può comunque concludere che tutti i pazienti si trovano in una situazione economica simile **(similarity attack)**.

Dagli esempi mostrati, si evince che nonostante si siano applicate tecniche per l'anonimizzazione dei dati che non consentono l'immediato riconoscimento dei record, si possono comunque dedurre informazioni sensibili attraverso l'ausilio di informazioni aggiuntive, che possono mettere a repentaglio il diritto fondamentale alla privacy delle persone nel dataset. Inoltre, l'utilizzo di uno pseudonimo che ha permesso di oscurare gli identificatori diretti **I** non è stato in grado di prevenire l'incrocio tra informazioni aggiuntive possedute dall'avversario con quelle pubbliche presumibilmente anonimizzate.

Tra gli attacchi appena discussi meritano un approfondimento particolare gli attacchi di collegamento. Questi, possono essere suddivisi in tre classi di complessità crescente:

- attacchi di corrispondenza esatta;
- attacchi di corrispondenza robusta;
- attacchi di profilazione.

Gli attacchi di corrispondenza esatta si riferiscono ai semplici attacchi di collegamento già evidenziati, applicabili grazie all'abbinamento dei record attraverso stessi attributi in due o più dataset.
Gli attacchi di collegamento robusti condividono lo stesso obiettivo dei precedenti, ma, in questo caso, si tenta di abbinare i record anche quando non dispongono degli stessi valori di attributo esatti, ad esempio a causa di incongruenze o rumore.
Gli attacchi di profilazione, infine, provano a ricollegare record attraverso dataset che non si riferiscono allo stesso periodo di tempo, ma a momenti differenti, come anni diversi.

Si evince chiaramente che la proprietà fondamentale che rende possibili tutti gli attacchi di collegamento è l'unicità: se il numero di attributi di un dataset è sufficiente, la maggior parte dei record è inevitabilmente unica nel database. Attraverso un'analisi statistica, si scopre facilmente che basta una quantità limitata di attributi per poter applicare attacchi di corrispondenza esatta in maniera efficace, poiché le combinazioni dei valori degli attributi sono molto maggiori rispetto al numero di record di un dataset. Pertanto, la conoscenza di attributi di una vittima target come informazione ausiliaria, consente con molta probabilità l'individuazione del record corrispondente, rivelandone l'identità originaria e le sue informazioni sensibili. È importante notare però un'aspetto cruciale: gli attacchi di collegamento basano la loro logica sull'assunzione iniziale che la vittima target si trovi nel dataset considerato (prosecutor risk). Se così non fosse, le informazioni ausiliari di cui si dispone potrebbero essere associate al record errato

Gli attacchi di collegamento non sono solamente un rischio teorico. La loro efficacia è stata già ampiamente dimostrata in numerosi dataset reali anonimizzati, anche pubblici. Nonostante siano stati proposti nel 1959 da Ivan Newcombe, raggiunsero popolarità solo grazie a Latanya Sweeney nel 1997. Come viene spiegato all'interno dell'articolo [7], Sweeney è stato in grado di re-identificare il governatore del Massachusetts all'interno di un dataset di assicurazione medica, collegando i suoi dati con i record di registrazione degli elettori. Sweeney è stato infatti capace di dimostrare che la data di nascita, il sesso e il codice postale di una persona contenuti nel dataset pubblico di registrazione degli elettori erano sufficienti per instaurare una connessione tra elettori e record medici, consentendo di risalire all'identità di ciascun individuo ed entrando in possesso dei loro dati medici sensibili.

A conclusione di ciò, durante il processo di anonimizzazione dei dati, dovrebbero essere anche considerate misurazioni interpretabili e realisticamente fattibili per stimare il livello di anonimato dei dati anonimizzati. Si ribadisce che non esiste un confine netto tra

pseudonimizzazione e anonimizzazione, poiché anche dati anonimizzati corrono i rischi di re-identificazione illustrati in questa sezione. Tuttavia, i dati pseudonimizzati rendono più semplice l'attuazione di questi attacchi a causa dell'integrità delle informazioni, che sono più facilmente confrontabili con le conoscenze ausiliare degli avversari. Ciò non esclude che la combinazione di entrambe le strategie sia spesso una scelta valida: l'opzione più sensata deve scaturire sia dalle intenzioni e dal beneficio che si vuole trarre dai dati in possesso, sia da un'attenta misurazione dei rischi di re-identificazione associati e dalle informazioni primarie che occorre proteggere.

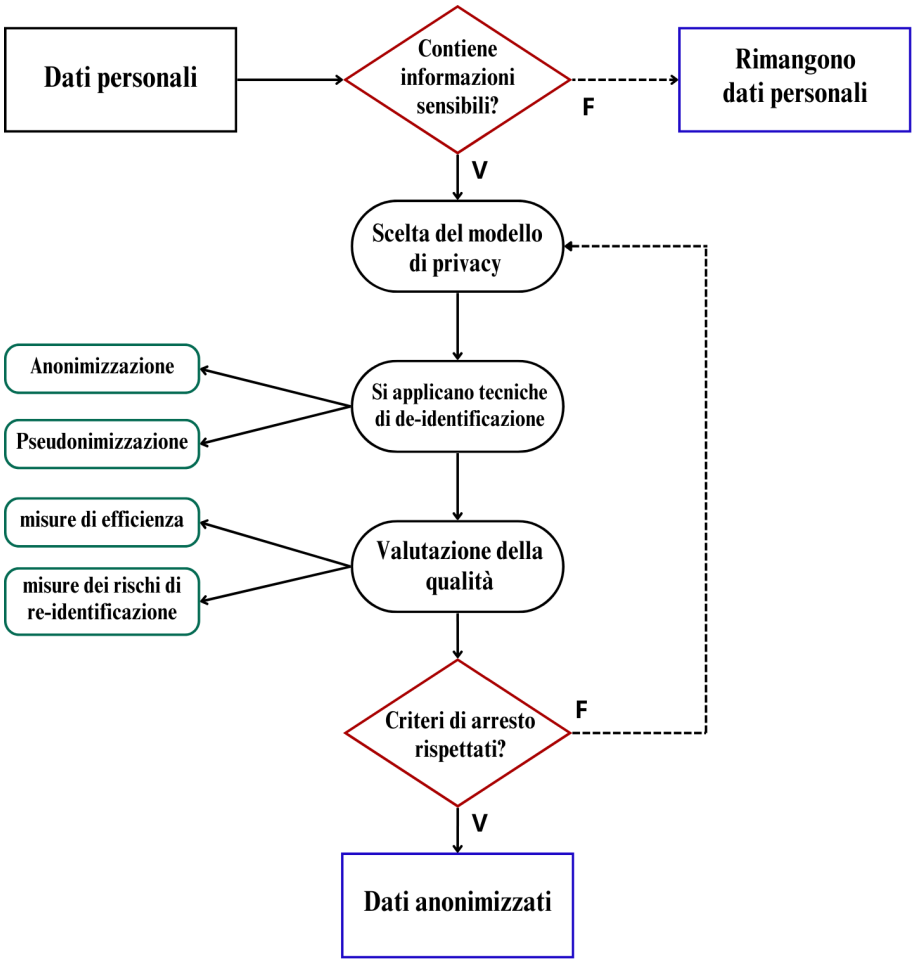


Figura 2.6: schema logico da seguire per anonimizzare efficacemente i dati. Nel caso in cui i dati personali all'interno di un dataset contengano informazioni sensibili, occorre scegliere un modello di privacy applicando le rispettive tecniche di de-identificazione. Dopo una valutazione della qualità dell'applicazione implementata, se i criteri di sicurezza sono rispettati, si arresta il processo e si restituiscono i dati anonimizzati; altrimenti, occorre procedere con la scelta di un modello più soddisfacente.

3 Anonimizzazione dei dati sensibili

L'anonimizzazione, come già ampiamente discusso, racchiude in sé l'insieme di meccanismi volti a modificare i dati sensibili ed è cruciale per garantire il rispetto dei principi di protezione della privacy già esistenti, ulteriormente fortificati dalle normative specifiche introdotte.
Sebbene sia un settore tuttora in sviluppo che auspica alla protezione totale delle informazioni sensibili tramite il raggiungimento dell'anonimato, molti studiosi sostengono che sia impossibile eliminare completamente i rischi per la privacy nei dati rilasciati pubblicamente utilizzando tecniche di anonimizzazione [21]. L'inevitabile pubblicazione di altri dataset contenenti informazioni correlate permetterebbe ,all'ipotetico aggressore, di collegare i due dataset portando alla re-identificazione degli individui del primo dataset. Al contrario, i sostenitori dell'anonimizzazione ribattono che nonostante sia vero che la possibilità di attacchi non sia solamente un rischio teorico ma dimostrato nella pratica, la probabilità di effettiva re-identificazione rimane minima per la maggioranza dei dataset. Usando quindi le tecniche di anonimizzazione già consolidate da tempo, molti dataset resteranno anonimi.
Nello specifico, i dati medici sono noti per contenere informazioni ad alto contenuto sensibile e accumulate da fonti eterogenee. Da ormai un po di anni, per migliorare l'efficienza dei servizi e per ridurre le spese associate, i dati sono stati digitalizzati e inseriti nelle EHR, contenenti non solo dati relativi all'identità e alle informazioni personali, ma anche dati biomedici sensibili, che sono stati la causa di molteplici attacchi informatici volti a recuperare informazioni personali. Per prevenire ciò, il nuovo sistema basato sull'uso delle EHR è stato costretto a implementare tecniche di anonimizzazione sempre più efficaci, in modo da prevenire l'uso improprio, l'abuso o la violazione della privacy dei pazienti.

3.1 Tecniche generali di anonimizzazione

La scelta della tecnica più adatta non è scontata. Per selezionare il metodo più efficace per ogni situazione, è necessario comprendere l'obiettivo specifico dell'anonimizzazione che si vuole effettuare nel caso considerato. Ogni tecnica presenta, infatti, caratteristiche uniche, che la rendono più o meno appropriata in base al contesto e alle esigenze specifiche. Pertanto, è di vitale importanza che colui che possiede i dati conosca e adoperi nel migliore dei modi le varie misure di protezione della privacy, al fine di controllare e tutelare le informazioni dei dati rilasciati prima della loro pubblicazione o condivisione. Si noti ancora una volta che oltre l'obbligo di perseguire la salvaguardia dei dati sensibili, l'importanza di preservare l'utilità delle informazioni in possesso è il discriminante principale di queste operazioni.

In letteratura, si discutono numerosissime tecniche di limitazione della divulgazione statistica (Statistical Disclosure Limitation, SDL), cosi denominate poiché in grado di prevenire la divulgazione involontaria di informazioni riservate, mantenendo al contempo l'utilità analitica dei dati. La loro complessità varia, ma sono tutte ugualmente utili e integrabili insieme per una protezione dei dati robusta. Per questa ragione verranno presentate quelle che sono state maggiormente riscontrate e analizzate da più fonti.

3.1.1 Soppressione

La soppressione, per quanto sia semplice, risulta tra le tecniche più efficaci e immediate nell'anonimizzazione dei dati. Essa consiste nella rimozione completa o parziale di colonne o tuple[1] da un determinato dataset, rendendo del tutto inutilizzabili le informazioni soppresse. Questa tecnica è particolarmente utile quando gli attributi rimossi non apportano alcun beneficio agli utilizzatori dei dati, ovvero nei casi in cui la loro presenza potrebbe solamente aumentare il rischio di attacchi di collegamento senza restituire informazioni valide.
Per garantire un'adeguata anonimizzazione, è quindi sicuramente indispensabile sopprimere gli identificatori diretti che collegano immediatamente il record all'identità della persona. Inoltre, come discusso in precedenza, uno dei principi che consente di discernere tra attributi identificatori diretti e quasi identificatori, è l'utilità dell'informazione trasmessa dall'attributo. Nel caso in cui un attributo non consenta la re-identificazione immediata del record, ma potrebbe comunque rivelare informazioni personali che ne agevoli il processo, si adopera una suddivisione: gli attributi che contengono informazioni rilevanti per l'analisi del dataset devono essere trattati come quasi identificatori **Q**, mentre quelli che non risultano utili per la ricerca, come il nome o informazioni di contatto, devono essere classificati come identificatori diretti **I** e rimossi tramite soppressione per proteggere la privacy.
L'intuibile vantaggio di questa tecnica sta nell'impossibilità di recupero delle informazioni soppresse, le quali sono state rimosse permanentemente. Di seguito, la Tabella 3.1 mostra un esempio di dati in versione originale, e nella Tabella 3.2 si presentano gli stessi dopo l'applicazione della tecnica di soppressione.

[1] Una tupla è una riga all'interno di una tabella di un database relazionale, rappresenta un singolo record composto da una serie di valori organizzati in colonne (attributi).

id	nome	età	indirizzo	cap	numero visite
147	Marco	39	84, Via Roma	35126	20
258	Giulia	36	19, Via Torino	35143	24
369	Matteo	31	53 ,Via Milano	30172	16

Tabella 3.1: versione originale di un minidataset. id e nome sono identificatori diretti **I**, età, indirizzo e cap sono quasi identificatori **Q**, il numero di visite è un attributo sensibile **S**

età	indirizzo	cap	numero visite
39	84, Via Roma	35126	20
36	19, Via Torino	35143	24
31	53 ,Via Milano	30172	16

Tabella 3.2: versione soppressa della Tabella 3.1. Le colonne ”id” e ”nome” contenenti identificatori diretti **I** sono state rimosse tramite soppressione

Gli attributi ”nome” e ”id”, rappresentano identificatori diretti **I** privi di informazioni utili e, per tali ragioni, la loro soppressione non solo semplifica l’analisi dei dati, ma basta per eliminare del tutto la possibilità di re-identificazione senza l’ausilio di informazioni aggiuntive da combinare con il dataset. È necessario considerare inoltre che se nel dataset sono presenti combinazioni dei valori degli attributi unici o quasi unici, l’identità di queste combinazioni possono essere facilmente dedotte e, di conseguenza, il record corrispettivo potrebbe essere soggetto a soppressione come possibile soluzione per il mantenimento della privacy.

In base a quanto detto, la soppressione può anche essere eseguita in forma parziale e spesso si traduce in un ulteriore sottocategoria: la **sostituzione dei caratteri**.
La soppressione definitiva non è efficace per alcuni tipi di attributo, come il cap, poiché comprometterebbe la funzionalità dei dati. Per questo motivo, in alcuni contesti sarebbe maggiormente indicata l’adozione di tecniche meno aggressive. Sostituire alcuni caratteri del valore di un attributo sembra essere un compromesso più vantaggioso, che permette di proseguire con una buona anonimizzazione senza rinunciare del tutto all’informazione utile. Di solito, si tende ad occultare questi caratteri con altri simboli privi di significato, come ad esempio ”*****” oppure ”**x**” (Tabella 3.3). La quantità di caratteri rimpiazzati dipende da quanto si vuole preservare l’informazione da anonimizzare: se questa è particolarmente utile, si può ridurre il numero di sostituzioni in favore di una maggiore accessibilità ai dati per analisi future.

età	indirizzo	cap	numero visite
39	84, Via Roma	3****	20
36	19, Via Torino	3****	24
31	53 ,Via Milano	3****	16

Tabella 3.3: versione soppressa e con sostituzione dei caratteri del cap della Tabella 3.1.

La soppressione e la relativa sostituzione dei caratteri sono due tecniche che puntano ad eliminare il rischio di re-identificazione immediata da parte di un avversario, ma non impediscono del tutto la possibilità di risalire alle informazioni sensibili attraverso strategie di attacco più sofisticate, come quelle esplorate nel capitolo precedente.

Un ulteriore metodo con caratteristiche simili alla sostituzione dei caratteri, è il **mascheramento**, il quale adopera la sostituzione in maniera differente. Nel caso del mascheramento, ogni numero dall'1 al 9 viene rimpiazzato col numero 1, ed ogni lettera minuscola o maiuscola dalla a alla z viene rimpiazzata rispettivamente con la z o la Z. Il primo carattere, il numero 0 e i caratteri speciali sono mantenuti al loro stato originale.
Poiché questa tecnica non permette di restituire informazioni utili per la ricerca, non viene applicata, ma persiste l'uso della semplice sostituzione, che restituisce risultati medesimi utilizzando al contempo meno risorse per controllare e cambiare ciascun carattere.

3.1.2 Generalizzazione

La generalizzazione è il processo secondo cui i valori di alcuni attributi vengono rimpiazzati con quantità meno specifiche ma semanticamente equivalenti. Questo meccanismo aumenta la difficoltà per l'attaccante di estrapolare i dati sensibili, poiché la generalizzazione dei quasi identificatori **Q** complica il confronto con le informazioni ausiliari che si possiedono.
Va considerato che questa tecnica non si può applicare a qualsiasi attributo contenuto nel dataset da anonimizzare, e la sua modalità di implementazione varia in base al tipo di attributo. Nel caso di attributi numerici si può spesso modificare definendo un intervallo di valori contenente la quantità esatta del dato originale. Da attributi come indirizzi o date, invece, si possono rimuovere le informazioni più specifiche, inserendo una versione generica del dato. Lo scopo di questa tecnica consiste nella creazione di confusione nei dati, in modo da ostacolare il processo di re-identificazione per un aggressore, che dovrà tenere conto di identificatori meno specifici che racchiudono in se maggiori possibilità per uno stesso dato.

id	nome	età	indirizzo	cap	numero visite
147	Marco	30-40	Via Roma	35126	20
258	Giulia	30-40	Via Torino	35143	24
369	Matteo	30-40	Via Milano	30172	16

Tabella 3.4: versione generalizzata della Tabella 3.1. L'età è rappresentata da intervalli e il numero civico è stato rimosso dall'indirizzo.

Nella Tabella 3.4 la generalizzazione è applicata in due modi differenti già esaminati: l'età è stata generalizzata introducendo uno spettro più ampio di possibilità, cosi da rendere indistinguibili le identità di ciascun individuo, mentre l'indirizzo è stato inserito senza il numero civico, impedendo la localizzazione esatta delle persone nel dataset. Anche il cap poteva subire una generalizzazione diversa rispetto alle altre, sostituendolo con la città, la provincia o la regione, a seconda del grado di anonimizzazione desiderato. Tuttavia, ciò poteva tradursi in una snaturazione dell'attributo, e per questo motivo, in alcuni casi si preferisce non adoperare questo meccanismo.
A causa della diversa forma con cui questo processo viene implementato per attributi differenti, non è possibile schematizzare la struttura di questo metodo in maniera univoca. Tuttavia, alcuni algoritmi tuttora ampiamente consolidati sfruttano le dinamiche di questa tecnica attraverso una struttura solida ed efficace, in grado di restituire un modello valido di dataset anonimizzato.

3.1.3 Distorsione

La distorsione si riferisce ad un processo di trasformazione dei dati che protegge la privacy mantenendo l'utilità statistica delle informazioni. I dati possono essere successivamente ripristinati utilizzando la loro versione originale [22]. Prendendo in esame l'equazione sottostante:

$$V_d = V_u + V_r \ , \tag{3.1}$$

V_u rappresenta il dato originale effettivo, al quale viene aggiunta una quantità V_r, giungendo al valore distinto V_d. Attraverso una semplice operazione inversa: $V_u = V_d - V_r$, si può ritornare al valore originale V_u. Questo meccanismo è particolarmente utile quando si vuole identificare la persona una volta ultimato il processo di elaborazione dei dati; tale è il caso dell'industria sanitaria. La distorsione dei dati può anche avvenire tramite l'utilizzo di funzioni di crittografia come quelle di hash. L'applicazione di funzioni di hash comporta l'irreversibilità dell'operazione: l'unico modo per risalire al valore originale è eseguire l'hash del valore originale e confrontarlo con il valore del database crittografato.
Questo tipo di tecnica può inoltre essere applicata sotto diverse forme, alcune molto più dinamiche, che ricalcano protezioni simili a quella fornita dalla differential privacy. L'articolo [23] presenta una variante che implementa la distorsione tramite funzione di probabilità. La sequenza di dati distorta probabilisticamente sembrerebbe fornire asintoticamente le stesse proprietà statistiche di quelle della sequenza originale, poiché entrambe al di sotto della stessa distribuzione. La distorsione probabilistica presenta un punto di forza non indifferente, ovvero la capacità di riuscire ad ovviare all'uso di query di carico e, grazie alla sua resistenza agli attacchi basati su query ripetute, preserva al massimo le informazioni statistiche utili.

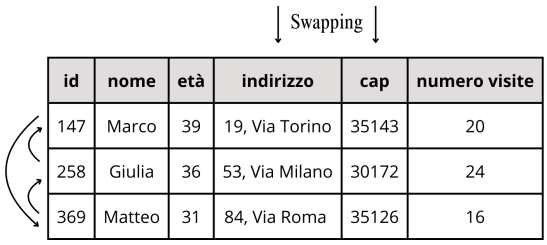
3.1.4 Perturbazione

La perturbazione tenta di salvaguardare la privacy attraverso modifiche ai dati originali che preservano tuttavia l'integrità statistica dei dati. Ciò significa che le variazioni applicate sono abbastanza piccole o sottili da non influenzare la distribuzione o le caratteristiche statistiche generali del dataset [9]. L'alterazione dei valori del dataset, pur costituendo una potenziale riduzione dell'accuratezza dei dati, contribuisce a ridurre la vulnerabilità ad alcuni tra gli attacchi più ostici, come quelli di collegamento, e per tali ragioni, la sua implementazione acquista rilevanza.
La perturbazione di un dataset comprende una molteplicità di meccanismi con cui può essere applicata, permettendo di adottare e combinare insieme le forme più opportune per l'esigenza che maggiormente si cerca di soddisfare. Di seguito si illustrano alcune tra quelle maggiormente riportate nella letteratura scientifica.

Swapping e shuffling

Il **data swapping** (scambio di dati), comporta un riarrangiamento delle variabili all'interno di ciascuna colonna. Fu proposto per la prima volta da Dalenius e Reiss nel 1982 [24], come tecnica SDL. In alcuni contesti, come nello studio di tabelle che mostrano conteggi e frequenze, lo scambio dei dati avviene in modo tale da mantenere i conteggi marginali della tabella, ovvero la somma dei valori per ciascuna riga e colonna deve continuare ad essere la medesima. In generale, questo procedimento non può essere applicato a tutti gli attributi presenti nel dataset, altrimenti l'accuratezza de dati verrebbe definitivamente a mancare.
Lo scambio di informazioni tra variabili di una stessa colonna, introduce associazioni errate tra quasi identificatori **Q** e dati sensibili **S**, consentendo di deviare il possibile avversario durante il processo di re-identificazione, nonostante si sia mantenuta la stessa struttura dei dati. Il rischio maggiore di questo procedimento sussiste nel caso di randomizzazione imperfetta, in cui si ottiene lo stesso valore originale per alcuni record del dataset. Bisogna anche porre attenzione alla natura degli attributi in cui viene applicata, poiché, dopo lo scambio, in alcuni casi potrebbero verificarsi delle combinazione senza senso.
Similmente opera il processo di **shuffling** (mescolamento), grazie al quale i dati contenuti in un attributo vengono mescolati casualmente attraverso un processo di randomizzazione più pronunciato che non dovrebbe permettere l'ottenimento di record uguali all'originale.
Come già detto in principio, la forza di queste tecnice scaturisce dalla capacità di conservare l'utilità statistica dei dati. Risulta particolarmente efficace dunque quando si vuole analizzare un attributo senza la necessità di metterlo in relazione con gli altri.
Ad esempio, nel caso della Tabella 3.1, se è di interesse per la ricerca stabilire il numero di visite effettuate dal centro sanitario, adoperare un mescolamento della colonna "visite", non intacca il

numero totale di visite che sono state effettuate, ma permette di dissociare ciascun informazione con il suo proprietario, rinforzando la sicurezza del dataset.



id	nome	età	indirizzo	cap	numero visite
147	Marco	39	19, Via Torino	35143	20
258	Giulia	36	53, Via Milano	30172	24
369	Matteo	31	84, Via Roma	35126	16

Tabella 3.5: versione della Tabella 3.1 dopo swapping della colonna ”indirizzo” e ”cap”

id	nome	età	indirizzo	cap	numero visite
147	Marco	39	84, Via Roma	35126	16
258	Giulia	36	19, Via Torino	35143	20
369	Matteo	31	53 ,Via Milano	30172	24

Tabella 3.6: versione della Tabella 3.1 dopo shuffling della colonna ”numero visite”

Anche se le due perturbazioni presentano dinamiche estremamente simili, nascondono sottili differenze. Nella Tabella 3.5 avviene uno scambio preciso delle coppie di valori degli attributi ”indirizzo” e ”cap” in senso circolare, mentre nel caso della Tabella 3.6, si verifica un mescolamento casuale della colonna ”visite”, anche se coincide con lo stesso scambio prodotto dallo swapping a causa del numero ristretto di record.
Tuttavia, questa tecnica non garantisce l’anonimizzazione completa dei dati, che potrebbero quindi essere riorganizzati nella loro forma originale.

Aggiunta di rumore e post-randomizzazione

L’**aggiunta di rumore** è una delle tecniche attualmente più in uso, e consiste nella leggera modificazione degli attributi che vengono disturbati da un rumore aggiunto che li rende meno accurati, come ad esempio l’aggiunta o la sottrazione di giorni a una data. La perdita di utilità, tuttavia, si traduce in una protezione più robusta dei dati, e la sua applicazione richiede, quindi, la comprensione del livello di rumore da applicare per non avere un impatto eccessivo sulla validità delle informazioni.
A seguito dell’aggiunta di rumore, i dati perturbati possono essere elaborati correttamente tenendo conto dell’incertezza derivante dal rumore aggiunto. Nel caso di dati continui, si fa riferimento all’aggiunta di rumore descritta, sul piano discreto, si evidenzia l’utilizzo del **metodo post-randomizzazione** (Post Randomization Method, PRAM).
Il PRAM è un metodo che perturba ciascun record di un dataset, introducendo una qualche forma di distribuzione di probabilità. In [24] si propone una distinzione tra risposta randomizzata e PRAM, dove la risposta randomizzata è una tecnica utilizzata nei sondaggi nei casi in cui si pongono domande particolarmente sensibili (come per qualche malattia delicata). L’idea è che la probabilità di ottenere una risposta sincera è p, mentre la probabilità di risposta non veritiera

è 1 - p: chi conduce il sondaggio non sa con certezza quale delle due alternative ha scelto il rispondente, e ciò contribuisce alla protezione della privacy dei rispondenti.
Infine, per ciascuna osservazione, il valore reale di un campo sensibile verrebbe rilasciato con una certa probabilità, mentre il suo opposto verrebbe rilasciato con la probabilità complementare. L'analisi successiva di questi dati richiede tuttavia la conoscenza del meccanismo con cui è avvenuta la randomizzazione.
Nella risposta randomizzata, il processo è casuale e indipendente dalla probabilità reale effettiva, mentre con il PRAM, il valore vero è noto, e dunque il meccanismo di probabilità usato per perturbare i dati è definito attenendosi a questo valore.

Nello stesso articolo, l'aggiunta di rumore viene riportata come un caso specifico del **matrix masking**, che è un altro metodo più generale di limitazione della divulgazione statistica. Si considera una matrice di dati X di dimensioni n per p, composta da n righe e p colonne. Invece di rilasciare i dati della matrice X al loro stato originale, questo metodo suggerisce di pubblicare una versione differente: $Y = AXB + C$, dove A,B, e C sono matrici conformi appropriate, che se definite correttamente includono casi particolari come l'aggiunta di rumore e il sampling (campionamento).
Nei casi di applicazioni di matrix masking, l'analista deve conoscere la procedura di mascheramento utilizzata e, pur conoscendola, la complessità dell'analisi dei dati richiede ugualmente l'uso di un software apposito.

Sampling

Il sampling consiste nell'operare una selezione di un sottoinsieme di osservazione dal dataset originale, rilasciando pubblicamente solo il campione prelevato sotto forma di microdati. La semplicità di questa tecnica nasconde un doppio vantaggio: da una parte riduce notevolmente la possibilità di re-identificazione, dall'altra risulta facile da implementare, semplificando anche la successiva analisi dei dati campionati.
Il principio di protezione che questo meccanismo sfrutta è che il campione non riflette necessariamente tutte le caratteristiche uniche del dataset originale, rendendo più difficile confermare l'unicità di un particolare record del dataset campionato. Per essere più precisi, questo sistema riesce a proteggere adeguatamente dagli attacchi di collegamento, poiché anche se un osservatore esterno riuscisse ad instaurare una corrispondenza tra i dati ausiliari che possiede e un record del dataset campionato, non potrebbe comunque avere la certezza che il record in questione sia l'unico del dataset originale a presentare quelle caratteristiche. Le informazioni aggiuntive di cui l'avversario dispone potrebbero quindi appartenere a qualcun altro presente nel dataset originale e a cui non ha accesso.

Taglio, microaggregazione e ricombinazione

Queste tre tecniche manipolano i record delle tabelle, così da altere le informazioni del dataset. L'applicazione della tecnica di **taglio** prevede la suddivisione di record completi in gruppi, ciascuno dei quali contiene un numero ridotto di variabili e viene rilasciato separatamente. Questo metodo mantiene livelli accettabili di riservatezza, anche se difficilmente attuabili per analisi più complesse.
Attraverso la **microaggregazione**, un valore designato viene sostituito con il valore medio calcolato su un piccolo gruppo di almeno tre unità, implicando la creazione di record differenti rispetto agli originali. Le unità dello stesso gruppo sono rappresentate dallo stesso valore nei dati anonimizzati, e i valori nelle colonne originali vengono modificati, sostituendo i corrispondenti valori microaggregati a quelli originali [25]. Questa tecnica è stata però considerata inadatta a causa della perdita sostanziale di utilità nei dati.
La **ricombinazione**, consiste nella divisione dei record in sottogruppi contenenti diverse variabili. Successivamente, si ricombinano i vari sottogruppi, utilizzando tecniche di abbinamento statistico fino a ricomporli tutti in ulteriori sottogruppi unendo informazioni provenienti da individui diversi. Tale riassemblamento si traduce nella creazione di record sintetici, che trasportano informazioni ricombinate e fuorvianti per gli avversari. L'uso della ricombinazione si è dimostrato il migliore tra i tre metodi valutati, e si presenta sicuro difronte ai tentativi più comuni di divulgazione che sfruttano le informazioni aggiuntive.

3.1.5 Dati sintetici

L'utilizzo de dati sintetici per proteggere la privacy delle persone, rappresenta una tecnica non comune che fonda la sua struttura su approcci che sfruttano l'imputazione multipla, riuscendo così a creare un dataset fittizio, ovvero costituito da record non reali, pur mantenendo le stesse caratteristiche e relazioni dei dati originali.
Il primo passo consiste nel considerare che alcuni tra questi dati sensibili siano "mancanti", e si sostituiscono con valori casuali generati attingendo a modelli statistici appropriati. I dati sintetici, per mantenere le stesse relazioni interne tra dati, cercano di replicare infatti la distribuzione di una popolazione più ampia. Successivamente, avviene un processo di sostituzione dei valori sensibili posti come "mancanti" con valori plausibili, il quale viene eseguito ripetutamente tramite l'utilizzo dell'imputazione multipla, creando ogni volta una popolazione sintetica.
Un'alternativa potrebbe essere anche quella di sostituire con imputazioni solo gli attributi sensibili, stabilendo quali variabili devono rimanere private poiché a maggior rischio di divulgazione.

Ogni set di dati sintetici viene analizzato con una tecnica per dati completi e le regole di combinazione utilizzate per abbinare le inferenze devono essere specifiche e appropriate.
Secondo quanto riportato in [24], i risultati delle elaborazioni statistiche svolte su dati sintetici sono virtualmente identici ai risultati che si otterrebbero con le medesime elaborazioni svolte su dataset originali. Ciò che discrimina la veridicità dell'affermazione precedente è il modello di imputazione specificato: se è accurato, molte delle analisi produrranno esiti coerenti, altrimenti, la stima dei parametri effettuata può differire anche in larga misura.
Nonostante l'ostacolo dettato dall'irrealità delle informazioni trattate, l'utilizzo dei dati sensibili rappresenta una tecnica ad elevata flessibilità, che si presta ad analisi importanti riducendo drasticamente la possibilità di attacchi di re-identificazione, grazie alla facoltà di scegliere i dati ritenuti maggiormente in pericolo e che occorre preservare.

3.2 Algoritmi per l'anonimizzazione

Le tecniche esplorate nella sezione precedente offrono spunti di riflessione differenti e ugualmente considerevoli, che proiettano la ricerca scientifica su un ventaglio di possibilità più vasto. Ciascuna istituzione, pubblica o privata, che desidera beneficiare dell'uso dei dati, ha a disposizione numerose risorse per conformarsi ai principi stabiliti dalle normative sulla protezione della privacy.
La scelta della soluzione più idonea dipende dal fine ultimo per cui le informazioni vengono usate, ed è compito del custode assicurarsi che la successiva esposizione dei dati anonimizzati non rappresenti un rischio di divulgazione di informazioni personali.
Oltre alle tecniche evidenziate, sono stati proposti in letteratura alcuni algoritmi che incorporano strategie operative la cui funzionalità è stata consolidata nel corso del tempo. La loro implementazione non è garanzia di assoluto anonimato, ma contribuisce sicuramente a ridurre in misura significativa i rischi che si corrono in merito.
L'utilizzo di algoritmi permette la standardizzazione di alcuni meccanismi, rendendo più fluidi i processi di anonimizzazione e più veloce la gestione dei flussi di dati. I dataset anonimizzati sfruttando questi algoritmi, come già detto, non sono esenti da attacchi di re-identificazione, poiché il loro funzionamento è ampiamente conosciuto e altrettanto diffusi e riconosciuti sono gli espedienti che consentono di contrastare queste misure di sicurezza.
Ciò non sminuisce la loro importanza, in quanto ogni algoritmo è capace di proteggere da alcuni degli attacchi di re-identificazione analizzati, presentando tuttavia debolezze nei confronti di assalti più sofisticati. Si obbligano in questo modo gli avversari a disporre di una mole maggiore di informazioni ausiliarie per re-identificare un dataset senza margini di errore.

Tra gli algoritmi studiati e sviluppati, si discutono di seguito quelli che sono stati maggiormente riconosciuti: k-anonymous, *l*-diversity e *t*-closeness. La praticità di implementazione, unita alla prospettiva di sicurezza che assicurano nei dataset in cui vengono applicati, ha permesso loro di affermarsi come soluzioni valide in numerosi contesti. Si noti che i tre algoritmi sono strettamente interconnessi e possono essere interpretati come stadi successivi di evoluzione. Nella Sezione 3.3, verrà inoltre presentata un'implementazione di questi algoritmi su alcuni dataset fittizi al fine di facilitarne la comprensione anche dal punto di vista pratico.

3.2.1 k-Anonymous

Uno dei metodi che più di tutti è stato in grado di accogliere le necessità di preservare le informazioni sensibili e che ancora oggi garantisce solidità e affidabilità ai sistemi di sicurezza è l'algoritmo k-anonymous.
Sviluppato dal già menzionato Latanya Sweeney nel 1998, sfrutta e intreccia i criteri della soppressione e della generalizzazione descritti nelle Sezioni 3.1.1 e 3.1.2, con il principio di k-anonimità stabilito dall'algoritmo, al fine di rivelare informazioni in maniera controllata. Entrando nel dettaglio, una tabella contenente attributi quasi identificatori **Q** si dice soddisfare la k-anonimità se ogni valore dei quasi identificatori contenuti nelle tuple della tabella si ripete almeno 'k' volte, eliminando così la possibilità di identificare unicamente ciascuna tupla all'interno della tabella.
Riformulando e contestualizzando: *se in un determinato dataset ciascun record presenta quasi identificatori indistinguibili da almeno altri (k-1) record della stessa tabella, allora il dataset è definito k-anonimo.*

Ad esempio, un dataset 3-anonimo indicherà la presenza di almeno 3 record indistinguibili l'uno dall'altro per ogni combinazione di quasi identificatori **Q**. I quasi identificatori **Q** sono gli attributi capaci di condurre a re-identificazioni se incrociati con informazioni ausiliarie conosciute da un avversario. Quando un dataset contiene più attributi quasi identificatori **Q**, ogni specifico insieme di valori per questi attributi può formare un accoppiamento unico, potenzialmente riconducibile a un individuo. Attraverso tecniche di generalizzazione e soppressione, tali accoppiamenti unici vengono inclusi in insiemi più ampi che definiscono le classi di equivalenza. Ogni classe è caratterizzata da una combinazione generalizzata di valori dei quasi identificatori **Q**, capace di rappresentare più individui. Ad esempio, un intervallo di età [30-35) combinato con una versione più generale di un cap in cui è stata soppressa l'ultima cifra (1012*) costituisce una rappresentazione più generica di quasi identificatori **Q**, associabile a più persone diverse presenti nel dataset (vedi Tabella 3.8). Nel caso di 3-anonimità, un individuo che presenta le stesse caratteristiche di un record a cui è riconducibile, sarà

ugualmente associabile ad altri (3-1) record, rendendo impossibile stabilire una corrispondenza certa con uno dei 3, che formeranno insieme una classe di equivalenza. Si deduce facilmente che la probabilità di identificazione decresce significativamente all'aumentare di k secondo la relazione 1/k, che nel caso specifico è uguale a 1/3.

Il punto di forza di quest'algoritmo sta nell'espressione di un concetto semplice ma particolarmente efficace, che consente di ottenere risultati ampiamente utilizzabili senza dover ricorrere a costi di implementazione eccessivamente elevati, richiesti invece da altri tipi di soluzione. Inoltre, per semplificare l'uso dell'algoritmo, è pratica comune attuare procedure di clustering, ovvero di raggruppamento nei cosiddetti "cluster", che corrispondono a gruppi di dati che presentano caratteristiche maggiormente simili tra loro. Attraverso la suddivisione delle informazioni in cluster, ricondursi a tabelle che soddisfano la k-anonimità, oltre ad essere più semplice, provoca anche una minore perdita di informazione. Questo accade poiché le informazioni all'interno di ciascun cluster sono più omogenee, e quindi le generalizzazioni necessarie per garantire l'anonimato possono essere meno drastiche rispetto a quelle richieste per dati non raggruppati.
Unendo il vantaggio ottenuto dal clustering con l'applicazione dell'algoritmo, si riescono a prevenire molte forme di attacchi di collegamento attraverso diversi dataset minimizzando al contempo il rischio di divulgazione dell'identità. Ciò contribuisce a giustificare la diffusione di questa tecnica tra qualsiasi istituzione che richiede la gestione di dati sensibili, fra cui ovviamente anche l'ambiente ospedaliero [26].

Sebbene k-anonymous affronti il problema della divulgazione dell'identità, non riesce ad estinguere del tutto gli altri rischi associati, che costituiscono limitazioni importanti alla sua applicazione. Persiste, infatti, il rischio di divulgazione degli attributi in seguito ad attacchi meno immediati rispetto a quelli di collegamento, come gli attacchi basati su conoscenze pregresse e gli attacchi di omogeneità. k-Anonymous non è in grado di escludere pienamente la eventuali deduzioni indirette derivanti da questi attacchi, poiché il meccanismo di controllo esercitato dall'algoritmo si focalizza sull'omogeneizzazione dei quasi identificatori **Q**, lasciando invece intatti gli attributi sensibili **S**. Per quanto questi attributi non contengano informazioni capaci di identificare una persona, un avversario con conoscenze più approfondite sulla vittima potrebbe comunque sfruttarli per effettuare associazioni sensate. Inoltre, la distribuzione di questi attributi nel dataset e nelle corrispettive classi di equivalenza, se eccessivamente omogenea, potrebbe suggerire involontariamente informazioni veritiere con elevata percentuale di correttezza. Infine, dataset con dimensioni notevoli, subiscono inevitabilmente una perdita di informazioni non indifferente a seguito del processo di k-anonimizzazione, in quanto,

comprendendo informazioni più complesse e variegate, è proporzionalmente plausibile che le generalizzazioni effettuate siano più incisive, riducendo di conseguenza la qualità dei dati.

3.2.2 l-Diversity

L'algoritmo l-diversity è stato proposto per superare le limitazioni riscontrate dal semplice utilizzo di k-anonymous. Lo scopo della sua introduzione era quello di riuscire a prevenire il rischio di divulgazione degli attributi parallelamente alla divulgazione dell'identità già ostacolata dalla condizione imposta dalla k-anonimità.
L'invenzione si deve ad Ashwin Machanavajjhala [27], che nel 2006, dopo aver dimostrato la debolezza del primo algoritmo attraverso due attacchi strutturati che quest'ultimo non era in grado di prevenire, ha proposto un nuovo tipo di algoritmo capace di contrastare tali attacchi, evidenziando la superiorità della nuova soluzione.
Prima di procedere con la formulazione della tecnica è importante definire il concetto di q*-block, ovvero blocchi o classi di equivalenza identificate da un unica combinazione di quasi identificatori. Si stabilisce quindi che un q*-block di quasi-identificatori è l-diverso se contiene almeno 'l' valori "ben rappresentati" per ogni attributo sensibile **S**. Se ogni q*-block di una tabella è l-diverso, allora la tabella soddisfa il concetto di l-diversity.
Per maggiore chiarezza, la nuova tecnica può essere formulata come segue: *posta una tabella k-anonima, si definisce anche l-diversa se ciascun q*-block di quasi identificatori nella tabella contiene almeno l valori "ben rappresentati" per ogni attributo sensibile.*

Il principio appena descritto, si può comprendere precisamente solo una volta chiarito il concetto di "ben rappresentati". Si definiscono tre principi in grado di interpretare questo concetto, ciascuno dei quali viene spiegato attraverso le tre istanze. Grazie a queste istanze si definiscono tre tipologie di l-diversità differenti, che costituiscono tre criteri non necessariamente mutualmente esclusivi:

- **l-diversità distinta:** all'interno della classe di equivalenza, esistono l valori "ben rappresentati" se si riscontra la presenza di l elementi distinti, ma non si esclude la maggiore frequenza di un valore rispetto agli altri. Con questo approccio, dunque, l'avversario può facilmente instaurare una relazione tra l'attributo e l'entità a cui fa riferimento la tabella, basandosi sulla probabilità dell'occorrenza dell'attributo.
- **l-diversità entropica:** per contenere l valori "ben rappresentati", l'entropia di ogni q*-block di una tabella deve essere almeno maggiore di $\log(l)$ (vedi Equazione 3.2), dove con log si intende il logaritmo naturale. Tramite questo principio si soddisfa il criterio di l-diversità entropica per ogni q*-block. Nel caso di bassa entropia, ovvero quando

la distribuzione dei valori sensibili è dominata da pochi valori ripetuti frequentemente, questo criterio può risultare troppo restrittivo. Sebbene la varietà dei valori distinti possa essere sufficiente, la loro distribuzione non garantisce un'entropia elevata, rendendo difficile soddisfare il requisito di entropia maggiore di $\log(l)$.

- (c, l)**-diversità ricorsiva:** una tabella contiene l valori "ben rappresentati" ed è conforme al principio di diversità ricorsiva, se i valori sensibili in ogni q*-block non si presentano né troppo frequentemente né troppo raramente. Questo concetto è più forte rispetto ai due precedenti menzionati, e coincide con la proposta di Machanavajjhala.

In particolare, la l**-diversità entropica** e la (c, l)**-diversità ricorsiva** sono quelle che sono state esplorate nel documento di Machanavajjhala.

Secondo quanto viene espresso, la l-diversità a entropia è stata inizialmente proposta da Ohrn e Ohno-Machado nel 1999 come strumento per la difesa da attacchi di omogeneità senza considerare però gli attacchi basati su conoscenze pregresse. Considerato $s \in S$ come valore di un attributo sensibile appartenente ad un q*-block, nel caso di una tabella, l'entropia viene definita come:

$$H(S) = -\sum_{s \in S} p_{(q^*,s)} \log(p_{(q^*,s)}), \tag{3.2}$$

in cui $p_{(q^*,s)} = \frac{n_{(q^*,s)}}{\sum_{s' \in S} n_{(q^*,s')}}$ è la probabilità di trovare il valore s dell'attributo sensibile in un q*-block e si calcola come la frazione di tuple che hanno il valore s, rispetto al totale delle tuple nel q*-block. $n_{(q^*,s)}$ rappresenta il numero di tuple nel q*-block con valore di attributo sensibile uguale a s, mentre la sommatoria $\sum_{s' \in S} n_{(q^*,s')}$ permette di calcolare il numero di tuple totali del q*-block contando il numero di tuple per ogni valore $s' \in S$ distinto all'interno dello stesso attributo sensibile del q*-block.
Il criterio di l-diversità entropica viene quindi stabilito attraverso la disuguaglianza:

$$-\sum_{s \in S} p_{(q^*,s)} \log(p_{(q^*,s)}) \geq \log(l), \tag{3.3}$$

che impone una diversità minima di l valori distinti all'interno del gruppo. Per ottenere l-diversità, la distribuzione dei valori sensibili deve essere "sufficientemente uniforme" e, di conseguenza, l'entropia deve essere maggiore di una soglia uguale a $\log(l)$ che garantisce la diversità del q*-block. Il concetto di gruppi "ben rappresentati" è racchiuso quindi dalla definizione di entropia, in funzione del fatto che questa aumenta quando le frequenze si distribuiscono più uniformemente: minore è la prevalenza di un attributo sensibile, maggiore è l'entropia. Se la condizione viene rispettata, la conseguenza dell'imposizione di questa disuguaglianza è che ogni q*-block ha almeno l valori distinti di attributi sensibili, che sono

distribuiti in modo sufficientemente omogeneo all'interno dei q*-block, rendendo così la tabella l-diversa.

La seconda istanza riguardante la (c, l)-diversità ricorsiva, che è quella maggiormente esplorata nel documento e della quale si è dimostrata l'efficacia di implementazione, stabilisce che: in un dato q^*-block contenente almeno l valori distinti, sia r_i il numero di volte in cui il i-esimo valore sensibile più frequente appare in quel q^*-block e, data una costante c, il q^*-block soddisfa la (c, l)-diversità ricorsiva se

$$r_1 < c(r_2 + r_3 + \cdots + r_m), \tag{3.4}$$

in cui $r_1 > r_2 > \cdots > r_m$. Una tabella T^* soddisfa la (c, l)-diversità ricorsiva se ogni q*-block soddisfa la diversità ricorsiva. Più semplicemente, questo principio stabilisce che l'elemento maggiormente presente in un q*-block, non sia "troppo presente" all'interno dello stesso q*-block, garantendo una maggiore omogeneità della distribuzione dei valori sensibili.
Per comprendere meglio l'istanza precedente, si consideri quanto segue: siano $s_1, \ldots, s_m$ i possibili valori dell'attributo sensibile S in un q*-block. Si considerano $n_{(q^*,s_1)}, \ldots, n_{(q^*,s_m)}$ i conteggi che esprimono il numero di tuple presenti nel q*-block per ciascun valore di attributo sensibile $s \in S$. Si supponga, quindi, di ordinare questi conteggi in ordine decrescente, denominando gli elementi della sequenza risultante $r_1, \ldots, r_m$, dove r_1 rappresenta il conteggio maggiore, ovvero il numero di apparizioni del valore sensibile più presente nel q*-block e, al contrario, r_m rappresenta il conteggio minore corrispondente al numero di apparizioni del valore sensibile meno presente nello stesso q*-block. Per $l = 2$, si stabilisce che un q^*-block è $(c, 2)$-diverso se soddisfa la disuguaglianza precedente $r_1 < c(r_2 + \ldots + r_m)$ per una costante specificata dall'utente c. Questa costante rappresenta un fattore di proporzionalità che permette di controllare il livello di disuguaglianza accettabile tra i conteggi, regolando quanto il valore più frequente (r_1) può dominare rispetto alla somma degli altri valori ($r_2, \ldots, r_m$).
Nel caso in cui sia $l > 2$, un q*-block si dice soddisfare la (c, l)-diversità ricorsiva se, eliminando un possibile valore sensibile nel q*-block, si continua ad ottenere un q*-block $(c, l-1)$-diverso. Questo significa che, in seguito alla rimozione di un valore sensibile, la distribuzione dei valori rimanenti deve continuare a rispettare i requisiti dettati dalla (c, l)-diversità ricorsiva.
Si può concludere che la 1-diversità è sempre soddisfatta, in quanto l'esistenza stessa di un q*-block implica la presenza di almeno un valore distinto, e che per controllare il soddisfacimento della l-diversità in caso di valori di $l > 2$, occorre procedere ricorsivamente fino a verificare il rispetto della $(c, 2)$-diversità ricorsiva. Conseguentemente, l'eliminazione di $l - 2$ valori sensibili deve restituire un q*-block che soddisfi la condizione di $(c, 2)$-diversità. L'avversario deve, quindi, eliminare almeno un valore sensibile in più ($l - 1$ possibili valori di

S) per essere in grado di inferire una rivelazione positiva.

L'algoritmo l-diversity, rappresenta un miglioramento dell'algoritmo k-anonymous, in quanto fonda i suoi principi di funzionamento solo dopo che la condizione iniziale di k-anonimità del dataset è stata soddisfatta. Una corretta implementazione di l-diversity può rafforzare significativamente la sicurezza del database, poiché fornisce una maggiore distribuzione degli attributi sensibili all'interno delle classi di equivalenza. In questo modo, garantendo un'esaustiva protezione sia dei rischi di divulgazione degli attributi che da quelli di divulgazione dell'identità, l-diversity ha successo dove k-anonymous fallisce.

Tuttavia, i vantaggi apportati da questo algoritmo nell'anonimizzazione dei dati non sono sufficienti a rispondere alle esigenze delle numerose istituzioni che gestiscono grandi quantità di dati, come ad esempio le aziende ospedaliere. l-Diversity non è ancora in grado di proteggere da tutti i tipi di assalti realizzabili: gli attacchi di asimmetria e di similarità, continuano a rappresentare un rischio insidioso, poiché l'algoritmo non riesce a evitare l'esposizione degli attributi, i quali potrebbero presentare relazioni semantiche non indifferenti.
Infine, l'utilizzo di un algoritmo che può risultare ridondante e laborioso da implementare, come l-diversity, potrebbe spingere gli utenti ad adottare strategie meno complesse ma altrettanto efficaci. Questo è il motivo per cui tecniche più semplici continuano ad essere utilizzate ampiamente in contesti reali.

3.2.3 t-Closeness

Dopo l'introduzione di l-diversity, nel 2007 è stato presentato un ulteriore algoritmo da parte di Ninghui Li, Tiancheng Li, e Suresh Venkatasubramanian [28], proposto per affrontare le limitazioni poste dai due metodi già esistenti: k-anonymity e l-diversity. L'applicazione di l-diversity, presuppone che l'avversario possa acquisire conoscenze su un attributo sensibile **S** se la distribuzione dell'attributo è nota, costituendo una limitazione per questo metodo. Il funzionamento dell'algoritmo t-closeness è unicamente rivolto agli attributi sensibili **S** di un dataset e, in maniera analoga ma sostanzialmente differente da l-diversity, preserva la divulgazione di questi pur mantenendo esposta l'identità delle persone, non ponendo alcuna condizione sui quasi identificatori **Q** che sono maggiormente coinvolti nella divulgazione dell'identità. Pertanto, un lavoro combinato di k-anonymity e t-closeness, può essere in grado di preservare la privacy dei dati pubblicati.

L'algoritmo t-closeness formalizza l'idea di una conoscenza globale di base, richiedendo che la distribuzione di un attributo sensibile **S** in una classe di equivalenza sia sufficientemente simile alla distribuzione dello stesso attributo nell'intera tabella.
Per chiarire meglio l'obiettivo di quest'algoritmo, si prenda in considerazione quanto segue: un osservatore possiede una conoscenza preliminare B_0 su un attributo sensibile **S** di una tabella. Al momento della pubblicazione di questa, la distribuzione *R* dell'attributo sensibile **S** nell'intera tabella fornisce delle informazioni all'osservatore, e la sua conoscenza su **S** si estende a B_1. In seguito al rilascio della tabella, grazie alle informazioni sui quasi identificatori **Q**, l'osservatore sarà in grado di identificare le classi di equivalenza dell'individuo. Dopo un'analisi della distribuzione *P* dell'attributo sensibile all'interno di una classe di equivalenza, l'osservatore acquisisce ulteriori informazioni che gli permettono di raggiungere il grado di conoscenza B_2.
Poiché l-diversity lavora sulle classi di equivalenza e non sull'intera tabella, il suo scopo è quello di limitare lo squilibrio tra B_0 e B_2 ($B_0 \simeq B_2$), imponendo che la distribuzione *P* delle classi di equivalenza abbia un livello di diversità tale da ridurre la differenza tra la conoscenza preliminari B_0 e la conoscenza B_2 derivante dalle nuove informazioni sulla distribuzione *P* nelle classi di equivalenza. In pratica si tenta di rendere la distribuzione *P* il più omogenea possibile, così da non lasciar trasparire informazioni in più sugli attributi sensibili **S** degli individui nelle classi di equivalenza rispetto a quelle di partenza già conosciute (B_0). Non si tiene conto, tuttavia, della distribuzione *R* relativa all'intera tabella, che è la causa della conoscenza B_1 acquisita dall'avversario, effettivamente utile per le analisi ma rischiosa se confrontata con B_2. Trasversalmente, t-closeness, cerca proprio di limitare significativamente la differenza tra B_1 e B_2 ($B_1 \simeq B_2$) insieme a questi rischi connessi, che si riflette in un avvicinamento della distanza tra la distribuzione *R* riferita all'intera tabella e la distribuzione *P* relativa alle classi di equivalenza (vedi Figura 3.1).

Un punto chiave sta nella conseguente assunzione che *R* sia trattato come informazione pubblica. Idealmente, la differenza tra B_0 e B_1, costituisce la conoscenza dovuta alle informazioni ottenute dall'intera tabella e rappresenta il guadagno di conoscenza effettivamente utile: più alto è il suo valore, maggiore è l'utilità delle informazioni in esso contenute. Poiché è la distribuzione *R* a fornire le informazioni che contribuiscono ad ottenere la conoscenza B_1 e che si riferiscono all'intera popolazione, la sua pubblicazione non intacca i singoli individui ma impreziosisce comunque il dataset. Per tali ragioni, il traguardo a cui si punta grazie a questo nuovo approccio, è l'assottigliamento della differenza tra *R* e *P* ($R \simeq P$), la quale potrebbe suggerire invece informazioni più specifiche sugli individui nelle rispettive classi di equivalenza.

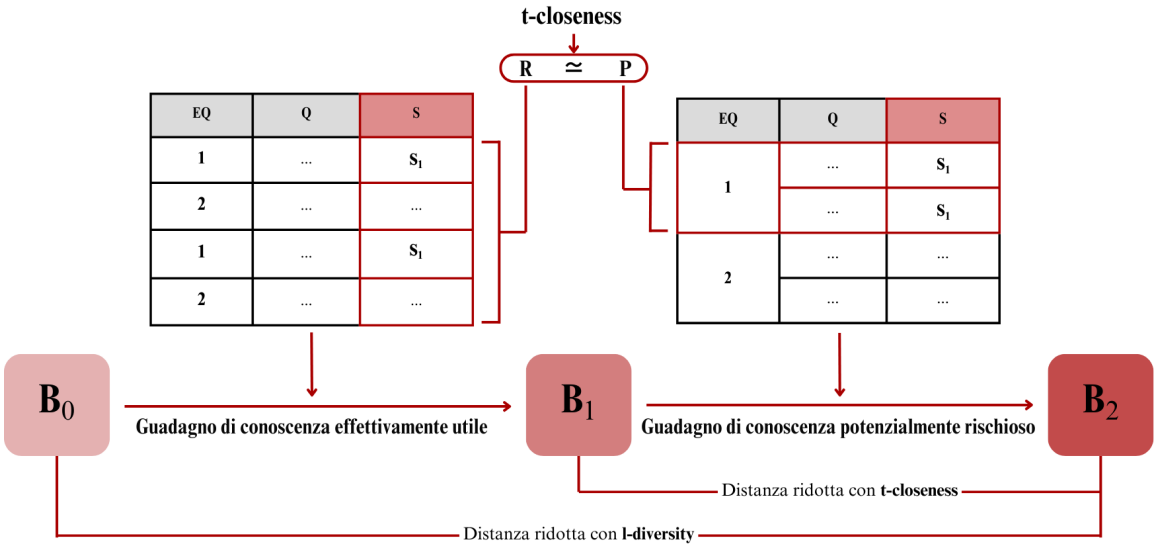


Figura 3.1: Schematizzazione del confronto tra l-diversity e t-closeness. I valori B_0, B_1 e B_2 rappresentano i tre stadi di conoscenza di un osservatore, mentre R e P rappresentano rispettivamente la distribuzione di un valore sensibile s_1 all'interno dell'intera tabella e la distribuzione dello stesso valore sensibile s_1 in una classe di equivalenza specifica. La quantità B_0 riflette le conoscenze preliminari di un osservatore circa il valore sensibile s_1.

Nel caso presentato in Figura 3.1, in seguito a valutazioni sulla distribuzione R nell'intera tabella, si ha un guadagno di conoscenza effettivamente utile che porta al grado di conoscenza B_1. Dopo un confronto tra R e P, la conoscenza si estende a B_2 ma, la differenza tra B_1 e B_2 rappresenta il guadagno di conoscenza potenzialmente rischioso nel caso in cui un avversario voglia inferire informazioni personali sugli individui. Le due tabelle in figura mostrano un esempio di anonimizzazione rischiosa. Questo perché, sebbene risulti omogenea la distribuzione R dell'attributo nell'intera tabella (probabilità di comparsa di s_1 = 50%), lo stesso non si può dire della distribuzione P dell'attributo nella prima classe di equivalenza (100%). t-Closeness cerca di diminuire questa differenza, riducendo, al contempo, la distanza tra B_1 e B_2 ($B_1 \simeq B_2$). Anche l-diversity contribuirebbe a risolvere il problema di omogeneità del caso presentato ma, in linea generale, non facendo nessuna ipotesi sulla distribuzione R, non può fare altro che ridurre la distanza tra B_0 e B_2 ($B_0 \simeq B_2$).

Il principio con il quale quest'algoritmo opera è dunque il seguente: *una classe di equivalenza si dice avere t-closeness se la distanza fra la distribuzione di un attributo sensibile in questa classe e la distribuzione dell'attributo nell'intera tabella non supera un valore soglia t, denominato anche valore di threshold. Si dice che una tabella ha t-closeness se tutte le classi di equivalenza appartenenti alla tabella possiedono t-closeness.*

Il fulcro del principio appena esposto risiede nella corretta comprensione dei meccanismi attraverso i quali questa distanza tra distribuzioni di attributi viene quantificata.
Esistono diversi modi per definire la distanza, tuttavia, quelli più comuni come la distanza variazionale e la distanza di Kullback-Leibler (KL), non riflettono la distanza semantica tra i valori. Nell'algoritmo, l'approccio impiegato per calcolare la distanza tra distribuzioni si fonda sull'Earth Mover's Distance (EMD). Concettualmente, l'EMD si basa sulla quantità minima di lavoro necessaria per trasformare una distribuzione in un'altra spostando la massa della distribuzione l'una verso l'altra. Per sfruttare questo concetto nella t-closeness, è necessario riuscire a calcolare l'EMD tra due distribuzioni. Nello specifico, esistono due diversi casi che occorre considerare nel calcolo dell'EMD, ossia il caso di attributi numerici e il caso di attributi categorici.
A scopo illustrativo, si riporta il calcolo per gli attributi numerici, nel quale si tiene conto della **distanza ordinata** tra due valori, che è basata sul numero di valori tra di essi nell'ordine totale, cioè:

$$\text{distanza ordinata}(v_i, v_j) = \frac{|i-j|}{m-1}, \tag{3.5}$$

dove v_i è il valore più piccolo e v_j il più grande tra quelli considerati. Successivamente, per quantificare la distanza tra le distribuzioni *P* e *R*, si introduce la variabile: $d_i = p_i - r_i$ (con $i = 1, 2, \ldots, m$), che rappresenta la differenza tra le probabilità associate ai valori v_i secondo la distribuzione *P* ed *R*. La distanza tra P e R può essere espressa formalmente come:

$$D[P, R] = \frac{1}{m-1}\left(|d_1| + |d_1 + d_2| + \ldots + |d_1 + d_2 + \ldots + d_{m-1}|\right), \tag{3.6}$$

dove si tiene conto delle differenze cumulative tra le due distribuzioni che vengono normalizzate rispetto al numero totale di valori distinti *m*.

Analogamente si può procedere con il calcolo nel caso di attributi categorici, ma poiché spiegare l'esatto funzionamento esula dallo scopo di questo lavoro, per ulteriori approfondimenti si consulti il documento [28], nel quale si illustrano i dettagli metodologici relativi a questa tipologia di attributi.

Sfruttare i principi definiti da t-closeness, consente di risolvere alcune delle problematiche più pericolose per la privacy delle persone, eliminando, ad esempio, il rischio di divulgazione degli attributi. I numerosi confronti tra gli algoritmi presenti nella letteratura, dimostrano la robustezza di questa nuova tecnica nel prevenire anche gli attacchi si asimmetria o di similarità, grazie alla capacità dell'algoritmo di rilevare la vicinanza semantica degli attributi, a differenza

di l-diversity che è in grado di assicurare solo la disomogeneità degli attributi nelle classi di equivalenza.

Nonostante l'elevata protezione dalla divulgazione degli attributi, t-closeness pecca nella protezione dai rischi di re-identificazione dell'identità, rendendo quasi necessaria la sua implementazione in combinazione con k-anonymous. Inoltre, calibrare un valore adeguato di t non è sempre intuitivo e potrebbe compromettere l'equilibrio tra privacy e utilità. Infine, l'uso dell'EMD, può richiedere un'elevata capacità computazionale, specialmente nei casi di attributi categorici o di dataset con dimensioni elevate.

3.3 Valutazioni e applicazioni degli algoritmi

Le tecniche e gli algoritmi analizzati nelle sezioni precedenti non impongono restrizioni specifiche da seguire né garantiscono certezza assoluta di sicurezza ma, se lavorano in maniera coesa sfruttando i vantaggi reciproci, si raggiunge un equilibrio tale da poter affermare che l'anonimizzazione sia avvenuta con successo. Ciò non significa che l'applicazione indiscriminata di più tecniche complicate contemporaneamente sia la soluzione ai problemi riguardanti la preservazione della privacy. Idealmente, se si combinassero i criteri stabiliti dai 3 algoritmi (k-anonymous, l-diversity, t-closeness), si garantirebbe privacy certa. Nella maggioranza dei casi, tuttavia, ciò risulta impossibile, poiché riuscire a far coesistere criteri differenti e complicati, oltre ad essere una sfida difficile e talvolta impossibile, richiederebbe costi computazionali così elevati da rendere impraticabile l'attuazione. Per tali ragioni, un uso sensato dei mezzi che si hanno a disposizione è essenziale, e un'analisi preliminare dei dataset da anonimizzare è necessaria per comprendere la strada migliore da seguire. Gli algoritmi e le tecniche che si vogliono adoperare, devono bilanciarsi sia ai benefici che bisogna ricavare dalle informazioni in possesso, sia alla delicatezza delle medesime. Informazioni con un grado elevato di sensibilità, richiedono un'attenzione proporzionalmente adeguata. L'analisi dei numerosi fattori che vanno considerati per ottenere un'accurata anonimizzazione sfocia nella ricerca del giusto equilibrio tra privacy ed utilità, che risulta essere il punto focale a cui è necessario convergere sempre.
A tal proposito, si consideri che quanto esplorato nelle sezioni precedenti non riflette la totalità delle strategie adottabili, ma racchiude comunque una sfera abbastanza ampia dei presupposti su cui si basano anche le tecniche più recenti.

L'anonimizzazione attraverso l'algoritmo k-anonymous, ad esempio, è stata ampiamente studiata grazie alla sua robustezza nella preservazione dai rischi di divulgazione dell'identità,

ed essendo ritenuta sufficientemente valida, viene utilizzata in numerosi ambiti, tra cui la sanità. La ricerca di una soluzione sempre più efficace, ha anche condotto all'introduzione di nuove varianti di quest'ultimo algoritmo, che tentano di facilitare e rafforzare l'applicazione del principio della k-anonimità. Posto infatti che l'anonimizzazione delle identità degli individui avvenga correttamente, ma assunto che non esiste un implementazione univoca della k-anonimizzazione, ciò che si cerca di ottenere è l'applicazione in grado di restituire il migliore guadagno di informazione anche dopo il processo de-identificativo.
Una tra le implementazioni di k-anonymous tuttora maggiormente utilizzate, è quella presentata da Kristen LeFevre et al. [29], definito come metodo **Mondrian**. Quest'ultimo utilizza un algoritmo di partizionamento per generare sottoinsiemi di dati che vengono in seguito protetti dalla presenza di almeno k record simili, in base a quanto stabilito dalla k-anonimità. Il partizionamento è il processo che permette di suddividere il dataset in gruppi o partizioni, puntando a raggruppare insieme i dati che condividono maggiori caratteristiche. Concettualmente simile al clustering, il partizionamento riesce ad agevolare la successiva applicazione di k-anonymous, contribuendo al contempo alla preservazione dell'utilità.
Generalmente questo procedimento avviene unidimensionalmente, ovvero prendendo in considerazione un attributo per volta, dal quale si procede con la suddivisione in partizioni. Il metodo Mondrian, invece, introduce un algoritmo di partizionamento ricorsivo, che si distingue per la capacità di gestire efficacemente la multidimensionalità dei dati. Effettuando le suddivisioni su più dati simultaneamente, si creano partizioni che soddisfano i requisiti per la k-anonimizzazione riducendo al minimo la perdita di informazioni. Per queste ragioni, il metodo Mondrian si è ampiamente affermato, rivelando un approccio determinante per una corretta esecuzione di k-anonymous.

3.3.1 Implementazioni pratiche

Per una migliore comprensione dei concetti teorici discussi in precedenza, si prosegue adesso con la presentazione di esempi pratici in grado di evidenziare i principi stabiliti dagli algoritmi di anonimizzazione e di mostrare le modifiche sostanziali che essi apportano ai dataset.

k-anonymous

Si consideri la Tabella 3.7, contenente i dati relativi a persone che hanno contratto un determinato virus e le patologie causate a seguito della contrazione. Il nome delle persone è un attributo identificatore diretto **I**, mentre età, cap e lo stato del virus, risultano essere quasi identificatori **Q** poiché in grado di suggerire informazioni rilevanti in combinazione con

dati ausiliari posseduti da un potenziale avversario. Infine, la patologia causata dal virus è l'informazione sensibile **S** che si desidera proteggere.

	ID	Quasi identificatori			Attributo sensibile
Tupla	**Nome**	**Età**	**Cap**	**Stato del Virus**	**Patologia causata**
1	Luca	31	10126	Sintomatico	Fibrosi polmonare
2	Alice	43	10143	Guarito	Nausea
3	Chiara	35	10152	Sintomatico	Raffreddore
4	Matteo	39	10156	Sintomatico	Febbre
5	Giulia	41	10145	Ricoverato	Polmonite
6	Simone	34	10123	Asintomatico	Nessuna
7	Paolo	37	10151	Sintomatico	Nausea
8	Carlo	32	10129	Ricoverato	Fibrosi polmonare
9	Sofia	43	10148	Sintomatico	Febbre

Tabella 3.7: contiene i dati riguardanti lo stato di salute a seguito della contrazione di un virus e la patologia causata associata.

Per procedere con un implementazione sensata occorre fare alcune considerazioni. Come spiegato nella Sezione 3.3, k-anonymous sfrutta i vantaggi ricavati dalla soppressione e dalla generalizzazione e combina queste due tecniche in modo che la tabella risponda dei requisiti per la k-anonimizzazione, in cui valori più alti di k implicano livelli di protezione maggiori. Nella Tabella 3.7, l'unico identificatore diretto **I** è il nome e, non restituendo informazioni significative per un'eventuale analisi del dataset, è sicuramente necessario procedere con la soppressione dell'intera colonna così da prevenire il riconoscimento diretto degli individui. I quasi identificatori **Q**, i quali contengono informazioni utili all'analisi, non possono essere del tutto eliminati: l'età dei partecipanti viene generalizzata in intervalli, nel cap si applica una sostituzione dei caratteri sopprimendo le cifre finali mentre l'attributo categorico, "stato del virus" subisce una generalizzazione che comporta l'inclusione di tutti gli attributi sotto una categoria meno specifica ma semanticamente equivalente.

La Tabella 3.8 rappresenta un esempio di versione 3-anonima della Tabella 3.7, dove EQ identifica le 3 classi di equivalenza generate. Si noti infatti che ogni classe di equivalenza è identificata specificamente da una combinazione tra l'intervallo dei valori di età e il cap, in cui è stato possibile sostituire solamente l'ultimo carattere. I valori dello "stato del virus" sono stati sostituiti dalla condizione generale di positività al virus, perdendo così l'utilità dell'informazione trasportata dall'attributo. Per affermare con certezza che la tabella sia 3-anonima, è necessario che ogni classe di equivalenza contenga almeno 3 elementi che condividono la stessa combinazione unica. Osservando la Tabella 3.8, ciascuna classe di

EQ		Quasi identificatori			Attributo sensibile
	Tupla	Età	Cap	Stato del Virus	Patologia causata
1	1	[30-35)	1012*	Positivo	Fibrosi polmonare
	6	[30-35)	1012*	Positivo	Nessuna
	8	[30-35)	1012*	Positivo	Fibrosi polmonare
2	2	[40-45)	1014*	Positivo	Nausea
	5	[40-45)	1014*	Positivo	Polmonite
	9	[40-45)	1014*	Positivo	Febbre
3	3	[35-40)	1015*	Positivo	Raffreddore
	4	[35-40)	1015*	Positivo	Febbre
	7	[35-40)	1015*	Positivo	Nausea

Tabella 3.8: versione 3-anonima della Tabella 3.7, EQ indica le classi di equivalenza.

equivalenza presenta esattamente 3 valori con la stessa combinazione di quasi identificatori **Q**, e il principio secondo cui ogni record deve essere indistinguibile da almeno altri (3-1) record della stessa tabella è soddisfatto. In base a questo criterio, la privacy dell'identità è tutelata, poiché anche se si conoscessero informazioni ausiliari riguardanti i quasi identificatori **Q** di una persona specifica della tabella, si avrebbero almeno tre opzioni plausibili di corrispondenza. Distinguere quale tra queste appartiene al bersaglio dell'avversario è impossibile, e la probabilità di individuare il record corretto è al massimo 1/k, nel caso in esame 1 su 3.

Sebbene l'algoritmo riesca ad occultare l'identità delle persone, rendendo ciascun individuo indistinguibile da altri 2 record della tabella, la sicurezza del dataset non è assoluta, e altri attacchi più ricercati riuscirebbero a trarre informazioni sensibili dalla tabella anonimizzata. k-Anonymous, non tiene conto di considerazioni poco evidenti che potrebbero ricondurre ad informazioni acquisite direttamente dagli attributi piuttosto che dalle identità individuali.
Attacchi di omogeneità o basati su conoscenze pregresse dell'avversario rappresentano quindi ancora una minaccia da controllare a causa dei rischi di divulgazione involontaria degli attributi che favoriscono. Ad esempio, se un avversario fosse a conoscenza del fatto che Carlo di 32 anni si trova all'interno del dataset, potrebbe facilmente ricondursi alla prima classe di equivalenza, senza però riuscire a distinguere quale tra i tre record corrisponda a quello di Carlo. Si supponga però, in aggiunta, che l'avversario sappia che Carlo, a causa del virus, ha sviluppato delle conseguenze **(background knowledge attack)**. Pur non sapendo quali, riesce a concludere facilmente che Carlo, in seguito al virus, ha sofferto di fibrosi polmonare. Ciò è dovuto al fatto che, nella suddetta classe di equivalenza, ci sono 2 valori per fibrosi polmonare e, escludendo la possibilità che la patologia causata sia "nessuna", l'unica alternativa è che il soggetto abbia sofferto di questa patologia. Anche non riuscendo a distinguere quale sia il

record corrispondente a Carlo tra l'1 e l'8, si può comunque dedurre l'informazione sensibile riferita alla sua patologia, a causa dell'omogeneità degli attributi sensibili **S** all'interno della prima classe di equivalenza, **(homogeneity attack)**.
Va inoltre considerato che alcune patologie, come la fibrosi polmonare, possono avere un grado di sensibilità maggiore rispetto ad altre patologie più lievi come un influenza e, ancor più, rispetto al non aver sofferto di nessuna patologia. La presenza in 2 casi su 3 del valore "fibrosi polmonare" nella prima classe di equivalenza, potrebbe comportare un rischio maggiore che il valore di quest'attributo più sensibile rispetto agli altri venga divulgato **(skewness attack)**.

La semplice k-anonimizzazione non è spesso sufficiente a realizzare una corretta anonimizzazione dei dati e, per questo motivo, quando i dataset contengono informazioni particolarmente sensibili o personali, va spesso abbinata ad altri meccanismi e algoritmi che ne rafforzino le capacità di protezione su più fronti.

l-diversity

L'algoritmo *l*-diversity, è in grado di proteggere le informazioni sugli attributi sensibili **S** imponendo, grazie al parametro *l*, un certo livello di diversità nella distribuzione degli attributi all'interno di ciascuna classe di equivalenza. Questa diversità, riesce a difendere i dati da attacchi che mirano alla divulgazione di informazioni sugli attributi, come quelli appena descritti. Introducendo la condizione imposta dalla *l*-diversità nella Tabella 3.8, si è in grado aumentare il grado di diversità degli attributi sensibili, distribuendoli omogeneamente nelle diverse classi di equivalenza. La Tabella 3.9 rappresenta una versione 3-diversa della Tabella 3.7. È importante effettuare alcune valutazioni su quest'ultima versione 3-diversa.

EQ		Quasi identificatori			Attributo sensibile
	Tupla	**Età**	**Cap**	**Stato del Virus**	**Patologia causata**
1	1	[30-40)	101**	Positivo	Fibrosi polmonare
	6	[30-40)	101**	Positivo	Nessuna
	3	[30-40)	101**	Positivo	Raffreddore
	8	[30-40)	101**	Positivo	Fibrosi polmonare
	4	[30-40)	101**	Positivo	Febbre
	7	[30-40)	101**	Positivo	Nausea
2	2	[40-45)	1014*	Positivo	Nausea
	5	[40-45)	1014*	Positivo	Polmonite
	9	[40-45)	1014*	Positivo	Febbre

Tabella 3.9: Versione 3-diversa della Tabella 3.7, con 2 classi di equivalenza

Per primo, bisogna ricordare che l-diversity risulta essere una estensione di k-anonymous, e presuppone la suddivisione della tabella in classi di equivalenza definiti q*-block. Nel caso specifico, si sono raggruppati i record in 2 q*-block di dimensioni diverse dove il primo contiene 6 record, mentre il secondo ne contiene 3: anche questa è una versione 3-anonima della tabella originale e, nonostante il primo q*-block contenga 6 record indistinguibili l'uno all'altro, la soglia di anonimità viene comunque definita dalla classe di equivalenza con meno record, che in questo caso contiene 3 record equivalenti.
Implementare l-diversity nella tabella 3.7, ha richiesto un'organizzazione differente delle classi di equivalenza rispetto alla sola implementazione di k-anonymous. Le soppressioni e le generalizzazioni effettuate in quest'ultima applicazione sono state più forti. Emerge, dunque, che riuscire ad implementare la condizione aggiuntiva dell'l-diversità oltre alla k-anonimità implica il rafforzamento delle tecniche di anonimizzazione effettuate, riducendo così ulteriormente l'utilità delle informazioni contenute nel dataset

Per dimostrare che la tabella sia anche 3-diversa, si può esaminare nuovamente quanto detto nella Sezione 3.3 sulla l-diversità, prendendo in considerazione alcune delle istanze descritte: ognuna delle classi di equivalenza contiene almeno 3 valori distinti secondo il concetto di l- diversità distinta. Si può verificare che la Tabella 3.9 rispecchia anche i requisiti della l-diversità entropica secondo cui l'entropia di ciascuna classe deve essere $\geq log(l)$, e si procede con il calcolo dell'entropia di ciascuna classe definita in precedenza nell'Equazione 3.2 come:

$$H(S) = -\sum_{s \in S} p_{(q^*,s)} \log(p_{(q^*,s)}) \quad \text{dove} \quad p_{(q^*,s)} = \frac{n_{(q^*,s)}}{\sum_{s' \in S} n_{(q^*,s')}}.$$

Il primo q*-block contiene 4 valori di attributi sensibili con frequenza di comparsa 1/6, e un valore (fibrosi polmonare) con frequenza 1/3. Si calcola dunque l'entropia come: $H(S) = -(4 \cdot \frac{1}{6} log(\frac{1}{6}) + \frac{1}{3} log(\frac{1}{3}))$ e attraverso un inversione della formula si ottiene che $l = e^{H(S)} = 4.76$. Nel secondo q*-block, sono presenti 3 valori diversi e si intuisce che l'entropia è massima e uguale a $log(3)$, da cui per la stessa formula si ricava un valore di l=3. Il valore del parametro l della tabella sarà quindi esattamente 3.

Introdurre il principio di l-diversità nella tabella aiuta a contrastare alcune problematiche sulla divulgazione degli attributi che una tabella anonimizzata solo attraverso k-anonymous non riesce a prevenire. Nello specifico, la 3-diversità impone un livello di diversità tale da rendere la distribuzione degli attributi sensibili **S** più omogenea all'interno di ciascuna classe di equivalenza, in modo tale che attacchi che sfruttano le conoscenze pregresse degli avversari o

che si basano sulla disomogeneità degli attributi, come quelli ipotizzati nell'esempio di tabella 3-anonima, siano difficili se non impossibili da applicare. Risulta chiaro che la protezione da questi possibili attacchi è proporzionale alla grandezza del valore l.

Infine, si può notare che, nonostante la combinazione dei due algoritmi, una tabella anonimizzata può comunque subire attacchi più complessi che sfruttano la similarità o l'asimmetria degli attributi. La presenza del valore "fibrosi polmonare" all'interno della stessa classe di equivalenza persiste, e il rischio di divulgare attributi con maggior grado di sensibilità non viene del tutto eliminato. Questa debolezza sarebbe potuta essere compensata implementando ad esempio i principi della t-closeness, secondo i quali la distribuzione locale degli attributi nelle rispettive classi di equivalenza deve tener conto della distribuzione totale all'interno dell'intera tabella. In questo modo, l'attributo più sensibile citato, si sarebbe potuto distribuire omogeneamente all'interno delle diverse classi di equivalenza e, impedendone la condensazione in un unica classe, si sarebbero ridotti i rischi dovuti alla similarità o alla asimmetria degli attributi.

t-closeness

La Tabella 3.10, tratta dal documento [28], racchiude un altro esempio di ipotetico dataset medico, il quale contiene le informazioni riguardanti i salari di alcuni pazienti e le malattie di cui soffrono. Codice ZIP ed età sono quasi identificatori **Q**, mentre stipendio e malattia sono attributi sensibili **S**.

Tabella 3.10: contiene i dati relativi agli stipendi e alle malattie di alcuni pazienti, tratto e adattato da [28].

Tupla	Codice ZIP	Età	Stipendio	Malattia
1	47677	29	3K	ulcera gastrica
2	47602	22	4K	gastrite
3	47678	27	5K	cancro allo stomaco
4	47905	43	6K	gastrite
5	47909	52	11K	influenza
6	47906	47	8K	bronchite
7	47605	30	7K	bronchite
8	47673	36	9K	polmonite
9	47607	32	10K	cancro allo stomaco

A causa della maggiore complessità del calcolo del parametro t, è interessante effettuare un'analisi dell'algoritmo t-closeness tramite un confronto diretto con l-diversity, cosi da mettere in risalto i punti di forza di quest'ultimo approccio nei confronti di un altro metodo consolidato.

Sia l-diversity che t-closeness, necessitano essere combinati insieme ad una k-anonimizzazione efficiente: il primo la richiede come presupposto per l'esecuzione dell'algoritmo, mentre il secondo la presuppone per garantire un livello efficiente di anonimizzazione. t-Closeness, infatti, riesce ad apportare vantaggi al dataset solamente dal punto di vista della protezione dai rischi di divulgazione degli attributi, senza garantire che la re-identificazione degli individui non avvenga se usato singolarmente. Di seguito si presentano le due versioni anonimizzate della Tabella 3.10, secondo i due algoritmi.

Tabella 3.11: versione 3-diversa della Tabella 3.10, tratto e adattato da [28].

Tupla	Codice ZIP	Età	Stipendio	Malattia
1	476**	2*	3K	ulcera gastrica
2	476**	2*	4K	gastrite
3	476**	2*	5K	cancro allo stomaco
4	4790*	≥ 40	6K	gastrite
5	4790*	≥ 40	11K	influenza
6	4790*	≥ 40	8K	bronchite
7	476**	3*	7K	bronchite
8	476**	3*	9K	polmonite
9	476**	3*	10K	cancro allo stomaco

Tabella 3.12: versione 0.167-closeness rispetto allo stipendio e 0.278-closeness rispetto alla malattia della Tabella 3.10, tratto e adattato da [28].

Tupla	Codice ZIP	Età	Stipendio	Malattia
1	4767*	≤ 40	3K	ulcera gastrica
3	4767*	≤ 40	5K	cancro allo stomaco
8	4767*	≤ 40	9K	polmonite
4	4790*	≥ 40	6K	gastrite
5	4790*	≥ 40	11K	influenza
6	4790*	≥ 40	8K	bronchite
2	4760*	≤ 40	4K	gastrite
7	4760*	≤ 40	7K	bronchite
9	4760*	≤ 40	10K	cancro allo stomaco

Le due tabelle, risultano essere versioni 3-anonime della tabella originale, poiché suddivise in tre classi di equivalenza con tre record ciascuna.
Si consideri dapprima la Tabella 3.11. Supponiamo che si sappiano informazioni su una vittima bersaglio che ci permettono di dedurre che la vittima in questione appartenga alla prima classe di equivalenza. In tal caso, si può dedurre che il suo stipendio rientri nell'intervallo [3K-5K], e che di conseguenza sia relativamente basso. Allo stesso modo, queste considerazioni

possono essere rivolte anche ad attributi categorici. Tenendo conto delle stesse supposizioni, si può ulteriormente dedurre che la vittima designata soffra anche di un qualche disturbo legato allo stomaco, poiché tutte le problematiche all'interno della classe di equivalenza sono correlate a quest'ultimo. Sebbene gli attributi sensibili **S**siano smistati equamente tra le classi di equivalenza, l-diversity non riesce a tenere conto della vicinanza semantica dei valori, lasciando trapelare di conseguenza informazioni sensibili importanti **(similarity attack)**.

Si procede calcolando la vicinanza tra gli attributi utilizzando l'EMD [28]. Per l'attributo numerico "stipendio", siano P_1 e P_2 due insiemi tali che $P_1 = \{3K, 4K, 5K\}$ e $P_2 = \{6K, 8K, 11K\}$, contenenti rispettivamente i valori della prima classe di equivalenza della Tabella 3.11 e i valori della seconda classe di equivalenza della Tabella 3.12, che rappresentano le distribuzioni dei valori sensibili all'interno delle classi di equivalenza corrispettive. Sia inoltre $R = \{3K, 4K, 5K, 6K, 7K, 8K, 9K, 10K, 11K\}$ la distribuzione complessiva, allora si ottengono $D[P1, R] = 0.375$, e $D[P2, R] = 0.167$ come valori della distanza tra distribuzioni locali e distribuzione complessiva. Per l'attributo categorico "malattia", utilizzando le appropriate considerazioni gerarchiche, si ottiene che la distanza tra la distribuzione della prima classe di equivalenza {ulcera gastrica, gastrite, cancro allo stomaco} e la distribuzione complessiva è 0.5. Se si confronta quest'ultimo dato con la distanza calcolata tra la distribuzione {ulcera gastrica, cancro allo stomaco, polmonite} della prima classe di equivalenza della Tabella 3.12 e quella complessiva, si ottiene il valore 0.278. Questo valore è decisamente inferiore rispetto a quello precedente, ciò implica che la vicinanza tra le informazioni della seconda tabella è minore rispetto a quella della prima. Se si assumessero le stesse supposizioni fatte per la tabella anonimizzata con l-diversity, in questo caso l'avversario non potrebbe dedurre nessun informazione di valore riguardante gli attributi sensibili **S**, perché entrambi presentano valori di t sufficientemente bassi da impedire correlazioni dei valori all'interno di classi di equivalenza dovute a similarità semantiche.

Come notato in precedenza, sebbene l'anonimizzazione effettuata con t-closeness, nel caso specifico, sia adeguatamente più efficace di quella di l-diversity, un maggior grado di anonimizzazione comporta spesso una minore utilità dei dati. Apportare una protezione più robusta agli attributi, ha costretto ad effettuare una generalizzazione maggiore del quasi identificatore **Q** "età", costituendo una perdita di informazioni per questo tipo di attributo. Conseguentemente, la scelta del metodo da implementare dipenderà sempre dai risultati che si vogliono ottenere dai dati a disposizione, sebbene i requisiti minimi di anonimizzazione vadano comunque garantiti per adempiere ai criteri stabiliti dalle normative attuali.

3.3.2 Misure di efficienza

Riuscire a definire opportunamente alcune metriche applicabili a tutti gli algoritmi agevolerebbe la comprensione delle migliori soluzioni per ciascun esigenza. Tramite un confronto diretto tra gli algoritmi, è possibile infatti valutare la qualità dell'anonimizzazione effettuata, così da individuare i punti di forza e le debolezze di ciascuna strategia. Tuttavia, stabilire quali metriche siano adeguate per questo scopo, risulta più complicato del previsto a causa della scarsità di misure standardizzate capaci di coprire la diversa casistica. Nella comunità statistica, si ricorre spesso a valutazioni della distribuzione dei dati, come calcolo di divergenze o della norma, o anche a misure dell'omogeneità del clustering. Altre metriche includono calcoli della distorsione dei dati che dipendono dalla gerarchia generalizzata dei valori, oppure misurazioni basate sull'utilità dei dati in un determinato scenario [30]. Queste ultime, in particolare, sono generalmente inadatte, poiché chi pubblica i dati non è spesso a conoscenza dell'ambito di applicazione di questi, altrimenti potrebbe eseguire il compito e condividere i risultati solamente con le parti interessate piuttosto che pubblicare l'intera tabella in forma anonima.
L'uso di metriche standardizzate rappresenta un ottimo passo per la valutazione oggettiva degli algoritmi di anonimizzazione, e permette inoltre di definire alcuni concetti che aiutano coloro che si occupano del processo di anonimizzazione. Grazie a queste misurazioni i risultati sono sempre più efficienti, sia in termini di sicurezza, sia come guadagno di informazione utile. Di seguito, si riportano alcune di queste metriche evidenziate in [30] in grado di restituire parametri particolarmente rilevanti, riflettendo al meglio il processo di standardizzazione.

Generalized Information Loss (GenILoss)

Questa metrica permette di quantificare la perdita generale di informazione dovuta all'adozione di tecniche di generalizzazione sul dataset. Quando si generalizza, si tende a trasformare un'informazione specifica in una meno specifica, e ciò comporta una perdita di valore dell'informazione misurabile come riduzione della precisione del dato: tanto più ampio è l'intervallo che contiene il valore specifico, maggiore sarà la perdita di informazione.
Più precisamente: siano L_i e U_i i limiti inferiori e superiori di un attributo i. Un valore per l'attributo i è generalizzato in un intervallo ij definito dai limiti inferiore L_{ij} e superiore U_{ij}. La perdita di informazione complessiva di una tabella anonimizzata T^* può essere calcolata come:

$$\text{GenILoss}(T^*) = \frac{1}{|T| \cdot n} \times \sum_{i=1}^{n} \sum_{j=1}^{|T|} \frac{U_{ij} - L_{ij}}{U_i - L_i}, \tag{3.7}$$

dove T è la tabella originale, n è il numero di attributi e $|T|$ è il numero di record nella tabella.

Valori più bassi di GenILoss(T^*) indicano minore perdita di informazione. Idealmente, 0 significa che nessuna trasformazione è stata applicata, mentre 1 indica livello di soppressione totale o di massima generalizzazione dei dati. Per calcolare la perdita per gli attributi categorici usando la formula, si può assegnare a ciascun attributo un valore numerico.
Ad esempio, prendendo in esame la Figura 3.2, per l'attributo stato del virus, a "sintomatico" viene assegnato 1, ad "asintomatico" viene assegnato 2, fino a "non vaccinato" a cui viene assegnato 6. Pertanto, lo stato "positivo" è rappresentato dall'intervallo [1-4], che copre gli stati da "sintomatico" a "guarito".

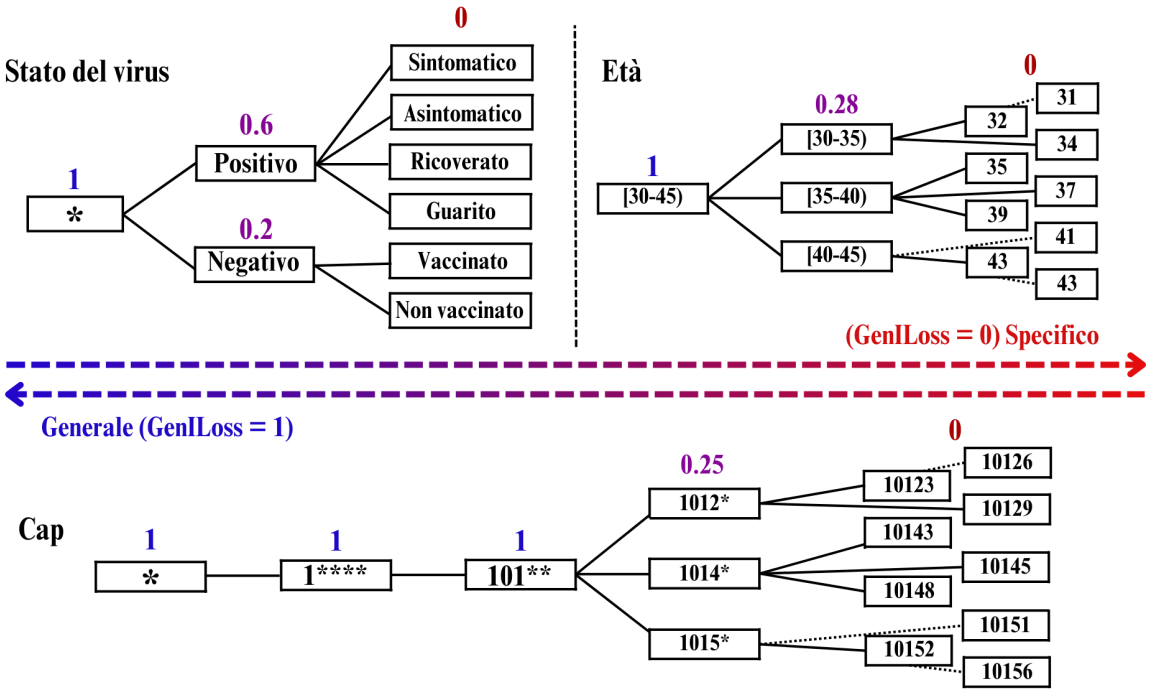


Figura 3.2: livello di gerarchia degli attributi numerici e categorici della Tabella 3.7; i valori a sinistra sono le generalizzazioni dei valori a destra. Gli attributi specifici hanno GenILoss uguale a 0 per definizione, poiché nessuna anonimizzazione è stata applicata, mentre gli attributi totalmente generalizzati hanno GenILoss uguale a 1, il che implica il non poter trarre più nessuna utilità da questi. Tratto e adattato da [30]

Per quanto riguarda il cap, dalla figura si evince che il GenILoss risulta essere uguale ad 1 dopo la soppressione di solo due caratteri: questo accade poiché il GenILoss è strettamente relazionato alla tabella per cui si calcola. Sebbene generalmente la soppressione di due caratteri sia meno incisiva rispetto alla soppressione del valore dell'intero attributo, nel caso in esame basta eliminare due cifre del cap per avere la perdita totale delle informazioni utili della Tabella 3.7. Tutte le informazioni specifiche sono, quindi, ugualmente riconducibili al valore di cap con due cifre soppresse.

Discernibility Metric (DM)

Questa metrica cerca di quantificare quanto un record sia indistinguibile dagli altri, assegnando a ciascuno di questi una penalità proporzionata alla classe di equivalenza a cui appartiene, essendo così in grado di quantificare quanto oppressiva sia stata l'anonimizzazione. Se un record è soppresso, gli viene assegnata una penalità pari alla dimensione della tabella originale. Il punteggio DM complessivo per una tabella T^* k-anonimizzata è definito da:

$$DM(T^*) = \sum_{\forall EQ\ s.t.\ |EQ| \geq k} |EQ|^2 + \sum_{\forall EQ\ s.t.\ |EQ| < k} |T| \cdot |EQ|, \tag{3.8}$$

dove T è la tabella originale, $|T|$ è il numero di record e $|EQ|$ è la dimensione delle classi di equivalenza create dopo aver eseguito l'anonimizzazione. L'idea è che EQ più grandi costituiscono una maggiore perdita di informazioni; anche in questo caso, quindi, valori più bassi del parametro indicano una minore anonimizzazione dei dati.

Average Equivalence Class Size Metric (C_{AVG})

Rappresenta una misurazione della dimensione media delle classi di equivalenza EQ, grazie alla quale si può stabilire se la loro creazione è stata fatta nel modo migliore, ovvero il caso in cui ogni record si trovi all'interno di una classe di equivalenza con k record. Un valore di 1 indica quindi una suddivisione ideale, in cui la dimensione delle EQ è pari e non maggiore al valore k. Il punteggio C_{AVG} complessivo per una tabella anonimizzata T^* è dato da:

$$C_{AVG}(T^*) = \frac{|T|}{|EQs| \cdot k} \tag{3.9}$$

dove T è la tabella originale, $|T|$ è il numero di record, $|EQs|$ è il numero totale di classi di equivalenza create e k è il requisito di privacy dettato dalla k-anonimizzazione.

Calcolo delle metriche nelle implementazioni effettuate

Come ultimo aspetto è interessante valutare i parametri descritti nelle tabelle implementate nella Sezione 3.3, così da rafforzare il confronto diretto tra gli algoritmi esaminati in questa tesi e trarre conclusioni più solide riguardo all'efficacia e all'applicabilità di ciascuno di essi.

Prendendo in esame la tabella 3.8, che è 3-anonima, e considerando la Figura 3.2 che illustra il livello di generalizzazione degli attributi, si calcola il GenILoss per ogni attributo in relazione al raggruppamento a cui è soggetto.

Per l'attributo numerico "Età" si calcola il GenILoss del valore [30-35), che è uguale a $\frac{34-30}{44-30} = \frac{4}{14}$, poiché l'intervallo totale di valori va dal valore massimo plausibile all'interno della tabella, che è 44, al valore minimo, che è 30. Per l'attributo "Cap" (che va trattato come attributo categorico), il valore del GenILoss di 1012* è uguale a $\frac{3-1}{9-1} = \frac{2}{8}$, infatti 1012* racchiude i primi 3 valori della tabella anonimizzata, ma i valori plausibili di cap sono in tutto 9. Infine per l'attributo categorico "Stato del virus", il GenILoss di "positivo" vale $\frac{4-1}{6-1} = \frac{3}{5}$, poiché per le stesse considerazioni di prima sulla Figura 3.2, i valori plausibili sono 6 in totale, ma nella tabella anonimizzata vengono tutti inclusi dal valore "positivo" che racchiude dal primo al quarto valore (4-1). Poiché tutti i valori di GenILoss calcolati per ogni tipo di attributo sono ugualmente calcolabili per tutti gli altri casi di attributi con gli stessi risultati, allora, in base all'Equazione 3.7, moltiplicando ciascun valore per il numero di casi simili si ottiene il valore finale di GenILoss(T^*) dalla formula corrispettiva:

$$\text{GenILoss}(T^*) = \frac{1}{9 \cdot 3} \cdot (9 \cdot \frac{5}{14} + 9 \cdot \frac{2}{8} + 9 \cdot \frac{3}{5}) = 0.3785.$$

Dato che ogni classe di equivalenza della Tabella 3.8 contiene 3 elementi, e considerato che il livello di k-anonimità è proprio 3, per calcolare il punteggio di DM basta considerare la prima sommatoria dell'Equazione 3.8 corrispettiva. Si ottiene:

$$DM(T^*) = 3^2 + 3^2 + 3^2 = 27.$$

Per calcolare infine il C_{AVG} si applica l'Equazione 3.9, e nel caso specifico, poichè la dimensione media delle classi è proprio uguale al livello di k-anonimizzazione (3), si ottiene:

$$C_{AVG}(T^*) = \frac{9}{3 \cdot 3} = 1,$$

ovvero il minimo valore possibile, che rappresenta il raggruppamento in classi di equivalenza ideale.

Attraverso i medesimi calcoli, si possono trovare i valori delle metriche descritte per ognuna delle tabelle anonimizzate nella sezione precedente. I risultati sono stati racchiusi nelle due tabelle sottostanti così da poter effettuare un confronto immediato tra i livelli di efficienza dei diversi algoritmi. La Tabella 3.13, compara la 3-anonimità della Tabella 3.8 con la 3-diversità della Tabella 3.9, mentre la Tabella 3.14, compara la 3-diversità della Tabella 3.11 con la 0.278-closeness della Tabella 3.12.

Ricordando che valori maggiori indicano una maggiore perdita di utilità del dataset, i risultati evidenziano alcuni aspetti interessanti delle implementazioni effettuate.

	GenILoss	DM	C_{AVG}
3-anonymous	0.378	27	1
3-diversity	0.541	45	1.5

Tabella 3.13: confronto delle metriche tra la Tabella 3.8 e la Tabella 3.9

	GenILoss	DM	C_{AVG}
3-diversity	0.416	27	1
0.278-closeness	0.391	27	1

Tabella 3.14: confronto delle metriche tra la Tabella 3.11 e la Tabella 3.12

Il confronto tra la 3-anonimità e la 3-diversità nella Tabella 3.13, dimostra che aver implementato l-diversity ha richiesto la perdita di parte di informazione, oltre che un aumento sia del DM che del C_{AVG}, mettendo in risalto la debolezza della suddivisione in classi di equivalenza effettuata.
Nella Tabella 3.14, il confronto tra la 3-diversità e la 0.278 t-closeness segnala una minore perdita di informazioni in seguito all'applicazione di t-closeness. Ciò ha confermato che, oltre alla maggiore protezione dai rischi di divulgazione degli attributi fornita dall'algoritmo t-closeness, complessivamente, l'anonimizzazione è stata meno incisiva, preservando più informazioni utili.

Le considerazioni effettuate non sono comunque generalizzabili a qualsiasi tipo di scenario ma, riuscire a identificare metriche standardizzate di valutazione dell'utilità dei dati, può essere un ausilio importante nella determinazione della strada da percorrere nei processi di anonimizzazione dei dati.

4 Conclusione e sviluppi futuri

Il presente elaborato si è posto l'obiettivo di analizzare i principi e il funzionamento delle tecniche e degli algoritmi per l'anonimizzazione dei dati sensibili, mettendo in evidenza i vantaggi e le limitazioni di ciascuna strategia in relazione ai rischi di re-identificazione che non possono mai essere completamente eliminati. L'introduzione alle normative ha consentito di contestualizzare lo scenario in cui tali tecniche si collocano, sottolineando la costante necessità di miglioramento sia delle leggi vigenti sulla protezione dei dati, sia delle metodologie impiegate. Queste ultime, infatti, garantiscono un significativo guadagno di informazioni, favorendo il progresso e lo sviluppo della ricerca scientifica. L'analisi dei rischi di re-identificazione e dei possibili attacchi alla privacy, che rappresentano minacce da contrastare anche nel caso di de-identificazioni presumibilmente accurate, ha fornito, inoltre, una solida giustificazione all'implementazione delle attuali normative da parte degli enti governativi.
Le descrizione delle tecniche generali ha dimostrato la varietà di approcci esistenti per l'anonimizzazione dei dati sensibili, mentre, attraverso l'analisi degli algoritmi, è stato possibile approfondire le problematiche concrete che qualsiasi istituzione, intenzionata a utilizzare tali dati, deve affrontare. k-Anonymous rappresenta il fondamento su cui si basa la validità degli altri algoritmi e, pur non soddisfacendo tutti i requisiti necessari per un'anonimizzazione ideale, costituisce comunque un passaggio cruciale nella protezione degli individui, specialmente in relazione al rischio di divulgazione della loro identità. Gli algoritmi l-diversity e t-closeness si configurano come evoluzioni di k-anonymous che, se implementate nei giusti termini, completano e rafforzano il processo di anonimizzazione, mitigando ulteriormente i rischi di re-identificazione.
Le analisi condotte e le implementazioni effettuate hanno contribuito a delineare meglio nella pratica i principi teorici esplorati, mettendo in risalto il binomio privacy-utilità come elemento chiave in tutti i processi de-identificativi: ogni approccio adottato deve garantire un livello di anonimizzazione tale da consentire un equilibrio tra l'utilità dei dati anonimizzati e la protezione della privacy. Poiché l'adozione di tecniche più rigorose tende a ridurre l'utilità dei dati, la scelta della soluzione più appropriata richiede la considerazione di un numero maggiore di fattori. Riuscire a quantificare l'utilità generale dei dataset anonimizzati attraverso parametri e metriche standardizzate sembra essere, infine, un ausilio importante per il processo di de-identificazione stesso.

Il motivo per cui le problematiche discusse emergono risiede nell'utilizzo di enormi quantità di dati, che consentono di effettuare confronti e analisi statistiche in grado di evidenziare tendenze

comuni e schemi ricorrenti. I benefici derivanti da queste informazioni risultano particolarmente significativi, aprendo nuove opportunità su diversi fronti, con un impatto rilevante soprattutto nel settore medico. Poiché la tutela della privacy è una necessità imprescindibile, l'obbiettivo primario da perseguire è lo sviluppo di sistemi sempre più efficienti in grado di compromettere solo in minima parte gli attributi dei dataset. In questo contesto, il miglioramento della standardizzazione di alcune metriche per la valutazione dei parametri di anonimizzazione rappresenta un passo importante verso il perfezionamento di tali processi.

Attualmente, una sfida significativa è legata all'ottimizzazione delle impostazioni interattive che, sebbene sfruttino meccanismi più complessi, mostrano risultati promettenti oltre ad una versatilità superiore. Tra queste, nonostante non sia stato possibile approfondirne i dettagli in questo elaborato, la differential privacy emerge come uno degli approcci più innovativi, grazie alla sua capacità di sfruttare pienamente le in formazioni contenute nel dataset, pur mantenendolo inalterato e protetto dal custode delle informazioni. Eliminando la necessità di pubblicare o condividere direttamente il dataset, si rimuove un ostacolo non indifferente, permettendo di concentrare gli sforzi esclusivamente sullo sviluppo di meccanismi sempre più efficienti, in grado di restituire le informazioni richieste dalle query senza lasciar trapelare dettagli sulla composizione del dataset.
Soluzioni contemporanee, inotre, integrano dati medici con tecniche di machine learning, sfruttando appieno il valore dei dati sanitari, pur incrementando il rischio di esposizione di questi. Per affrontare il problema, vengono adottati meccanismi che combinano la differential privacy con l'uso dell'albero decisionale (differential privacy and decision tree, DPDT), che constiste in un modello di apprendimento automatico capace di creare una classificazione con struttura ramificata.

Va sottolineato che, in ambito medico, i dati da anonimizzare si presentano sotto molteplici forme. Ad esempio, anche le immagini derivanti dai vari apparecchi biomedici, come ecografie o risonanze magnetiche (RM), possono contenere informazioni estremamente utili per la ricerca, la cui tutela va garantita in egual misura. La differential privacy offre una soluzione per bilanciare la privacy e l'utilità delle immagini, applicando rumore ai pixel in modo controllato, al fine di alterare i dettagli sensibili preservando al contempo le caratteristiche globali dell'immagine. Tuttavia, le immagini biomediche rappresentano ancora una sfida notevole per l'anonimizzazione, e sviluppi futuri in tal senso potrebbero concentrarsi sull'impiego di tecniche avanzate di deep learning.

Bibliografia

[1] European Commission, *Data Protection in the EU*, https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_it, Accessed: 2024-10-09, 2024.

[2] RAND Corporation, «Review of the European Data Protection Directive,» *RAND Europe*, 2009, Accessed: 2024-10-09. indirizzo: https://afyonluoglu.org/PublicWebFiles/Reports/PDP/international/2009%20RAND%C2%A0Review%20of%20the%20European%20Data%20Protection%20Directive.pdf.

[3] A. Agaidis, «GDPR Case Study: Skelleftea School Board,» *Brown University GDPR Case Studies*, 2020, Accessed: 2024-10-09. indirizzo: https://cs.brown.edu/courses/csci2390/2020/assign/gdpr/agaidis-skelleftea-school-board.pdf.

[4] NDO3, «GDPR Case Study: Austrian Post,» *Brown University GDPR Case Studies*, 2021, Accessed: 2024-10-09. indirizzo: https://cs.brown.edu/courses/csci2390/2021/assign/gdpr/ndo3-austrian-post.pdf.

[5] S. Mbonihankuye, A. Nkunzimana e A. Ndagijimana, «Healthcare Data Security Technology: HIPAA Compliance,» *Wireless Communications and Mobile Computing*, vol. 2019, p. 7, 2019, Accessed: 2024-10-09. doi: 10.1155/2019/1927495. indirizzo: https://doi.org/10.1155/2019/1927495.

[6] C. S. Harris, R. A. Pozzar, Y. Conley et al., «Big Data in Oncology Nursing Research: State of the Science,» *Seminars in Oncology Nursing*, vol. 39, n. 3, p. 151 428, 2023, Epub 2023 Apr 19. doi: 10.1016/j.soncn.2023.151428.

[7] A. Gadotti e et al., «Anonymization: The imperfect science of using data while preserving privacy,» *Science Advances*, vol. 10, n. 6, eadn7053, 2024. doi: 10.1126/sciadv.adn7053.

[8] L. Sweeney, «Unmasking Patients in the Medical Data Privacy System,» *Data Privacy Lab*, 2000, Accessed: 2024-10-09. indirizzo: https://dataprivacylab.org/projects/identifiability/paper1.pdf.

[9] Z. Zuo, M. Watson, D. Budgen, R. Hall, C. Kennelly e N. Al Moubayed, «Data Anonymization for Pervasive Health Care: Systematic Literature Mapping Study,» *JMIR Medical Informatics*, vol. 9, n. 10, e29871, 2021. doi: 10.2196/29871.

[10] K. El Emam, *Guide to the De-Identification of Personal Health Information*, Illustrated. CRC Press, 2013, p. 413, isbn: 9781482218800.

[11] B. Riedl, V. Grascher, S. Fenz e T. Neubauer, «Pseudonymization for Improving the Privacy in E-Health Applications,» in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, 2008, pp. 255–255. doi: 10.1109/HICSS.2008.366.

[12] R. Tinabo, F. Mtenzi e B. O'Shea, «Anonymisation vs. Pseudonymisation: Which One is Most Useful for Both Privacy Protection and Usefulness of E-Healthcare Data,» in *2009 International Conference for Internet Technology and Secured Transactions (ICITST)*, 2009, pp. 1–6. doi: 10.1109/ICITST.2009.5402501.

[13] J. Rumbold e B Pierscionek, «The Effect of the General Data Protection Regulation on Medical Research,» *Journal of Medical Internet Research*, vol. 19, n. 2, e47, 2017. doi: 10.2196/jmir.7108.

[14] A. Pfitzmann e M. Köhntopp, «Anonymity, Unobservability, and Pseudonymity — A Proposal for Terminology,» in *Designing Privacy Enhancing Technologies*, ser. Lecture Notes in Computer Science, H. Federrath, cur., vol. 2009, Berlin, Heidelberg: Springer, 2001, pp. 1–9. doi: 10.1007/3-540-44702-4_1.

[15] «Practicing Differential Privacy in Health Care: A Review,» *Transactions on Data Privacy*, vol. 5, pp. 35–67, 2013.

[16] C. Dwork, F. McSherry, K. Nissim e A. Smith, «Calibrating Noise to Sensitivity in Private Data Analysis,» in *Theory of Cryptography*, ser. Lecture Notes in Computer Science, S. Halevi e T. Rabin, cur., vol. 3876, Berlin, Heidelberg: Springer, 2006, pp. 265–284. doi: 10.1007/11681878_14.

[17] Z. Sun, Y. Wang, M. Shu, R. Liu e H. Zhao, «Differential Privacy for Data and Model Publishing of Medical Data,» *IEEE Access*, vol. 7, pp. 146 074–146 087, 2019. doi: 10.1109/ACCESS.2019.2947295.

[18] A. Alnemari, C. J. Romanowski e R. K. Raj, «An Adaptive Differential Privacy Algorithm for Range Queries over Healthcare Data,» in *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 2017, pp. 397–402. doi: 10.1109/ICHI.2017.49.

[19] J. F. Marques e J. Bernardino, «Analysis of Data Anonymization Techniques,» in *Proceedings of the 9th International Conference on Data Science, Technology and Applications*, SCITEPRESS - Science e Technology Publications, 2020, pp. 252–259. doi: 10.5220/0010142302520259.

[20] K. M. Chong, «Privacy-preserving healthcare informatics: a review,» *ITM Web of Conferences*, vol. 36, p. 10, 2021, Published online: 26 January 2021. doi: 10.1051/itmconf/20213604005. indirizzo: https://doi.org/10.1051/itmconf/20213604005.

[21] I. S. Rubinstein e W. Hartzog, «Anonymization and risk,» *Wash. L. Rev.*, vol. 91, p. 703, 2016.

[22] S. Murthy, A. A. Bakar, F. A. Rahim e R. Ramli, «A comparative study of data anonymization techniques,» in *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, IEEE, 2019, pp. 306–309.

[23] C. K. Liew, U. J. Choi e C. J. Liew, «A data distortion by probability distribution,» *ACM Trans. Database Syst.*, vol. 10, n. 3, pp. 395–411, set. 1985, issn: 0362-5915. doi: 10.1145/3979.4017. indirizzo: https://doi.org/10.1145/3979.4017.

[24] G. J. Matthews e O. Harel, «Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy,» *Statistics Surveys*, vol. 5, pp. 1–29, 2011, issn: 1935-7516. doi: 10.1214/11-SS074.

[25] D. Defays e M. Anwar, «Masking Microdata Using Micro-Aggregation,» *Journal of Official Statistics*, vol. 14, n. 4, pp. 449–461, 1998. indirizzo: https://www.proquest.com/scholarly-journals/masking-microdata-using-micro-aggregation/docview/1266844123/se-2.

[26] K. Rajendran, M. Jayabalan e M. E. Rana, «A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data,» *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 17, n. 12, pp. 172–177, 2017, Manuscript received December 5, 2017, revised December 20, 2017.

[27] A. Machanavajjhala, D. Kifer, J. Gehrke e M. Venkitasubramaniam, «L-diversity: Privacy beyond k-anonymity,» *ACM Trans. Knowl. Discov. Data*, vol. 1, n. 1, 3–es, mar. 2007, issn: 1556-4681. doi: 10.1145/1217299.1217302. indirizzo: https://doi.org/10.1145/1217299.1217302.

[28] N. Li, T. Li e S. Venkatasubramanian, «t-Closeness: Privacy Beyond k-Anonymity and l-Diversity,» in *2007 IEEE 23rd International Conference on Data Engineering*, 2007, pp. 106–115. doi: 10.1109/ICDE.2007.367856.

[29] K. LeFevre, D. DeWitt e R. Ramakrishnan, «Mondrian Multidimensional K-Anonymity,» in *22nd International Conference on Data Engineering (ICDE'06)*, 2006, pp. 25–25. doi: 10.1109/ICDE.2006.101.

[30] V. Ayala-Rivera, P. McDonagh, T. Cerqueus e L. Murphy, «A Systematic Comparison and Evaluation of k-Anonymization Algorithms for Practitioners,» *Transactions on Data Privacy*, vol. 7, n. 3, pp. 337–370, 2014, Available online: http://www.tdp.cat/issues11/abs.a169a14.php, issn: 1888-5063. indirizzo: http://hdl.handle.net/10197/9109.