UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA BIOMEDICA

"Pseudonimizzazione e anonimizzazione per il trattamento dei dati sanitari in accordo alla legislazione vigente"

Relatore: Prof. Giovanni Sparacino

Laureanda: Mariasole Pasinato

ANNO ACCADEMICO 2021 – 2022

Data di laurea 22/09/2022

Indice

Sommario

L'avvento della telemedicina e di servizi come il dossier sanitario elettronico ha permesso un grande passo avanti verso una gestione sempre più veloce ed efficiente dei pazienti e del processo di diagnosi e di cura. Oltre a ciò, la digitalizzazione dei dati sanitari e la loro raccolta hanno consentito la realizzazione di studi farmacologici ed epidemiologici molto complessi e su grande scala, dando un enorme slancio a nuove terapie e campagne di prevenzione. Tra le sfide più grandi che questa rivoluzione nella sanità ha portato con sé, vi è di certo quella di trovare metodologie sicure ed efficaci per il trattamento dei dati personali digitalizzati, la cui mole risulta sempre più imponente e il cui valore è ormai incommensurabile.
Questo elaborato si propone di valutare come la legislazione in vigore si esprime a proposito di privacy e gestione di informazioni sensibili, indagando, in particolare, il significato di *pseudonimizzazione* e *anonimizzazione*. La trattazione che segue vuole dare un'ampia panoramica dei suddetti temi, non entrando in dettagli eccessivamente tecnici (se non dove necessario), ma cercando comunque di toccare con il dovuto grado di approfondimento tutti i punti cardine della materia.

In particolare, nel Capitolo 1 verrà presa in esame la normativa vigente a livello europeo sul tema della privacy (*General Data Protection Regulation*), per poi proseguire con una breve descrizione dell'evoluzione della normativa italiana in merito, fino al suo stato attuale. Il Capitolo 2 si occuperà di definire il concetto di *pseudonimizzazione*, affrontando le principali tecniche e strategie di implementazione in questo ambito. Infine, nel Capitolo 3 si cercherà di spiegare cosa significa *anonimizzare* un documento, quali sono le metodologie più diffuse ed efficaci per farlo e in quali contesti questa strategia risulta maggiormente utile.

Capitolo 1:
La legislazione vigente in materia di privacy

Già nel 1950, con l'adozione, nell'ambito del Consiglio d'Europa, della *Convenzione europea per la salvaguardia dei diritti dell'uomo e delle libertà fondamentali*, l'Unione Europea riconosceva il rispetto delle informazioni personali come un diritto inderogabile. Si legge, infatti, nell'art. 8:

> *"Ogni persona ha diritto al rispetto della propria vita privata e familiare, della propria casa e della propria corrispondenza"* [1].

Benché il concetto di *dato personale* risalga a tempi ben più recenti, è interessante notare come, fin dagli albori della legislazione europea, la questione della privacy sia sempre stata di grande interesse dal punto di vista giuridico.

In questa sezione verrà dunque fatta una breve panoramica dello stato attuale della normativa in materia di protezione dei dati personali, prima a livello europeo e successivamente a livello nazionale, con riferimento anche a pietre miliari della legislazione meno recente.

1.1 Il GDPR

Il *General Data Protection Regulation* (*Regolamento Generale sulla Protezione dei Dati*) è la normativa europea di riferimento in materia di gestione e protezione dei dati personali. Entrato in vigore il 24 maggio 2016 e divenuto operativo a partire dal 25 maggio 2018, il GDPR sostituisce, ampliandola, la *Direttiva 95/46/CE del Parlamento e del Consiglio d'Europa*.

In quanto Regolamento, una volta entrato in vigore, le sue norme sono vincolanti per tutti coloro che sono soggetti al rispetto del diritto dell'UE (a differenza, ad esempio, di ciò che accade con le Direttive); gli Stati dell'Unione, dunque, non possono in alcun modo prendere provvedimenti per limitarne l'applicazione.

Il Regolamento è nato principalmente con i seguenti obiettivi:

1. armonizzare l'applicazione della protezione dei dati personali all'interno dell'Unione (fatto che la Direttiva 95/46/CE non aveva il potere di garantire);

2. alimentare la fiducia dei cittadini nei confronti della società e dei servizi digitali, così da favorire lo sviluppo di un Mercato Unico Digitale;
3. rispondere alle sfide derivanti dallo sviluppo delle nuove tecnologie.

Il contenuto del GDPR si snoda attraverso 99 articoli e 173 "considerando". L'analisi dell'intero testo esula dagli scopi di questo elaborato; verranno quindi esposte le sole sezioni il cui contenuto risulta "ingegneristicamente" utile alla definizione e all'implementazione di tecniche di anonimizzazione e pseudonimizzazione, nonché a fornire un resoconto dei principi generali da applicare per la protezione dei dati [2] [3].

1.1.1 Articolo 4 – Definizioni

Si riportano, complete o parziali, solo alcune delle definizioni presentate nell'art. 4 del Regolamento.

Innanzitutto, si definisce *dato personale* una *"qualsiasi informazione riguardante una persona fisica identificata o identificabile («interessato»)"*. Con questa definizione, si sottintende dunque che, nel momento in cui l'*interessato* non sia più identificabile, non è più possibile parlare di *dato personale*. In particolar modo, *"si considera identificabile la persona fisica che può essere identificata, direttamente o indirettamente, con particolare riferimento a un identificativo come il nome, un numero di identificazione, dati relativi all'ubicazione, un identificativo online o a uno o più elementi caratteristici della sua identità fisica, fisiologica, genetica, psichica, economica, culturale o sociale"*. Vedremo come una raccolta di dati sottoposti ad *anonimizzazione* (per la definizione, si faccia riferimento alla *sezione 3.1*) non consti più di dati personali e, di conseguenza, non sia soggetta alle norme che regolano questi ultimi.

In riferimento ai dati personali, l'articolo riporta la definizione di casi particolari di questi, come i *dati relativi alla salute* e i *dati genetici*. Ogniqualvolta nel resto dell'elaborato si parlerà di dati personali, si tenga dunque presente che i dati sanitari rientrano a tutti gli effetti all'interno di questo macrogruppo.

Si considerino ora alcune definizioni significative dal punto di vista giuridico:

- *trattamento*: *"qualsiasi operazione o insieme di operazioni, compiute con o senza l'ausilio di processi automatizzati e applicate a dati personali o insiemi di dati personali […]";*

- *titolare del trattamento*: *"la persona fisica o giuridica, l'autorità pubblica, il servizio o altro organismo che, singolarmente o insieme ad altri, determina le finalità e i mezzi del trattamento di dati personali […]"*;
- *responsabile del trattamento*: *"la persona fisica o giuridica, l'autorità pubblica, il servizio o altro organismo che tratta dati personali per conto del titolare del trattamento"*.

In merito al ruolo del titolare del trattamento, sarà spiegato brevemente nelle prossime sezioni come la detenzione di dati personali sia sempre subordinata all'esistenza di un fine, che deve assolutamente essere dichiarato prima della raccolta degli stessi.

Per concludere, il GDPR riporta anche una prima definizione del termine *pseudonimizzazione* come *"trattamento dei dati personali in modo tale che i dati personali non possano più essere attribuiti a un interessato specifico senza l'utilizzo di informazioni aggiuntive, a condizione che tali informazioni aggiuntive siano conservate separatamente e soggette a misure tecniche e organizzative intese a garantire che tali dati personali non siano attribuiti a una persona fisica identificata o identificabile"*. In altre parole, una raccolta di dati pseudonimizzati può ancora essere ricondotta ad un interessato (quindi si tratta ancora di dati identificabili e, di conseguenza, di dati personali), ma il processo di identificazione può essere svolto solo in determinate circostanze e con l'ausilio di informazioni terze mantenute separate dai dati di interesse. In sostanza, la pseudonimizzazione consiste nel sostituire i dati direttamente identificativi, come cognome e nome, con dati indirettamente identificativi (e.g. alias, numero di classificazione…).

1.1.2 Articolo 5 – Principi applicabili al trattamento di dati personali

Il GDPR espone le regole da applicare nella raccolta di dati personali. I principi guida che vengono indicati sono

1. *liceità, correttezza e trasparenza;*
2. *limitazione della finalità:* i dati personali possono essere raccolti solo per finalità determinate e trattati solo compatibilmente a tale finalità;
3. *minimizzazione dei dati:* devono essere raccolti solo i dati che risultano necessari e pertinenti con la finalità dichiarata;
4. *esattezza;*

5. *limitazione della conservazione:* i dati personali possono essere detenuti solo per il periodo necessario al conseguimento della finalità con cui sono stati raccolti;
6. *integrità e riservatezza:* i dati raccolti devono essere trattati in modo tale da garantire un'adeguata sicurezza e protezione, mediante strategie e strumenti tecnologici all'avanguardia.

1.1.3 Articolo 9 – Trattamento di categorie particolari di dati personali

Il trattamento di dati personali quali i dati relativi alla salute e i dati genetici è concesso esclusivamente in determinati casi (se ne elencano alcuni esempi):

- l'interessato ha espresso un consenso esplicito per una o più finalità specifiche;
- il trattamento risulta necessario per la tutela di un interesse vitale dell'interessato;
- il trattamento risulta necessario per finalità di medicina preventiva, medicina del lavoro, diagnosi, assistenza, terapia sanitaria o sociale;
- il trattamento risulta utile in termini di sanità pubblica;
- il trattamento risulta utile per il progresso della ricerca scientifica o a fini statistici.

1.1.4 Articolo 25 – Protezione dei dati fin dalla progettazione e protezione dei dati per impostazione predefinita

Tenendo conto dello stato dell'arte, al fine di attuare efficacemente i principi di protezione dei dati di cui all'art. 5, il titolare del trattamento deve garantire la messa in atto di *"misure tecniche e organizzative adeguate, quali la pseudonimizzazione"*. Questi provvedimenti devono essere previsti già a partire dalla progettazione dei mezzi del trattamento e la protezione dei dati deve avvenire per impostazione predefinita, cioè a prescindere (si parla di *privacy by design* e *privacy by default*).

1.1.5 Articolo 32 – Sicurezza del trattamento

Tra le misure tecniche che il titolare del trattamento è tenuto ad implementare, il GDPR annovera

1. la pseudonimizzazione e la cifratura dei dati;

2. la capacità di assicurare permanentemente che i sistemi e i servizi di trattamento siano riservati, integri, disponibili e resilienti;
3. in caso di incidente, la capacità di ripristinare tempestivamente la disponibilità e l'accesso dei dati personali;
4. una procedura per testare e valutare su base regolare l'efficacia delle misure stesse.

1.2 La legislazione italiana

1.2.1 Decreto Legislativo 30 giugno 2003, n. 196 – Codice in materia di protezione dei dati personali

Il *D.Lgs. 196/03* [4] è una delle pietre miliari della legislazione italiana per quanto riguarda il diritto alla privacy.

Per dare attuazione alla Direttiva 95/46/CE, era già stata emanata la *Legge 31 dicembre 1996 n. 675*, a cui però, nel corso degli anni, si erano affiancate ulteriori leggi in merito a specifici scenari e applicazioni della protezione dei dati personali. A ciò andava inoltre ad aggiungersi la giurisprudenza della Suprema Corte di Cassazione in materia di privacy. Il Decreto Legislativo del 2003 venne dunque redatto per riordinare interamente la normativa fino a quel momento esistente [5].

Poiché sia il GDPR che il D.Lgs. 196/03 hanno come fondamento la Direttiva 95/46/CE, esistono alcuni punti di contatto tra i due testi, a partire dalle definizioni presentate, fino ad arrivare agli obblighi di progettazione e ai principi guida da seguire. Di interesse notare come già in questo Decreto si trattasse il tema della protezione dei dati tramite tecniche di cifratura (cfr. *art. 34 lettera h)* del testo).

1.2.2 Decreto Legislativo 10 agosto 2018, n. 101 – Disposizioni per l'adeguamento della normativa nazionale al regolamento UE 2016/679

Il *D.Lgs. 101/18* [6] è stato redatto, a scopo attuativo, in seguito all'emanazione del GDPR.

Importante porre l'attenzione sul fatto che il nuovo Decreto non sia totalmente abrogativo nei confronti del vecchio codice del 2003 (alcuni articoli infatti risultano comunque compatibili con il nuovo Regolamento europeo).

A partire dal 19 settembre 2018, data di entrata in vigore del testo, la legislazione italiana in materia di privacy risulta quindi totalmente allineata con il GDPR e, dunque, con la normativa degli altri Paesi dell'Unione.

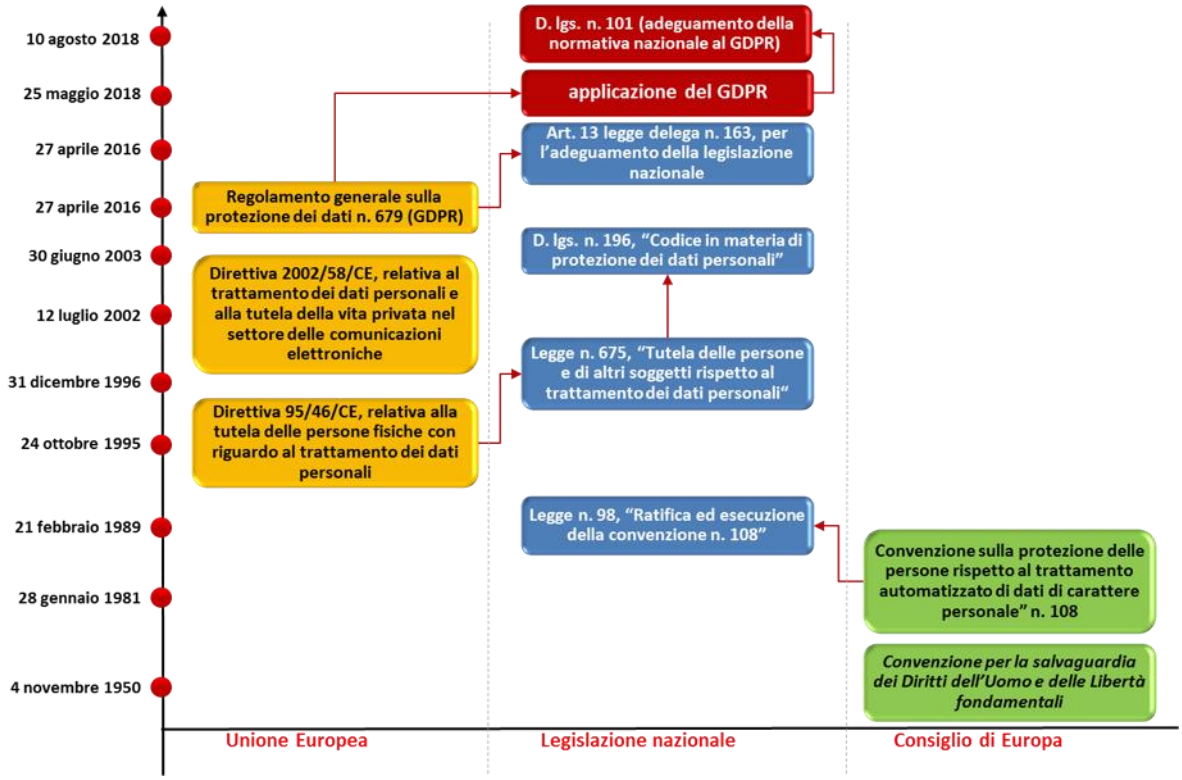

Figura 1: Cronologia riassuntiva della legislazione europea e nazionale [7]

1.3 La questione dell'anonimizzazione

Finora abbiamo sempre parlato di protezione di *dati personali*, così come definiti nell'art. 4 del GDPR. Come già anticipato, dunque, la normativa fin qui esaminata non si applica a dati anonimizzati, a partire dai quali è del tutto impossibile risalire all'interessato. Si legge infatti nel considerando 26 del GDPR: *"[...] I principi di protezione dei dati non dovrebbero pertanto applicarsi a informazioni anonime, vale a dire informazioni che non si riferiscono a una persona fisica identificata o identificabile o a dati personali resi sufficientemente anonimi da impedire o da non consentire più l'identificazione dell'interessato. Il presente regolamento non si applica pertanto al trattamento di tali informazioni anonime, anche per finalità statistiche o di ricerca"* [3]. Non esiste, dunque, una vera a propria legislazione in materia di anonimizzazione, quanto più delle linee guida pubblicate dai singoli Paesi sulle basi giuridiche e sugli ambiti di applicazione di tale tecnica. Per un'analisi più pratica e approfondita, si rimanda alla sezione apposita nel seguito di questo elaborato.

Capitolo 2:
La pseudonimizzazione

La pseudonimizzazione ha attirato una sempre crescente attenzione nell'ambito della gestione della privacy a seguito della pubblicazione del GDPR, all'interno del quale, come riportato nel capitolo precedente, viene esplicitamente definita e citata come tecnica da adottare nel trattamento dei dati personali.

Nei paragrafi che seguono, dopo aver fornito la terminologia fondamentale per la trattazione del tema, verranno presentate alcune delle tecniche di pseudonimizzazione di base ad oggi più utilizzate, cercando di valutare vantaggi e svantaggi di ognuna.

2.1 Definizioni

Una definizione di pseudonimizzazione sufficientemente esaustiva è stata riportata al *paragrafo 1.2.1* dell'elaborato. Si forniscono ora le definizioni di altri termini propri dell'argomento, che risulteranno utili per una miglior comprensione della trattazione [8]:

- *identificativo*: valore che identifica un elemento all'interno di uno schema (tabella) di identificazione. Ad esempio, il cognome può essere considerato l'identificativo di una persona (o più) all'interno di un elenco. Un *identificativo univoco* è associato ad un solo elemento (quindi, nell'esempio precedente, il cognome è identificativo univoco solo nell'ipotesi in cui non vi siano nell'elenco due persone con lo stesso cognome);

- *pseudonimo* (o *nome in codice*): informazione associata all'identificativo o ad altri dati personali di un individuo. Si tratta di una sorta di alias, un modo per mascherare l'identificativo di un soggetto. Gli pseudonimi possono presentare diversi gradi di associabilità all'identificativo originale;

- *funzione di pseudonimizzazione*: funzione che sostituisce un identificativo con uno pseudonimo;

- *chiave di pseudonimizzazione*: parametro che permette il "calcolo" della funzione di pseudonimizzazione. La chiave è quell'*informazione aggiuntiva* che, da definizione di pseudonimizzazione, permette la re-identificazione dei dati;

- *tabella di mappatura di pseudonimizzazione*: associa ciascun identificativo allo pseudonimo corrispondente. Può coincidere del tutto o in parte con la chiave di pseudonimizzazione.

2.2 Tecniche di pseudonimizzazione

Come già detto, una funzione di pseudonimizzazione associa uno pseudonimo a ciascun identificativo di un particolare dataset. Si tenga presente che una funzione di pseudonimizzazione deve sempre garantire che a due diversi identificativi vengano associati due diversi pseudonimi, onde evitare ambiguità nell'inversione della funzione. In caso contrario, si dice che la funzione è soggetta a *collisioni*. È invece accettabile l'eventualità in cui, per un motivo qualsiasi, ad uno stesso identificativo vengano associati più pseudonimi, sempre a patto che la funzione rimanga invertibile (si vedano, a proposito, le *sezioni 2.4.2* e *2.4.3*).

Avendo in mente queste regole generali, passiamo quindi in rassegna quelle che, ad oggi, vengono considerate le tecniche di pseudonimizzazione maggiormente in uso secondo l'*Agenzia dell'Unione Europea per la cybersicurezza* (*ENISA*) [8].

2.2.1 Contatore

Si tratta della strategia più semplice di pseudonimizzazione: gli identificativi vengono sostituiti da un numero dato da un generatore monotono (il valore dell'incremento tra un numero e il successivo è a libera scelta dell'implementatore, ma deve essere sempre fisso). Per garantire l'unicità dell'identificativo associato ad un certo pseudonimo, è necessario che i numeri generati non si ripetano mai.

Si tratta di una soluzione valida per set di piccole dimensioni e costituiti da dati non molto complessi; in caso di set molto grandi, il problema della memorizzazione della mappa di pseudonimizzazione potrebbe risultare non trascurabile. In termini di protezione, l'unico modo di risalire agli identificativi è conoscere la mappa, ma la natura sequenziale del contatore potrebbe comunque fornire informazioni sull'ordine dei dati all'interno del set.

2.2.2 Generatore di numeri casuali

Questo approccio è molto simile al precedente, con la differenza che lo pseudonimo viene scelto all'interno di un set di numeri che presentano tutti la stessa probabilità di essere selezionati.

Per creare la mappatura, vi sono due opzioni [9]:

1. generatore di numeri casuali vero e proprio (nello specifico, si parla di *generatore hardware di numeri casuali*);
2. *generatore di numeri pseudo-casuali crittograficamente sicuri* (*CSPRNG*), che, oltre alle caratteristiche di base dei normali algoritmi per la generazione di numeri pseudo-casuali, deve assicurare che non sia possibile ricostruire l'intera sequenza osservandone solo una parte.

La probabilità di incorrere in collisioni dipende dal noto *paradosso del compleanno* (si veda [10] e [11]) e si rende minima generando numeri in un range molto ampio.

Come nel caso precedente, questo metodo necessita la memorizzazione dell'intera tabella di mappatura per permettere l'inversione. Rispetto al contatore, tuttavia, si ha il vantaggio di riuscire a celare qualsiasi informazione riguardante l'organizzazione del set e l'ordine dei dati al suo interno.

2.2.3 Funzione crittografica di hash

Una funzione di hash prende in input una stringa di lunghezza arbitraria e la associa ad output di lunghezza fissa chiamati *digest* [12]. Dato in input l'identificativo, dunque, la funzione di hash restituisce in output lo pseudonimo.

Essa presenta le seguenti proprietà:

1. *non-invertibilità;*
2. *resistenza alle collisioni*: è computazionalmente molto difficile trovare due input distinti che si associno al medesimo output;
3. *unidirezionalità* o *resistenza alle contro-immagini*: è computazionalmente molto difficile trovare input che diano particolari output specificati in precedenza;
4. *resistenza alla correlazione*: anche piccole variazioni del messaggio originale comportano evidenti variazioni nel digest.

Sebbene la loro non-invertibilità dovrebbe garantire l'impossibilità di derivare il messaggio originale a partire dal digest, le funzioni di hash sono considerate una strategia di pseudonimizzazione piuttosto debole, in quanto sono particolarmente soggette ad *attacchi esaustivi* o *a forza bruta*: noti la funzione e il dominio dell'identificativo, riuscire a ricavare quest'ultimo provando tutti i possibili identificativi risulta fattibile in tempi ragionevoli. Per questo motivo, sono più spesso utilizzate, ad esempio, come tecnica di firma digitale.

Esempio – SHA-1

Secure Hash Algorithm 1 è una funzione crittografica di hash sviluppata dalla *National Security Agency* degli Stati Uniti nel 1993. Essa non è più considerata sicura dal 2005, in quanto ormai violata più volte.

SHA-1 prende in input un determinato messaggio M di lunghezza $l < 2^{64}$ bit e produce un digest di 160 bit.
Dopo aver convertito il messaggio in forma binaria, è necessario procedere con un *padding* (ovvero l'aggiunta in coda di ulteriori cifre binarie), in modo tale da raggiungere un numero totale di bit che sia multiplo di 512. Ciò si ottiene mediante i seguenti passaggi:

1. aggiungere un "1";
2. aggiungere un numero k di "0", con k più piccolo intero non negativo tale che $k+l+1$ sia pari a 448 in modulo 512;
3. aggiungere un blocco di 64 bit che corrisponda al valore l in forma binaria (formato big endian).

Dopo il padding, si procede con la suddivisione del nuovo messaggio in N blocchi da 512 bit ($M_0, M_1, \ldots, M_{N-1}$); questo procedimento è detto *parsing*.

Ciascuna funzione di hash del tipo SHA prevede dei valori iniziali $H^{(0)}$ (*Initial Values)* fissati e universalmente definiti. Per SHA-1 il valore iniziale si compone di cinque parole a 32 bit, così espresse in forma esadecimale:

$$IV = (H_0^{(0)}, H_1^{(0)}, H_2^{(0)}, H_3^{(0)}, H_4^{(0)})_{16}$$

con

$$H_0^{(0)} = 67452301,$$

$$H_1^{(0)} = efcdab89,$$

$$H_2^{(0)} = 98badcfe,$$

$$H_3^{(0)} = 10325476,$$

$$H_4^{(0)} = c3d2e1f0.$$

L'elaborazione vera e propria consiste nel processare iterativamente tutti gli N blocchi in cui è stato suddiviso il messaggio, mediante quella che viene chiamata *funzione di compressione*. Questa, a grandi linee, si compone di una serie di operazioni di XOR (OR esclusivo) tra singole parole di 32 bit ottenute partizionando il blocco di volta in volta in esame. Ciascuna delle N iterazioni produce in output 5 parole da 32 bit che nel complesso prendono il nome di *Intermediate Hash Values*.

Per $1 \leq i \leq N$, vale, dunque,

$$IHV_i = Compress(IHV_{i-1}, M_{i-1}) =$$

$$= (H_0^{(i)}, H_1^{(i)}, H_2^{(i)}, H_3^{(i)}, H_4^{(i)}).$$

Chiaramente, la prima iterazione prende in ingresso IHV_0 che coincide con IV.

Al termine delle N iterazioni, il digest è dato dall'ultimo valore di hash intermedio (IHV_N) espresso come concatenazione delle cinque parole binarie a 32 bit che lo compongono, per un totale, dunque, di 160 bit. In alcuni casi, prima della concatenazione, le parole binarie vengono convertite in base sedici, risultando così in una sequenza di 40 cifre esadecimali [13] [14].

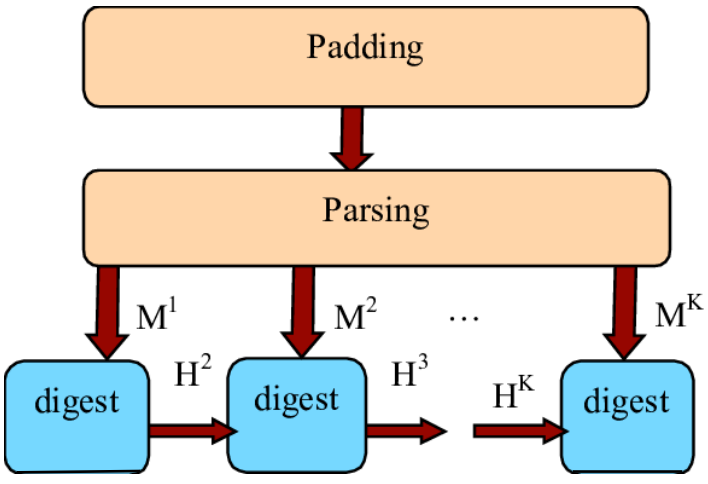

Figura 2: Schema riassuntivo del funzionamento di una funzione del tipo SHA [15]

2.2.4 Codice di autenticazione del messaggio

Detta anche *MAC* (*Message Authentication Code*), si tratta di una funzione simile a quella di hash, ma che prevede l'utilizzo di una chiave segreta. Un algoritmo MAC accetta in ingresso una chiave di criptazione (si veda la prossima sezione per approfondire il concetto) e un messaggio di lunghezza arbitraria (nel nostro caso, l'identificativo), per restituire un cosiddetto *tag* (lo pseudonimo) di lunghezza fissa [16].

Come è facile intuire, i MAC garantiscono un grado di sicurezza di gran lunga superiore alle funzioni di hash, in quanto, a meno che la chiave non sia stata compromessa, è impossibile decodificare lo pseudonimo.

Come esempio notevole, si cita *HMAC*, di gran lunga la più diffusa strategia di codice di autenticazione del messaggio impiegata nei protocolli Internet.

2.3 Crittografia

La crittografia è la branca della *crittologia* che si occupa dei metodi utilizzati per trasformare un certo messaggio (*testo in chiaro*) in un altro messaggio (*testo cifrato*), che, in generale, risulta incomprensibile a chiunque non conosca tutti i dettagli della tecnica di trasformazione [17]. Nell'ambito della pseudonimizzazione, è possibile ottenere gli pseudonimi attraverso la cifratura degli identificativi. La *decriptazione* avviene poi con un algoritmo "inverso" a quello di cifratura [12].

La crittografia si basa su un algoritmo e su una *chiave crittografica*. È possibile classificare le tecniche crittografiche in due categorie principali: *crittografia a chiave simmetrica* e *a chiave asimmetrica*. Una terza tipologia di crittografia, la *crittografia quantistica*, ha ricevuto un grande impulso a partire dalla seconda metà del secolo scorso, ma il suo campo di applicazione non comprende, almeno per ora, la pseudonimizzazione dei dati.

2.3.1 Crittografia a chiave simmetrica

I sistemi di crittografia a chiave simmetrica sfruttano la stessa chiave sia per cifrare che per decifrare un messaggio; ciò consente di realizzare algoritmi di cifratura molto performanti e facili da implementare.

Il requisito che entrambe le parti (mittente e destinatario del messaggio) siano a conoscenza della chiave presuppone che questa debba essere in qualche modo condivisa prima della comunicazione; questo scenario risulta particolarmente critico, in quanto gran parte della sicurezza degli algoritmi a chiave simmetrica dipende dalla segretezza della chiave stessa e, nel caso in cui questa venisse in qualche modo intercettata, la comunicazione risulterebbe irrimediabilmente compromessa.

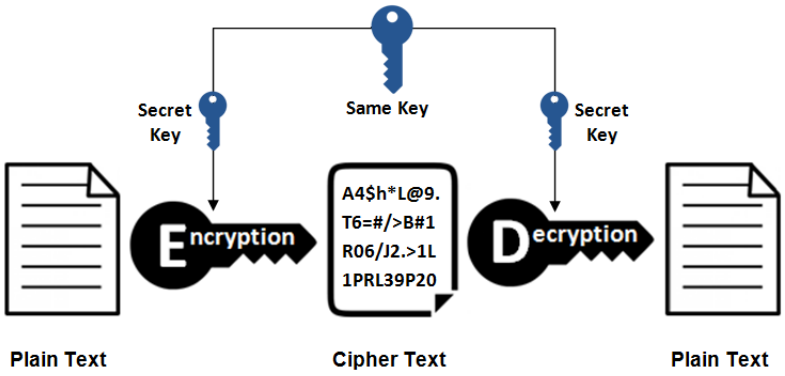

Figura 3: Funzionamento generale della crittografia simmetrica [18]

La sicurezza di un sistema crittografico non deve dipendere dal tenere celato l'algoritmo, ma solo la chiave (*principio di Kerckhoffs*); ciò significa che, nel progettare un sistema di cifratura, è buona norma considerare la possibilità che il "nemico" ne conosca il funzionamento. Per questo motivo è necessario che la chiave sia abbastanza complessa da scongiurare la buona riuscita di un attacco esaustivo. Oltre a ciò, è necessario garantire l'impossibilità di ricavare la chiave dal messaggio criptato, sempre conoscendo l'algoritmo, attraverso, ad esempio, valutazioni statistiche sulla frequenza delle lettere in una determinata lingua (si veda, per esempio il *cifrario a permutazione* in [12]).

I sistemi simmetrici sono in assoluto i più utilizzati come strategia di pseudonimizzazione; solitamente, infatti, l'ente che opera la cifratura non è chiamato a comunicare la chiave a terzi (eventualità in cui, come già detto, la crittografia simmetrica sarebbe molto vulnerabile), per cui è possibile optare per questo tipo di tecnica, che peraltro risulta computazionalmente davvero poco dispendiosa, senza alcun considerevole rischio aggiuntivo.

Alcuni algoritmi di cifratura simmetrica ben consolidati sono, ad esempio, *DES*, *3DES*, *AES*, *Twofish*, *Serpent*, *IDEA*…

Esempio – DES

Il *Data Encryption Standard* è un algoritmo di cifratura a chiave simmetrica con chiave a 64 bit (di cui, come vedremo, solo 56 utili). Questo sistema è stato scelto come standard per il governo degli Stati Uniti d'America nel 1976, per poi raggiungere una vastissima diffusione in tutto il mondo. Ad oggi, il DES è considerato poco sicuro, soprattutto a causa della brevità della chiave di cifratura (è possibile rompere l'algoritmo in poche ore con attacco esaustivo); come alternativa valida, è possibile operare la cifratura iterando l'intero procedimento tre volte (*3DES*), utilizzando due o tre chiavi diverse.

Il DES è un *algoritmo di cifratura a blocchi*: a differenza degli *algoritmi a flusso*, che processano un singolo elemento per volta, questi cifrano nello stesso momento l'intero input previsto. Il DES lavora su stringhe di 64 bit; ciò significa che, nel caso in cui il messaggio da cifrare superi tale lunghezza, è necessario dividerlo in porzioni della giusta dimensione, eventualmente procedendo con un padding. In output viene dato un blocco della stessa lunghezza di quello in input, quindi nuovamente 64 bit.

Come detto, benché nominalmente la chiave sia di 64 bit, otto di questi vengono utilizzati come *bit di parità* (codice di controllo utilizzato per prevenire errori nella trasmissione o nella memorizzazione dei dati); pertanto, i bit "liberi" sono solo 56. Si procede ora a descrivere più nel dettaglio il funzionamento del DES, tenendo a mente che quanto presentato è da intendersi applicato a ciascun blocco di 64 bit in cui il messaggio da cifrare è stato precedentemente diviso.

L'algoritmo si sviluppa in sedici fasi identiche dette *round* o *cicli*. Prima di questi sedici passaggi, l'input passa attraverso una mappa di permutazione (*permutazione iniziale*, *IP*); dopo l'ultimo round, il blocco risultante viene nuovamente permutato utilizzando una mappa inversa a quella inziale (*permutazione finale*, $FP = IP^{-1}$).

All'inizio di ogni round, la stringa binaria in ingresso viene suddivisa in due metà

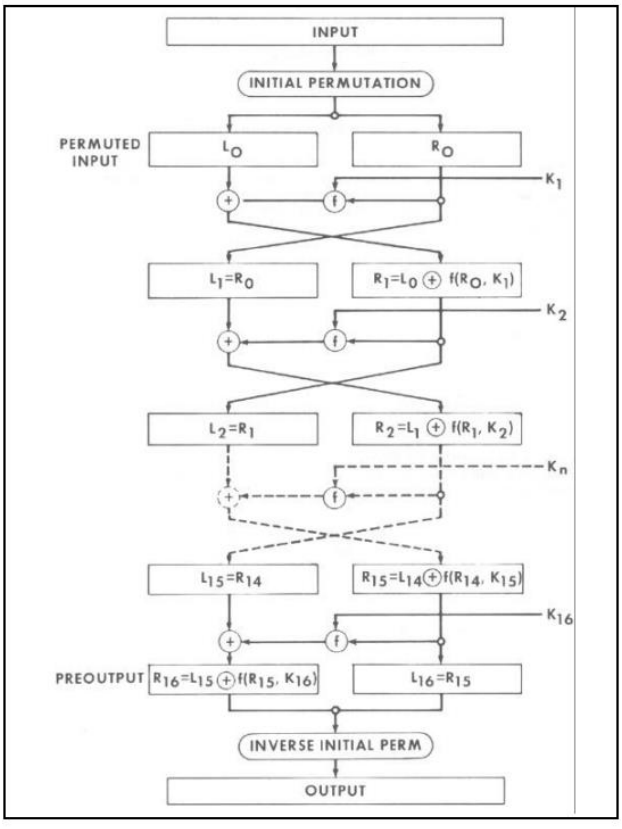

Figura 4: Schema riassuntivo algoritmo DES [19]

(*L* e *R*) di 32 bit; di queste due parti, una sola viene processata mediante una funzione detta *funzione di Feistel* (descritta in seguito) che restituisce un nuovo blocco da 32 bit. La metà da processare viene scelta alternativamente nel susseguirsi dei cicli: se nell'i-esimo ciclo viene processata la metà di sinistra, nell'(i+1)-esimo quella di destra e viceversa. Il risultato della funzione di Feistel viene combinato con l'altra metà della stringa tramite un'operazione di XOR e le due metà vengono invertite prima di entrare nel ciclo successivo. L'unica eccezione a quest'ultimo passaggio è il ciclo finale, al termine del quale non avviene lo scambio.

La funzione di Feistel (*Figura 6*) si snoda attraverso quattro passaggi:

1. *espansione*: il mezzo blocco di 32 bit passa attraverso una *permutazione di espansione* (*E*) che duplica alcuni bit, fino a raggiungere un nuovo blocco da 48 bit;
2. *miscelazione con la chiave*: il risultato del passo precedente viene combinato, tramite operazione di XOR, con una sottochiave da 48 bit. Le sedici sottochiavi (una per ogni round) vengono derivate dalla chiave di cifratura principale attraverso il *gestore della chiave* (esposto più avanti);
3. *sostituzione*: il nuovo blocco viene suddiviso in otto parti da sei bit. Ognuna delle porzioni viene processata da una *substitution box* (*S-box*), che sostituisce i sei bit in input con quattro in output mediante una trasformazione non lineare;
4. *permutazione*: i 32 bit risultanti vengono riordinati in base a permutazioni fisse regolate dalla *permutation box* (*P-box*).

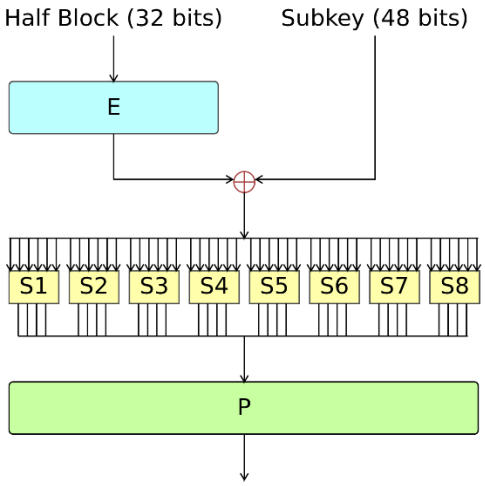

Figura 6: Funzione di Feistel [35]

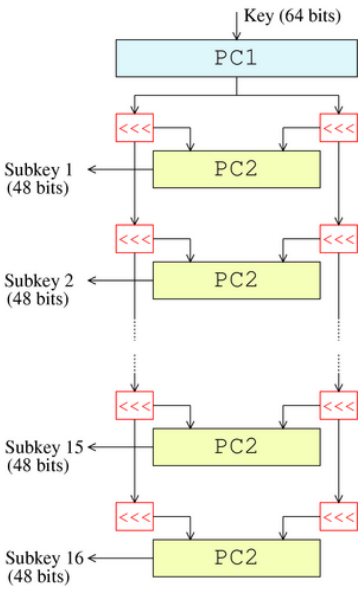

Figura 5: Gestore della chiave [35]

Il gestore della chiave, mostrato in *Figura 5* qui sopra, si occupa della generazione delle sottochiavi a partire dalla chiave principale. Per prima cosa, i 56 bit utili di quest'ultima vengono suddivisi in due metà di 28 bit; ogni metà viene poi trattata in maniera indipendente. Nei vari round, entrambe le stringhe vengono fatte slittare verso sinistra di uno o due bit (a seconda del ciclo in cui ci si trova). Mediante la funzione *Permuted Choice 2* (*PC2*), da ciascuna delle due metà shiftate vengono estratti 24 bit; in conclusione, dunque, si ottiene, per ogni round, una chiave di 48 bit.

Ogni operazione descritta ha proprietà di simmetria; questo vuol dire che, per decifrare un messaggio, è sufficiente riapplicare l'algoritmo prendendo i blocchi in ordine inverso e generando le sottochiavi con uno shift verso destra invece che verso sinistra [19].

2.3.2 Crittografia a chiave asimmetrica

Questo sistema di cifratura prevede che ogni utente disponga di due chiavi: la *chiave pubblica*, conoscibile da tutti e quindi trasmettibile anche in chiaro, e la *chiave privata*, che invece è strettamente personale e deve rimanere segreta. Le due chiavi sono in rapporto di "alias", ovvero, se con una delle due chiavi si codifica il messaggio, allora questo sarà decifrabile solo con l'altra chiave [12].

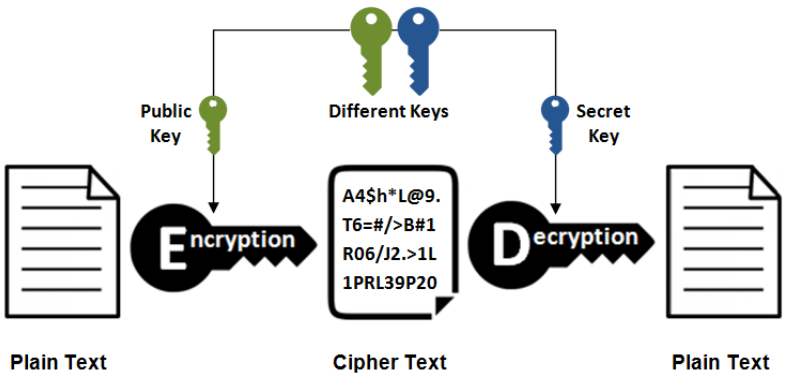

Figura 7: Funzionamento generale della crittografia asimmetrica [18]

L'efficacia di questa tecnica si basa sul fatto che le due chiavi assegnate all'utente sono costruite in maniera tale da rendere computazionalmente irrealizzabile la determinazione della chiave segreta a partire da quella pubblica.

Ci sono due tipi di funzioni che possono essere realizzate tramite questo sistema:

1. *firma digitale*: il messaggio viene cifrato ("firmato") dal mittente/autore tramite la sua chiave segreta; in questo modo chiunque voglia controllare la paternità del documento

deve solamente verificare che la chiave pubblica del mittente riesca effettivamente a decodificarlo;

2. *scambio di messaggi "segreti"*: il mittente deve cifrare il messaggio utilizzando la chiave pubblica del destinatario; il messaggio, dunque, sarà decifrabile solo da quest'ultimo tramite la sua chiave privata. Questa strategia viene, in alcuni casi, impiegata come tecnica di pseudonimizzazione.

Il più celebre metodo di cifratura asimmetrica è senza dubbio l'algoritmo *RSA*.

Come già detto, i sistemi asimmetrici non vengono utilizzati molto spesso come tecnica di pseudonimizzazione, motivo per cui non verrà portata la descrizione dettagliata di un algoritmo di esempio.

2.4 Strategie di implementazione

Questa sezione tratterà la questione più generale della pseudonimizzazione di uno o più dataset. A tale scopo, si consideri un identificativo *ID* che appare più volte nei set *A* e *B*. Le strategie con cui è possibile implementare la pseudonimizzazione in questi schemi sono tre: *pseudonimizzazione deterministica*, *randomizzata al documento* e *completamente randomizzata* [8].

2.4.1 Pseudonimizzazione deterministica

Ogni volta che in un dataset appare *ID*, esso viene sempre sostituito con il medesimo pseudonimo *pseudo*: uniformemente, dunque, all'interno del dataset stesso, così come nei due dataset in esame *A* e *B*.

Per implementare questa strategia, è necessario per prima cosa estrarre l'elenco degli identificativi *univoci* (identificativi che compaiono più volte vengono contati come identificativo unico) contenuti nei diversi set di dati. A questo punto, l'elenco viene associato agli pseudonimi, che vanno poi a sostituire gli identificativi negli schemi.

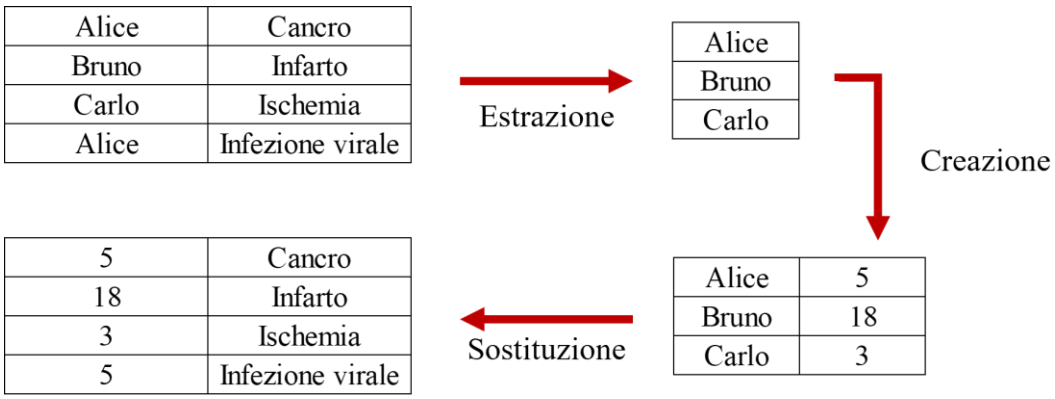

Figura 8: Pseudonimizzazione deterministica riferita ad un unico set

2.4.2 Pseudonimizzazione randomizzata al documento

Tutte le volte in cui *ID* appare all'interno dello stesso dataset, viene sostituito con uno pseudonimo differente (*pseudo*$_1$, *pseudo*$_2$, ...). Tuttavia, *ID* viene sempre associato alla medesima raccolta di pseudonimi nei dataset *A* e *B*.

La tabella di mappatura viene realizzata estraendo tutti gli identificativi presenti nel set di dati: ogni occorrenza di un determinato identificativo (*Alice* in figura sotto) viene trattata indipendentemente.

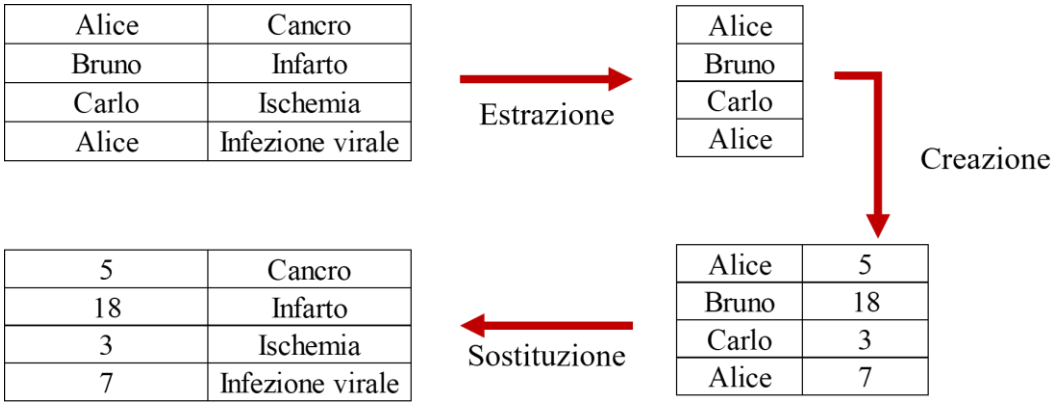

Figura 9: Pseudonimizzazione randomizzata al documento riferita ad unico set

2.4.3 Pseudonimizzazione completamente randomizzata

Ogniqualvolta appare *ID* all'interno di uno schema *A* o *B*, questo viene sostituito con uno pseudonimo differente; non esiste dunque una relazione univoca tra *ID* e un particolare insieme di *pseudo*, come invece succedeva nel caso precedente. Si tenga dunque presente che, nel caso in cui lo stesso documento venga pseudonimizzato due volte attraverso una pseudonimizzazione completamente randomizzata, si ottengono due output diversi.

2.5 Valutazioni sulla scelta della tecnica e della strategia di implementazione

La scelta di una particolare tecnica dipende da molti fattori, tra cui la tipologia di dati, la finalità con cui sono stati raccolti, la probabilità e il rischio associati ad un ipotetico attacco.

Come emerso nelle sezioni precedenti, le tecniche più efficaci sono il generatore di numeri casuali, il codice di autenticazione del messaggio e la crittografia, in quanto appositamente mirati a contrastare attacchi esaustivi. In alcuni scenari l'entità di pseudonimizzazione potrebbe decidere di optare per una combinazione di diversi approcci o per una variante di uno di questi.

È molto importante valutare anche i requisiti in termini di formato dell'identificativo richiesto, così come le dimensioni dello pseudonimo ottenuto in output.

Tabella 1: Confronto tra diverse tecniche in termini di dimensioni di identificativo e pseudonimo, con k numero degli identificativi nel dataset [8]

Metodo	Dimensione dell'identificativo	Dimensione *m* dello pseudonimo in bit
Contatore	Qualsiasi	$m = log_2 k$
Generatore di numeri casuali	Qualsiasi	$m >> 2log_2 k$
Funzione di hash	Qualsiasi	Fissa o $m >> 2log_2 k$
MAC	Qualsiasi	Fissa o $m >> 2log_2 k$
Crittografia	Qualsiasi o fissa	Fissa o uguale all'*ID*

Per identificativi di dimensione variabile, è possibile adottare la maggior parte delle soluzioni, eccezion fatta per determinate scelte in caso di crittografia. Si tenga poi presente che i generatori di numeri casuali, le funzioni di hash e i codici di autenticazione del messaggio sono a rischio collisioni; pertanto, la scelta della dimensione dello pseudonimo è molto critica.

Per quanto riguarda la strategia di implementazione, la pseudonimizzazione completamente randomizzata garantisce la migliore protezione, ma risulta proibitiva in termini di confronti tra dataset. Le funzioni di pseudonimizzazione randomizzata al documento e deterministica risultano quindi più funzionali, ma si portano appresso l'associabilità tra diverse righe degli schemi.

Capitolo 3:
L'anonimizzazione

All'inizio di questo elaborato, si è visto come la legislazione vigente in materia di privacy si occupi esclusivamente di dati personali, categoria all'interno della quale non ricadono i dati *anonimi*. In questo capitolo, dunque si fornirà innanzitutto una definizione esaustiva del termine *anonimizzazione*, per poi esporne le basi giuridiche e gli ambiti di applicazione e, infine, presentare le principali tecniche di implementazione di tale strategia e la loro robustezza.

3.1 Cosa significa *anonimizzazione*?

La *Legge 675/96*, primo testo normativo dedicato alla privacy in Italia, all'art. 2 specificava come per *dato anonimo* si dovesse intendere *"il dato che, in origine o a seguito di un trattamento, non può essere associato ad un interessato identificato o identificabile"* [20]; la stessa definizione venne poi traslata nel D.Lgs. 196/03. Quando poi il D.Lgs. 101/18 emendò la normativa italiana per armonizzarla al GDPR, il riferimento ai dati anonimi scomparve, in quanto non contemplati nella legislazione europea.

Ad oggi, si può prendere per univoca la definizione di anonimizzazione fornita nel 2011 dall' *Organizzazione internazionale per la normazione* (*ISO*) affiancata dalla *Commissione elettrotecnica internazionale* (IEC), secondo cui è il *"processo mediante il quale i dati personali vengono modificati in modo irreversibile così che il titolare del trattamento, da solo o in collaborazione con altre parti, non possa più identificare direttamente o indirettamente l'interessato"* [21]. L'anonimizzazione, dunque, non è altro che il processo di rimozione o manipolazione irreversibile degli identificativi diretti (come il nome o il codice fiscale) e indiretti (come gli pseudonimi a seguito di un processo di pseudonimizzazione) che potrebbero portare all'identificazione di un individuo.

3.1.1 Incomprensioni legate all'anonimizzazione

Non esistendo una normativa univoca in materia, il concetto di anonimizzazione risulta spesso nebuloso e vittima di incomprensioni.

Innanzitutto, non di rado si percepisce il termine *anonimizzato* come un concetto binario: si pensa che un set di dati possa essere etichettato semplicemente come anonimo o meno. In realtà, l'anonimizzazione non riduce sempre a zero la probabilità di re-identificazione di un dataset. Sebbene un'anonimizzazione totale sia l'obiettivo idealmente desiderabile, un sistema robusto si può limitare a ridurre il rischio di re-identificazione sotto una certa soglia, stabilita in base a diversi fattori (impatto sulla privacy in caso di identificazione, rischio di attacco…). Il grado di anonimità di un dataset è quindi un concetto abinario e misurabile.

Oltre a ciò, è necessario tenere presente che l'anonimizzazione non è sempre realizzabile: potrebbe non essere possibile abbassare il rischio di re-identificazione sotto la soglia stabilita mantenendo l'utilità del set di dati per lo scopo previsto. Questo accade, ad esempio, quando l'universo dei soggetti è esiguo (e.g. studenti in una classe).

È importante considerare, inoltre, che, benché, come vedremo, sia possibile fare delle considerazioni di stampo generale sul grado di validità e sicurezza di una determinata tecnica di anonimizzazione, non è detto che seguire un processo che qualcuno ha usato con successo porterà altri a risultati equivalenti. Le procedure devono essere modellate sulla base della natura dei dati, del contesto e delle finalità di trattamento, nonché dei rischi di varia probabilità e gravità per i diritti e le libertà personali. Per questo motivo, è errato credere che l'anonimizzazione possa essere completamente automatizzata: è sempre necessario l'intervento di un esperto umano che esegua un'approfondita analisi del processo in conformità allo specifico caso.

Da sfatare, infine, la convinzione che l'anonimizzazione sia per sempre: esiste il rischio che, con gli sviluppi tecnici e la raccolta nel tempo di informazioni aggiuntive, alcuni processi divengano reversibili. Risulta necessario, dunque, un continuo controllo e aggiornamento sui dataset anonimizzati.

Tabella 2: "10 misunderstandings related to anonymisation" [22] [23]

Pseudonimizzazione e anonimizzazione coincidono	Crittografia e anonimizzazione coincidono
L'anonimizzazione dei dati è sempre possibile	L'anonimizzazione è per sempre
L'anonimizzazione riduce sempre a zero la probabilità di re-identificazione di un set di dati	L'anonimizzazione è un concetto binario che non può essere misurato
L'anonimizzazione può essere completamente automatizzata	L'anonimizzazione rende i dati inutili
Seguire un processo di anonimizzazione che altri hanno usato con successo porterà tutti a risultati equivalenti	Non c'è nessun rischio e nessun interesse a scoprire a chi si riferiscono questi dati

Per completezza, si riportano in *Tabella 2* qui sopra, tutti i *"10 misunderstandings related to anonymisation"* individuati in un paper congiunto del *Garante della privacy spagnolo* (*AEPD*) e dell'*European Data Protection Supervisor* (*EDPS*) [22]; i più significativi sono già stati approfonditi all'interno di questa sezione.

3.2 Liceità del trattamento e ambiti di applicazione

Benché i dati sottoposti ad un processo di anonimizzazione non siano più considerabili dati personali, prima dell'elaborazione lo erano senza dubbio. Ciò significa che è necessario garantire che la raccolta dei dati sia stata effettuata secondo i principi descritti nel GDPR e presentati nel primo capitolo di questo elaborato; in particolare, al momento della raccolta e per tutta la durata della detenzione di quei dati in forma "personale", deve sussistere una finalità ben specificata.

Il principale motivo per cui si ricorre all'anonimizzazione è la volontà di mantenere il possesso di quei dati anche quando lo scopo per cui sono stati acquisiti è stato raggiunto o non sussiste più. I dati de-personalizzati possono infatti ancora essere utili per fini statistici o storici, oppure con lo scopo di monitorare o migliorare i servizi forniti da un'azienda [23].

In campo sanitario, l'anonimizzazione dei dati è necessaria, ad esempio, nell'ambito della ricerca (e.g. studi epidemiologici o farmaceutici), soprattutto nel momento in cui i risultati dovessero essere resi pubblici [24].

3.3 Robustezza dell'anonimizzazione

Ci sono tre particolari scenari critici a cui è necessario far riferimento nel momento in cui si voglia verificare che il rischio di re-identificazione dei propri dati sia abbastanza remoto:

1. *singling out*: possibilità di isolare uno o più record[1] e di associarli ad un particolare individuo;

[1] Si tenga presente che una *base di dati* (*database*) è organizzata in *tabelle* o *dataset*; ciascun dataset si compone di diversi *campi* o *attributi* (e.g. nome, cognome) ogni riga della tabella si dice *record* o *entry* e ogni record non è altro che un insieme di *valori* che corrispondono agli attributi di quel particolare dataset (e.g. Mario, Rossi; l'insieme dei due valori è un record). Infine, con *query* si indica l'interrogazione di un database per estrarre dati che soddisfino un determinato criterio di ricerca.

2. *linkability*: possibilità di stabilire che due o più record, appartenenti allo stesso dataset o a dataset differenti, riferiscono allo stesso individuo o gruppo di individui, senza necessariamente identificarli;
3. *inference*: possibilità di dedurre, con una sicurezza significativa, il valore di un attributo di un particolare record a partire dai valori di un set di altri attributi.

Nei paragrafi seguenti, verranno presentate diverse tecniche di anonimizzazione valutandone la validità in questi tre scenari [25].

3.4 Tecniche di anonimizzazione: Masking

Consiste nell'eliminare totalmente gli identificativi più "ovvi", come nomi e cognomi, da ogni record, creando un dataset privo di identificatori personali.

Esistono due principali varianti di questa tecnica:

1. *rimozione parziale dei dati*: vengono eliminati solo alcuni degli identificatori, mantenendone altri; la scelta viene effettuata mediante valutazioni di tipo costo-beneficio, considerando il rischio di attacco e l'utilità dei dati;
2. *mantenimento di quasi-identificatori*: vengono mantenuti nel dataset dei riferimenti univoci per ogni record (e.g. codice di riferimento) in modo tale da poter ancora distinguere dati che appartengono alla stessa persona, ma sia impossibile risalire all'identità di questa.

Questa tecnica garantisce l'ottenimento di un dataset anonimizzato ancora molto ricco in termini di contenuto informativo. Per contro, si tratta di una strategia molto rischiosa in quanto, come detto, non impedisce affatto la linkability dei dati e, avendo accesso ad informazioni terze, l'eventualità di una re-identificazione non è per nulla remota [24].

3.5 Tecniche di anonimizzazione: Randomizzazione

Si tratta di una famiglia di tecniche che modificano la veridicità dei dati, così da rimuovere lo stretto collegamento che li lega ai soggetti. Queste tecniche non risultano particolarmente efficaci per quanto riguarda la singolarità dei record (ognuno di questi continua chiaramente a far riferimento ad un solo individuo), ma possono essere molto potenti contro l'inferenza [25].

3.5.1 Aggiunta di rumore

Consiste nel modificare i valori di un particolare attributo nel dataset in modo tale che risultino meno accurati, ma mantenendone la distribuzione generale e, più o meno precisamente, le proprietà statistiche. Solitamente si sceglie di sommare alla serie di dati un rumore stocastico a distribuzione normale, con media nulla e deviazione standard che dipende dall'ampiezza del dataset e dal range di valori; si parla di *rumore additivo*.

Per esempio, se sono stati raccolti in una tabella i punteggi in centesimi relativi ad un particolare test, il dataset anonimizzato potrebbe contenere i punteggi con un errore nel range di ±5 punti. In questo modo, dovrebbe risultare più difficile collegare i soggetti ai rispettivi punteggi, anche con l'aggiunta di informazioni esterne riguardanti, ad esempio, la media complessiva dei risultati dei vari individui.

L'aggiunta di rumore è una tecnica valida per rendere più ardua la re-identificazione dei dati o la ricerca di un loro legame con altri set, ma rimane una strategia a tratti molto debole, soprattutto se i valori dell'attributo in questione sono pochi e distribuiti in un range molto ampio; questo, pertanto, non è altro che un processo accessorio, che deve sempre (o quasi) essere affiancato da altre elaborazioni.

3.5.2 Permutazione

Implementare questa tecnica significa mescolare i valori di uno o più attributi all'interno di un dataset, distruggendo così i legami tra quegli attributi e gli identificatori. In questo modo rimangono però completamente immutati il range e la distribuzione dei dati.

Se due (o più) attributi sono legati da una relazione logica o una correlazione statistica, è bene permutarli entrambi: effettuare il mescolamento dei valori di uno solo di loro risulterebbe poco efficiente, poiché i valori dell'altro fornirebbero una guida sufficiente a ristabilire l'ordine.

Si prenda a titolo esemplificativo la tabella sottostante:.

Tabella 3: Esempio di dataset con permutazione di un attributo

Anno di nascita	Occupazione (permutata)	Introito annuo
1968	Ingegnere	100k
1972	Disoccupato	45k
1970	Manager	43k
1980	CEO	5k
1978	Ingegnere	70k

Il legame tra occupazione e introito annuo è molto stretto; pertanto, è facile dedurre che il CEO (a cui possiamo suppore con buona certezza corrisponda il valore *100k* del terzo attributo) è nato nel 1968.

Così come per l'aggiunta di rumore, spesso la sola permutazione non garantisce un grado di anonimizzazione adeguato e deve quindi essere affiancata da altre tecniche.

3.5.3 Differential privacy

Propriamente parlando, con *differential privacy* non si fa riferimento ad un particolare tipo di algoritmo di protezione dei dati, ma piuttosto ad una definizione matematica del concetto stesso di gestione della privacy. Nell'applicazione di questa tecnica, si utilizza di nuovo la strategia dell'aggiunta di rumore (per questo parleremo di *output randomizzati*), ma in maniera sostanzialmente diversa da quanto visto in precedenza.

Si supponga di avere due dataset *D* e *D'*, che differiscono solo per un record. L'idea di differential privacy si applica di norma a set di dati da cui si vogliono estrarre proprietà statistiche; dunque, si supponga di analizzare i dataset con un algoritmo che computi questo tipo di proprietà (media, mediana, varianza…). L'algoritmo si dice *differentially private* se per i due set fornisce output randomizzati che seguono distribuzioni di probabilità quasi identiche. In altre parole, conoscendo anche a fondo il background di una delle due raccolte, risulta impossibile ricavare informazioni sul record che la differenzia dall'altra a partire dagli output. Si tratta, quindi, di riuscire a minimizzare a tal punto il contributo di una singola entry, da renderla minimamente significativa dal punto di vista statistico, così da impossibilitarne l'identificazione. Lo stesso concetto si può applicare considerando un singolo dataset, che è possibile interrogare con diverse query, comprendendo più o meno elementi dello schema; in ogni caso, gli output devono risultare praticamente indistinguibili [26] [27].

Ovviamente, il contributo che un singolo individuo dà alla query dipende in buona parte dal numero totale di individui che sono coinvolti in suddetta interrogazione: se la richiesta coinvolge una sola persona, i dati riferiti a questa contribuiranno al 100%; se invece sono coinvolti dati relativi a 100 persone, ciascuna di queste contribuirà per l'1%. Il cardine della privacy differenziale sta nel fatto che è necessario calibrare la quantità di rumore che viene aggiunto ai dati a seconda del numero di individui coinvolti nella query, in modo tale da raggiungere, in ogni caso, output statisticamente molto simili. Il compito di un algoritmo *differentially private* è quello di stabilire, di volta in volta, quanto rumore applicare [28].

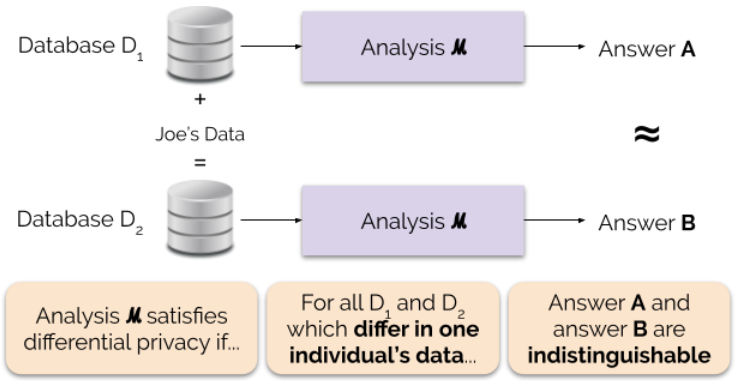

Figura 10: Spiegazione del concetto di algoritmo differentially private [29]

La differenza principale rispetto alle tecniche già viste, dunque, risiede nel fatto che il rumore non viene applicato sistematicamente all'intero database: il controllore dei dati continua a mantenere una versione non anonimizzata di questo, al quale viene applicata una quantità di rumore analiticamente scelta da un algoritmo solo nel momento in cui viene fatta una specifica query da una terza parte. Ciò porta anche il vantaggio di poter costantemente monitorare con precisione il tipo di richieste e di consultazione a cui viene sottoposto il database e, volendo, tenerne un registro [25].

La somiglianza degli output risultanti da diverse query è regolata da un parametro ε stabilito in partenza in base al grado di anonimizzazione che risulta necessario; minore è il valore di ε, più gli output saranno indistinguibili e minore sarà la probabilità di una re-identificazione, con il rischio però che i risultati risultino sempre meno rappresentativi della realtà. Diventa dunque necessario trovare il giusto equilibrio tra protezione dei dati e utilità degli stessi.

3.6 Tecniche di anonimizzazione: Generalizzazione

Tale approccio consiste nel modificare la scala o l'ordine di grandezza dei valori di un determinato attributo (e.g. sostituire l'indicazione della città di residenza con la regione, oppure una precisa altezza con un range di altezze). Questa strategia fornisce una buona protezione contro il *singling out*, ma necessita di analisi quantitative molto sofisticate per prevenire l'inferenza e la *linkability* [25].

3.6.1 Aggregazione e k-anonimity

Con il termine *aggregazione* si intende il processo di raggruppamento dei valori degli identificatori di due o più soggetti, in modo tale che non risultino più distinguibili.

Si prenda il seguente esempio:

Tabella 4: Esempio di dataset con attributi aggregati

Anno di nascita	Genere	CAP	Diagnosi
1957	M	37*	Infarto
1957	M	37*	Ischemia
1957	M	37*	Ischemia
1964	M	37*	Infarto
1964	M	37*	Infarto

In questo caso, gli identificatori sono la data di nascita, il genere e l'indirizzo di residenza. Si sceglie di indicare la data di nascita solo tramite l'anno e la residenza solo tramite le prime due cifre del CAP (che indicano la provincia, in questo caso Verona). Così facendo, sapendo tramite informazioni terze, ad esempio, che un uomo nato il 22/02/1957 e residente a Villafranca di Verona è stato inserito nel dataset sopra esposto, è impossibile sapere se la sua diagnosi sia di infarto o di ischemia.

Attraverso questo processo di generalizzazione, vengono a formarsi delle *classi di equivalenza*, ovvero dei record che presentano gli stessi valori per tutti gli attributi identificativi (nello schema sopra esposto, i primi tre record compongono una classe di equivalenza, così come gli ultimi due).

Una tipologia particolare di aggregazione è la *k-anonimity*. Si dice che un set di dati possiede la proprietà di k-anonimato quando ciascun individuo presente nel set risulta indistinguibile da almeno *k-1* altri individui rispetto ad ogni combinazione di identificatori. Nell'esempio sopra, il dataset risulta 2-anonimo.

Questo tipo di anonimizzazione è particolarmente soggetta ad inferenza: prendendo nuovamente la tabella riportata come esempio, se in qualche modo si venisse a conoscenza, sempre sfruttando informazioni aggiuntive, che uno specifico individuo è nato nel 1964 e fa parte del dataset, allora si ricaverebbe in automatico la sua diagnosi di infarto.

3.6.2 l-diversity e t-closeness

La proprietà di *l-diversity* estende la k-anonimity al fine di minimizzare il rischio di attacchi di inferenza; ciò è ottenibile assicurandosi che, in ogni classe di equivalenza, ogni attributo sensibile (nell'esempio del paragrafo precedente, la diagnosi) assuma almeno *l* valori diversi.

Aumentando *l*, aumenta la variabilità dei valori, per cui un attaccante che ha a disposizione una certa conoscenza di background su un soggetto specifico viene lasciato comunque ad un'identificazione incerta.

Tabella 5: Esempio di dataset con due classi di equivalenza 3-diverse

Età	CAP	Nazionalità	Diagnosi
≤ 40	37*	*	Infarto
≤ 40	37*	*	Infezione virale
≤ 40	37*	*	Cancro
≤ 40	37*	*	Cancro
> 40	35*	*	Cancro
> 40	35*	*	Infarto
> 40	35*	*	Infezione virale
> 40	35*	*	Infezione virale

La *t-closeness* rappresenta un ulteriore raffinamento del processo: un dataset presenta la proprietà di t-closeness se, per ogni classe di equivalenza che lo compone, la differenza tra la distribuzione di un attributo sensibile all'interno della classe stessa e la distribuzione dell'attributo all'interno dell'intero dataset è minore di una determinata soglia *t*. Poiché spesso la distribuzione statistica complessiva di un set di dati è pubblica (o comunque facilmente derivabile), facendo in modo che quella stessa distribuzione sia rispecchiata anche nelle singole classi, si limitano le informazioni che un attaccante può ricavare sulle proprietà statistiche di queste, minimizzando il rischio che si possano ricavare collegamenti utili tra identificatori di un particolare gruppo e determinati attributi sensibili.

I concetti di l-diversity e t-closeness risultano comunque molto complessi e di recente introduzione nel panorama della protezione dei dati. Andare ulteriormente nei dettagli di tali tecniche esula dagli scopi di questa trattazione; si lasciano in bibliografia i riferimenti ad alcune pubblicazioni di particolare interesse che sviluppano il tema in maniera più tecnica e approfondita [30] [31].

Conclusioni

Come anticipato nell'introduzione, lo scopo con cui questo elaborato è stato redatto è quello di fornire un resoconto "ad alto livello", ma abbastanza ampio, dei concetti di pseudonimizzazione e anonimizzazione e di come queste si inseriscono nel panorama della gestione della privacy.

A chiusura della trattazione, si vuole innanzitutto mettere in guardia dal rischio di non riservare la giusta attenzione alla sezione riguardante la normativa, argomento spesso vissuto dai più come ostico, ma che in realtà *deve* essere il punto di partenza per qualsiasi progetto. La legislazione definisce lo spazio di manovra all'interno del quale si deve muovere l'ingegnere che si trova a dover affrontare il tema della privacy: prima di cimentarsi nello sviluppo di un sistema di gestione dei dati, è fondamentale conoscere i doveri a cui è obbligatorio sottostare in quanto progettisti.

Nelle sezioni successive, si è cercato, poi, di chiarire, in termini più tecnici, le caratteristiche delle due strategie in esame e di descrivere quelle che, ad oggi, sono le tecniche più comuni nell'applicarle. Non sono stati portati esempi veramente concreti di implementazione, in quanto ciascuna delle metodologie presentate porta con sé un bagaglio di pubblicazioni davvero imponente (e.g. [32], [33], [34]); si è valutato essere molto più efficace esporre una descrizione dell'argomento varia e generale, che di certo fornisce le conoscenze e gli strumenti di base necessari per navigare nella letteratura in materia, lasciando invece al lettore l'approfondimento delle singole strategie che più gli risultano interessanti.

Concludendo, si vuole ricordare che, come emerso più volte anche nel corpo dell'elaborato, il tema della gestione dei dati personali (e, dunque, di quelli sanitari) è in continua mutazione ed evoluzione. L'invito è quindi a mantenere sempre gli occhi puntati sulle nuove frontiere della tecnologia in materia, perché ciò che oggi risulta pionieristico e all'avanguardia potrebbe presto diventare obsoleto, costituendo un pericolo per uno dei beni più preziosi che possediamo, la nostra privacy.

Bibliografia

[1] Consiglio d'Europa, *European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14*, Nov. 4, 1950, ETS 5 Available: https://www.refworld.org/docid/3ae6b3b04.html (accessed July 20, 2022).

[2] B. Saetta, «Regolamento generale per la protezione dei dati, » *Protezione dati personali - Data Protection*, June 2, 2018. [Online].
Available: https://protezionedatipersonali.it/regolamento-generale-protezione-dati (accessed Aug. 10, 2022).

[3] Parlamento e Consiglio d'Europa, *Regolamento (UE) n. 2016/679 del Parlamento europeo e del Consiglio del 27 aprile 2016 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95,* 2016.

[4] *D. Lgs. 30 giugno 2003 n. 196 - Codice in materia di protezione dei dati personali,* 2003.

[5] «Codice in materia di protezione dei dati personali, » *Wikipedia.it*. [Online].
Available: https://it.wikipedia.org/wiki/Codice_in_materia_di_protezione_dei_dati_personali (accessed Aug. 12, 2022).

[6] *D. Lgs. 10 agosto 2018 n. 101 - Disposizioni per l'adeguamento della normativa nazionale alle disposizioni del regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio,* 2018.

[7] G. Nucci, «Data Protection e GDPR: un sistema complesso o complicato?, » *Risk & Compliance*, Dec. 3, 2018. [Online].
Available: https://www.riskcompliance.it/news/data-protection-e-gdpr-un-sistema-complesso-o-complicato/ (accessed Aug. 4, 2022).

[8] European Union Agency for Cybersecurity (ENISA), *Tecniche di pseudonimizzazione e migliori pratiche - Raccomandazioni per sviluppare tecnologie conformi alle disposizioni in materia di protezione dei dati e privacy,* 2019.
Available: https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices_it/at_download/file

[9] J. von Neumann, «Various Techniques Used in Connection with Random Digits,» in *Monte Carlo Method*, Washington (DC), USA: Government Printing Office, 1951, pp. 36-38.

[10] M. Cortina Borja e J. Haigh, *The Birthday Problem*, 2007, pp. 124-127, doi: 10.1111/j.1740-9713.2007.00246.x .

[11] L. Demir, A. Kumar, M. Cunche e C. Lauradoux, «The Pitfalls of Hashing for Privacy», *IEEE Communication Surveys and Tutorials*, IEEE Communications Society, 2018, pp. 551-565, doi: 10.1109/COMST.2017.2747598ff .

[12] G. Sparacino, *Slide dell'insegnamento "Informatica Medica" (Corso di Laurea Triennale in Ingegneria Biomedica, Università degli Studi di Padova),* 2022.

[13] *Secure Hash Standard, Federal Information Processing Standards Publication FIPS PUB 180-4*, National Institute of Standards and Technology, 2015, doi: 10.6028/NIST.FIPS.180-4 .

[14] M. Stevens, *Attack on Hash Functions and Applications,* Amsterdam, Netherlands: Ipskamp Drukkers, 2012, ISBN: 978-94-6191-317-3.

[15] D. Toma, A. Perez, D. Borrione e E. Bergeret, «Design of a Proven Correct SHA Circuit,» 2004, doi: 10.1109/ICEEC.2004.1374373 .

[16] R. Shirey, «Internet Security Glossary, Version 2, » *RFC editor*, Aug., 2007. [Online]. Available: https://www.rfc-editor.org/rfc/rfc4949 (accessed Aug. 20, 2022).

[17] «Crittografia,» *Enciclopedia del Novecento - Treccani*, [Online]. Available: https://www.treccani.it/enciclopedia/crittografia_(Enciclopedia-del-Novecento) (accessed Aug. 15, 2022)

[18] SSL Information, «Symmetric vs. Asymmetric Encryprion - What are the differences?,» *SSL2Buy*, [Online]. Available: https://www.ssl2buy.com/wiki/symmetric-vs-asymmetric-encryption-what-are-differences (accessed Aug. 23 2022).

[19] *Data Encryption Standard (DES), Federal Information Processing Standard Publication FIPS PUB 46-3,* National Institute of Standards and Technology, 1999.

[20] *L. 31 dicembre 1996 n. 675 - Tutela delle persone e di altri soggetti rispetto al trattamento dei dati personali,* 1996.

[21] ISO e IEC, «ISO/IEC 29100 - Privacy Framework,» *ISO.org,* 2011, [Online]. Available: https://www.iso.org/obp/ui/es/#iso:std:iso-iec:29100:ed-1:v1:en (accessed Aug. 27, 2022)

[22] AEPD e EDPS, «10 misunderstandigs related to anonymisation,» *EDPS.europa.eu*, 2021, [Online]. Available: https://edps.europa.eu/data-protection/our-work/publications/papers/aepd-edps-joint-paper-10-misunderstandings-related_en (accessed Aug. 27, 2022)

[23] M. Massimini, «Anonimizzazione dei dati personali: significato, benefici e dubbi in ottica GDPR,» *Privacy.it*, May 11, 2021. [Online]. Available: https://www.privacy.it/2021/05/11/anonimizzazione-gpdr-massimini/ (accessed July 20, 2022).

[24] Information Commissioner's Office (ICO), *Anonymisation managing data protection risk code of practice,* 2012. [Online] Available: https://ico.org.uk/media/1061/anonymisation-code.pdf

[25] Article 29 Data Protection Working Party, *Opinion 05/2014 on Anonymisation Techniques,* 2014. [Online] Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[26] F. K. Dankar e K. El Emam, «Practicing Differential Privacy in Health Care: A Review,» 2013, doi: 10.5555/2612156.2612159 .

[27] «Harvard University Privacy Tools Project,» *Harvard.edu* [Online]. Available: https://privacytools.seas.harvard.edu/differential-privacy (accessed Sept. 6, 2022).

[28] C. Dwork, F. McSherry, K. Nissim e A. Smith, «Calibrating Noise to Sensitivity in Private Data Analysis,» *Theory of Cryptography*, Springer, 2006, doi: 10.1007/11681878_14 .

[29] J. Near, D. Darais e K. Boeckl, «Differential Privacy for Privacy-Preserving Data Analysis: An Introduction to our Blog Series,» *NIST.gov*, July 27, 2020. [Online]. Available: https://www.nist.gov/blogs/cybersecurity-insights/differential-privacy-privacy-preserving-data-analysis-introduction-our (accessed Sept. 6, 2022).

[30] A. Machanavajjhala, J. Gehrke, D. Kifer e M. Venkitasubramaniam, «L-diversity: Privacy beyond k-anonymity,» presented at the *IEEE 22nd Int. Conf. on Data Engineering*, 2006, pp. 24-24, DOI: 10.1109/ICDE.2006.1 .

[31] N. Li, T. Li e S. Venkatasubramanian, «T-closeness: Privacy beyond k-anonymity and l-diversity,» presented at the *IEEE 23rd Int. Conf. on Data Engineering*, 2007, pp. 106-115, DOI: 10.1109/ICDE.2007.367856.2007 .

[32] T. Neubauer e J. Heurix, «A methodology for the pseudonymisation of medical data,» *International Journal of Medical Informatics,* 2010, pp. 190-204 doi: 10.1016/j.ijmedinf.2010.10.016 .

[33] B. Riedl, V. Grascher, S. Fenz e T. Neubauer, «Pseudonymisation for improving the Privacy in e-Health Applications,» presented at the *41st Hawaii Int. Conf. on System Sciences*, 2008, doi: 10.1109/HICSS.2008.366 .

[34] M. Al-Zubadie, Z. Zhang e J. Zhang, «PAX: Using Pseudonymisation and Anonymisation to Protect Patients' Identities and Data in the Healthcare System,» *International Journal of Environmental Research and Public Health,* 2019, doi: 10.3390/ijerph16091490 .

[35] «Data Encryption Standard,» *Wikipedia.it*, [Online]. Available: https://it.wikipedia.org/wiki/Data_Encryption_Standard (accessed Aug. 1, 2022).