# LINEE GUIDA PER L'UTILIZZO DI TECNICHE DI PSEUDONOMIZZAZZIONE DI DATASET DELL'ENTE LOCALE

# Generalità

Qui di seguito vengono indicati alcune metodologie, strategie e best practice da utilizzare come tecniche di pseudonomizzazione nei dataset dell'ente locale e su tutte le tipologie di dati trattati secondo i principi di accountability, compliance, privacy by default e privacy by design previsti dal vigente Regolamento UE 679/2016 (GDPR).

Tali tecniche sono peraltro state suggerite e supportate da ENISA nel 2019

**TERZE PARTI COME ENTITÀ DI PSEUDONIMIZZAZIONE** - In questo scenario, la pseudonimizzazione viene eseguita da una terza parte (non dal responsabile del trattamento), che va poi a inoltrare i dati al titolare del trattamento. In questo caso il titolare del trattamento non può accedere agli identificativi degli interessati (in quanto la terza parte non ricade sotto il suo controllo). La sicurezza e la protezione dei dati risultano in tal modo potenziate a livello del titolare del trattamento conformemente al principio della minimizzazione dei dati.

- **TECNICHE DI DIFESA DA ATTACCANTI INTERNI ED ESTERNI-** Un attaccante interno dispone di specifiche conoscenze, capacità o autorizzazioni (rapportate al suo obiettivo). Nel contesto della pseudonimizzazione, ciò implica che l'attaccante sia in grado di ottenere informazioni sulla chiave di pseudonimizzazione e/o relative informazioni di rilievo (per es., il dipendente malintenzionato di un contraente). Contrariamente agli attaccanti interni, un attaccante esterno non ha accesso diretto alla chiave di pseudonimizzazione o ad altre informazioni pertinenti. Tuttavia, questo tipo di attaccante può avere accesso a un set di dati pseudonimizzato, oltre a essere in grado di eseguire l'operazione di pseudonimizzazione su valori arbitrari dei dati di ingresso da lui selezionati. L'obiettivo di un attaccante esterno è di accrescere le sue informazioni riguardo al set di dati pseudonimizzato, venendo per es. a conoscenza dell'identità associata a un determinato pseudonimo (ricavando ulteriori informazioni su tale identità dai dati aggiuntivi contenuti nel set di dati connesso a tale pseudonimo).
- **PRINCIPALI TECNICHE DI ATTACCO**- Esistono tre principali tecniche per decodificare una funzione di pseudonimizzazione: attacchi a forza bruta (ricerca esaustiva), ricerca in un dizionario e inferenze.
    - **Attacco a forza bruta**. La funzionalità di tale tecnica di attacco è subordinata alla capacità dell'attaccante di calcolare la funzione di pseudonimizzazione (ovvero, non vi è alcuna chiave di pseudonimizzazione) o dal suo accesso a un 'implementazione black box della funzione di pseudonimizzazione. A seconda dell'obiettivo dell'attacco, potrebbero essere applicabili ulteriori condizioni.
    - **Ricerca in un dizionario**- La ricerca nel dizionario è un'ottimizzazione dell'attacco a forza bruta, in quanto può consentire di risparmiare sui costi computazionali. L'attaccante infatti, per effettuare una completa reidentificazione o discriminazione, deve confrontarsi con una grande quantità di pseudonimi. Di conseguenza, calcola preliminarmente un (enorme) insieme di pseudonimi e salva il risultato in un dizionario.
    - **Inferenze** Questo tipo di attacco si basa su alcune conoscenze generali (quali la distribuzione di probabilità o qualsiasi altra informazione aggiuntiva) di cui l'attaccante può disporre relativamente ad alcuni (o tutti) i titolari di pseudonimo, alla funzione di pseudonimizzazione o al set di dati. La ricerca esaustiva e la ricerca in un dizionario presuppongono implicitamente che tutti gli identificativi presentino la stessa probabilità o frequenza di occorrenze.

**TECNICHE DI PSEUDONIMIZZAZIONE** ad oggi più comuni.

- **A) PSEUDONIMIZZAZIONE DI UN SINGOLO IDENTIFICATORE** - Partendo dalla pseudonimizzazione di un singolo identificativo, vengono di seguito elencati alcuni possibili approcci, con relativi vantaggi e limiti. Il **contatore** è la più semplice forma di pseudonimizzazione. Gli identificativi sono sostituiti da un numero scelto da un contatore monotono. I vantaggi del contatore derivano dalla sua semplicità, che lo rende un buon candidato per set di dati non complessi e di piccole dimensioni**. In termini di protezione dei dati, il contatore fornisce pseudonimi che non sono associabili agli identificativi iniziali (sebbene il carattere sequenziale del contatore possa comunque fornire informazioni sull'ordine dei dati all'interno di un set).** Tale soluzione può tuttavia presentare problemi di implementazione e scalabilità in caso di set di dati più sofisticati e di grandi dimensioni, poiché occorrerebbe in tal caso archiviare la tabella completa di mappatura della pseudonimizzazione.
- **B) Generatore di numeri casuali** Il generatore di numeri casuali è un meccanismo che produce, all'interno di un set, valori che presentano tutti la stessa probabilità di essere selezionati, risultando pertanto imprevedibili . Questo approccio è simile al contatore, con la differenza che all'identificativo viene assegnato un numero casuale**. Il generatore di numeri casuali fornisce una solida protezione dei dati (poiché, a differenza del contatore, per creare ogni pseudonimo si va a utilizzare un numero casuale, rendendo difficile l'estrazione di informazioni riguardanti l'identificativo iniziale, a meno che la tabella di mappatura non sia stata compromessa).**
- **C)Funzione crittografica di hash-** Una funzione crittografica di hash prende stringhe di input di lunghezza arbitraria e le associa ad output di lunghezza fissa . Essa presenta le proprietà riportate di seguito:
  - 1) **Unidirezionale**: è computazionalmente impraticabile trovare input che si associno a output specificati in precedenza.
  - 2) **Senza collisioni**: è computazionalmente impraticabile trovare due input distinti che si associno al medesimo output. Si applica una funzione crittografica di hash direttamente all'identificativo, così da ottenere lo pseudonimo corrispondente.

**D) Codice di autenticazione del messaggio -** Questa primitiva può essere considerata come una funzione di hash con chiave. È molto simile alla soluzione precedente, salvo per l'introduzione di una chiave segreta atta a generare lo pseudonimo. Se non si è a conoscenza di tale chiave, non è possibile associare gli identificativi agli pseudonimi. HMAC è di gran lunga la più diffusa modalità di codice di autenticazione del messaggio impiegata nei protocolli Internet. **Il Codice di autenticazione del messaggio è generalmente considerato una tecnica di pseudonimizzazione solida dal punto di vista della protezione dei dati poiché, a meno che la chiave non sia stata compromessa, è impossibile decodificare lo pseudonimo**.

**E) Crittografia-** la crittografia simmetrica (deterministica) e, in particolare, le cifrature a blocchi come l'AES, insieme alle loro modalità operative. Si utilizza una cifratura a blocchi per crittografare un identificativo servendosi di una chiave segreta, che è sia chiave di pseudonimizzazione sia la chiave da impiegare per il recupero. L'uso di cifrature a blocchi ai fini della pseudonimizzazione deve misurarsi con la dimensione del blocco. Gli identificativi possono avere dimensioni minori o maggiori rispetto alla dimensione del blocco di input della cifratura a blocchi.

**Inoltre nel nostro caso è necessario stabilire delle strategie di pseudonomizzazione che possono essere: a) pseudonimizzazione deterministica, b) randomizzata al documento e c ) completamente randomizzata**.

- **Pseudonimizzazione deterministica**. In tutti i database e ogniqualvolta appare, *ID* viene sempre sostituito con lo stesso pseudonimo. Esso è uniforme all'interno di un database e tra

database differenti. Per implementare tale modalità, occorre anzitutto estrarre l'elenco degli identificativi univoci contenuti nel database. In secondo luogo, l'elenco viene associato agli pseudonimi e gli identificativi sono infine sostituiti agli pseudonimi nel database.

- **Pseudonimizzazione randomizzata al documento**. Ogniqualvolta $ID$ appare in un database, viene sostituito con un differente pseudonimo e così via).
- **Pseudonimizzazione completamente randomizzata**. Infine, per ogni occorrenza di $ID$ all'interno di un database $A$ o $B$, $ID$ viene sostituito con uno pseudonimo differente.

Inoltre la a scelta di una tecnica e di una strategia di pseudonimizzazione dipende da diversi parametri, in primo luogo dal livello di protezione dei dati e dalla funzionalità del set di dati pseudonimizzato (che l'entità di pseudonimizzazione intende raggiungere). **In termini di protezione dati , l'RNG (random number generator), i codici di autenticazione del messaggio e la crittografia rappresentano le tecniche più efficaci, appositamente mirate a contrastare ricerche esaustive, ricerche nel dizionario e congetture**.

- **RECUPERO** Poichè, per definizione, l'uso di informazioni aggiuntive è fondamentale per la pseudonimizzazione, l'entità di pseudonimizzazione è chiamata a implementare un meccanismo di recupero. Questo meccanismo può essere più o meno complesso, a seconda della funzione di pseudonimizzazione. Esso consiste in genere nell'utilizzare uno pseudonimo e un segreto di pseudonimizzazione $S$, per recuperare l'identificativo corrispondente $ID$. **Questo meccanismo di recupero potrebbe rivelarsi necessario anche per consentire agli interessati di esercitare i propri diritti (ai sensi degli articoli 12-21 del RGPD).**
- **PROTEZIONE DELLA CHIAVE DI PSEUDONIMIZZAZIONE**- In primo luogo, occorre isolare la chiave di pseudonimizzazione dal set di dati, non dovendo mai, per es., essere gestiti in uno stesso file (in caso contrario, un attaccante potrà facilmente recuperare gli identificativi). In secondo luogo, occorre eliminare in modo sicuro la chiave di pseudonimizzazione da qualsiasi supporto non sicuro (memoria e sistemi). Terzo, bisogna far sì che solide politiche di controllo dell'accesso assicurino che solo le entità autorizzate possano accedere a tale chiave.
- **TECNICHE AVANZATE DI PSEUDONIMIZZAZIONE**. Oltre al semplice hashing di dati, strutture più avanzate quali gli alberi Merkle utilizzano hash di set di hash, per es. h3=hash(h1,h2), per ottenere pseudonimi ben articolati, che possono essere solo parzialmente identificati. **Tra le altre soluzioni degne di interesse, figurano la dimostrazione a conoscenza zero e il più vasto ambito delle credenziali basate su attributi**.
- **PSEUDONIMIZZAZIONE DEGLI INDIRIZZI IP**- **Lo status giuridico degli indirizzi IP è stato dibattuto dalla Corte di giustizia dell'Unione europea nell'ambito della causa C-582/14 Breyer contro la Repubblica Federale di Germania. Che siano statici o dinamici, gli indirizzi IP vengono comunque considerati dati personali. Ciò ha trovato peraltro conferma nel parere 4/2007 sul concetto di dati personali espresso dal Gruppo dell'articolo 29 per la tutela dei dati. Le tracce di database o di rete contenenti indirizzi IP devono essere pertanto protette, e la pseudonimizzazione rappresenta ovviamente una funzione di protezione, che da un lato consente l'uso di indirizzi IP, e dall'altro impedisce la loro associabilità a individui specifici. Ciò detto, per scegliere la tecnica di pseudonimizzazione per gli indirizzi IP più opportuna, occorre trovare un buon compromesso tra funzionalità e protezione dei dati. In effetti, anche in questo caso il titolare del trattamento potrebbe aver bisogno di calcolare statistiche o rilevare modelli (in caso di errata configurazione di un dispositivo o per la qualità dei servizi) nel database pseudonimizzato.**

- **PSEUDONIMIZZAZIONE E LIVELLO DI PROTEZIONE DEI DATI**. Il principale problema connesso alla pseudonimizzazione di indirizzi IP è rappresentato dalla dimensione dello spazio di input (dominio dell'identificativo), dal momento che ci sono solo 232 indirizzi IP possibili. **Per la protezione dei dati vanno dunque preferite altre funzioni di pseudonimizzazione, come il codice di autenticazione del messaggio, la crittografia con chiave segreta ad hoc o il generatore di numeri casuali**. Come in precedenza affermato, un attaccante non può eseguire gli stessi attacchi, in quanto questi metodi utilizzano una chiave segreta (Codice di autenticazione del messaggio e crittografia) oppure una fonte di casualità (per l'RNG). Può essere impiegato anche un contatore, ma occorre prestare attenzione alle possibili previsioni (dovute alla sua natura sequenziale).
- **PSEUDONIMIZZAZIONE E LIVELLO DI FUNZIONALITÀ**- la possibilità di minimizzare il livello/campo applicativo della pseudonimizzazione degli indirizzi IP e la scelta della strategia di pseudonimizzazione (modalità).
- **PSEUDONIMIZZAZIONE DEGLI INDIRIZZI E-MAIL**- Quando gli indirizzi e-mail vengono utilizzati come identificativi, è particolarmente importante proteggerli. Nel nostro caso gli indirizzi e-mail sono considerati identificativi. Generalmente il processo di pseudonimizzazione è eseguito da un'entità di pseudonimizzazione (nel nostro caso il titolare del trattamento dei dati)
- **FUNZIONE CRITTOGRAFICA DI HASH** All'interno delle funzioni crittografiche di hash, è opportuno sottolineare che i fornitori di servizi spesso condividono indirizzi e-mail con terze parti, semplicemente eseguendo l'hashing. I valori degli hash crittografici sono utili in determinate condizioni, per es. per la codifica interna degli indirizzi e-mail (come nel caso di attività di ricerca) e come meccanismo di convalida/integrità per un titolare del trattamento dei dati. Le funzioni di hash possono anche essere utilizzate per pseudonimizzare parti di un indirizzo e-mail.
- **CODICE DI AUTENTICAZIONE DEL MESSAGGIO** **Rispetto al semplice hashing, un codice di autenticazione del messaggio offre notevoli vantaggi in termini di protezione dei dati anche per la pseudonimizzazione dell'indirizzo e-mail, purché la chiave segreta sia archiviata in modo sicuro**. **Inoltre, l'entità di pseudonimizzazione può utilizzare chiavi segrete diverse, per settori diversi, per generare per es. pseudonimi di settore diversi per lo stesso indirizzo e-mail. È possibile utilizzare un Codice di autenticazione del messaggio per impedire al titolare del trattamento di accedere agli indirizzi e-mail nei casi in cui l'accesso agli pseudonimi sia sufficiente per lo scopo specifico del trattamento. Usare un MAC per generare indirizzi e-mail pseudonimizzati con alcune funzionalità**
- **CRITTOGRAFIA CON CONSERVAZIONE DEL FORMATO (FPE)** Uno schema di database può prevedere un determinato tipo di dati per campi specifici. **Se il titolare del trattamento dei dati non ha bisogno di conservare gli indirizzi e-mail iniziali, ma è tenuto comunque a conservare un elenco pseudonimizzato mantenendo la struttura del database, la crittografia con protezione del formato si rivela una procedura adeguata**. Esistono diverse implementazioni note relative alla crittografia con conservazione del formato, basate su schemi crittografici noti.