UNIVERSITÀ DEGLI STUDI DI CATANIA

Dipartimento di Matematica e Informatica

Corso di Laurea Triennale in Informatica

Giulio Pedicone

Pseudonimizzazione (Attacco)

Progetto Internet Security

Prof.re: Giampaolo Bella

Anno Accademico 2023 - 2024

Indice

Capitolo 1

Definizione di Pseudonimizzazione

La **pseudonimizzazione** è una tecnica di trattamento dei dati che mira a **proteggere la privacy degli individui** sostituendo i dati identificativi con pseudonimi. In questo modo, le informazioni originali non possono essere attribuite a una specifica persona senza l'uso di ulteriori informazioni che sono conservate separatamente.

Questa tecnica è utile per ridurre i rischi associati al trattamento dei dati personali, in quanto limita la possibilità di identificare gli individui senza avere accesso alle informazioni aggiuntive necessarie per invertire il processo di pseudonimizzazione.

La pseudonimizzazione è un metodo che permette di **sostituire i dati originali** (ad esempio, un indirizzo e-mail o un nome) **con un alias** o **pseudonimo**. È un **processo reversibile** che **de-identifica i dati**, consentendo la **re-identificazione** in seguito, se necessario. Questa tecnica è altamente raccomandata dal Regolamento Generale sulla Protezione dei Dati (GDPR) come uno dei metodi di protezione dei dati.

1.1 Utilizzo della Pseudonimizzazione nella Protezione dei Dati

La pseudonimizzazione facilita il trattamento dei dati personali, riducendo il rischio di esposizione di dati sensibili a personale e dipendenti non autorizzati.

1.1.1 Esempio di utilizzo

Ad esempio, quando si inviano fogli Excel contenenti dati sensibili via e-mail. Sebbene il mittente e il destinatario delle e-mail siano autorizzati ad accedere a tali informazioni, il supporto IT ha anche accesso a quelle e-mail. Ora immagina che si trattasse di bonus per il top management o informazioni sui salari aziendali. Quando i dati sono pseudonimizzati, c'è meno possibilità di esporre dati personali, poiché i record dei dati diventano non identificabili, rimanendo comunque adatti per l'elaborazione e l'analisi dei dati.

1.1.2 Cos'è un Pseudonimo?

In questo contesto, un pseudonimo è un **identificatore associato a un individuo**. Proprio come gli scrittori usano pseudonimi per nascondere la loro identità e proteggere la loro privacy, gli pseudonimi vengono utilizzati per lo stesso scopo nella **protezione dei dati**. Un pseudonimo può essere un numero, una lettera, un carattere speciale o una qualsiasi combinazione di questi legati a un dato personale specifico o a un individuo, rendendo quindi i **dati più sicuri** da usare in un contesto aziendale.

Capitolo 2

Il GDPR e la Pseudonimizzazione

2.1 Pseudonimizzazione nel GDPR

L'**European Union Agency for Cybersecurity** (ENISA) lavora dal 2004 per rendere l'Europa sicura dal punto di vista cibernetico. ENISA collabora con l'Unione Europea, i suoi Stati membri, il settore privato e i cittadini europei per sviluppare consigli e raccomandazioni sulle buone pratiche in materia di sicurezza delle informazioni. Assiste gli Stati membri dell'UE nell'implementazione della legislazione europea pertinente e lavora per migliorare la resilienza delle infrastrutture e delle reti di informazione critiche in Europa. ENISA cerca di potenziare l'esperienza esistente negli Stati membri dell'UE supportando lo sviluppo di comunità transfrontaliere impegnate a migliorare la sicurezza delle reti e delle informazioni in tutta l'UE. Dal 2019, l'ENISA ha elaborato schemi di certificazione della cybersecurity.Lo scorso 3 dicembre, l'European Union Agency for Cybersecurity (ENISA, precedentemente denominata European Union Agency for Network and Information Security) ha pubblicato un importante documento sulla pseudonimizzazione dal titolo *"**Pseudonymisation techniques and best practices**"* in cui vengono proposti alcuni possibili scenari di attacco a cui sono esposti i nostri dati e le migliori tecniche di difesa oggi in circolazione.

2.2 Definizione di Pseudonimizzazione nel GDPR

Nell'Articolo 4(5) del GDPR, i**l processo di pseudonimizzazione** è definito come:

> "Il 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 siano soggette a misure tecniche e organizzative atte a garantire che i dati personali non siano attribuiti a una persona fisica identificata o identificabile."

2.2.1 Benefici della Pseudonimizzazione secondo il GDPR

Se sei un Responsabile della Protezione dei Dati (DPO), puoi vedere l'attrattiva e i benefici della pseudonimizzazione. Permette di **identificare i dati** se necessario, ma li rende **inaccessibili agli utenti non autorizzati** e consente ai responsabili e agli incaricati del trattamento dei dati di ridurre il rischio di una potenziale violazione dei dati e proteggere i dati personali.

Il GDPR richiede di adottare tutte le misure tecniche e organizzative appropriate per proteggere i dati personali, e la pseudonimizzazione può essere un metodo appropriato se si desidera mantenere l'utilità dei dati.

2.2.2 I Dati Pseudonimizzati sono ancora Dati Personali secondo il GDPR?

Un pseudonimo è ancora **considerato un dato personale** secondo il GDPR poiché il **processo è reversibile** e, con una chiave appropriata, è **possibile identificare l'individuo**. Il Considerando 26 spiega:

> "...i dati personali che hanno subito pseudonimizzazione, che potrebbero essere attribuiti a una persona fisica utilizzando informazioni aggiuntive, dovrebbero essere considerati informazioni su una persona fisica identificabile."

Inoltre, durante una violazione dei dati, una chiave di crittografia potrebbe essere esposta, mettendo a rischio anche i dati pseudonimizzati.

2.3 I Dati Anonimizzati sono ancora considerati Dati Personali?

Il GDPR si preoccupa solo del trattamento dei dati personali relativi a una persona fisica che consente l'identificazione di un individuo direttamente o indirettamente.

Se i **dati** sono **anonimizzati** in modo che gli individui non possano più essere identificati, il **GDPR** semplicemente **non li considera più dati personali**. Tuttavia, l'anonimizzazione dei dati può spesso distruggere il valore che i dati hanno per la tua organizzazione.

Capitolo 3

Anonimizzazione VS Pseudonimizzazione

Anonimizzazione e **pseudonimizzazione** sono tecniche utilizzate per proteggere l'identità degli individui nei dati, ma non sono sinonimi. In questo documento, esploreremo come funzionano entrambe le tecniche e come vengono trattate dal GDPR e dalle raccomandazioni dell'ENISA.

3.1 Pseudonimizzazione

Con la pseudonimizzazione, se sei autorizzato ad accedere a tali informazioni, avrai la chiave che permetterà di **de-identificare i dati**. La pseudonimizzazione è una tecnica che altera **reversibilmente** i dati in modo che possano essere riconosciuti in un secondo momento, se necessario.

3.2 Anonimizzazione

L'anonimizzazione è una tecnica che **altera irreversibilmente i dati** in modo che un individuo non possa più essere identificato direttamente o indirettamente.

3.3 Perché Optare per la Pseudonimizzazione?

Nelle operazioni quotidiane di qualsiasi azienda, molti dati sensibili passano attraverso i dipartimenti HR, marketing o IT, e la pseudonimizzazione può aiutare a **ridurre il rischio** e **prevenire eventuali violazioni dei dati**.

3.3.1 Recital 28 del GDPR

> "L'applicazione della pseudonimizzazione ai dati personali può ridurre i rischi per gli interessati e aiutare i responsabili e gli incaricati del trattamento a soddisfare i loro obblighi di protezione dei dati."

La pseudonimizzazione non solo protegge i dati, ma supporta anche la conformità generale al GDPR di qualsiasi organizzazione.

3.4 Raccomandazioni sull'Uso della Pseudonimizzazione e Anonimizzazione

3.4.1 Anonimizzazione

Si raccomanda vivamente di anonimizzare i dati personali negli ambienti non di produzione, utilizzati per lo sviluppo, il testing e la formazione.

3.4.2 Pseudonimizzazione sui Sistemi di Produzione

Quando si progetta la protezione dei dati per i sistemi di produzione live, si consiglia di utilizzare la pseudonimizzazione.

3.4.3 Automatizzazione

Sia la pseudonimizzazione che l'anonimizzazione dovrebbero essere automatizzate, così come le convalide dei dati, per ridurre al minimo gli errori umani.

3.4.4 Scelta della Tecnica Appropriata

Le tecniche utilizzate devono essere applicabili a uno specifico caso d'uso o sistema.

3.5 Raccomandazioni dell'ENISA per la Pseudonimizzazione

L'European Union Agency for Cybersecurity (ENISA) ha pubblicato un rapporto su "Pseudonymisation Techniques and Best Practices", in risposta alle sfide dell'implementazione della pseudonimizzazione nella pratica.

3.5.1 Criteri per la Scelta delle Tecniche di Pseudonimizzazione

La guida discute i criteri per la scelta delle tecniche di pseudonimizzazione appropriate, come la protezione dei dati, la scalabilità e il recupero.

3.5.2 Conclusioni del Rapporto

Il rapporto ha concluso che non esiste una soluzione unica che funzioni per tutte le industrie o tutti gli scenari.

> "... non esiste una soluzione unica e facile alla pseudonimizzazione che funzioni per tutti gli approcci in tutti i possibili scenari. Al contrario, richiede un alto livello di competenza per applicare un processo di pseudonimizzazione robusto, possibilmente riducendo la minaccia di discriminazione o attacchi di re-identificazione, mantenendo il grado di utilità necessario per il trattamento dei dati pseudonimizzati."

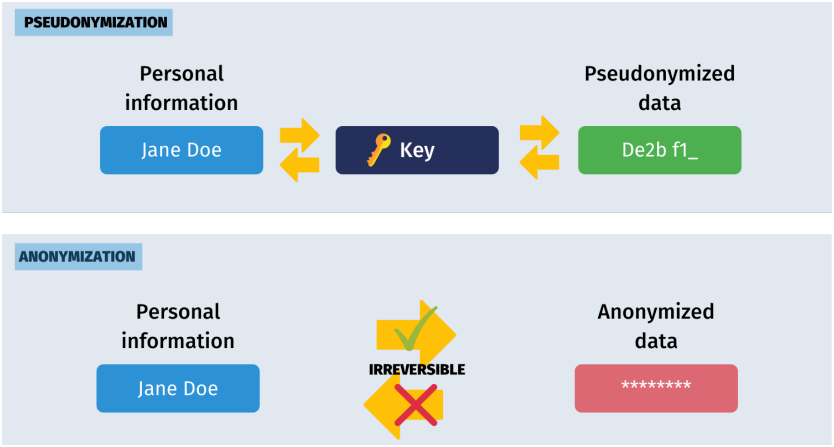

Figura 3.1: Pseudonimizzazione VS Anonimizzazione

Capitolo 4

Modello di attacco

Come detto in precedenza, l'obiettivo principale della pseudonimizzazione è limitare la collegabilità tra un dataset pseudonimizzato e i detentori dei pseudonimi al fine di proteggere l'identità dei soggetti dati. Questo tipo di protezione è generalmente pensato per contrastare i tentativi di un avversario di eseguire un **attacco di re-identificazione**.

4.1 Avversari Interni ed Avversari Esterni

Questo capitolo considera i possibili modelli di attacco e **diversi tipi di attacchi di re-identificazione** che sono importanti per la pseudonimizzazione. A tale scopo, vengono esaminati i concetti di avversari interni ed esterni, analizzando i loro possibili ruoli nei scenari di pseudonimizzazione discussi in precedenza nel rapporto.

4.1.1 Avversari Interni

Secondo la definizione comune nel campo della sicurezza informatica, un **avversario interno** è una **persona con conoscenze specifiche**, capacità o permessi (riguardo all'obiettivo dell'avversario).

Nel contesto della pseudonimizzazione, ciò implica che l'avversario sia in grado di **ottenere informazioni sul segreto di pseudonimizzazione** e/o altre informazioni significative rilevanti.

Ad esempio, un avversario interno potrebbe essere un dipendente che lavora per il responsabile del trattamento, oppure potrebbe essere situato all'interno di una terza parte fidata (agendo in questo caso come entità di pseudonimizzazione). Per default, le terze parti che potrebbero legittima-

mente avere accesso ai dati personali (come un'autorità di vigilanza o di forze dell'ordine) non vengono considerate avversarie.

4.1.2 Avversari Esterni

A differenza dell'avversario interno, un avversario esterno **non ha accesso diretto al segreto di pseudonimizzazione** o ad altre informazioni rilevanti. Tuttavia, questo tipo di avversario potrebbe **avere accesso a un dataset pseudonimizzato** e potrebbe essere in grado di **eseguire il processo di pseudonimizzazione** su valori di input arbitrari scelti dall'avversario.

L'obiettivo di un avversario esterno è aumentare le proprie conoscenze sul dataset pseudonimizzato, come ad esempio scoprire l'identità dietro un determinato pseudonimo e acquisire ulteriori informazioni su tale identità dai dati aggiuntivi presenti nel dataset associati al pseudonimo fornito.

4.2 Obiettivi degli Attacchi alla Pseudonimizzazione

A seconda del contesto e del metodo di pseudonimizzazione utilizzato, l'avversario può avere diversi obiettivi che intende raggiungere nei confronti dei dati pseudonimizzati, come il **recupero del segreto di pseudonimizzazione**, la **re-identificazione completa** o la **discriminazione.**

4.2.1 Segreto di Pseudonimizzazione

L'avversario si concentra sullo scoprire il segreto di pseudonimizzazione (cioè quando viene utilizzato il segreto di pseudonimizzazione). Questo tipo di attacco è il più grave, poiché **con il segreto di pseudonimizzazione** l'avversario è in grado di **re-identificare qualsiasi pseudonimo** nel dataset (re-identificazione completa o discriminazione), nonché di eseguire ulteriori processi di pseudonimizzazione sul dataset.

4.2.2 Re-identificazione Completa

Quando l'obiettivo dell'attacco è la re-identificazione completa, l'avversario desidera **collegare uno o più pseudonimi all'identità dei detentori degli pseudonimi.**

Il più grave attacco di re-identificazione completa consiste nella re-identificazione di tutti gli pseudonimi.

L'avversario può utilizzare due strategie per raggiungere questo obiettivo:

- Recuperare ogni identificatore dal corrispondente pseudonimo indipendentemente
- Recuperare il segreto di pseudonimizzazione

La forma meno grave degli attacchi di re-identificazione completa coinvolge un avversario che può solo re-identificare un sottoinsieme di pseudonimi nel dataset.

Ad esempio, consideriamo un dataset pseudonimizzato dei voti degli studenti di un corso universitario. Ogni voce del dataset contiene un pseudonimo corrispondente all'identità dello studente (nome e cognome) e un secondo pseudonimo sul genere dello studente (ad esempio, mappando gli studenti maschi a numeri dispari e le studentesse a numeri pari). Un avversario ha successo in un attacco di re-identificazione completa se riesce a recuperare il nome, il cognome e il genere di uno studente.

4.2.3 Discriminazione

L'obiettivo dell'attacco di discriminazione è **identificare le proprietà di un detentore di pseudonimo** (almeno una). Queste proprietà potrebbero non portare direttamente alla scoperta dell'identità del detentore dello pseudonimo, ma possono essere sufficienti per discriminare in qualche modo. È importante capire che l'avversario **non apprende l'identità del detentore dello pseudonimo** in questo caso, ma **solo alcune proprietà**.L'avversario **non è in grado di individuare l'esatto record** di dati di un determinato detentore di pseudonimo. Tuttavia, le informazioni aggiuntive acquisite possono già essere sufficienti per scopi di discriminazione che l'avversario intende eseguire, o possono essere utilizzate in un successivo attacco di conoscenza di fondo per scoprire l'identità dietro uno pseudonimo.

Capitolo 5

Tecniche di pseudonimizzazione

In linea di principio, una funzione di pseudonimizzazione mappa identificatori in pseudonimi. C'è un requisito fondamentale per una funzione di pseudonimizzazione. Consideriamo due identificatori differenti Id_1 e Id_2 e i loro pseudonimi corrispondenti pseudo_1 e pseudo_2. Una funzione di pseudonimizzazione deve verificare che pseudo_1 sia diverso da pseudo_2. Altrimenti, il recupero dell'identificatore potrebbe essere ambiguo: l'entità di pseudonimizzazione non può determinare se pseudo_1 corrisponde a Id_1 o Id_2. Tuttavia, un singolo identificatore Id può essere associato a più pseudonimi $(\mathrm{pseudo}_1, \mathrm{pseudo}_2, \ldots)$ purché sia possibile per l'entità di pseudonimizzazione invertire questa operazione.

5.1 Principali tecniche di Pseudonimizzazione

5.1.1 Contatore

Il contatore è la **funzione di pseudonimizzazione più semplice**, dove gli identificatori vengono sostituiti da un numero incrementato in modo monotono. È adatto per dataset piccoli e non complessi, ma può presentare problemi di implementazione e scalabilità per dataset più grandi.

5.1.2 Generatore di numeri casuali (RNG)

Il generatore di numeri casuali assegna un **numero casuale agli identificatori**, garantendo che **ogni pseudonimo sia imprevedibile**. Tuttavia, possono verificarsi collisioni, influenzate dal noto paradosso del compleanno.

5.1.3 Funzione hash crittografica

Una funzione hash crittografica **mappa stringhe di lunghezza variabile** in output di **lunghezza fissa**. Tuttavia, è considerata **debole** per la pseudonimizzazione a causa della vulnerabilità a **attacchi di forza bruta** e **dizionario.**

5.1.4 Codice di autenticazione del messaggio (MAC)

Il **MAC** è una **funzione di hash** con chiave che genera pseudonimi **utilizzando una chiave segreta**. È robusto dal punto di vista della protezione dei dati, a condizione che la chiave non venga compromessa.

5.2 Meccanismi di recupero

In conformità alla definizione, l'uso di informazioni aggiuntive è fondamentale per la pseudonimizzazione, pertanto l'entità di pseudonimizzazione deve implementare un **meccanismo di recupero.** Questa operazione può essere necessaria, ad esempio, quando l'entità di pseudonimizzazione rileva un'anomalia nel sistema e deve contattare le entità designate. Tale "anomalia" potrebbe essere una **violazione dei dati**, obbligando l'entità di pseudonimizzazione a **notificare i soggetti dei dati** in base al GDPR. Inoltre, il meccanismo di recupero potrebbe essere necessario per consentire l'esercizio dei diritti dei soggetti dei dati (ai sensi degli articoli 12-21 del GDPR).

Metodo	**Recupero basato su pseudonimo**
Counter	Tabella di mappatura
Generatore di numeri casuali	Tabella di mappatura
Funzione di hash crittografica	Tabella di mappatura
Codici di autenticazione del messaggio	Tabella di mappatura
Crittaggio	Decrittazione

Tabella 5.1: Confronto tra meccanismi di recupero

Capitolo 6

Principali tecniche di attacco

Ci sono **tre principali tecniche** generiche per rompere una funzione di pseudonimizzazione: **attacchi brute force**, ricerca tramite **dizionario** e **tentativi** (guesswork). L'efficacia di questi attacchi dipende da diversi parametri, tra cui:

- La **quantità di informazioni** sul titolare del pseudonimo (soggetto dei dati) contenuta nel pseudonimo.
- La **conoscenza pregressa dell'avversario**.
- La **dimensione del dominio** dell'**identificatore**.
- La **dimensione del dominio** del **pseudonimo**.
- La scelta e l**a configurazione della funzione di pseudonimizzazione** utilizzata (che include anche la dimensione del segreto di pseudonimizzazione).

6.1 Attacco brute force

La praticità di questa tecnica di attacco dipende dalla capacità dell'avversario di **calcolare la funzione di pseudonimizzazione**. A seconda dell'obiettivo dell'attacco, possono applicarsi condizioni aggiuntive. Se l'attacco brute force è utilizzato per ottenere il ripristino dell'identità originale, il dominio dell'identificatore deve essere finito e relativamente piccolo. Per ogni pseudonimo incontrato dall'avversario, questi può tentare di recuperare l'identificatore originale applicando la funzione di pseudonimizzazione su ogni valore del dominio dell'identificatore fino a trovare una corrispondenza.

Consideriamo la **pseudonimizzazione del mese di nascita in un dataset**. La dimensione del dominio dell'identificatore è **12**, quindi un avversario può enumerare rapidamente tutte le possibilità. I pseudonimi associati a ciascun mese sono calcolati in questo caso come la somma del codice ASCII delle prime tre lettere del mese di nascita (con la prima lettera maiuscola). Supponiamo che un avversario incontri il pseudonimo 301. Questo può applicare la funzione di pseudonimizzazione su ogni mese di nascita fino a trovare quello che corrisponde al valore 301. La Tabella 1 mostra i calcoli effettuati dall'avversario per riconoscere il pseudonimo 301, risultando nella tabella di mappatura della funzione di pseudonimizzazione.

Mese di nascita	Pseudonimo
Gen.	281
Feb.	269
Mar.	288
Apr.	291
Mag.	295
Giu.	301
Lug.	299
Ago.	285
Sett.	296
Ott.	294
Nov.	307
Dic.	268

Tabella 6.1: Pseudonimizzazione del mese di nascita

6.2 Attacco Dizionario

La ricerca tramite dizionario è un'**ottimizzazione dell'attacco brute force**, che può risparmiare costi computazionali. Ogni voce nel dizionario contiene un **pseudonimo** e l'identificatore o l'informazione corrispondente. Ogni volta che l'avversario ha bisogno di riconoscere nuovamente un pseudonimo, cerca nel dizionario. La ricerca tramite dizionario consiste essenzialmente nel **calcolo** e nel **salvataggio della tabella di mappatura**. Sono possibili compromessi tra tempo e memoria utilizzando tavole di Hellman o tabelle arcobaleno per estendere ulteriormente il range.

6.3 Guesswork

Questo tipo di attacco **utilizza conoscenze pregresse** (come distribuzioni di probabilità o altre informazioni secondarie) che l'avversario può avere su alcuni (o tutti) i titolari di pseudonimi. Sfruttare le **caratteristiche statistiche** degli identificatori è noto come guesswork ed è ampiamente utilizzato nella comunità di cracking delle password. L'avversario **non ha** necessariamente **bisogno di accedere alla funzione di pseudonimizzazione** (poiché la discriminazione è possibile semplicemente eseguendo un'analisi della frequenza dei pseudonimi osservati).

Consideriamo un caso che riguarda i pseudonimi corrispondenti ai "nomi propri". Esplorare completamente il dominio dei "nomi propri" è difficile. Tuttavia, l'avversario sa quali "nomi propri" sono i più popolari. L'avversario può eseguire una ricerca esaustiva o una ricerca tramite dizionario nel dominio dei "nomi propri" più popolari e ottenere la discriminazione.

Nomi propri più popolari
Bob
Alice
Charlie
Eve
Robert
Marie

Tabella 6.2: Lista di nomi propri più popolari

A seconda della quantità di informazioni di background o metadati di cui dispone l'avversario e della quantità di informazioni collegabili trovate nel dataset pseudonimizzato, questo tipo di attacco può portare a scoprire l'identità di un singolo titolare di pseudonimo, una frazione di essi o l'intero dataset. Specialmente per dataset piccoli, tali attacchi possono essere fattibili con alti tassi di successo.

Capitolo 7

Principali meccanismi di difesa

7.1 Cos'è un attacco di re-identificazione?

Gli attacchi di re-identificazione sono metodi per de-anonimizzare i dati utilizzando informazioni aggiuntive, come database esterni, metadati o analisi statistica, per inferire le identità dei soggetti dei dati. Gli attacchi di re-identificazione possono essere classificati in tre tipi principali:

- **Attacchi di collegamento (linking attack):** Coinvolgono il match di dati anonimizzati con altre fonti di dati che contengono informazioni identificative, come nomi, indirizzi o numeri di telefono.
- **Attacchi di inferenza (inference attack):** Basati sull'uso di metodi statistici o di apprendimento automatico per dedurre informazioni sensibili dai dati anonimizzati, come genere, età o stato di salute.
- **Attacchi di ricostruzione:** I più avanzati, utilizzano dati anonimizzati da più fonti o interrogazioni per ricostruire i dati originali o una loro approssimazione.

7.2 Come prevenire gli attacchi di collegamento?

Utilizzare **tecniche robuste di anonimizzazione** per garantire che ogni record anonimizzato sia indistinguibile dagli altri e che ogni gruppo di record abbia una diversità sufficiente di valori sensibili.

7.3 Come prevenire gli attacchi di inferenza?

Utilizzare l'**iniezione di rumore** per aggiungere **errori casuali o controllati ai dati**, riducendo così la loro precisione e utilità per gli attaccanti. Applicare la privacy differenziale per garantire che la presenza o l'assenza di un individuo nei dati non influenzi l'esito delle analisi. Limitare la **quantità** e la **granularità dei dati condivisi** e utilizzare meccanismi di controllo degli accessi e crittografia.

7.4 Come prevenire gli attacchi di ricostruzione?

Utilizzare la computazione sicura multiparte per consentire a più parti di eseguire calcoli sui loro dati senza rivelarli tra loro o a terzi. Adottare la crittografia omomorfica per eseguire operazioni su dati crittografati senza decifrarli. Sfruttare il learning federato per apprendere da dati locali in modo distribuito senza centralizzarli.

Capitolo 8

Sperimentazione di attacco di linking

La dimostrazione in questione simula un **attacco di linking** tra un **dataset pseudonimizzato** e un dataset contenente dati non anonimizzati. L'attacco riesce perché nei due dataset sono presenti **corrispondenze tra soggetti** riguardo attributi come:

- Età
- Genere
- Gruppo sanguigno

Tramite l'utilizzo di **Python** e **Jupyter Notebook** ho creato il seguente esempio di attacco.

8.1 Dataset di Informazioni Mediche sui Cittadini Americani

In questa fase iniziale, procediamo alla visualizzazione di un dataset contenente informazioni mediche di cittadini americani. Il dataset è gratuito ed è stato scaricato dalla piattaforma Kaggle. Può essere trovato al seguente link: https://www.kaggle.com/datasets/prasad22/healthcare-dataset?resource=download e include le seguenti informazioni sui pazienti:

- **Name**: Nome del paziente associato al record sanitario.
- **Age**: Età del paziente al momento dell'ammissione, espressa in anni.
- **Gender**: Genere del paziente, può essere "Maschio" o "Femmina".
- **Blood Type**: Gruppo sanguigno del paziente, come "A+", "O-", eccetera.
- **Medical Condition**: Condizione medica primaria o diagnosi associata al paziente, come "Diabete", "Ipertensione", "Asma", e altre.
- **Date of Admission**: Data di ammissione del paziente presso la struttura sanitaria.
- **Doctor**: Nome del medico responsabile delle cure durante l'ammissione del paziente.
- **Hospital**: Identifica l'ospedale o la struttura sanitaria dove il paziente è stato ammesso.
- **Insurance Provider**: Fornitore dell'assicurazione del paziente, che può essere tra diversi opzioni come "Aetna", "Blue Cross", "Cigna", "UnitedHealthcare" e "Medicare".
- **Billing Amount**: Importo fatturato per i servizi sanitari forniti al paziente durante l'ammissione, espressa come numero decimale.
- **Room Number**: Numero della stanza dove il paziente è stato alloggiato durante l'ammissione.
- **Admission Type**: Specifica il tipo di ammissione, come "Emergenza", "Elettiva" o "Urgente", riflettendo le circostanze dell'ammissione.

- **Discharge Date**: Data di dimissione del paziente dalla struttura sanitaria, basata sulla data di ammissione e un numero casuale di giorni entro un intervallo realistico.
- **Medication**: Identifica una medicazione prescritta o somministrata al paziente durante l'ammissione, come "Aspirina", "Ibuprofene", "Penicillina", "Paracetamolo" e "Lipitor".
- **Test Results**: Descrive i risultati di un test medico eseguito durante l'ammissione del paziente. Possibili valori includono "Normale", "Anormale" o "Inconcludente".

Il dataset contiene un totale di 55,500 record.

8.2 Visualizzazione del Dataset

Il seguente codice Python mostra come leggere e visualizzare i primi cinque record del dataset utilizzando la libreria `pandas`:

```python
import pandas as pd

pd.set_option("display.max_columns", None)
file_path = "Allegati/healthcare_dataset.csv"
df = pd.read_csv(file_path)

display(df.head())  # Stampa solo i primi 5 record
```

Codice 8.1: Visualizzazione dataset

	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date	Medication
0	Bobby JacksOn	30	Male	B-	Cancer	2024-01-31	Matthew Smith	Sons and Miller	Blue Cross	18856.281306	328	Urgent	2024-02-02	Paracetamol
1	LesLie TErRy	62	Male	A+	Obesity	2019-08-20	Samantha Davies	Kim Inc	Medicare	33643.327287	265	Emergency	2019-08-26	Ibuprofen
2	DaNnY sMitH	76	Female	A-	Obesity	2022-09-22	Tiffany Mitchell	Cook PLC	Aetna	27955.096079	205	Emergency	2022-10-07	Aspirin
3	andrEw waTtS	28	Female	O+	Diabetes	2020-11-18	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.782410	450	Elective	2020-12-18	Ibuprofen
4	adrIENNE bEll	43	Female	AB+	Cancer	2022-09-19	Kathleen Hanna	White-White	Aetna	14238.317814	458	Urgent	2022-10-09	Penicillin

Figura 8.1: Dataset originale

8.3 Pseudonimizzazione del Dataset

In questa fase, si vogliono proteggere le informazioni sensibili del dataset come il nominativo del paziente per rendere il dataset pseudo-anonimizzato. Utilizziamo la tecnica di sostituire il nominativo della persona creando una funzione che sostituisce il nome tramite la seguente operazione matematica:

$$\text{newName} = \sum_{i=1}^{n} \text{ASCII}(c_i) + \text{Room Number}$$

dove:

- c_i sono i caratteri di *oldName*
- $\text{ASCII}(c_i)$ rappresenta il valore ASCII del carattere c_i

Tronchiamo inoltre il dataset a soli 1000 valori per velocizzare le operazioni successive.

```
output_file = "Allegati/truncated_dataset.csv"
df.head(1000).to_csv(output_file, index=False)

new_dataframe = pd.read_csv(output_file)

# Somma dei valori ASCII dei caratteri in 'name'
def pseudonimize(name, room):
    name_ascii_sum = sum(ord(char) for char in name)
    return name_ascii_sum + room

for index, row in new_dataframe.iterrows():
    new_name = pseudonimize(row["Name"], row["Room Number"])
    new_dataframe.loc[index, "Name"] = new_name

display(new_dataframe.head())
```

Codice 8.2: Visualizzazione dataset

	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date	Medication
0	1535	30	Male	B-	Cancer	2024-01-31	Matthew Smith	Sons and Miller	Blue Cross	18856.281306	328	Urgent	2024-02-02	Paracetamol
1	1341	62	Male	A+	Obesity	2019-08-20	Samantha Davies	Kim Inc	Medicare	33643.327287	265	Emergency	2019-08-26	Ibuprofen
2	1164	76	Female	A-	Obesity	2022-09-22	Tiffany Mitchell	Cook PLC	Aetna	27955.096079	205	Emergency	2022-10-07	Aspirin
3	1590	28	Female	O+	Diabetes	2020-11-18	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.782410	450	Elective	2020-12-18	Ibuprofen
4	1551	43	Female	AB+	Cancer	2022-09-19	Kathleen Hanna	White-White	Aetna	14238.317814	458	Urgent	2022-10-09	Penicillin

Figura 8.2: Dataset pseudonimizzato

8.4 Creazione di un meccanismo di recupero

Quando pseudonimizziamo un dataset e abbiamo bisogno successivamente di risalire al nominativo originale del paziente, è necessario creare una **tabella di mappatura** tra il nominativo del paziente e il suo pseudonimo. Questa tabella deve essere conservata in modo sicuro e protetto poiché contiene tutte le coppie tra nominativo e pseudonimo presenti nel dataset pseudonimizzato.

```
df_nomi = df['Name'].head(1000)
new_df_nomi = new_dataframe['Name'].head(1000)

# Creiamo un nuovo dataframe con i nomi ottenuti
nomi_dataframe = pd.DataFrame({
    'Nome': df_nomi,
    'Nome Pseudonimizzato': new_df_nomi
})

# Scriviamo il dataframe in un file CSV
nomi_dataframe.to_csv('Allegati/tabellaMappatura.csv', index=
    False)

display(nomi_dataframe.head())
```

Codice 8.3: Creazione Tabella di Mappatura

	Nome	Nome Pseudonimizzato
0	Bobby JacksOn	1535
1	LesLie TErRy	1341
2	DaNnY sMitH	1164
3	andrEw waTtS	1590
4	adrIENNE bEll	1551

Figura 8.3: Dataset pseudonimizzato

8.5 Inizio dell'attacco di linking

Durante questa fase, l'attaccante dispone di due dataset cruciali:

- Il file `pseudonymized.csv` che contiene informazioni sanitarie di vari pazienti, con i loro nominativi pseudonimizzati.
- Il file `blood_donation_dataset.csv` che include dettagli su donatori di sangue, specificamente:
  - **Nome**: Il nome completo del donatore di sangue.
  - **Età**: L'età del donatore al momento della donazione.
  - **Genere**: Il genere del donatore, indicato come "Maschio" o "Femmina".
  - **Gruppo Sanguigno**: Il tipo di sangue del donatore.
  - **Data della Donazione**: La data in cui è stata effettuata la donazione.

L'obiettivo dell'attacco è ricondurre all'identificazione dei pazienti all'interno del dataset pseudonimizzato, sfruttando le informazioni disponibili nel dataset delle donazioni di sangue che contiene i nomi completi in chiaro.

```
file_path = "Allegati/blood_donation_dataset.csv"
blood_donation = pd.read_csv(file_path)

display(blood_donation.head())
```

Codice 8.4: Dataset Donatori di sangue

	Name	Age	Gender	Blood Type	Date of Donation
0	AMANDa DURhAm	46	Female	O+	2021-06-16
1	Erin oRTEga	43	Male	AB-	2023-05-24
2	BETh sChwaRTZ	58	Male	AB-	2024-03-29
3	JASmine sHort	40	Female	O-	2021-02-23
4	dAnIElle lOPeZ	37	Male	O-	2020-11-20

Figura 8.4: Dataset donatori di sangue

Supponiamo adesso di voler trovare il nominativo associato alla entry **n°546** del nostro dataset anonimizzato

```
file_path = "Allegati/pseudonymized.csv"
pseudonymized = pd.read_csv(file_path)

# Modificare questo valore per cercare un altro utente
indiceDaCercare = 546

display(pseudonymized.iloc[indiceDaCercare].to_frame().T)
```

Codice 8.5: Ricerca pseudonimo

	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date	Medication	Test Results
546	1623	32	Female	B+	Asthma	2020-02-20	Joseph Miller	Snyder-Perry	Medicare	17862.69225	482	Elective	2020-02-21	Paracetamol	Inconclusive

Figura 8.5: Dataset donatori di sangue

Dalla seguente voce, possiamo estrarre informazioni utili come:

- Età: 32 anni
- Genere: Femmina
- Gruppo Sanguigno: B+

Queste informazioni sono fondamentali per condurre un attacco di collegamento tramite il nostro dataset `blood_donation_dataset.csv`.

Ora cerchiamo un record che soddisfi tutte e tre queste condizioni.

```
def search(age, gender, bloodType):
    for index, row in blood_donation.iterrows():
        if (
            row["Age"] == age
            and row["Gender"] == gender
            and row["Blood Type"] == bloodType
        ):
            return index

age = pseudonymized.iloc[indiceDaCercare]["Age"]
gender = pseudonymized.iloc[indiceDaCercare]["Gender"]
bloodType = pseudonymized.iloc[indiceDaCercare]["Blood Type"]

result = search(age, gender, bloodType)
display(blood_donation.iloc[result].to_frame().T)
```

Codice 8.6: Attacco di linking

	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date	Medication	Test Results
546	1623	32	Female	B+	Asthma	2020-02-20	Joseph Miller	Snyder-Perry	Medicare	17862.69225	482	Elective	2020-02-21	Paracetamol	Inconclusive

Figura 8.6: Output attacco di linking

8.6 Conclusione dell'attacco di linking

L'attacco di linking ha identificato un record di una donatrice chiamata "Connor BARTOn" che sembra corrispondere al paziente cercato. Supponiamo che l'attaccante conosca l'algoritmo utilizzato per la pseudonimizzazione.

Per verificare la corrispondenza, sottraiamo il numero della stanza dal nome pseudonimizzato del paziente, ottenendo la somma dei caratteri ASCII del nome. Confrontiamo poi questa somma con quella del nome trovato nel dataset originale.

```
# Somma dei caratteri ASCII del nome nel database pseudonimizzato
check = pseudonymized.iloc[indiceDaCercare]["Name"] - pseudonymized.iloc[indiceDaCercare]["Room Number"]

# Nome trovato nel DataFrame blood_donation
nomeTrovato = blood_donation.iloc[result]["Name"]

# Calcolo dell'ASCII dei nomi trovati
nomeTrovatoASCII = sum(ord(char) for char in nomeTrovato)

# Confronto degli ASCII
if nomeTrovatoASCII == check:
    print("Corrispondenza trovata")
else:
    print("Corrispondenza non trovata")
```

Codice 8.7: Conclusione attacco di linking

Conclusione

L'attacco di *linking* effettuato su un database contenente informazioni mediche pseudonimizzate ha rivelato vulnerabilità significative nella protezione dei dati sensibili. Il processo di pseudonimizzazione, che prevede la sostituzione di informazioni identificative dirette con pseudonimi, è stato compromesso attraverso l'utilizzo di un dataset esterno contenente informazioni personali dei donatori con nominativi in chiaro. Questo attacco ha dimostrato come, nonostante le misure di protezione implementate, i dati pseudonimizzati possano essere reidentificati con sufficiente accuratezza mediante tecniche di correlazione tra dataset.

Il dataset di partenza conteneva informazioni mediche sensibili pseudonimizzate. Questo approccio è comunemente utilizzato per ridurre il rischio di esposizione di dati personali in scenari di trattamento e analisi dei dati. Tuttavia, la presenza di un secondo dataset, contenente dati identificativi dei donatori, ha permesso di effettuare un collegamento tra le informazioni pseudonimizzate e le identità reali. In particolare, l'attacco ha sfruttato attributi comuni tra i due dataset, come **età**, **genere** e **gruppo sanguigno**, per stabilire connessioni e identificare i soggetti nel database pseudonimizzato.

L'attacco di *linking* si è articolato nelle seguenti fasi:

1. **Preparazione dei dataset**: Sono stati raccolti due dataset distinti. Il primo contenente dati medici pseudonimizzati e il secondo contenente informazioni identificative in chiaro. Entrambi i dataset includevano attributi demografici comuni che sono stati utilizzati come chiavi di collegamento.

2. **Validazione dei risultati**: Le corrispondenze identificate dall'algoritmo sono state validate manualmente per verificare l'accuratezza del collegamento. Questo processo ha confermato la capacità dell'attacco di reidentificare correttamente una porzione significativa dei record pseudonimizzati.

I risultati ottenuti hanno evidenziato la necessità di **rafforzare le misure di protezione** dei dati anche quando vengono applicate tecniche di pseudonimizzazione. In particolare, è emerso che la pseudonimizzazione, sebbene riduca il rischio di esposizione diretta, non è sufficiente a prevenire la reidentificazione quando esistono dataset esterni con informazioni sovrapponibili. È quindi essenziale adottare ulteriori misure di sicurezza, quali:

- **Minimizzazione dei dati**: Limitare la quantità di informazioni personali raccolte e trattate nei dataset.
- **Aggiunta di rumore**: Implementare tecniche di anonimizzazione più avanzate, come l'aggiunta di rumore ai dati demografici, per ridurre la possibilità di collegamento.
- **Controlli di accesso rigorosi**: Garantire che solo personale autorizzato abbia accesso ai dati sensibili e che vengano effettuati controlli di accesso periodici.
- **Valutazioni di rischio**: Effettuare valutazioni di rischio periodiche per identificare e mitigare potenziali vulnerabilità nei sistemi di gestione dei dati.

In conclusione, l'attacco di *linking* evidenziato in questo studio serve come monito per le organizzazioni che trattano dati sensibili. È fondamentale comprendere che la protezione dei dati non può fare affidamento esclusivamente sulla pseudonimizzazione, ma deve essere integrata in un quadro di sicurezza complessivo che consideri potenziali minacce interne ed esterne. Solo attraverso un approccio alla sicurezza dei dati sarà possibile garantire la privacy e la protezione delle informazioni sensibili.

Glossario

Dati Personali

Si riferisce a qualsiasi informazione relativa a una persona fisica identificata o identificabile (interessato); una persona fisica identificabile è colui che può essere identificato, direttamente o indirettamente, in particolare mediante un identificativo come nome, numero di identificazione, dati di localizzazione, identificativo online o uno o più fattori specifici della identità fisica, fisiologica, genetica, mentale, economica, culturale o sociale di quella persona (GDPR, art. 4(1)).

Responsabile del Trattamento

Persona fisica o giuridica, autorità pubblica, agenzia o altro ente che, da solo o congiuntamente ad altri, determina le finalità e i mezzi del trattamento dei dati personali (GDPR, art. 4(7)).

Responsabile del Trattamento

Persona fisica o giuridica, autorità pubblica, agenzia o altro ente che tratta dati personali per conto del titolare del trattamento (GDPR, art. 4(8)).

Pseudonimizzazione

Il trattamento dei dati personali in modo tale che i dati personali non possano più essere attribuiti a un soggetto specifico senza l'uso di informazioni aggiuntive, a condizione che tali informazioni aggiuntive siano conservate separatamente e siano soggette a misure tecniche e organizzative per garantire che i dati personali non siano attribuibili a una persona fisica identificata o identificabile (GDPR, art. 4(5)).

Anonimizzazione

Un processo mediante il quale i dati personali vengono alterati in modo irreversibile in modo che un soggetto non possa essere più identificato direttamente o indirettamente, né dal titolare del trattamento da solo né in collaborazione con altre parti (ISO/TS 25237:2017).

Identificativo

Un valore che identifica un elemento all'interno di uno schema di identificazione. Un identificativo univoco è associato a un solo elemento.

Pseudonimo

Conosciuto anche come criptonimo o semplicemente nym, è un pezzo di informazione associato a un identificativo di un individuo o a qualsiasi altro tipo di dato personale (ad esempio, dati di localizzazione). I pseudonimi possono avere diversi gradi di collegabilità agli identificativi originali.

Avversario

Un ente che cerca di rompere la pseudonimizzazione e collegare un pseudonimo (o un dataset pseudonimizzato) al detentore del pseudonimo.

Attacco di Re-identificazione

Un attacco alla pseudonimizzazione eseguito da un avversario che mira a ri-identificare il detentore di un pseudonimo.

Bibliografia

[1] Data Privacy Manager, *Pseudonymization according to the GDPR*, [Online]. Available: https://dataprivacymanager.net/pseudonymization-according-to-the-gdpr/.

[2] ENISA, *Pseudonymisation Techniques and Best Practices*, [Online]. Available: https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices.

[3] LinkedIn, *How Can You Protect Against Re-identification Attacks*, [Online]. Available: https://www.linkedin.com/advice/0/how-can-you-protect-against-re-identification-attacks.

[4] Wikipedia, *Data Re-Identification*, [Online]. Available: https://en.wikipedia.org/wiki/Data_re-identification.

[5] NewSchool, *Guidelines Anonymization and Pseudonymization*, [Online]. Available: https://ispo.newschool.edu/guidelines/anonymization-pseudonymization/.

[6] Research Gate, *Develop Privacy Friendly Software*, [Online]. Available: https://www.researchgate.net/publication/336043425_Privacy_by_Evidence_A_Methodology_to_Develop_Privacy-Friendly_Software_Applicationsn.

[7] Kaggle, *Healthcare Dataset*, [Online]. Available: https://www.kaggle.com/datasets/prasad22/healthcare-dataset?resource=download.