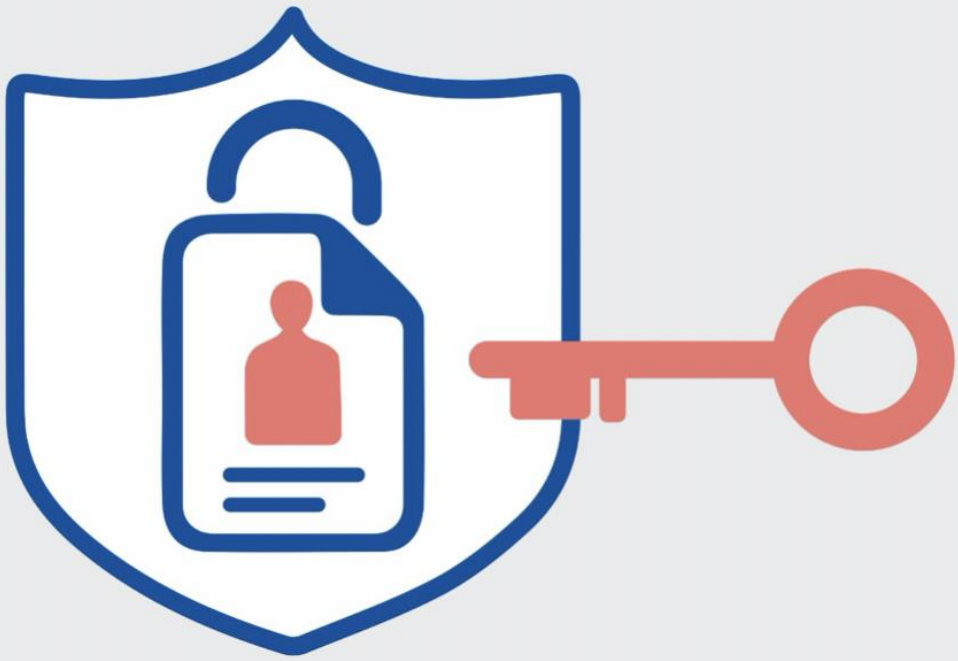
enisa

EUROPEAN UNION AGENCY
FOR CYBERSECURITY



Tecniche di pseudonimizzazione e buone pratiche

Raccomandazioni sulla creazione della tecnologia ai sensi della normativa in tema di privacy e protezione dei dati personali

NOVEMBRE 2019

SU ENISA

L'Agenzia europea per la sicurezza delle reti e dell'informazione (ENISA), sin dal 2004, si occupa di rendere l'Europa cyber-sicura. ENISA collabora con l'Unione Europea, gli stati membri, il settore privato e i cittadini europei per riuscire a sviluppare pareri e raccomandazioni sulle buone pratiche in ambito di sicurezza delle informazioni. Inoltre, assiste gli stati membri dell'Unione Europea nell'attuazione della legislazione comunitaria rilevante e si occupa di migliorare la resistenza delle informazioni critiche delle infrastrutture e delle reti Europee. ENISA mira a consolidare, all'interno degli stati membri, la conoscenza esistente, supportando lo sviluppo integrato intrapreso dalle comunità per migliorare la sicurezza nelle reti e informazione nell'Unione Europea. Dal 2019 è stato stilato uno schema di certificazione nell'ambito della cyber-sicurezza. Per ulteriori approfondimenti su ENISA e il suo lavoro si rimanda alla consultazione del sito www.enisa.europa.eu

CONTATTI

E' possibile contattare gli Autori all'indirizzo mail: isdp@enisa.europa.eu

Per informazioni mediatiche sul trattato, contattare: press@enisa.europa.eu

CONTRIBUTI

Meiko Jensen (Kiel University), Cedric Lauradoux (INRIA), Konstantinos Limniotis (HDPA)

EDITORI

Athena Bourka (ENISA), Prokopios Drogkaris (ENISA), Ioannis Agrafiotis (ENISA)

RICONOSCIMENTI

Si ringrazia Giuseppe D'Acquisto (Garante), Nils Gruschka (University of Oslo) e Simone Fischer-Hübner (Karlstad University) per la revisione del presente trattatto e il valido contributo fornito.

NOTE LEGALI

Tale pubblicazione non rappresenta necessariamente lo stato dell'arte e ENISA si riserva di apportare, nel tempo, le opportune integrazioni.

Le fonti e citazioni di terze parti sono state accuratamente riportate.

ENISA non è responsabile per il contenuto delle fonti esterne, inclusi i riferimenti riportati in siti web esterni, riportate nella presente pubblicazione.

La presente pubblicazione si intende esclusivamente a scopo informativo. La stessa sarà consultabile gratuitamente.

Né ENISA né ogni altro suo rappresentante è responsabile per l'uso che potrà farsi con le informazioni contenute all'interno della presente pubblicazione.

NOTE DI COPYRIGHT

ISBN 978-92-9204-307-0, DOI 10.2824/247711

DISCLAIMER

Il presente documento è riservato ad un uso strettamente privato. Esso costituisce una traduzione non ufficiale della pubblicazione “Pseudonymisation techniques and best practices” elaborato dall’Enisa (European Union Agency for Cybersecurity), al quale sono riservati tutti i diritti. Non si fornisce alcuna garanzia in merito all’affidabilità ed all’esattezza delle notizie riportate, ovvero della esattezza delle traduzioni.

Si declina pertanto ogni responsabilità per qualsiasi danno, diretto, indiretto, incidentale e consequenziale legato all’uso, proprio o improprio delle informazioni contenute in questo documento, ivi inclusi, senza alcuna limitazione, la perdita di profitto, l’interruzione di attività aziendale o professionale, la perdita di programmi o altro tipo di dati ubicati sul sistema informatico dell’utente o altro sistema, e ciò anche qualora gli Autori del documento fossero stati espressamente messi al corrente della possibilità del verificarsi di tali danni.

Curatori del progetto di traduzione

(in ordine alfabetico)

Coordinamento e Revisione

Rosario Mauro Catanzaro – *Presidente dell’Associazione Nazionale per la Protezione dei Dati*

Manuela Sforza – *Responsabile Cybersecurity Associazione Nazionale per la Protezione dei Dati*

Comitato Scientifico Associazione Nazionale per la Protezione dei Dati

Alessandro Del Ninno

Pierluigi Perri

Massimo Simbula

Giovanni Ziccardi

Progetto

Rino Cannizzaro – *Adfor Socio Sostenitore Associazione Nazionale per la Protezione dei Dati*

Traduzione

Cecilia Brodu - Revisione linguistica

Lucrezia Croce

Paolo Cucchi

Vincenzo Tarantini

INDICE

EXECUTIVE SUMMARY

Sotto la vigenza del GDPR[1], la sfida di una corretta applicazione della pseudonimizzazione ai dati personali sta gradualmente diventando un argomento assai dibattuto in diversi settori: a partire dalla ricerca e dal mondo accademico, passando per la giustizia e le forze dell'ordine, fino ad arrivare al settore della compliance di diverse organizzazioni Europee. Sulla base del precedente lavoro[2] di ENISA nel settore, il presente trattato esamina le nozioni base della pseudonimizzazione, così come le soluzioni tecniche che possono supportarne l'attuazione pratica.

In particolare, partendo da diverse ipotesi di pseudonimizzazione, il trattato identifica, in primis, i principali attori coinvolti nel processo di pseudonimizzazione così come il loro possibile ruolo. Secondariamente, analizza i diversi modelli e le diverse tecniche di attacco alla pseudonimizzazione da parte di intrusi, quali l'attacco di forza bruta, la ricerca nel dizionario e la congettura. Inoltre, presenta le maggiori tecniche di pseudonimizzazione (quali, ad esempio, il contatore, il generatore casuale di numeri, la funzione crittografica di Hash, il message authentication code e la crittografia) attualmente disponibili.

Soprattutto, fornisce i criteri in grado di influenzare, nella pratica, la scelta della tecnica o del modello di pseudonimizzazione, quali la protezione dei dati, la funzionalità, la scalabilità e il ripristino. Inoltre, vengono menzionate alcune tecniche avanzate di pseudonimizzazione.

Sulla base delle suddette descrizioni, il presente trattato riporta due casi pratici di pseudonimizzazione degli indirizzi IP, analizzando le particolarità derivanti dagli specifici tipi di identificativi utilizzati. Inoltre, esamina un caso pratico più complesso di pseudonimizzazione di numerosi record di dati, esaminando la possibilità di una re-identificazione.

Uno dei principali obiettivi del presente trattato è quello di sottolineare come non vi sia una soluzione facile e univoca di pseudonimizzazione in grado di adattarsi ad ogni approccio o singola ipotesi.

Viceversa, è necessario un elevato livello di competenza capace di elaborare un processo di pseudonimizzazione resistente che riduca al minimo le minacce di attacchi di discriminazione o di re-identificazione, garantendo al contempo il livello di funzionalità necessaria ai fini della pseudonimizzazione dei dati.

In conclusione, il presente trattato rassegna le seguente conclusioni e raccomandazioni utili a tutte le parti interessate, con particolare riguardo all'attuazione e adozione pratica della pseudonimizzazione dei dati.

VERSO LA PSEUDONIMIZZAZIONE ATTRAVERSO UN APPROCCIO BASATO SUL RISCHIO

Nonostante tutte le tecniche di pseudonimizzazione conosciute abbiano le loro proprie, ben chiare, caratteristiche intrinseche, ciò non rende, nella pratica, la scelta del giusto approccio un compito banale. A tal fine è necessario un attento esame del contesto in cui applicare la pseudonimizzazione, considerando

[1] Il testo integrale del GDPR è reperibile al presente link https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

[2] Tale lavoro è reperibile al sito https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-gdpr-provisions

tutti gli obiettivi specifici della pseudonimizzazione (da chi proteggere le identità, qual è la funzionalità che si desidera raggiungere dagli pseudonimi ottenuti, e così via), così come la semplicità nella realizzazione. Pertanto, per quanto attiene la scelta della tecnica di pseudonimizzazione più appropriata, occorre adottare un approccio basato sul rischio, in modo da valutare e limitare adeguatamente le relative minacce alla privacy. Infatti, la semplice protezione dei dati aggiuntivi necessari per la re-identificazione, pur essendo un prerequisito, non garantisce necessariamente l'eliminazione di tutti i rischi.

I titolari del trattamento e i responsabili del trattamento dovrebbero considerare attentamente l'attuazione della pseudonimizzazione secondo un approccio basato sul rischio, tenendo conto delle finalità e del contesto generale del trattamento dei dati personali, così come i livelli di funzionalità e di scalabilità che si intendono raggiungere.

I creatori di prodotti, servizi e applicazioni dovrebbero fornire ai titolari del trattamento e ai responsabili del trattamento informazioni adeguate in merito al proprio utilizzo delle tecniche di pseudonimizzazione e ai livelli di sicurezza e di protezione dei dati che forniscono.

Le autorità di regolamentazione (ad esempio le Autorità di controllo e il Comitato Europeo per la Protezione dei Dati) dovrebbero fornire linee guida pratiche ai titolari e ai responsabili del trattamento dei dati personali circa la valutazione del rischio, promuovendo, al contempo, le buone pratiche in materia di pseudonimizzazione.

LA DEFINIZIONE DELLO STATO DELL'ARTE

Per supportare un approccio basato sul rischio nell'ambito della pseudonimizzazione, è essenziale definire lo stato dell'arte nel settore. Infatti, come mostrato in questo rapporto, se da un lato sono disponibili diverse tecniche di pseudonimizzazione, d'altro lato l'applicazione pratica delle stesse può variare, ad esempio, a seconda della tipologia di identificativi o di set di dati. A tal fine, è importante lavorare sui casi d'uso e sugli esempi specifici, ampliando il ventaglio delle opzioni di applicazioni tecniche possibili nel campo della pseudonimizzazione.

La Commissione europea e le Istituzioni europee competenti dovrebbero sostenere la definizione e la diffusione di informazioni sullo stato dell'arte della pseudonimizzazione, in collaborazione con le comunità di ricerca e l'industria del settore.

Le Autorità di regolamentazione (ad esempio le Autorità di controllo e il Comitato Europeo per la Protezione dei Dati) dovrebbero promuovere la pubblicazione delle buone pratiche nel campo della pseudonimizzazione.

L'AVANZAMENTO DELLO STATO DELL'ARTE

Il presente trattato ha focalizzato l'attenzione sulle tecniche basilari di pseudonimizzazione che, al giorno d'oggi, i titolari e i responsabili del trattamento hanno a disposizione. Tuttavia, nelle ipotesi più complesse

(che, come visto, sono assai frequenti nella pratica), sarà sempre più indispensabile ricorrere all'uso di tecniche più avanzate (e salde), come quelle provenienti dal settore dell'anonimizzazione. Anzi, la nozione stessa di anonimizzazione dovrebbe essere rivisitata, in quanto i modelli sono in continua evoluzione (e, quindi, l'anonimizzazione diventa, nei casi reali, sempre più impegnativa).

La comunità scientifica dovrebbe lavorare per sviluppare le attuali tecniche di pseudonimizzazione in modo che diventino soluzioni più avanzate in grado di affrontare efficacemente le particolari sfide sorte nell'era dei big data. La Commissione europea e le istituzioni europee competenti dovrebbero sostenere e diffondere tali sforzi.

1. INTRODUZIONE

1.1 PREMESSA

La pseudonimizzazione è un conosciuto processo di re-identificazione (anonimizzazione) che ha avuto una notevole crescita di attenzione in seguito all'adozione del GDPR, nel quale viene indicata, all'interno di un sistema di progettazione, sia quale meccanismo di sicurezza che si protezione dei dati. Inoltre, nell'ambito del GDPR, la pseudonimizzazione, ove correttamente applicata, è in grado di garantire un certo grado sicurezza in relazione alle prescrizioni di legge imposte ai titolari del trattamento dei dati.

Vista la sua crescente importanza, sia per i titolari del trattamento che per gli interessati, ENISA ha pubblicato nel 2018 [1] un trattato sulla nozione e sulle tecniche di pseudonimizzazione in relazione al suo ruolo nell'ambito del GDPR.

In particolare, partendo dalla definizione di pseudonimizzazione (nonché dalla differenza con le altre tecnologie quali l'anonimizzazione e la crittografia), il trattato evidenzia soprattutto i vantaggi che la pseudonimizzazione è in grado di apportare alla protezione dei dati.

Proseguendo nella trattazione, vengono indicate alcune tecniche di pseudonimizzazione, quali l'hashing, l'hashing mediante l'utilizzo di una chiave o di un sale (sequenza casuale di bit), la crittografia, la tokenizzazione, e ogni altro approccio rilevante.

Inoltre, il richiamato lavoro dell'ENISA, tratta alcuni punti chiave rispetto alle problematiche della pseudonimizzazione, evidenziando come, ai fini di rafforzare il concetto di pseudonimizzazione quale misura di sicurezza (ai sensi dell'art. 32 del GDPR) e di delineare il suo ruolo nell'ambito della protezione dei dati si dalla fase di progettazione (ai sensi dell'art. 25 del GDPR), siano necessarie ulteriori ricerche e analisi sul tema.

Infatti, come anche riconosciuto nel trattato dell'ENISA, vi è una particolare esigenza di promuovere, nell'ambito della pseudonimizzazione, delle buone pratiche e fornire un prontuario di casi pratici in grado di offrire una panoramica sullo stato dell'arte nel settore.

Proprio a tal fine, ENISA ha, inoltre, elaborato un programma di lavoro per il 2019 sulle applicazioni pratiche della pseudonimizzazione dei dati[3].

1.2 FINI E OBIETTIVI

Il fine complessivo del presente trattato è quello di fornire una guida e le buone pratiche sulle tecniche di attuazione della pseudonimizzazione dei dati.

In particolare, l'obiettivo che il trattato perseguire è quello di:

- Discutere le diverse ipotesi di pseudonimizzazione e i principali soggetti coinvolti.

[3] Poiché ENISA si occupa di fornire indicazioni in materia di politica di sicurezza delle reti e delle informazioni dei dati in ambito UE, è logico aspettarsi che il proprio lavoro attraversi altre aree tematiche affini, quali la privacy, così da conciliare tutte le posizioni delle varie parti coinvolte. Infatti, ai sensi dell'art. 32 del GDPR, nell'ottica della protezione dei dati, un importante elemento è dato dalle attuazioni pratiche della pseudonimizzazione

- Presentare le possibili tecniche di pseudonimizzazione applicabili per scongiurare le principali intrusioni e attacchi.
- Analizzare l'applicabilità della pseudonimizzazione a specifici tipi di identificativi, in particolare gli indirizzi IP, gli indirizzi e-mail e altri tipi di set di dati strutturati (e i relativi casi pratici).
- Rassegnare conclusioni rilevanti e offrire raccomandazioni utili ai fini di ulteriori studi e lavori in materia.

Si noti che la selezione di casi pratici si basa sulla circostanza per cui, nella quotidianità, alcune specifiche tipologie di identificativi (indirizzi IP, indirizzi mail, identificativi in data set strutturati, e così via) sono assai utilizzati.

Allo stesso tempo, i casi pratici selezionati riflettono anche le diverse caratteristiche della pseudonimizzazione, quali ad esempio la struttura più complessa degli indirizzi IP, quella più flessibile degli indirizzi e-mail, nonché quella assai più imprevedibile dei grandi set di dati.

Il presente trattato si rivolge ai titolari e ai responsabili del trattamento dei dati, a fornitori di prodotti, servizi e applicazioni, alle autorità di protezione dei dati (DPAS), nonché ad ogni altra parte interessata nel processo di pseudonimizzazione dei dati.

Il presente trattato, inoltre, presuppone un livello basilare si conoscenza dei principi che regolano la protezione dei dati personali e il processo di pseudonimizzazione. Per una visione completa sulla pseudonimizzazione dei dati come regolata dal GDPR, si rimanda alla lettura dei precedenti lavori di ENISA in materia [1].

Le tematiche e gli esempi riportati nel presente trattato si concentrano sulle soluzioni tecniche in grado di favorire la privacy e la protezione dei dati; senza che ciò implichi una visione normativa dei casi maggiormente rilevanti.

1.3 COMPENDIO

Il compendio del presente trattato è il seguente:

- Il Capitolo 2 fornisce un dettaglio sulle terminologie impiegate all'interno del trattato e, ove necessario, le opportune spiegazioni.
- Il Capitolo 3 riporta le ipotesi più comuni di pseudonimizzazione.
- Il Capitolo 4 descrive le possibili intrusioni e tipologie di attacco alla pseudonimizzazione (in relazione ai casi pratici di cui al precedente Capitolo).
- Il Capitolo 5 presenta le principali tecniche e metodi di pseudonimizzazione.
- I Capitoli 6, 7 e 8 analizzano l'applicazione delle diverse tecniche di pseudonimizzazione degli indirizzi IP, degli indirizzi mail, e dei più complessi data set (e i relativi casi pratici).
- Il Capitolo 8 elenca i dibattiti sul tema e riporta le principali conclusioni e raccomandazioni utili a tutte le parti interessate.

Il presente trattato, riconducibile al lavoro di ENISA nell'ambito della privacy e della protezione dei dati[4], si concentra sull'analisi delle soluzioni tecniche di attuazione del GDPR, sulla privacy by design e sulla sicurezza del trattamento dei dati personali.

[4] https://www.enisa.europa.eu/topics/data-protection

2. TERMINOLOGIA

Il presente Capitolo propone una serie di termini utilizzati all'interno del trattato ed essenziali ai fini di una corretta e migliore comprensione e lettura. Alcuni di questi termini derivano dal GDPR, mentre altri si riferiscono agli standard tecnici o sono stati appositamente coniati nell'ambito del presente trattato.

In particolare, vengono utilizzati i seguenti termini:

Dato personale: qualsiasi informazione riguardante una persona fisica identificata o identificabile («interessato»); si considera identificabile la persona fisica che può essere identificata, direttamente o indirettamente, con particolare riferimento ad un identificativo come il nome, un numero di identificazione, dati relativi all'ubicazione, un identificativo online o ad uno o più elementi caratteristici della sua identità fisica, fisiologica, genetica, psichica, economica, culturale o sociale (art. 4(1) GDPR)

Titolare del trattamento o, semplicemente, titolare: 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; quando le finalità e i mezzi di tale trattamento sono determinati dal diritto dell'Unione o degli Stati membri, il titolare del trattamento o i criteri specifici applicabili alla sua designazione possono essere stabiliti dal diritto dell'Unione o degli Stati membri; (art. 4(7) GDPR).

Responsabile del trattamento o, semplicemente, responsabile: la persona fisica o giuridica, l'autorità pubblica, il servizio o altro organismo che tratta dati personali per conto del titolare del trattamento (art. 4(8) GDPR).

Pseudonimizzazione: il trattamento dei dati personali in modo tale che i dati personali non possano più essere attribuiti ad 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 (art. 4(5) GDPR)[5].

Anonimizzazione: il processo attraverso il quale il dato personale viene irreversibilmente alterato in modo tale che l'interessato non possa più essere identificato, direttamente o indirettamente, dal titolare del trattamento, né da solo, né in collaborazione con altre parti (ISO/TS 25237:2017)[6].

Identificativo: il valore in grado di indentificare un elemento attraverso uno schema di identificazione[7]. Un unico identificativo viene associato a un solo elemento. Nel presente trattato è spesso impiegato nel senso che viene usato un unico identificativo, in quanto associato al dato personale.

Pseudonimo: altrimenti detto criptonimo o, semplicemente, nomio; il pezzo di informazione associata a un identificativo di un individuo o a un altro qualsiasi tipo di dato personale (quali, ad esempio, i dati di residenza). Lo pseudonimo può avere diversi gradi di collegamento all'identificativo originario[8]. Il differente

[5] si vedano altri definizioni tecniche di pseudonimizzazione rilevanti al punto [1] della bibliografia.
[6] Per ulteriori approfondimenti sulla differenza tra pseudonimizzazione e anonimizzazione al punto [1] della bibliografia.
[7] Il Gruppo dell'articolo 29 per la tutela dei dati [31] definisce l'identificativo quale pezzo di informazione particolarmente importante e vicino all'individuo cui si riferisce, tale da consentirne l'identificazione. La misura in cui tali identificativi sono, di per sé, sufficienti a consentire l'identificazione di un soggetto, dipende dal contesto dello specifico trattamento dei dati personali. Pertanto, l'identificativo rappresenta, al contempo, sia un singolo pezzo di informazione (quale, ad esempio, il nome, l'indirizzo email, il numero di previdenza sociale, e così via), sia un dato assai più complesso.
[8] A tal fine, si potrebbe affermare che lo pseudonimo rappresenta una sorta di "mascheramento" dell'identificativo di un individuo che, a seconda del contesto, è in grado di rendere un individuo più o meno identificabile.

livello di collegamento dei diversi tipi di pseudonimo è importante ai fini di effettuare una valutazione sull'affidabilità degli pseudonimi impiegati, così come ai fini di creare un sistema di pseudonimizzazione che abbia il livello di collegamento desiderato (come, ad esempio, nell'ipotesi in cui si analizzano i file di pseudonimizzazione, ovvero nell'ipotesi in cui si analizza la reputazione di un sistema)[9].

Funzione di pseudonimizzazione: indicata come P, la funzione che sostituisce un identificativo Id con uno pseudonimo *pseudo*.

Segreto di pseudonimizzazione: indicato come S, il parametro (opzionale) della funzione di pseudonimizzazione P. La funzione P non può essere risolta/calcolata se non si conosce S.

Funzione di ripristino: indicata come R, la funzione che sostituisce lo pseudonimo *pseudo* con l'identificativo IP, utilizzando il segreto di pseudonimizzazione S. Inverte la funzione di pseudonimizzazione P.

Tabella di mappatura della pseudonimizzazione: la rappresentazione dell'operazione della funzione di pseudonimizzazione. Associa ogni identificativo al suo corrispondente pseudonimo. A seconda della funzione di pseudonimizzazione P, tale tabella può corrispondere al segreto di pseudonimizzazione o a una parte di esso.

Ente di pseudonimizzazione: l'ente responsabile del processo di conversione dell'identificativo nello pseudonimo attraverso l'utilizzo della funzione di pseudonimizzazione. Può essere, a seconda dei casi, il titolare del trattamento, il responsabile del trattamento, (il quale applica la pseudonimizzazione congiuntamente al titolare), una parte terza autorizzata, ovvero lo stesso interessato. Si sottolinea che, sulla base della presente definizione, il ruolo dell'ente di pseudonimizzazione assume una stretta rilevanza in ordine alle attuazioni pratiche della pseudonimizzazione nelle diverse ipotesi elencate di seguito[10]. Ad ogni modo, per quel che concerne la presente trattazione, la responsabilità dell'intero processo di pseudonimizzazione (e, in generale, per ogni operazione riguardante i dati personali) resta sempre in capo al titolare del trattamento.

Dominio dell'identificativo / Dominio dello pseudonimo: i domini dai i quali vengono ricavati l'identificativo e lo pseudonimo. Possono essere impiegati domini differenti, ovvero il medesimo. Possono essere domini finiti o infiniti.

Intruso: il soggetto che cerca di violare la pseudonimizzazione e collegare uno pseudonimo (o un set di dati pseudonimizzato) al suo titolare/i.

Attacco di re-identificazione: l'attacco alla pseudonimizzazione compiuto da un intruso con l'obiettivo di re-identificare il titolare dello pseudonimo.

[9] Per un maggiore approfondimento sul livello di collegabilità degli pseudonimi di rimanda al punto [4] della bibliografia.
[10] Note that under the definition of pseudonymisation in GDPR (article 4(5)), there is no reference as to who holds the additional information.

3. IPOTESI DI PSEUDONIMIZZAZIONE

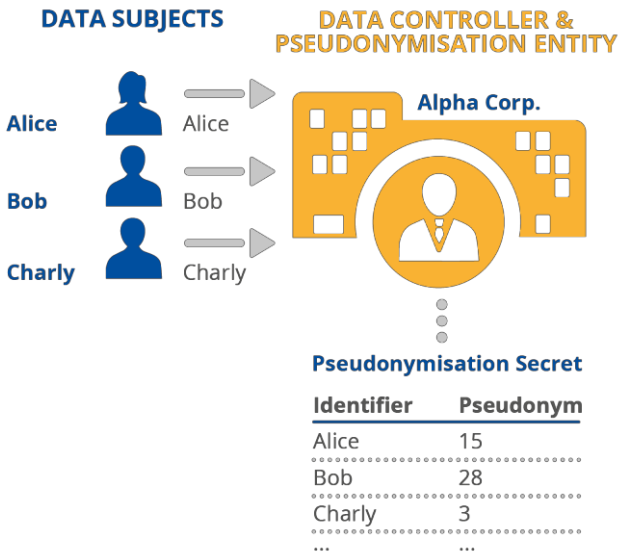
Come già accennato al punto [1], la pseudonimizzazione riveste un ruolo importante all'interno del GDPR, sia come misura di sicurezza (art. 32 del GDPR), sia nell'ambito della protezione dei dati personali sin dalla progettazione (art. 25 GDPR). Il maggiore e ovvio vantaggio della pseudonimizzazione è quello di nascondere a terze parti (diverse dall'ente di pseudonimizzazione), nell'ambito di una specifica elaborazione di dati, l'identità dell'interessato. Peraltro, la pseudonimizzazione può andare ben oltre il mero occultamento delle identità, raggiungendo – nell'ottica di accrescimento della protezione dei dati - l'obiettivo della incollegabilità [2], così riducendo il rischio che i dati personali vengano collegati attraverso la lavorazione dei dati dei domini. Inoltre, la pseudonimizzazione (la quale rappresenta una tecnica di minimizzazione dei dati) può contribuire a favorire, ai sensi del GDPR, i principi di minimizzazione, come, ad esempio, nei casi in cui il titolare non abbia necessità di accedere alle reali identità degli interessati, ma esclusivamente al loro pseudonimo. Infine, un ulteriore importante vantaggio della pseudonimizzazione che non può essere sottovalutato è rappresentato dalla veridicità dei dati (per un'analisi dettagliata sul ruolo della pseudonimizzazione si rimanda al punto [1] della bibliografia).

Tenendo in considerazione i suddetti vantaggi, il presente Capitolo si concentrerà sulle diverse ipotesi di pseudonimizzazione, analizzando i vari soggetti e gli obiettivi specifici di ogni singolo caso.

3.1 IPOTESI 1: PSEUDONIMIZZAZIONE PER USO INTERNO

L'ipotesi più comune di pseudonimizzazione dei dati si ha quando i dati vengono forniti direttamente dall'interessato e pseudonimizzati dal titolare del trattamento mediante una sequenza di lavorazione interna.

Figura 1 - Esempio di pseudonimizzazione di cui



Come mostrato nella Figura 1, il titolare del trattamento (Alpha Corp.) assume il ruolo di ente di pseudonimizzazione, in quanto provvede all'assegnazione degli pseudonimi ai diversi identificativi. Va sottolineato come gli interessati non debbano necessariamente conoscere, né imparare, il loro pseudonimo; in tal modo, il segreto di pseudonimizzazione (quale, come nel presente esempio, la tabella di mappatura) è conosciuta solo da Alpha Corp. In questo caso, il ruolo della pseudonimizzazione è quello di accrescere il livello di sicurezza del trattamento dei dati personali, sia per un

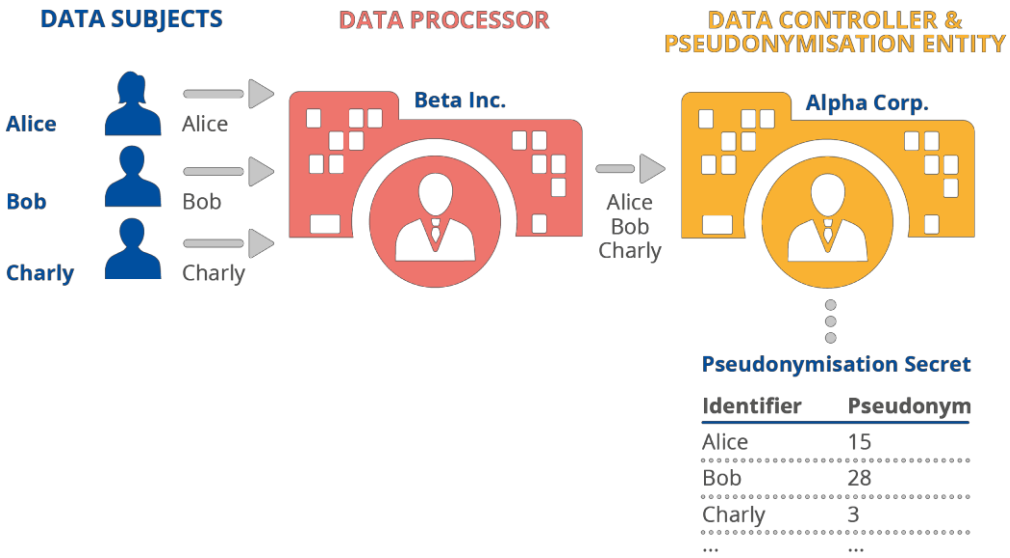
uso interno (come, ad esempio, la condivisione tra diverse unità del titolare del trattamento)[11], sia in caso di violazione della sicurezza.

[11] Cfr. Anche il considerando (29) GDPR sulla nozione di "analisi generale" per uso interno.

3.2 IPOTESI 2: COINVOLGIMENTO DEL RESPONSABILE DEL TRATTAMENTO NELLA PSEUDONIMIZZAZIONE

Tale ipotesi rappresenta una variante dell'Ipotesi 1 e si verifica allorquando, all'interno del processo, venga coinvolto anche un responsabile del trattamento, ai fini di ottenere (per conto del titolare del trattamento) gli identificativi degli interessati. Ad ogni modo, la pseudonimizzazione è sempre effettuata dal titolare del trattamento.

Figura 2 - Esempio di pseudonimizzazione di cui all'ipotesi 2

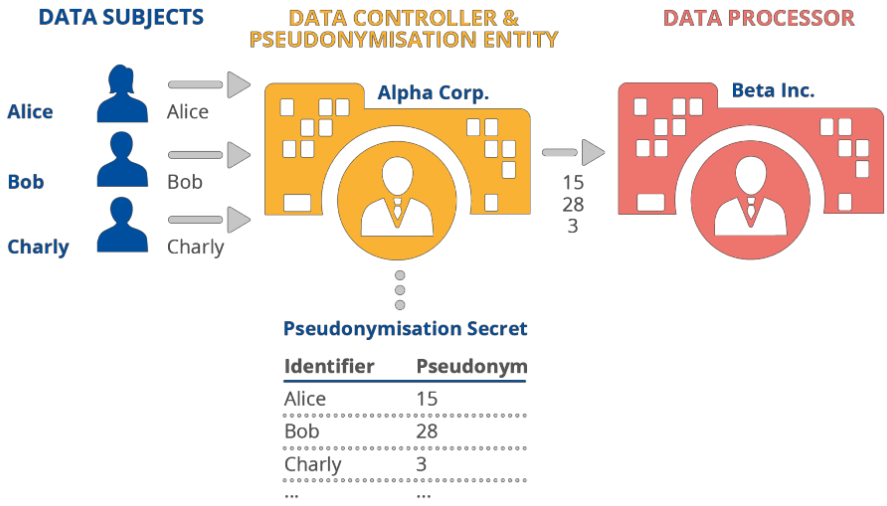


Come mostrato nella Figura 2, un dato responsabile del trattamento dei dati (Beta Inc.) si occupa di raccogliere gli identificativi dagli interessati e inviare tali informazioni al titolare del trattamento (Alpha Corp.), il quale – da ultimo – esegue la pseudonimizzazione. Il titolare è, anche in tale ipotesi, l'ente di pseudonimizzazione. Si pensi al caso di un fornitore di servizi cloud che ospita un servizio di raccolta di dati per conto del titolare del trattamento. In tale ipotesi, il titolare permane nel suo ruolo di responsabile dell'applicazione della pseudonimizzazione dei dati prima di ogni successiva elaborazione. Gli obiettivi di tale pseudonimizzazione sono i medesimi di cui all'ipotesi 1 (ma in tal caso viene coinvolto nel processo anche il responsabile del trattamento dei dati).

3.3 IPOTESI 3: INOLTRO DEI DATI PSEUDONIMIZZATI AL RESPONSABILE DEL TRATTAMENTO

Contrariamente a quanto visto nell'ipotesi precedente, in tal caso il titolare del trattamento si occupa della pseudonimizzazione, ma il responsabile non viene coinvolto nel processo, limitandosi a ricevere i dati già pseudonimizzati dal titolare.

Figura 3 - Pseudonimizzazione - Esempio Scenario 3



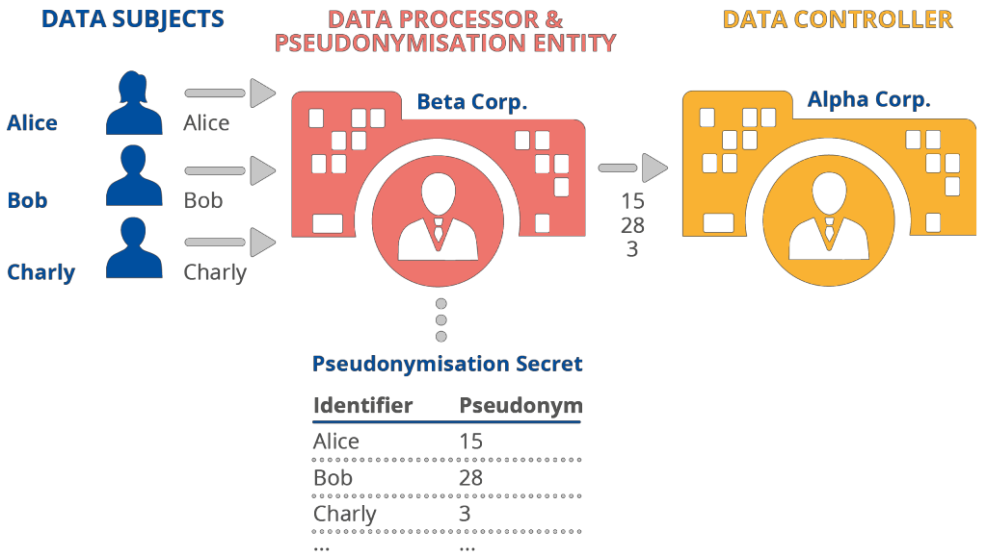
Come mostrato nella Figura 3 il titolare del trattamento (Alpha Corp.) si occupa della raccolta e della pseudonimizzazione dei dati (nella sua qualità di ente della pseudonimizzazione). La differenza rispetto all'ipotesi di cui al punto che precede risiede nella circostanza che, in tal caso, il titolare del trattamento invia i dati già pseudonimizzati a un successivo responsabile (Beta Inc.), ai fini, ad esempio, delle analisi statistiche, ovvero della conservazione dei dati. In tale ipotesi, l'obiettivo della protezione dei dati data dalla pseudonimizzazione sarà garantito dal fatto che Beta Inc. non è in grado di risalire agli identificativi degli interessati (assumendo che non vi siano altre caratteristiche che Beta Inc. possa utilizzare ai fini di compiere una re-identificazione). In questo caso, la pseudonimizzazione protegge la sicurezza dei dati persino dal responsabile del trattamento.

Una variante di tale ipotesi può aversi nel caso in cui i dati pseudonimizzati non vengano inviati al responsabile del trattamento, bensì a un altro titolare (come, ad esempio, nei casi in cui il titolare sia a ciò legalmente obbligato o da qualche altro vincolo di legge).

3.4 IPOTESI 4: RESPONSABILE DEL TRATTAMENTO QUALE ENTE DI PSEDONIMIZZAIZONE

Un'altra possibile ipotesi si ha quando il titolare del trattamento affida il compito della pseudonimizzazione al responsabile del trattamento (quale, ad esempio, un servizio cloud che gestisce il segreto di pseudonimizzazione e/o applica i mezzi tecnici rilevanti).

Figura 4 - Esempio di pseudonimizzazione di cui all'ipotesi 4



Come mostrato nella Figura 4, i dati personali vengono inviati dall'interessato al responsabile del trattamento (Beta Inc.), il quale, successivamente, effettuerà la pseudonimizzazione, agendo in qualità di ente di pseudonimizzazione per conto del titolare del trattamento (Alpha Corp). In seguito, i dati pseudonimizzati vengono inoltrati al titolare. In tale particolare ipotesi, solo i dati pseudonimizzati vengono raccolti presso il titolare del trattamento. Cosicché, la sicurezza da parte del titolare è rafforzata dalla necessaria re-identificazione dei dati (ad. esempio, nel caso di una violazione di dati presso il titolare).

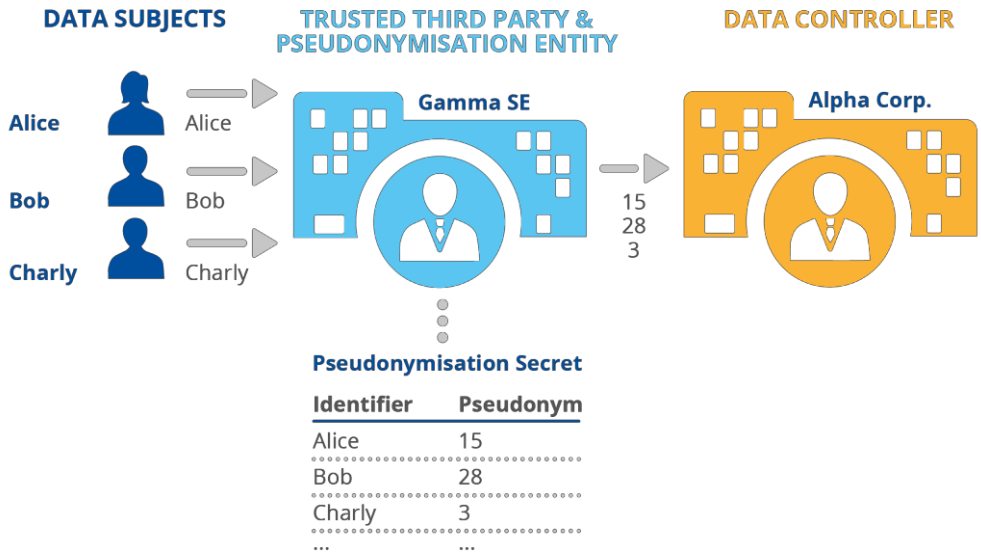
E anche in tutti quei casi in cui il titolare del trattamento sia in grado di re-identificare i dati dell'interessato attraverso il responsabile del trattamento. Dunque, la sicurezza da parte del responsabile del trattamento assume una importanza cruciale.

Una variante di tale ipotesi è rappresentata dal caso in cui siano coinvolti nel processo di pseudonimizzazione numerosi responsabili differenti, all'interno di una sequenza di enti di pseudonimizzazione (c.d. catena di responsabili).

3.5 IPOTESI 5: TERZE PARTI QUALI ENTI DI PSEUDONIMIZZAZIONE

In tale ipotesi la pseudonimizzazione è eseguita da una terza parte (diversa dal responsabile) che, successivamente, fornirà i dati al titolare. Diversamente a quanto visto nell'Ipotesi 4, in questo caso il titolare del trattamento non ha accesso agli identificativi degli interessati (poiché la terza parte non è assoggettata al controllo del titolare).

Figura 5 - Esempio di pseudonimizzazione di cui all'ipotesi 5



Come mostrato nella Figura 5 i dati personali sono inviati a una terza parte (Gamma SE), la quale, successivamente, eseguirà la pseudonimizzazione, in qualità di ente di pseudonimizzazione. I dati pseudonimizzati vengono, poi, inviati al titolare del trattamento (Alpha Corp). In tale ipotesi, il titolare non è, da solo, in grado, direttamente o indirettamente, di collegare i singoli record di dati ai rispettivi interessati. In tal caso, la sicurezza e la protezione dei dati da parte del titolare sono garantiti dal rispetto del principio di minimizzazione. Una tale ipotesi potrà trovare applicazione nei casi in cui il titolare non ha necessità di accedere agli identificativi degli interessati (bastando l'accesso ai soli pseudonimi).

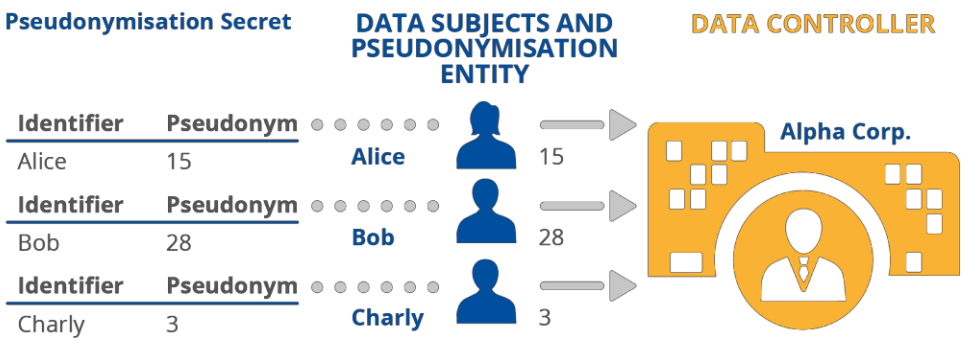
Tale ipotesi risulta assai rilevante nei casi di titolarità condivisa, ove uno dei titolari esegue la pseudonimizzazione (agendo in qualità di terza parte fidata - TTP nella Figura 5), mentre l'altro riceve i dati pseudonimizzati per una ulteriore lavorazione.

Una variante interessante di tale ipotesi (la quale meriterebbe ulteriori approfondimenti) può essere rappresentata dal caso in cui TTP venga consegnato a più di un solo ente, i quali possono creare o ripristinare congiuntamente gli pseudonimi (o possono basarsi sulla condivisione dello schema del segreto), in modo tale da ripartire l'affidabilità tra più enti.

3.6 IPOTESI 6: INTERESSATO QUALE ENTE DI PSEDONIMIZZAZIONE

Tale ipotesi rappresenta un caso particolare di pseudonimizzazione nel quale gli pseudonimi sono creati dall'interessato stesso quale parte integrante dell'intero processo di pseudonimizzazione.

Figura 6 - Esempio di pseudonimizzazione di cui all'ipotesi 6



Come mostrato nella Figura 6, ogni individuo genera il proprio pseudonimo e, successivamente, inoltra i propri dati pseudonimizzati[12].

Un esempio di tale sistema di pseudonimizzazione dei dati potrebbe consistere nell'utilizzo di chiavi pubbliche della coppia di chiavi in un sistema blockchain (quale, ad esempio, Bitcoin) per produrre pseudonimi. L'obiettivo della pseudonimizzazione è dato, in questo caso, dal fatto che il titolare non è in grado di conoscere[13] gli identificativi degli interessati, cosicché gli stessi mantengono il controllo del processo di pseudonimizzazione; ovviamente, la responsabilità dell'intero schema di pseudonimizzazione resta, comunque, nella sfera del titolare del trattamento[14]. Ancora una volta siamo difronte un'applicazione pratica del principio di minimizzazione, il quale può essere applicato nei casi in cui il titolare del trattamento non necessita degli identificativi originari (per cui, ad esempio, gli pseudonimi sono sufficienti ai fini dello specifico trattamento dei dati).

[12] Si sottolinea che possono essere utilizzati, a seconda dei diversi servizi/applicazioni, gli stessi pseudonimi ovvero pseudonimi differenti (di cui meglio al Capitolo 5)
[13] Nel senso che il titolare non acquisisce alcun segreto di pseudonimizzazione in grado di consentire una diretta re-identificazione.
[14] Si veda, tra l'altro, l'art. 11 del GDPR.

4. MODELLI DI INTRUSIONE

Come riportato nel Capitolo 3, l'obiettivo principale della pseudonimizzazione è quello di limitare la riconducibilità di un set di dati pseudonimizzato ai titolari degli pseudonimi, così proteggendo l'identità degli interessati. Questo tipo di protezione è, generalmente, volto a contrastare gli sforzi di un intruso tesi a compiere un attacco di re-identificazione.

Il presente Capitolo prende in considerazione i possibili esempi di intrusione e i diversi tipi di attacchi di re-identificazione, maggiormente rilevanti nell'ambito della pseudonimizzazione. A tal fine, vengono fornite le nozioni di intrusi interni ed esterni, alla luce del possibile ruolo ricoperto da ognuno nei diversi casi di pseudonimizzazione già discussi in precedenza. La comprensione di tali argomenti è un elemento essenziale ai fini di una corretta analisi sull'uso delle tecniche di pseudonimizzazione riportate nei capitoli seguenti.

4.1 INTRUSIONI INTERNE

Secondo la comune accezione dei termini correnti nel linguaggio di sicurezza IT, un intruso interno è un addetto ai lavori con conoscenze, capacità o autorizzazioni specifiche, in relazione all'obiettivo finale[15]. Nel contesto della pseudonimizzazione, ciò implica che l'intruso è in grado di ottenere informazioni sul segreto di pseudonimizzazione e/o altre informazioni rilevanti. Si prendano, ad esempio, in considerazione le ipotesi 1, 2, 3 e 4 di cui al Capitolo 3, nelle quali un addetto ai lavori potrebbe trovarsi dalla parte del titolare del trattamento (come nel caso di un dipendente che lavora per il titolare). Potrebbe anche ricoprire il ruolo di responsabile (come nel caso di un impiegato malintenzionato di un appaltatore), come visto nelle ipotesi 2 e 4. Infine, come nell'ipotesi 5, l'intruso interno potrebbe essere una terza parte fidata (agendo in questo caso quale ente di pseudonimizzazione). Per definizione le terze parti che potrebbero legittimamente avere accesso ai dati personali (quali, ad esempio, un'autorità di controllo) non sono considerate intrusi.[16]

4.2 INTRUSIONI ESTERNE

Contrariamente all'intruso interno, quello esterno non ha accesso diretto al segreto di pseudonimizzazione o ad altre informazioni pertinenti. Tuttavia, questo tipo di intruso può avere accesso a un set di dati pseudonimizzati e può anche essere in grado di eseguire una pseudonimizzazione basandosi sui valori dei dati di input scelti arbitrariamente dall'intruso (come, ad esempio, grazie all'accesso ad un'implementazione in black box della funzione di pseudonimizzazione, oppure obbligando l'ente di pseudonimizzazione a pseudonimizzare input arbitrari). L'obiettivo di un intruso esterno è quello di aumentare le proprie

[15] Secondo il CERT Insider Threat Center presso il Software Engineering Institute (SEI) della Carnegie Mellon University, la minaccia da parte di un intruso interno viene definita come la possibilità per un individuo che ha, o a ha avuto, un accesso autorizzato alle risorse di un'organizzazione di utilizzare il proprio accesso, più o meno intenzionalmente, ai fini di agire in modo tale da influire negativamente sull'organizzazione, https://insights.sei.cmu.edu/insider-threat/2017/03/cert-definition-of-insider-threat---updated.html

[16] 16 Va notato, tuttavia, che la legittimità di tale accesso potrebbe essere messa in discussione nei casi in cui non venga rispettato il principio di minimizzazione (come, ad esempio, nei casi di un'autorità di controllo che ottiene l'accesso al segreto di pseudonimizzazione piuttosto che ricevere esplicitamente solo i dati personali). Queste ipotesi rientrerebbero nei modelli di intrusione interna, poiché la terza parte, analogamente all'ente di pseudonimizzazione, è legittimata all'accesso interno.

informazioni sul set di dati pseudonimizzato, acquisendo, ad esempio, l'identità dietro ad un determinato pseudonimo (e, altresì, acquisendo ulteriori informazioni su tale identità mediane l'utilizzo di dati aggiuntivi trovati nel set di dati relativo lo pseudonimo di riferimento).

Si prendano in considerazione le ipotesi di cui al Capitolo 3, ove è possibile evincere come, per definizione, qualsiasi soggetto che agisce in modo fraudolento, senza essere parte dell'ente di pseudonimizzazione, né lavorando per conto dello stesso, dovrebbe essere sempre considerato un intruso esterno. Pure un titolare del trattamento dei dati (malintenzionato) potrebbe assumere il ruolo di intruso esterno, nei casi di cui alle ipotesi 5 o 6. Un responsabile del trattamento dei dati (malintenzionato) potrebbe assumere tale ruolo anche nell'ipotesi 3.

4.3 OBIETTIVI DEGLI ATTACCHI ALLA PSEUDONOMIZZAZIONE

A seconda del contesto e del metodo di pseudonimizzazione impiegato, l'intruso può voler raggiungere, con riguardo ai dati pseudonimizzati, obiettivi diversi, quali il recupero del segreto di pseudonimizzazione, la completa re-identificazione o la discriminazione. Mentre la maggior parte degli esempi descritti nei prossimi paragrafi si concentra sulla scoperta della vera identità degli interessati, va sottolineato che un attacco riuscito non è (solo) una questione di ingegneria inversa, ma attiene più alla capacità di individuare, all'interno di un gruppo, un utente specifico (anche qualora la vera identità non venga rivelata).

4.3.1 Il Segreto di Pseudonimizzazione

In questo caso (ovvero nelle ipotesi in cui viene usato il segreto di pseudonimizzazione), l'intruso ha come obiettivo la scoperta del segreto di pseudonimizzazione. Questo è il più grave tipo di attacco, poiché attraverso l'utilizzo del segreto di pseudonimizzazione, l'intruso è in grado di re-identificare, all'interno del set di dati, qualsiasi pseudonimo (re-identificazione completa o discriminazione), nonché di eseguire ulteriori processi di pseudonimizzazione sul set di dati.

4.3.2 La re-identificazione completa

Quando l'obiettivo dell'attacco è la completa re-identificazione, l'intruso ha l'obiettivo di ricondurre uno o più pseudonimi all'identità dei titolari dello pseudonimo. Tale tipo di intrusione è stata ampiamente discussa in letteratura (si veda, in merito, i punti [3] [4] [5] della bibliografia).

Il più grave attacco di re-identificazione completa consiste nella re-identificazione di tutti gli pseudonimi. A tal fine l'intruso può utilizzare due strategie: recuperare autonomamente ogni identificativo dallo pseudonimo corrispondente; oppure recuperare il segreto di pseudonimizzazione (di cui si è detto in 4.3.1). Gli attacchi di re-identificazione completa di minore entità riguardano, invece, i casi in cui l'intruso può solo re-identificare, all'interno del set di dati, un sottoinsieme di pseudonimi. Si prenda in considerazione il caso di un set di dati pseudonimizzato relativo ai voti degli studenti di un corso universitario. Ogni voce del set di dati contiene uno pseudonimo che corrisponde all'identità di uno studente (nome e cognome) e un secondo pseudonimo che corrisponde al genere dello studente (così attribuendo, ad esempio, agli studenti di genere

maschile i numeri dispari e a quelli di genere femminile i numeri pari). Un intruso sarà in grado di completare l'attacco di re-identificazione solo recuperando nome, cognome e genere di uno studente.

4.3.3 La discriminazione

L'obiettivo dell'attacco di discriminazione è quello di identificare le caratteristiche (o almeno una) del titolare dello pseudonimo. Tali caratteristiche potrebbero non essere in grado, da sole, di rivelare l'identità del titolare dello pseudonimo, ma potrebbero essere sufficienti a discriminarlo in qualche modo.

Si prenda in considerazione l'esempio dei voti degli studenti di cui al punto precedente; il set di dati relativo i voti studenteschi può contenere, quali pseudonimi, due soli numeri pari tra molti numeri dispari. Come visto, i numeri pari corrispondono agli studenti di genere femminile, mentre i numeri dispari corrispondono agli studenti di genere maschile. Si pensi all'ipotesi in cui i numeri pari abbiano ottenuto il 100% come risultato all'esame finale. Si supponga, inoltre, che, all'interno del set di dati pseudonimizzato, non vi siano altri studenti che abbiano ottenuto il 100%. Qualora un intruso fosse, ulteriormente, in grado di conoscere che un dato studente ha ottenuto il 100% in quel corso, automaticamente saprà che quello studente appartiene al genere femminile. Viceversa, qualora l'intruso venisse a conoscenza che uno studente di quel corso è di genere femminile, verrà automaticamente a sapere che lo stesso avrà totalizzato il 100%. E' importante sottolineare che, in tali ipotesi, l'intruso non apprende l'identità del titolare dello pseudonimo, riuscendo a ottenere solo alcune delle sue caratteristiche (come, negli esempi di cui sopra, il voto o il genere). Infatti, poiché diversi studenti condividono la caratteristica del medesimo voto, l'intruso non sarà in grado di individuare l'esatto record di dati appartenente a uno specifico titolare dello pseudonimo. Tuttavia, l'acquisizione di tali informazioni aggiuntive può, di per sé, essere sufficiente ai fini della discriminazione che l'intruso intende compiere, ovvero può essere utilizzata in un successivo attacco finalizzato a scoprire l'identità nascosta dietro uno pseudonimo.

4.4 PRINCIPALI TECNICHE DI ATTACCHI

Esistono tre principali tecniche generali in grado di interrompere una funzione di pseudonimizzazione: gli attacchi di forza bruta (ricerca esaustiva), la ricerca nel dizionario e la congettura[17]. L'efficacia di questi attacchi dipende da diversi parametri, tra cui:

- La quantità di informazioni che uno pseudonimo è in grado di fornire sul suo titolare (interessato).
- Le conoscenze di base dell'intruso.
- La dimensione del dominio dell'identificativo.
- La dimensione del dominio dello pseudonimo.
- La scelta e la configurazione della funzione di pseudonimizzazione utilizzata (ciò include anche la dimensione del segreto di pseudonimizzazione).

Dette tecniche di attacco verranno, brevemente, descritte di seguito.

[17] Va notato che, come evidenziato precedentemente nel trattato, anche altre caratteristiche (oltre allo pseudonimo e ai dati pseudonimizzati) possono essere utilizzate per identificare l'interessato. Per ulteriori approfondimenti sul tema, si rimanda al capitolo 8.

4.4.1 L’attacco brute force

La praticabilità di questa tecnica di attacco è condizionata dall’abilità dell'intruso nel calcolare la funzione di pseudonimizzazione (in questi casi non vi è alcun segreto di pseudonimizzazione) o alla sua capacità di accedere ad un'implementazione "scatola nera" della funzione di pseudonimizzazione. A seconda dell'obiettivo dell'attacco, la riuscita dell’attacco potrebbe essere subordinata ad ulteriori condizioni. Qualora l'attacco di forza bruta venga utilizzato per ottenere una completa re-identificazione (ovvero il ripristino dell'identità originaria), è necessario che il dominio dell’identificativo sia finito e relativamente piccolo. Per ogni pseudonimo incontrato, l'intruso potrà tentare di recuperare l'identificativo originario applicando la funzione di pseudonimizzazione su ciascun valore del dominio dell'identificativo, fino a quando non troverà una corrispondenza.

Tabella 1: Pseudonimizzazione del mese di nascita

Gen.	281	**Lug.**	299
Feb.	269	**Ago.**	285
Mar.	288	**Set.**	296
Apr.	291	**Ott.**	294
Mag	295	**Nov.**	307
Giu	301	**Dic.**	268

Si prenda in considerazione la pseudonimizzazione di un set di dati composto dai mesi di nascita. La dimensione del dominio dell’identificativo è 12, quindi un intruso potrà enumerare rapidamente tutte le possibilità. Gli pseudonimi associati ad ogni mese vengono calcolati in questo caso con la somma del codice ASCII delle prime tre lettere del mese di nascita (con la prima lettera maiuscola). Si consideri, allora, che un intruso incontri lo pseudonimo 301. A questo punto, potrà ricercare per ogni mese di nascita la funzione di pseudonimizzazione fino a quando non troverà il mese che corrisponde al valore 301. La Tabella 1 mostra i calcoli effettuati dall'intruso per re-identificare lo pseudonimo 301 risultante nella tabella di mappatura della funzione di pseudonimizzazione.

Ovviamente, per eseguire correttamente tale tipo di attacco, la dimensione del dominio degli identificativi risulta essere fondamentale. Per domini dell’identificativo di piccole dimensioni, come nell'esempio di cui sopra, un attacco di forza bruta è facilmente attuabile. Se la dimensione del dominio dell’identificativo è infinita, l'attacco di forza bruta diventa, solitamente, impossibile. Se la dimensione del dominio dell’identificativo è molto grande, la completa re-identificazione diventa estremamente difficile, lasciando tuttavia agli intrusi la possibilità di compire un attacco di discriminazione.

Infatti, in tal caso l'intruso può considerare un sottodominio del dominio dell’identificativo, per il quale può calcolare tutti gli pseudonimi. Tornando all'esempio della Tabella 1, che tratta un dominio di piccole dimensioni, si può supporre che l'intruso voglia discriminare le persone nate in un mese che inizia con la

lettera J dalle persone nate in un mese che inizia con una lettera diversa. Questo sottodominio contiene gennaio (Jennuary), giugno (June) e luglio (July). L'intruso può intraprendere una ricerca esaustiva su questo sottodominio calcolando gli pseudonimi corrispondenti a gennaio, giugno e luglio. Qualora trovasse uno pseudonimo diverso da 281, 301 e 299, allora sarà in grado di sapere che il mese di nascita non inizia con la lettera J.

Tuttavia, nel caso in cui venga utilizzato il segreto di pseudonimizzazione, anche in presenza di un dominio dell'identificativo di piccole dimensioni, potrebbe non verificarsi un tale attacco (quantomeno sin quando l'intruso non sia in grado di calcolare la funzione di pseudonimizzazione e non abbia accesso a un'implementazione a "scatola nera" di tale funzione).

In tal caso, un attacco di forza bruta può essere applicato su tutto lo spazio relativo i segreti di pseudonimizzazione – vale a dire che l'intruso controlla esaustivamente tutti i possibili segreti e, per ciascuno di essi, calcola la funzione di recupero. Questo attacco avrà successo se l'intruso sarà in grado di identificare correttamente il segreto di pseudonimizzazione, indipendentemente dalle dimensioni del dominio dell'identificativo. Pertanto, per contrastare un simile attacco, il numero dei possibili segreti di pseudonimizzazione dovrebbe essere sufficientemente elevato in modo tale da rendere pressoché impossibile l'attacco.

4.4.2 La ricerca nel dizionario

La ricerca nel dizionario è un'ottimizzazione dell'attacco di forza bruta, che permette di economizzare sui costi di calcolo. In tale ipotesi l'intruso, per effettuare una completa re-identificazione o discriminazione, deve confrontarsi con una grande quantità di pseudonimi. Pertanto, predispone un (consistente) insieme di pseudonimi e salva il risultato in un dizionario. Ogni voce del dizionario contiene uno pseudonimo e l'identificativo o le informazioni corrispondenti. Ogni volta che l'intruso deve re-identificare uno pseudonimo, eseguirà una ricerca nel dizionario. Tale ricerca ha un costo di pre-calcolo che equivale a quello di una ricerca esaustiva e archivia il risultato in una memoria di grandi dimensioni. Di contro, la successiva re-identificazione di uno pseudonimo ha il solo costo della ricerca nel dizionario. La ricerca nel dizionario è essenzialmente il calcolo e l'archiviazione della tabella di mappatura. Gli scambi memoria/tempo sono persino possibili usando le tabelle di Hellman [6] o le tabelle arcobaleno [7] in grado di ampliare ulteriormente la gamma. Tuttavia, esistono varianti specifiche di questo tipo di attacco basate su ulteriori conoscenze relative il funzionamento della funzione di pseudonimizzazione. Tali attacchi sono in grado persino di operare su infiniti domini di input.

4.4.3 La congettura

Questo tipo di attacco utilizza alcune conoscenze di base (quale la distribuzione di probabilità o qualsiasi altra informazione marginale) che l'intruso può avere a disposizione su alcuni (o tutti) i titolari di uno pseudonimo, la funzione di pseudonimizzazione o l'insieme di dati. Implicitamente, la ricerca esaustiva e la ricerca nel dizionario presuppongono che tutti gli identificativi abbiano la stessa probabilità o frequenza di verificarsi. Tuttavia, alcuni identificativi potrebbero verificarsi con una maggiore frequenza. L'impiego delle caratteristiche statistiche degli identificativi è noto come congettura [8] [9] [10] ed è ampiamente adottato nella comunità di "password-cracking". Si sottolinea che la congettura può comunque essere applicata

anche quando il dominio degli identificativi è molto ampio. L'intruso non deve necessariamente avere accesso alla funzione di pseudonimizzazione (poiché la discriminazione è possibile semplicemente eseguendo un'analisi di frequenza degli pseudonimi monitorati).

Si prenda in considerazione il caso in cui gli pseudonimi corrispondono a “nomi propri". Il dominio dei "nomi propri" è difficile da esaminare nella sua interezza. Tuttavia, l'intruso conosce quali "nomi propri" sono i più popolari (Tabella 2). A questo punto, l'intruso può avviare una ricerca esaustiva o una ricerca nel dizionario utilizzando come parametro i "nomi propri" più popolari, così ottenendo una discriminazione.

Tabella 2 - Lista dei nomi propri più diffusi

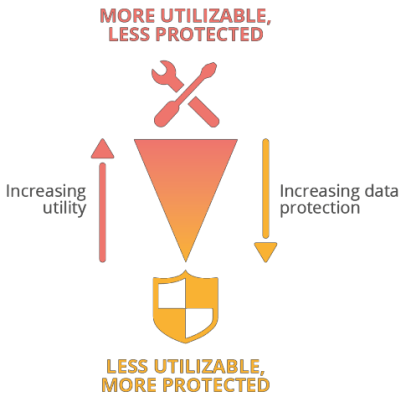
Most popular first names					
Bob	Alice	Charlie	Eve	Robert	Marie

Prendiamo in considerazione un caso analogo, ma con un dominio di identificativi di dimensione infinita. E’ possibile definire, all’interno del set di dati, un sottodominio finito di identificativi. Qualora l'intruso riuscisse a identificare tale sottodominio, sarà in grado di eseguire una ricerca esaustiva (si rimanda al Capitolo 6 per il pertinente caso pratico relativo la pseudonimizzazione dell'indirizzo e-mail). A seconda della quantità di informazioni di base o di metadati posseduti dall'intruso e della quantità di informazioni collegabili trovate nel set di dati pseudonimizzato, questo tipo di attacco può portare a scoprire l'identità del singolo utente, l'intero set di dati o una frazione di essi. Soprattutto per set di dati di piccole dimensioni, tali attacchi possono essere concretizzabili con alti tassi di successo.

4.5 FUNZIONALITA’ E PROTEZIONE DEI DATI

A seconda della funzione di pseudonimizzazione che si sceglie di utilizzare, uno pseudonimo può contenere alcune informazioni sull'identificativo originale. Pertanto, ogni pseudonimo implica il rischio di subire un attacco di re-identificazione come quelli sopra descritti. Ad esempio, un intruso con sufficienti conoscenze di base potrebbe essere in grado di ricollegare lo pseudonimo al suo identificativo utilizzando il metodo della congettura.

Figura 7 - Utilità e Protezione dei dati



Ad ogni modo, in molti casi, le informazioni aggiuntive contenute nello pseudonimo sull'identificativo originario vengono conservate per permettere il collegamento tra gli stessi pseudonimi, il quale deve essere eseguito dal successivo titolare del trattamento dei dati validamente nominato. Ad esempio, uno pseudonimo può conservare al suo interno l'anno di nascita di una persona come parte integrante dello pseudonimo (ad esempio “AAAA-1999”). In tal modo, è possibile classificare gli pseudonimi in base all’anno di nascita, ad esempio con riferimento all’età, allo status giuridico (minore o adulto), alle condizioni di vita (studente/impiegato/pensionato) e altro. Tale caratteristica della funzione di pseudonimizzazione può essere utilizzata intenzionalmente, al fine di consentire ai titolari del trattamento di classificare i dati pseudonimizzati.

Chiaramente, la scelta della funzione di pseudonimizzazione, pur tenendo conto della potenziale perdita di protezione causata da siffatta tecnica di pseudonimizzazione, può favorire la funzionalità degli pseudonimi creati. Pertanto, è possibile prendere in considerazione una via di mezzo tra funzionalità e protezione dei dati (v. Figura 7). Nei casi di applicazione pratica della pseudonimizzazione, un tale compromesso dovrebbe essere attentamente analizzato, in modo da raggiungere la massima funzionalità con riferimento agli scopi prefissati, mantenendo il più intatta possibile la protezione degli interessati pseudonimizzati.

5. TECNICHE DI PSEUDONIMIZZAZIONE

Partendo dai modelli di intrusione e i tipi di attacco di cui al Capitolo 4, il presente Capitolo presenterà, brevemente, le tecniche e i metodi di pseudonimizzazione attualmente maggiormente impiegati. Per una analisi più dettagliata sulle primitive crittografiche si rimanda al punto [1] della bibliografia.

In linea di principio, una funzione di pseudonimizzazione associa gli identificativi agli pseudonimi. Nella funzione di pseudonimizzazione vi è un requisito fondamentale. Si prendano in considerazione due diversi identificativi Id_1 e Id_2 corrispondenti agli pseudonimi $pseudo_1$ e $pseudo_2$. Una funzione di pseudonimizzazione deve verificare che $pseudo_1$ sia diverso da $pseudo_2$. Diversamente, il recupero dell'identificativo potrebbe risultare ambiguo, poiché l'ente di pseudonimizzazione non è in grado di determinare se $pseudo_1$ corrisponde a Id_1 o Id_2. Tuttavia, un singolo identificativo Id può essere associato a più pseudonimi ($pseudo_1$, $pseudo_2$ ...), purché sia possibile per l'ente di pseudonimizzazione invertire tale operazione. In ogni caso, a seconda della definizione di pseudonimizzazione (di cui al Capitolo 2), esistono alcune informazioni aggiuntive che consentono di associare gli pseudonimi con gli identificativi originali: ovvero il segreto di pseudonimizzazione. Il caso più semplice è la tabella di mappatura della pseudonimizzazione.

Nelle sezioni seguenti, vengono definite primariamente le principali opzioni disponibili per pseudonimizzare un singolo identificativo. Verranno quindi confrontate le diverse tecniche di pseudonimizzazione disponibili e le loro caratteristiche di attuazione. Verranno inoltre riferiti i principali criteri a cui il titolare del trattamento potrà ricorrere per scegliere validamente quale tecnica di pseudonimizzazione impiegare. Infine, verranno discusse le opzioni di ripristino della pseudonimizzazione da parte dell'ente di pseudonimizzazione.

5.1 PSEUDONIMIZZAZIONE DI UN SINGOLO IDENTIFICATIVO

Di seguito si riporta un elenco di possibili tecniche di pseudonimizzazione, unitamente ai vantaggi e ai limiti del caso, partendo dalla pseudonimizzazione di un singolo identificativo.

5.1.1 Il Contatore

Il Contatore è la funzione di pseudonimizzazione più semplice. In tale ipotesi, gli identificativi vengono sostituiti da un numero scelto da un contatore monotono. In primis, un seme s viene (ad esempio) impostato su 0 e, successivamente, viene incrementato. Al fine di evitare ambiguità, è fondamentale che i valori prodotti dal contatore non vengano mai ripetuti.

I vantaggi del contatore risiedono nella sua semplicità, che lo rende un buon candidato per set di dati ridotti e non complessi. In termini di protezione dei dati, il contatore fornisce pseudonimi senza alcuna connessione con gli identificativi iniziali (sebbene il carattere sequenziale del contatore possa comunque fornire informazioni sull'ordine dei dati all'interno di un set di dati). Tale soluzione, tuttavia, può presentare problemi

di implementazione e scalabilità nei casi di set di dati di grandi dimensioni e più sofisticati, poiché è necessario archiviare la tabella completa di mappatura della pseudonimizzazione.

5.1.2 Il generatore casuale di numeri (RNG)

L'RNG è un meccanismo che produce un insieme di valori i quali presentano, all'interno dell'intera selezione dell'insieme, la medesima probabilità di essere selezionati e sono, pertanto, imprevedibili[18]. Tale approccio è simile a quello del contatore, con la differenza che all'identificativo viene attribuito un numero causale. Ai fini della creazione di tale mappatura, sono disponibili due opzioni: l'utilizzo di un vero generatore di numeri casuali, ovvero di un generatore crittografico pseudo-casuale (per le definizioni specifiche si rimanda al punto [11] della bibliografia). Va evidenziato che, senza la dovuta cura, in entrambi i casi possono verificarsi delle collisioni[19]. Una collisione si verifica allorquando due identificativi vengono associati al medesimo pseudonimo. La probabilità che si verifichi una tale ipotesi collisiva è legata al verificarsi del cosiddetto paradosso del compleanno [12].

L'RNG fornisce una solida protezione dei dati (poiché, contrariamente al contatore, ogni pseudonimo viene creato impiegando un numero casuale, così risultando difficile l'estrazione di informazioni sull'identificativo iniziale, a meno che la tabella di mappatura non sia compromessa). A seconda del caso d'applicazione, come detto, le collisioni possono rappresentare un problema, così come la scalabilità (ragion per cui la tabella completa di mappatura della pseudonimizzazione deve essere archiviata).

5.1.3 La funzione crittografica di Hash

La funzione crittografica di Hash è progettata per prendere in input una stringa di qualsiasi lunghezza e produrre in output un valore di Hash di lunghezza fissa [13] [14]. Tale funzione di Hash soddisfa le seguenti proprietà:

- Unidirezionalità: è computazionalmente impossibile trovare un qualsiasi input che corrisponda a un qualsiasi output pre-specificato.
- Resistenza alla collisione: è computazionalmente impossibile trovare due input distinti associati allo stesso output.

Una funzione crittografica di Hash viene applicata direttamente all'identificativo in modo da ottenere lo pseudonimo corrispondente: $Pseudo = H\ (Id)$.

Il dominio dello pseudonimo dipende dalla lunghezza del digest prodotto dalla funzione.

Come menzionato nella nota bibliografica [1], se da un lato la funzione di HASH può essere considerata quale valido strumento in grado di preservare l'integrità dei dati, dall'altro lato viene generalmente considerata quale tecnica di pseudonimizzazione debole, poiché soggetta sia ad attacchi di forza bruta che di ricerca nel dizionario. Nei successivi Capitoli 6, 7 e seguenti, verranno trattati esempi specifici di tale vulnerabilità.

[18] Si noti che al posto del numero è possibile utilizzare anche una sequenza casuale di caratteri.
[19] Il rischio di collisioni può essere reso trascurabile se vengono generati pseudo numeri di grandi dimensioni (ad esempio di lunghezza di 100 cifre).

5.1.4 Il Message authentication code (MAC)

Questa tipologia di primitiva crittografica (HMAC) può essere vista come una modalità per l'autenticazione di messaggi (message authentication code) basata su una funzione di HASH. Sebbene simile, differisce dalla precedente tecnica poiché, per la generazione dello pseudonimo, viene impiegata una chiave segreta, senza la quale non è possibile mappare gli identificativi e gli pseudonimi. Il MAC [15] [16] rappresenta la tipologia di progettazione dei sistemi per l'autenticazione di messaggi maggiormente impiegato nei protocolli Internet.

Come riportato nella nota bibliografica [1], il MAC è generalmente considerato, dal punto di vista della protezione dei dati, una solida tecnica di pseudonimizzazione, poiché il ripristino dello pseudonimo, se la chiave non è stata compromessa, risulta impossibile. Possono essere impiegate, peraltro, diverse varianti di tale modalità crittografica, che differiscono per i diversi requisiti di funzionalità e scalabilità richiesti dall'ente di pseudonimizzazione (si rimanda ai seguenti Capitoli 6 e 7 per la trattazione pratica).

5.1.5 La crittografia

Il presente trattato prende in considerazione principalmente la crittografia simmetrica (deterministica) e in particolare i cifrari a blocchi come l'AES e le loro modalità operative [11]. La cifratura a blocchi viene utilizzata per crittografare un identificativo utilizzando una chiave segreta, la quale racchiude sia il segreto di pseudonimizzazione che il segreto di ripristino. L'uso dei cifrari a blocchi per la pseudonimizzazione richiede una distribuzione corretta della dimensione dei vari blocchi. Infatti, nella pratica, la dimensione del blocco degli identificativi può essere più piccola o più grande rispetto alla dimensione del blocco di input della cifratura a blocchi. Se la dimensione degli identificativi è inferiore, bisognerà considerare il riempimento [11]. Nel caso in cui la dimensione degli identificativi sia superiore rispetto alla dimensione del blocco possono essere adoperate due soluzioni: gli identificativi possono essere compressi fino a raggiungere la dimensione del blocco; oppure, laddove la compressione non fosse un'opzione disponibile, si può utilizzare una modalità operativa (come la modalità counter, CTR). Tuttavia, quest'ultima opzione richiede la gestione di un ulteriore parametro, il vettore di inizializzazione.

Come menzionato al punto [1] della bibliografia, la crittografia può anche essere una solida tecnica di pseudonimizzazione, con diverse proprietà simili al MAC. Si rimanda ai Capitoli 6 e 7 per alcuni esempi pratici.

Si sottolinea che, sebbene il presente trattato si concentri principalmente sugli schemi di crittografia deterministica, pure la crittografia probabilistica rappresenta un'alternativa valida da impiegare, specie nei casi in cui è necessario ottenere pseudonimi diversi per uno stesso identificativo (di seguito, si vedrà anche il metodo di pseudonimizzazione totalmente casuale). Per ulteriori approfondimenti sul tema, si rimanda al punto [1] della bibliografia.

5.2 METODI DI PSEUDONIMIZZAZIONE

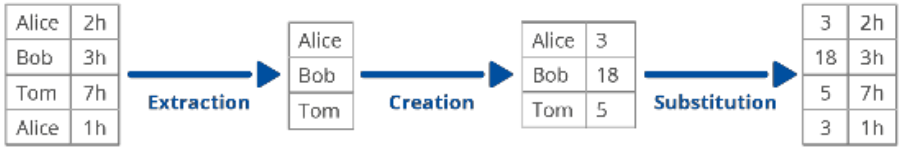
Così come risulta essenziale la scelta della tecnica di pseudonimizzazione da impiegare, pure il metodo (o modalità) di implementazione della pseudonimizzazione è altrettanto importante per la sua attuazione pratica.

Il presente paragrafo tratta il più ampio problema della pseudonimizzazione di un database o di un qualsiasi documento che contenga k identificativi. Si prenda in considerazione un identificativo Id che appare più volte in due set di dati A e B. In seguito alla pseudonimizzazione, l'identificativo Id viene sostituito attraverso uno dei seguenti metodi: pseudonimizzazione deterministica, pseudonimizzazione casuale di documento e pseudonimizzazione totalmente casuale.

5.2.1 La pseudonimizzazione deterministica

Con tale metodo, ogni qual volta si presenti, l'identificativo Id viene sostituito, in tutti i database, dallo stesso pseudonimo $pseudo$. In questo modo si ha la totale corrispondenza sia all'intero di un database che tra database diversi. Il primo passo per attuare tale metodo è quello di estrarre l'elenco degli identificativi univoci contenuti nel database. All'esito, l'elenco viene collegato agli pseudonimi e, da ultimo, gli identificativi vengono sostituiti agli pseudonimi all'interno del database (Figura 8).

Figura 8 - Pseudonimizzazione deterministica



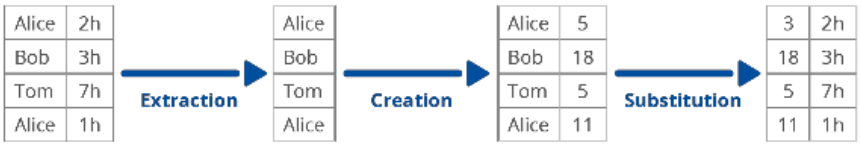
Tutte le tecniche già menzionate nel capitolo 5.1 possono essere utilizzate direttamente ai fini di realizzare la pseudonimizzazione deterministica.

5.2.2 La pseudonimizzazione casuale di documenti

Ogni qual volta l'identificativo id appaia nel database, viene sostituito da pseudonimi differenti ($pseudo_1$, $pseudo_2$…).

In ogni caso, l'identificativo id viene sempre associato, all'interno del nel set di dati A e B, alla stessa raccolta di pseudonimi ($pseudo_1$, $pseudo_2$).

Figura 9 - Pseudonimizzazione randomizzata di documenti



In tale ipotesi, la pseudonimizzazione risulta coerente solo tra database diversi. In questo caso, infatti, la tabella di mappatura viene creata utilizzando tutti gli identificativi contenuti all'interno del database. Per cui, la ripetizione di un determinato identificativo (come Alice, nell'esempio riportato nella Figura 9) viene trattata in maniera autonoma.

5.2.3 La pseudonimizzazione totalmente casuale

Infine, ogni qualvolta un identificativo Id appaia all'interno di un database A o B, l'Id viene sostituito da uno pseudonimo diverso ($pseudo_1$, $pseudo_2$). Trattasi, in tale ipotesi, di una pseudonimizzazione totalmente casuale. Tale metodo può essere visto come un ulteriore sviluppo della pseudonimizzazione casuale di documenti. In effetti, i due metodi, quando vengono applicati ad un singolo documento, agiscono alla stessa maniera. La differenza risiede nella circostanza per cui, se lo stesso documento viene pseudonimizzato due volte attraverso la pseudonimizzazione totalmente casuale, si ottengono due risultati diversi. Di contro, attraverso la pseudonimizzazione casuale di documenti, si ottiene due volte il medesimo risultato. In altre parole, nella pseudonimizzazione casuale di documento la casualità è selettiva (nell'esempio precedente, solo per l'identificativo Alice), mentre nella pseudonimizzazione totalmente casuale la casualità è generalizzata (si applica, cioè, ad ogni dato).

5.3 COME SCEGLIERE UNA TECNICA E UN METODO DI PSEUDONIMIZZAZIONE

La scelta di una tecnica e di un metodo di pseudonimizzazione dipende da diversi fattori, principalmente il livello di protezione dei dati e la funzionalità dell'insieme di dati pseudonimizzato che l'ente di pseudonimizzazione desidera raggiungere. In termini di protezione, come discusso nelle sezioni precedenti, l'RNG, il Message Authentication Code e la crittografia sono tecniche più efficaci per il contrasto degli attacchi operati attraverso le ricerche esaustive, le ricerche nel dizionario e le congetture. Tuttavia, in base al requisito di funzionalità, l'ente di pseudonimizzazione potrebbe optare per una combinazione di diverse tecniche o variazioni di una data metodologia. Analogamente, per quanto riguarda i metodi di pseudonimizzazione, la pseudonimizzazione totalmente casuale offre il miglior livello di protezione, ma impedisce qualsiasi confronto tra database. Le funzioni casuali e quelle deterministiche garantiscono la funzionalità, ma consentono la collegabilità dei dati. Pertanto, in base agli identificativi che devono essere pseudonimizzati, possono essere impiegate specifiche soluzioni (per l'approfondimento si rimanda ai Capitoli 6 e 7).

Inoltre, l'ente di pseudonimizzazione potrebbe conservare qualche riserva in ordine alla complessità di un determinato schema in termini di implementazione e scalabilità, confrontandosi con i seguenti quesiti: è semplice applicare la pseudonimizzazione agli identificativi? La pseudonimizzazione influisce sulla dimensione del database?

Tabella 3 - Confronto di diverse tecniche in termini di flessibilità (formato identificativo e dimensioni dello pseudonimo)

Metodologia	Identificativo di bits	Pseudonimo calcolato in m
Contatore	Qualsiasi	$m=log2\mathrm{k}$
Generatore casuale di numeri	Qualsiasi	$m \gg 2log2k$
Funzione di Hash	Qualsiasi	Fisso o $m \gg 2log2k$
Message Authentication Code	Qualsiasi	Fisso o $m \gg 2log2\mathrm{k}$
Crittografia	fisso[20]	Fisso o uguale all'identificativo

La maggior parte di dette soluzioni può essere applicata su identificativi di dimensioni variabili, ad eccezione di determinate scelte impiegate nel caso di utilizzo della crittografia. La dimensione dello pseudonimo dipende da k, dal numero degli identificativi contenuti nel database. Per le tecniche del generatore casuale di numeri, della funzione di Hash e del MAC, esiste una probabilità di collisione: la dimensione dello pseudonimo deve essere scelta con cura (vedi il paradosso del compleanno). La funzioni di Hash e il MAC sono opportunamente progettati in modo da garantire che la dimensione del digest prevenga qualsiasi rischio di collisione. Infine, la dimensione degli pseudonimi prodotti da uno schema di crittografia può essere fissa o uguale alla dimensione dell'identificativo originario. La Tabella 3 rappresenta la scalabilità degli approcci summenzionati per quanto riguarda la funzione di ripristino.

5.4 IL RIPRISTINO

Poiché, per definizione, l'uso di informazioni aggiuntive è fondamentale ai fini della pseudonimizzazione, l'ente di pseudonimizzazione deve attuare un meccanismo di ripristino. Tale meccanismo può essere più o meno complesso a seconda della funzione di pseudonimizzazione impiegata. In generale, consiste nell'uso di uno pseudonimo $pseudo$ e di un segreto di pseudonimizzazione S ai fini di recuperare il corrispondente identificatore Id. Ciò può verificarsi, ad esempio, quando l'ente di pseudonimizzazione rileva un'anomalia nel proprio sistema e deve contattare l'ente preposto. L'"anomalia" può consistere, ad esempio, in una violazione di dati per la quale l'ente di pseudonimizzazione, ai sensi del GDPR, deve informare gli interessati. Peraltro, il meccanismo di ripristino potrebbe rivelarsi necessario ai fini di consentire agli interessati l'esercizio dei propri diritti ai sensi degli articoli 12-21 GDPR.

[20] La crittografia mediante un codice a blocchi funziona con input di dimensioni fisse. Tuttavia, alcune modalità operative (come il CTR) possono consentire di lavorare su input di qualsiasi dimensione.

Tabella 4 - Confronto di diverse tecniche in ordine ai meccanismi di ripristino

Metodo	Recupero basato sullo pseudonimo
Contatore	Mappatura
Generatore Casuale di	Mappatura
Funzione di Hash	Mappatura
Message Authentica	Mappatura
Crittografia	Decifratura

La maggior parte dei metodi precedentemente descritti richiede che l'ente di pseudonimizzazione conservi, ai fini di eseguire il ripristino dell'identificatore e ad eccezione dei casi di impiego della crittografia, la tabella di mappatura tra identificatori e pseudonimi (Tabella 4). Infatti, la crittografia può essere applicata direttamente sull'identificatore.

5.5 PROTEZIONE DEL SEGRETO DI PSEUDONIMIZZAZIONE

Affinché la pseudonimizzazione sia efficace, l'ente di pseudonimizzazione deve sempre proteggere il segreto di pseudonimizzazione mediante adeguate misure tecniche e organizzative. Ciò dipende chiaramente dalle specifiche ipotesi di pseudonimizzazione (di cui al Capitolo 3).

In primo luogo, il segreto di pseudonimizzazione deve essere separato dal set di dati, di modo che il segreto di pseudonimizzazione e il set di dati non siano mai trattati all'interno dello stesso file (diversamente, per un intruso sarebbe assai agevole recuperare gli identificativi). In secondo luogo, il segreto di pseudonimizzazione deve essere eliminato in modo sicuro da ogni supporto (di memoria e di sistema) non sicuro. In terzo luogo, rigorose procedure di controllo d'accesso devono garantire che solo i soggetti autorizzatati abbiano accesso a tale segreto. Un sistema di registrazione sicuro deve tenere traccia di tutte le richieste di accesso eseguite. Infine, il segreto di pseudonimizzazione, qualora venga archiviato su un computer, deve essere crittografato, con conseguente necessità di archiviare e gestire correttamente le chiavi di crittografia.

5.6 TECNICHE AVANZATE DI PSEUDONIMIZZAZIONE

Oltre alle tecniche di pseudonimizzazione di cui sopra, esistono una serie di altre tecniche di pseudonimizzazione più avanzate, adatte a molteplici contesti differenti. Di seguito, per permettere ai lettori un'agevole lettura, si riportano brevemente alcune di queste tecniche, poiché soffermarsi sui dettagli di ognuna andrebbe oltre lo scopo di questo lavoro.

Oltre al comune hashing di dati, anche strutture più avanzate come gli alberi di Merkle [17, 18] utilizzano funzioni di Hash o set di funzioni di Hash, come ad esempio h_3 = Hash (h_1, h_2), per ottenere pseudonimi complessi che possono essere rivelati, anziché completamente, solo parzialmente. Allo stesso modo, le catene di Hash [19] utilizzano la ripetuta applicazione di una funzione crittografica di Hash a un dato valore, per cui ad esempio $h_4 = h_3 (h_2 (h_1 (x)))$, in modo da generare un valore che richieda diverse inversioni di Hash ai fini di re-identificare il dato originario di un determinato pseudonimo. Un esempio di tale tecnica di Hashing è dato da una catena di pseudonimizzazione che coinvolge diversi enti di pseudonimizzazione che, di seguito, utilizzano gli pseudonimi creati dal precedente ente di pseudonimizzazione come input per creare nuovi pseudonimi (applicando, ad esempio, un ulteriore livello di Hashing). Tale metodo a catena sarà efficace anche laddove un intruso riuscisse a scoprire tutte le pseudonimizzazioni dell'intera catena ad eccezione di una, così rendendola una tecnica di pseudonimizzazione assai salda. Tanto da essere, ad esempio, comunemente utilizzata negli studi clinici.

Se il dominio di input copre più dimensioni (si rimanda al capitolo 8 per gli esempi), i filtri bloom [20], oltre ad essere utilizzati come tecnica di anonimizzazione, possono essere utilizzati per eseguire in modo efficiente una pseudonimizzazione computazionalmente fattibile su tutte le possibili combinazioni di valori di input su domini diversi, nonostante il c.d. "state explosion problem".

Un altro valido approccio può essere costituito anche dall'utilizzo di pseudonimi di transazioni collegabili e/o dalla collegabilità degli pseudonimi controllati mediante l'opzione di re-identificazione [21].

Infine, tutte le tecniche che possono essere utilizzate con successo per incrementare l'anonimizzazione, si rivelano utili anche ai fini della pseudonimizzazione, così come pure le comuni tecniche del K-anonimato [3, 22, 23] o della privacy differenziale [24] e altri [25]. Si rimanda anche, per ulteriori approfondimenti sul tema, al punto [2] della bibliografia. Soluzioni interessanti possono essere fornire pure dalla c.d. dimostrazione a conoscenza zero [26] e dal più ampio ambito delle credenziali basate sulle proprietà [2].

6. PSEUDONIMIZZAZIONE DELL' INDIRIZZO IP

Partendo dalle tecniche e dalle informazioni sin qui trattate nel documento, in questo capitolo viene presentato un caso pratico specifico sulla pseudonimizzazione degli indirizzi IP.

Si prenda in considerazione il caso di un indirizzo IP che viene utilizzato per identificare in modo univoco un dispositivo su una rete IP. Vi sono due tipi di indirizzi IP: IPv4 [27] e IPv6 [28]. Ai fini del presente caso pratico, si prenderà in considerazione esclusivamente l'indirizzo IPv4, poiché, allo stato, è quello maggiormente in uso, mentre l'estendere i concetti trattati in precedenza anche all'indirizzo IPv6 sarebbe assai complesso ed esulerebbe dall'ambito del presente documento. Un indirizzo IPv4 è composto da 32 bit (mentre l'IPv6 è composto da 128 bit) divisi in un prefisso di rete (composto da un certo numero di bit significativi) e in un identificativo host (composto dal numero di bit rimanenti) attraverso l'ausilio di una maschera di sottorete. Questi vengono spesso rappresentati mediante l'utilizzo di un formato decimale puntato composto da 4 numeri decimali compresi tra 0-255 separati da punti, ad esempio 127.0.0.1. La dimensione del prefisso di rete e dell'identificativo host dipendono dalla dimensione del blocco CIDR (Classless Inter-Domain Routing [29]). Inoltre, vi sono alcuni indirizzi IP speciali, ad esempio 127.0.0.1 (localhost) o 224.0.0.1 (multicast). Tali ultimi indirizzi speciali, distinti in 15 classi, sono tutti meglio descritti nella nota bibliografica di cui al punto [30].

Attualmente, l'intero spazio degli indirizzi IP è gestito dall'Internet Assigned Numbers Authority (IANA), con l'aiuto di cinque registri Internet regionali (RIR). Gli stessi assegnano a organizzazioni locali come gli Internet Service Provider sottoinsiemi di indirizzi IP, che a loro volta assegnano gli indirizzi ai dispositivi degli utenti finali. Ogni assegnazione dell'indirizzo IP è documentata dal RIR corrispondente nel cosiddetto database WHOIS[21]. L'assegnazione può essere statica o dinamica (ad esempio utilizzando il protocollo Dynamic Host Configuration Protocol - DHCP).

Da un punto di vista giuridico, lo status degli indirizzi IP è stato discusso dalla Corte di Giustizia dell'Unione Europea nella causa C-582/14 Breyer contro Bundesrepublik Deutschland[22]. Gli indirizzi IP, sia statici che dinamici, sono considerati dati personali. Ciò è stato confermato anche dal parere 4/2007 sul concetto di dati personali espresso dal Gruppo di lavoro dell'articolo 29 per la tutela dei dati [31]. Pertanto, i database o le tracce di rete contenenti indirizzi IP devono essere protette e la pseudonimizzazione rappresenta una logica funzione di protezione in grado di consentire l'uso di indirizzi IP, impedendo al contempo la loro riferibilità a individui specifici. Fermo quanto detto, la scelta di un'appropriata tecnica di pseudonimizzazione per gli indirizzi IP consiste nel trovare un buon compromesso tra funzionalità e protezione dei dati. In effetti, il titolare del trattamento potrebbe ancora aver bisogno di calcolare statistiche o rilevare modelli all'interno del database pseudonimizzato (nei casi, ad esempio, dell'errata configurazione di un dispositivo o ai fini della valutazione della qualità dei servizi). Nella pratica, funzionalità e protezione dei dati non possono essere trattate in modo indipendente, tuttavia per una migliore comprensione verranno, di seguito, analizzate separatamente.

[21] Per ulteriori informazioni, si rimanda al seguente sito: https://whois.icann.org
[22] Maggiori dettagli sono disponibili al sito: https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=CELEX%3A62014CJ0582

6.1 PSEUDONIMIZZAZIONE E LIVELLO DI PROTEZIONE DEI DATI

La principale caratteristica del problema di pseudonimizzazione di un indirizzo IP è la dimensione dello spazio di input (dominio identificativo): ci sono infatti solo 2^{32} possibili indirizzi IP. Ciò rende possibile per un intruso, se la funzione di pseudonimizzazione non è stata scelta correttamente, compiere sia le ricerche esaustive che le ricerche nel dizionario, così portando a termine attacchi completi di re-identificazione o discriminazione.

Tenendo presenti le caratteristiche già descritte, in questo specifico caso pratico, le funzioni crittografiche di Hash risultano particolarmente vulnerabili. Si prenda in considerazione il caso di un indirizzo IP pseudonimizzato con la funzione di Hash SHA-256. Un intruso con uno pseudonimo/digest può avvalersi degli strumenti esistenti[23] per eseguire una ricerca esaustiva. La Tabella 5 mostra sia la durata di tale ricerca su un singolo comune laptop che lavora con un processore Intel (R) Core (TM) i7-8650U CPU @ 1.90GHz (8 core), sia le dimensioni del dizionario. Pure nella peggiore delle ipotesi, sono necessari appena 2 minuti circa per recuperare l'indirizzo IP appartenente a un determinato pseudonimo.

Tabella 5 - Costi pratici degli attacchi contro la pseudonimizzazione della funzione hash

Classe IP	Numero di IP possibili	Tempo di ricerca esaustiva	Dimensione dizionario
145.254.160.X	256	200ms	8KB
145.254.X.X	65536	200ms	2MB
145.X.X.X	16777216	2s	512MB
X.X.X.X	4294967296	2min16s	128GB

Inoltre, supponiamo che l'intruso voglia individuare se uno pseudonimo corrisponde a un indirizzo IP speciale [30]. Questo attacco di discriminazione non deve essere eseguito sui 2^{32} possibili indirizzi IP, ma solo sui 588,518,401 possibili indirizzi IP speciali.

Il comune caso di cui sopra dimostra come la pseudonimizzazione degli indirizzi IP attraverso le sole funzioni crittografiche di Hash non sia adeguata. Pertanto, per la protezione dei dati devono essere preferiti altri tipi di pseudonimizzazione, come il MAC (Code Authentication Message), la crittografia mediante una chiave segreta generata ad hoc o il generatore casuale di numeri (RNG). Come precedentemente discusso nel presente trattato, un intruso non può lanciare gli stessi attacchi, perché tali metodi utilizzano una chiave segreta (come nel caso del MAC e della crittografia) oppure una combinazione casuale (come nel caso del RNG). Pure il contatore può essere validamente impiegato, ma bisogna essere cauti rispetto ai possibili esiti (connessi, chiaramente, alla natura sequenziale del contatore).

[23] Ad esempio, un software di decifratura di password come "John The Ripper" o altri.

6.2 PSEUDONIMIZZAZIONE E LIVELLO DI FUNZIONALITA'

Come già accennato, la funzionalità rappresenta, nel caso degli indirizzi IP, un requisito essenziale per l'ente di pseudonimizzazione, ad esempio per il calcolo delle statistiche o della sicurezza della rete. Pertanto, il metodo utilizzato (indipendentemente dalla tecnica prescelta) dovrebbe consentire un'adeguata protezione, preservando al contempo alcune basilari informazioni utili (connessi agli indirizzi IP). Nella presente sezione, il problema viene in considerazione sotto due differenti profili: in primo luogo, la possibilità di ridurre al minimo il livello/ambito di pseudonimizzazione dell'indirizzo IP; in secondo luogo, la scelta del metodo di pseudonimizzazione da adottare (ovvero la modalità).

6.2.1 Livello Di Pseudonimizzazione

Nel precedente paragrafo, si è preso in considerazione il caso in cui la pseudonimizzazione venisse applicata sull'indirizzo IP completo (32 bit). Tuttavia, al fine di aumentarne la funzionalità, è possibile applicarla limitatamente ai bit meno significativi dell'indirizzo (ad esempio l'identificativo host), così preservando il prefisso di rete. Questa tecnica è chiamata pseudonimizzazione in grado di preservare i prefissi [32]. Consente l'identificazione dell'origine complessiva di un pacchetto (una rete) senza però sapere quale dispositivo all'interno della rete lo abbia effettivamente inviato. È necessario capire quanti dispositivi sussistono per un determinato prefisso. La Tabella 5 mostra diverse dimensioni di un prefisso. Questa tecnica è già ampliamente utilizzata da diversi fornitori di servizi per pseudonimizzare gli indirizzi IP (si rimanda all'esempio riportato al punto [33] della bibliografia).

6.2.2 Scelta della modalità di pseudonimizzazione

La scelta della modalità di pseudonimizzazione ha un forte impatto sulla funzionalità e sul livello di protezione dei dati, indipendentemente dalla scelta della tecnica di pseudonimizzazione.

In questo paragrafo, tale relazione viene ulteriormente approfondita attraverso un esempio specifico.

Tabella 6 - Origine e destinazione di una richiesta HTTP

	Fonte	Destinazione
Pacchetto 1	145.254.160.237	65.208.228.223
Pacchetto 2	65.208.228.223	145.254.160.237
Pacchetto 3	145.254.160.237	65.208.228.223
Pacchetto 4	145.254.160.237	65.208.228.223
Pacchetto 5	65.208.228.223	145.254.160.237

Nell'esempio sopra citato, applichiamo la pseudonimizzazione deterministica usando un RNG. Ogni indirizzo IP viene, così, associato a uno pseudonimo univoco. In questo caso, la tabella di mappatura così ottenuta

è riportata nella Tabella 7. Mentre, una volta applicata la pseudonimizzazione deterministica, si ottiene la Tabella 8.

Tabella 7 - Tabella di mappatura della pseudonimizzazione deterministica

Indirizzo IP	Pseudonimo
145.254.160.237	238
65.208.228.223	47

Tabella 8 - Indirizzi di origine e di destinazione trasformati attraverso la pseudonimizzazione deterministica

Numero del pacchetto	Fonte	Destinazione
Pacchetto 1	238	47
Pacchetto 2	47	238
Pacchetto 3	238	47
Pacchetto 4	238	47
Pacchetto 5	47	238

Mettendo a confronto le informazioni ottenute dalla traccia originale del network (di cui alla Tabella 6) con quelle della Tabella 8, si può notare, da entrambe le tracce (originale e pseudonimizzata), come sia possibile dedurre il numero totale di indirizzi IP coinvolti e quanti pacchetti siano stati inviati da ciascun indirizzo durante la comunicazione.

Ne consegue che, nella pseudonimizzazione degli indirizzi IP di cui alla Tabella 8, è possibile ottenere, sugli stessi, lo stesso livello di analisi statistica (e, quindi, di funzionalità).

Si consideri, ora, il caso della pseudonimizzazione casuale di documenti mediante l'RNG. Ogni volta che si incontra il medesimo indirizzo IP, allo stesso viene associato uno pseudonimo diverso. Ad esempio, l'indirizzo IP 145.254.160.237 è associato a 5 pseudonimi, ovvero 39, 71, 48, 136 e 120 (Tabella 9). Una volta concluso il procedimento di pseudonimizzazione casuale del documento, si ottiene il risultato di cui alla Tabella 10.

Tabella 9 - Tabella di mappatura della pseudonimizzazione casuale di documenti

Indirizzo IP	Pseudonimo
145.254.160.237	39,71,48,136,120
65.208.228.223	23,30,60,160,231

Tabella 10 - Indirizzi di origine e destinazione ottenuti utilizzando il documento di pseudonimizzazione casuale

Numero del pacchetto	Fonte	Destinazione
Pacchetto 1	39	23
Pacchetto 2	30	71
Pacchetto 3	48	60
Pacchetto 4	136	160
Pacchetto 5	231	120

Come mostrato nella Tabella 10, a differenza della Tabella 6 e della Tabella 8 in cui si ottengono 2 indirizzi IP, nel caso della presente Tabella si ottengono sostanzialmente 10 indirizzi IP. Pertanto, il livello di funzionalità è stato ridotto (seppur il livello di protezione risulta maggiore). Ovviamente, l'applicazione della pseudonimizzazione totalmente casuale ha un impatto ancora maggiore sulla funzionalità. A tal proposito, la Tabella 11 mette a confronto le diverse modalità di pseudonimizzazione dell'indirizzo IP.

Tabella 11 - Modalità di pseudonimizzazione e utilità

	Modalità di pseudonimizzazione		
Utilità	Deterministica	Documento casuale	Totalmente casuale
Statistiche (conti ...)	SI	NO	NO
Protocolli semantici	SI	NO	NO
Confronto tra le diverse tracce	SI	YES	NO

Chiaramente, non esiste una soluzione univoca a questo problema e la scelta finale spetterà sempre alle esigenze di funzionalità e protezione che l'ente di pseudonimizzazione vuole ottenere.

7. PSEUDONIMIZZAZIONE DELL' INDIRIZZO E-MAIL

Nel presente capitolo, si prende in considerazione la pseudonimizzazione degli indirizzi e-mail quale più specifico caso pratico delle tecniche già precedentemente illustrate nel documento.

Un indirizzo di posta elettronica (e-mail) costituisce un identificativo tipico di un individuo. Un indirizzo e-mail ha il formato local@dominio, dove la parte locale corrisponde all'utente che possiede l'indirizzo e il dominio corrisponde al fornitore del servizio di posta. Gli indirizzi e-mail sono generalmente utilizzati in diversi modi; ad esempio, possono rappresentare l'identificativo principale di un soggetto che si registra a un servizio elettronico. Inoltre, gli indirizzi di posta elettronica sono in genere presenti in molti database, in cui possono comparire pure altri identificativi, come i nomi delle persone.

Gli utenti tendono a utilizzare lo stesso indirizzo e-mail per diversi impieghi, condividendolo con varie organizzazioni, come, ad esempio, quando effettuano una registrazione per gli account online. Inoltre, gli indirizzi e-mail sono spesso pubblicati in rete, oltreché è stato dimostrato che possono essere facilmente reperiti o indovinati[24]. Proprio in ragione di tali circostanze, laddove gli indirizzi e-mail vengano utilizzati come identificativi, la loro protezione è particolarmente importante.

Nel presente caso pratico, nell'analizzate le loro diverse modalità di pseudonimizzazione, gli indirizzi e-mail vengono in considerazione quali identificativi (ad esempio all'interno di un database o in un servizio online). Si prenda in considerazione il caso in cui il processo di pseudonimizzazione venga sempre eseguito da un ente di pseudonimizzazione (quale il titolare del trattamento dei dati) nell'ambito dell'operazione / fornitura di un servizio.

7.1 IL CONTATORE E IL GENERATORE CASUALE DI NUMERI

Tenendo presenti le descrizioni di cui al Capitolo 5, sia il contatore che l'RNG possono essere impiegati per la pseudonimizzazione delle e-mail mediante l'utilizzo di una tabella di mappatura, come quella mostrata nell'esempio di cui alla Tabella 12. Chiaramente, la pseudonimizzazione è efficace fintanto che la tabella di mappatura rimane protetta e archiviata separatamente dai dati pseudonimizzati.

Tabella 12 - Esempio di pseudonimizzazione dell'indirizzo email attraverso l'RNG o il contatore (pseudonimizzazione completa)

Indirizzo Email	Pseudonimo (RNG)	Pseudonimo (contatore)
alice@abc.eu	328	10
bob@wxyz.com	105	11
eve@abc.eu	209	12
john@qed.edu	83	13
alice@wxyz.com	512	14
mary@clm.eu	289	15

Come mostrato nell'esempio della Tabella 12, sia il contatore che l'RNG generano pseudonimi che non rivelano alcuna informazione sugli identificativi iniziali (gli indirizzi e-mail) e non consentono ulteriori analisi sugli pseudonimi (quali l'analisi statistica).

[24] In effetti, è stato dimostrato che anche il recupero di una semplice informazione di base, quale i nomi degli utenti di un social network, consentono di raccogliere in modo efficace milioni di indirizzi e-mail [38].

Per aumentare la funzionalità, è possibile applicare la pseudonimizzazione solo a una parte dell'indirizzo e-mail, ad esempio la parte locale (lasciando inalterata la parte del dominio, come mostrato nella Tabella 13).

Tabella 13 - Esempio di pseudonimizzazione dell'indirizzo e-mail attraverso l'RNG o il contatore (pseudonimizzazione della sola parte locale)

Indirizzo Email	Pseudonimo (RNG)	Pseudonimo (contatore)
alice@abc.eu	328@abc.eu	10@abc.eu
bob@wxyz.com	105@wxyz.com	11@wxyz.com
eve@abc.eu	209@abc.eu	12@abc.eu
john@qed.edu	83@qed.edu	13@qed.edu
alice@wxyz.com	512@wxyz.com	14@wxyz.com
mary@clm.eu	289@clm.eu	15@clm.eu

Nell'esempio di cui alla Tabella 13, per quanto le e-mail siano pseudonimizzate, è ancora possibile conoscere il dominio e, di conseguenza, condurre le analisi pertinenti (ad esempio sul numero di utenti che utilizzano lo stesso dominio e-mail). Come precedentemente discusso nel documento, il contatore può essere considerato, in termini di protezione, maggiormente inefficace poiché, a causa della sua natura sequenziale, consente di effettuare previsioni (ad esempio, prendendo in considerazione i casi in cui gli indirizzi e-mail provengano dallo stesso dominio, l'uso del contatore può rivelare informazioni riguardanti la sequenza dei diversi utenti di posta elettronica all'interno del database).

Partendo da questo semplice esempio, a seconda del livello di protezione dei dati e funzionalità che l'ente di pseudonimizzazione vuole raggiungere, possono essere applicate diverse variazioni, mantenendo negli pseudonimi diversi livelli di informazione (ad esempio, sui domini identici, parti locali, ecc.).

Tabella 14 - Esempi di pseudonimizzazione dell'indirizzo e-mail attraverso l'RNG - vari livelli di funzionalità

Indirizzo Email	Pseudonimo (RNG) che conserva le informazioni su domini identici	Pseudonimo (RNG) che conserva anche le informazioni su paese / estensione identici	Pseudonimo (RNG) che conserva le informazioni su parti e domini locali identici	Pseudonimo (RNG) che conserva le informazioni su paese / estensione, domini e parti locali identici
alice@abc.eu	328@1051	328@1051.3	328@1051	328@1051.3
bob@wxyz.com	105@833	105@833.7	105@833	105@833.7
eve@abc.eu	209@1051	209@1051.3	209@1051	209@1051.3
john@qed.edu	83@420	83@420.8	83@420	83@420.8
alice@wxyz.com	512@833	512@833.7	328@833	328@833.7
mary@clm.eu	289@2105	289@2105.3	289@2105	289@2105.3

Le principali insidie sia del contatore che dell'RNG risiedono, nelle ipotesi di set di dati di grandi dimensioni, nella scalabilità, soprattutto nei casi in cui è necessario che lo stesso pseudonimo venga sempre assegnato allo stesso indirizzo (cioè nelle ipotesi di pseudonimizzazione deterministica, di cui alla Tabella 12). In effetti, in tal caso, l'ente di pseudonimizzazione è costretto a eseguire un controllo incrociato nell'intera tabella di pseudonimizzazione ogni qual volta venga inserita una nuova voce da pseudonimizzare. La complessità

aumenta in casi di implementazioni più sofisticate, come quelli mostrati nella Tabella 14 (ad esempio, nei casi in cui l'ente di pseudonimizzazione debba compiere una classificazione degli indirizzi e-mail con il medesimo dominio o il medesimo paese senza, però, rivelare quale sia il dominio / paese).

7.2 LA FUNZIONE CRITTOGRAFICA DI HASH

Come affermato nella nota bibliografica di cui al punto [34], si stima che il numero totale di account e-mail di tutto il mondo sia approssimativamente di 4,7 miliardi - 2^{32} (poiché, nonostante le dimensioni potenzialmente infinite dello spazio dedicato agli indirizzi e-mail validi, gli indirizzi realmente attivi occupano uno spazio assai più ristretto). Questo fatto, come già precedentemente accennato nel presente capitolo, rende gli indirizzi e-mail facilmente reperibili o indovinabili[25], rendendo così l'utilizzo delle funzioni crittografiche di Hash una tecnica inefficace ai fini della pseudonimizzazione [34]. In effetti, per un qualsiasi intruso, sia interno che esterno, che abbia accesso all'elenco di indirizzi e-mail pseudonimizzati, risulterà agevole eseguire un attacco di ricerca nel dizionario (Figura 10). Questa osservazione risulta valida anche per tutte le ipotesi di pseudonimizzazione di cui al Capitolo 3 (indipendentemente dal fatto che l'ente di pseudonimizzazione sia il titolare del trattamento, il responsabile oppure una terza parte autorizzata).

Figura 10 - Procedimento di inversione della funzione di Hash di un indirizzo e-mail

Nonostante le funzioni crittografiche di Hash presentito le criticità appena descritte, è opportuno comunque sottolineare che, come indicato nella nota bibliografica di cui al punto [35], i fornitori di servizi spesso condividono gli indirizzi e-mail con terze parti, semplicemente eseguendo l'hashing. Un esempio concreto di tale ipotesi è dato dal funzionamento dei c.d. elenchi pubblici dei clienti, che offre alle aziende la possibilità di confrontare i valori di hash degli indirizzi e-mail dei clienti per stabilire e creare elenchi di clienti comuni tra le aziende[26] .

Nonostante i rischi significativi in ordine alla protezione dei dati, di cui si è detto, i valori crittografici di Hash, a date condizioni, potrebbero comunque risultare funzionali, ad esempio nei casi di codifica interna degli indirizzi e-mail (come nel contesto delle attività di ricerca) o quale meccanismo di convalida / integrità per il titolare del trattamento dei dati (si rimanda alla nota bibliografica di cui al punto [1]). Le funzioni Hash potrebbero anche essere utilizzate per pseudonimizzare singole parti di un indirizzo e-mail (quali la sola parte del dominio), consentendo in tal modo una maggiore funzionalità ai relativi pseudonimi; inoltre, se la

[25] Teoricamente, se un intruso avesse a disposizione tutti gli indirizzi potenzialmente esistenti, anche un attacco di forza bruta sarebbe concretamente fattibile; in ogni caso, tuttavia, lo spazio (relativamente) ridotto dedicato agli indirizzi e-mail indica che una ricerca casuale sugli indirizzi e-mail potrebbe realmente avere successo. Ancora peggio, nell'era dei grandi dati, le ricerche casuali potrebbero non essere nemmeno necessarie poiché gli indirizzi di posta elettronica validi sono spesso disponibili pubblicamente o possono essere facilmente desumibili in contesti specifici (ad esempio nei casi in cui il dominio e il formato di una specifica organizzazione siano noti).
[26] Si rimanda al sito: https://www.facebook.com/business/help/112061095610075?helpref=faq_content

restante parte dell'identificativo viene pseudonimizzata attraverso un metodo più efficace (quale il MAC), allora il rischio di risalire all'intero indirizzo e-mail originario è notevolmente ridotto.

7.3 IL MESSAGE AUTHENTICATION CODE

Rispetto al semplice Hashing, il Message Authentication Code (MAC) offre, anche nell'ambito della pseudonimizzazione dell'indirizzo e-mail, notevoli vantaggi in termini di protezione dei dati; purché la chiave segreta venga archiviata in modo sicuro. Inoltre, l'ente di pseudonimizzazione può utilizzare, per settori diversi, differenti chiavi segrete, così generando, ad esempio, diversi pseudonimi di settore per lo stesso indirizzo di posta elettronica. Il MAC può anche essere utilizzato, nei casi in cui l'accesso agli pseudonimi risulti sufficiente al particolare scopo di elaborazione, ai fini di impedire al titolare del trattamento di accedere agli indirizzi e-mail (ad esempio, nelle ipotesi 5 e 6 di cui al Capitolo 3). Tale caso, potrebbe essere rappresentato, ad esempio, dalla visualizzazione della pubblicità mirata, per la quale gli inserzionisti devono associare un unico pseudonimo ad ogni individuo, pur tuttavia senza essere in grado di scoprire l'identità originaria dell'utente [36].

Come già visto per le tecniche precedenti, al fine di aumentare la funzionalità degli pseudonimi, nella pratica, potrebbero essere adoperate diverse ipotesi di attuazione. Ad esempio, un possibile approccio potrebbe essere quello di applicare il MAC separatamente a diverse parti dell'indirizzo e-mail (quali, ad esempio, le parti locali e il dominio), utilizzando la stessa chiave segreta. Un esempio tipico è mostrato nella Figura 11: l'utilizzo della stessa chiave per ogni MAC comporta, ogni qual volta i domini dell'indirizzo e-mail siano identici, la generazione dei medesimi sotto-pseudonimi relativi le parti di dominio corrispondenti (evidenziati in colore verde). Tuttavia, poiché l'output del MAC ha una dimensione fissa, generalmente molto più grande della dimensione dell'indirizzo e-mail originario[27], gli pseudonimi ottenuti possono essere di dimensioni assai notevoli (ulteriormente aumentate se le parti differenti vengono pseudonimizzate separatamente).

Figura 11 - Utilizzo del MAC per generare indirizzi e-mail pseudonimizzati con qualche funzionalità

Per quanto riguarda l'applicazione pratica del MAC, un aspetto importante è dato dal ripristino. Va sottolineato che neppure l'ente di pseudonimizzazione è in grado di invertire direttamente gli pseudonimi, nonostante detenga la chiave segreta; tale inversione può essere ottenuta solo indirettamente, riproducendo gli pseudonimi di ciascun indirizzo di posta elettronica noto, al fine di confrontare le corrispondenze con l'elenco pseudonimizzato. Chiaramente, se è disponibile una tabella di mappatura di pseudonimizzazione, l'inversione degli pseudonimi risulta banale, ma, in tale ipotesi, sono richiesti dei metodi di archiviazione più efficaci. Per tali ragioni, il MAC non è probabilmente la tecnica di pseudonimizzazione più pratica per tutti

[27] La dimensione tipica dell'output di una funzione di Hash (con o senza chiave) è di 256 bit, ovvero 32 caratteri

quei casi in cui il titolare del trattamento dei dati abbia l'esigenza di ricollegare agevolmente gli pseudonimi agli indirizzi e-mail (come, ad esempio, in alcuni casi pratici di cui al Capitolo 3.1 e 3.2).

7.4 LA CRITTOGRAFIA

Un'alternativa al MAC è la crittografia, applicata soprattutto in modo deterministico, ovvero utilizzando una chiave segreta ai fini di generare uno pseudonimo per ogni indirizzo di posta elettronica (crittografia simmetrica). In tale ipotesi, l'applicazione è più agevole, poiché non è necessario prevedere una tabella di mappatura della pseudonimizzazione: il ripristino avviene direttamente mediante il processo di decrittazione [37].

Si noti che, sebbene alcuni algoritmi crittografici asimmetrici (come la chiave pubblica) possano essere utilizzati in modo deterministico[28], l'utilizzo degli stessi non è raccomandabile nell'ambito della pseudonimizzazione di indirizzi e-mail (o di altri tipi di dati, come approfondito nella nota bibliografica di cui al punto [1]). Supponiamo, ad esempio, che l'ente di pseudonimizzazione debba generare, per ciascun indirizzo di posta elettronica, pseudonimi diversi per utenti/destinatari diversi (sia interni che esterni), con il presupposto che ciascun destinatario potrà identificare nuovamente i propri dati, ma non i dati pseudonimizzati degli altri destinatari.

Per raggiungere tale obiettivo si potrebbe, in primo luogo, crittografare le e-mail con la chiave pubblica di ciascun destinatario, consentendo quindi solo al destinatario specifico di eseguire la decodifica. Tuttavia, supponendo che le chiavi pubbliche siano in linea di principio disponibili a chiunque, qualsiasi intruso potrà assestare un attacco di ricerca sul dizionario basato su indirizzi di posta elettronica noti (o indovinati) (come quello mostrato in Figura 10, in cui viene utilizzata la crittografia mediante una chiave pubblica con una chiave pubblica nota, anziché una funzione Hash).

La natura della crittografia, per sua impostazione predefinita, non tiene conto della funzionalità dei dati pseudonimizzati. La crittografia separata delle parti di un indirizzo e-mail può essere sufficiente per sopperire a tale mancanza, analogamente all'utilizzo del Message Authentication (di cui alla Figura 11), in cui il MAC può essere sostituito da un algoritmo di crittografia. In generale, per consentire agli pseudonimi di trasportare alcune informazioni utili, è possibile utilizzare specifiche tecniche crittografiche; di seguito viene fornito un esempio illustrativo insieme alla crittografia con la conservazione del formato.

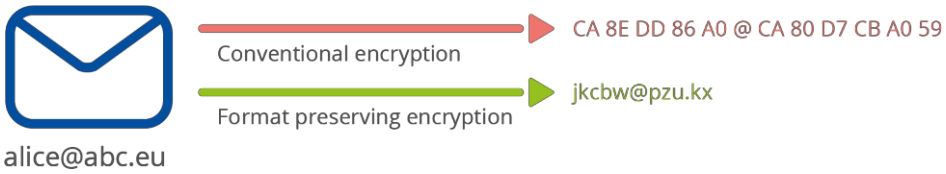
LA CRITTOGRAFIA CON LA CONSERVAZIONE DEL FORMATO (FPE)

La struttura di un database potrebbe prevedere, per campi specifici, un particolare tipo di dato. Ad esempio, un indirizzo e-mail dovrebbe contenere una parte locale (informazioni), seguita dal simbolo @, a sua volta seguito da un dominio. Se per il titolare del trattamento dei dati non vi è l'obbligo di conservare gli indirizzi di posta elettronica originari, ma è comunque necessario tenere un elenco pseudonimizzato pur preservando la struttura del database, a tale scopo la crittografia con la conservazione del formato è un valido candidato. Di tale crittografia (FPE) esistono diverse note applicazioni, basate su altrettanto note

[28] Nonostante ciò, per ragioni di sicurezza, l'argoritmo di una chiave pubblica deve essere, già in principio, probabilistica [1].

strutture crittografiche[29]. In ogni caso, qualsiasi sostituzione (pseudo) casuale di caratteri[30] con altri caratteri dello stesso alfabeto - ovvero l'insieme di caratteri alfanumerici arricchiti da caratteri speciali che compaiono in parti locali di indirizzi e-mail – è sufficiente a garantire che lo pseudonimo derivato abbia la forma richiesta. La differenza tra FPE e la crittografia convenzionale è illustrata nella Figura 12.

Figura 12 - Confronto tra crittografia convenzionale e crittografia che conserva il formato ai fini di ottenere lo pseudonimo dall'indirizzo e-mail



Nell'esempio di cui alla Figura 12, nel caso della crittografia convenzionale viene utilizzato un codice di flusso simmetrico, in modo da garantire che lo pseudonimo ottenuto abbia la stessa lunghezza dell'indirizzo originario (i caratteri dello pseudonimo così ottenuto sono non alfanumerici e, quindi, vengono indicati in forma esadecimale).

Si sottolinea che, a seconda del caso, potrebbe essere necessario progettare opportune implementazioni del FPE, al fine di evitare l'insorgere di configurazioni in grado di divulgare informazioni sull'identità degli individui.

[29] Si veda, ad esempio https://csrc.nist.gov/publications/detail/sp/800-38g/rev-1/draft, la stesura corrente del NIST sulla crittografia che conserva il formato in grado di risolvere potenziali vulnerabilità quando la dimensione del dominio è troppo piccola.

[30] La sostituzione dei caratteri è un caso particolare di crittografia (sebbene possa presentari problemi di sicurezza nei casi in cui non venga correttamente applicata).

8. LA PSEUDONIMIZZAZIONE IN PRATICA: UN'IPOTESI PIU' COMPLESSA

Come si può vedere dai due casi illustrati precedentemente nei Capitoli 6 e 7, la pseudonimizzazione anche dei dati più semplici, come gli indirizzi IP o gli indirizzi e-mail, è un compito impegnativo e soggetto a errori. Tuttavia, quando si tratta di sistemi che utilizzano parole reali, spesso la maggior parte dei problemi non sono causati dalla scelta della tecnica di pseudonimizzazione utilizzata per uno o due identificativi specifici; quanto piuttosto dal collegamento implicito tra un insieme di pseudonimi e altri valori di dati inseriti all'interno di una struttura di dati più complessa. L'esempio più comune è quello di un servizio online che, al momento della registrazione, crea dei profili utente e, ogni volta che sono disponibili nuovi dati, li arricchisce con le informazioni personali dell'utente. In questo caso, anche se, come visto in precedenza, l'indirizzo e-mail dell'utente e tutti gli indirizzi IP che si trovano nei log di accesso dell'utente sono rigorosamente pseudonimizzati, vi è pur sempre la grande minaccia di una possibile re-identificazione o discriminazione, persino sulla stessa struttura dei dati pseudonimizzati. Pertanto, in questa sezione vengono discussi i casi più complessi di pseudonimizzazione dei dati.

8.1 UN ESEMPIO ILLUSTRATIVO

Si consideri, ad esempio, un'ipotesi assai simile ai casi che si verificano normalmente nella pratica: un social network online. L'operatore immaginario, Social Network Inc. (di seguito denominato SN), agisce come titolare del trattamento e permette ai suoi utenti (che si ipotizza essere solo individui umani) di registrarsi per ottenere un account che viene memorizzato nel centro dati di SN. Con tale account, gli utenti possono utilizzare una serie di funzioni che consentono, ad esempio, di mettersi in contatto con altri utenti, organizzazioni o argomenti di interesse. Al momento della registrazione, gli utenti di SN devono fornire il loro vero nome e cognome, un nickname, la data di nascita e il genere di appartenenza, oltreché una serie di informazioni personali facoltative (ubicazione, interessi, biometria, ecc.), e un indirizzo e-mail valido. Ogni volta che gli utenti accedono a uno qualsiasi dei servizi di SN, la loro interazione viene registrata e aggiunta al loro profilo utente – compresi la registrazione temporale (il cosiddetto timestamp) e l'indirizzo IP da cui avviene l'accesso.

Al fine di migliorare la conformità al GDPR, la Direzione della SN ha deciso di pseudonimizzare gli indirizzi IP dei log di accesso utilizzando le tecniche di cui al capitolo 6. Tuttavia, le restanti informazioni vengono conservate in chiaro, in modo tale che – ove necessario – l'utente possa visualizzarle nei siti web della SN, oppure possano essere utilizzate per effettuare controlli e verifiche (ad esempio, la data di nascita è necessaria per calcolare l'età e verificare che l'utente abbia più di 16 anni quando accede a servizi speciali). La pseudonimizzazione dell'indirizzo e-mail in questo caso non è fattibile, in quanto la SN deve essere in grado di inviare agli utenti delle e-mail con notifiche (e altri contenuti).

Si prenda in considerazione, poi, una seconda organizzazione immaginaria, la Online Security Services Corp. (di seguito denominata OSS), che agisce in qualità di responsabile del trattamento dei dati per conto di SN, con il compito di curare i servizi di archiviazione e di sicurezza di alcune sezioni della banca dati degli

utenti di SN. In virtù del proprio ruolo, OSS ha accesso ai file di log pseudonimizzati di SN, quali gli indirizzi IP e le marche temporali pseudonimizzati di tutti gli accessi al sito web, ma non anche agli indirizzi IP originari. Con le proprie funzioni, OSS non è in grado di re-identificare gli utenti associati ad un certo indirizzo IP, poiché tali dati sono memorizzati in un database differente tenuto da SN e non accessibile a OSS. Quindi, ai fini della pseudonimizzazione, l'ipotesi rappresentata nel Capitolo 3.3 vedrà SN quale titolare del trattamento dei dati e OSS quale successivo responsabile del trattamento.

8.2 LE INFORMAZIONI SUI DATI

A prima vista, supponendo che SN abbia utilizzato una funzione di pseudonimizzazione sufficientemente salda, OSS non sarà in grado di violare la pseudonimizzazione degli indirizzi IP effettuata da SN. Tuttavia, a seconda della funzione di pseudonimizzazione e, soprattutto, della tecnica di pseudonimizzazione (di cui al Capitolo 5.2) impiegate, OSS potrebbe comunque essere in grado di comprendere se un dato pseudonimo ricorra frequentemente, raramente, una sola volta, o non sia affatto presente all'interno della banca dati. Ciò potrebbe, di per sé, non essere sufficiente a risalire a un'identità, ma potrebbe essere utilizzato per identificare gli utenti che accedono più frequentemente. Se all'interno di una registrazione d'accesso vi è uno pseudonimo che ricorre con un'alta frequenza, OSS sarà in grado di dedurre che si tratti di un utente molto attivo su SN. Viceversa, se uno pseudonimo compare per la prima volta nel dataset, molto probabilmente tale utente si è appena registrato a SN ed è entrato per la prima volta nel suo account utente, oppure l'indirizzo IP di un utente già registrato è cambiato (ipotesi frequente e che rende tali osservazioni meramente probabilistiche).

Siffatto tipo di informazioni sui dati può, già di per sé, essere utile a OSS, ad esempio, per sapere quanti degli utenti di SN sono utenti abituali e quanti si registrano ma poi non accedono nuovamente una seconda volta (con un certo margine di errore probabilistico basato sulla modifica degli indirizzi IP). Tali informazioni possono assumere un valore rilevante nelle relazioni commerciali tra SN e OSS.

Oltre le dette informazioni sui dati, la circostanza che OSS abbia continuo accesso al database di SN permette ad OSS un altro tipo di raccolta di informazioni: infatti, attraverso il continuo monitoraggio del dataset archiviato per conto di SN, OSS è in grado di conoscere ogni cambiamento del dataset stesso. Ciò include, certamente, il numero totale di accessi al sito web di SN, ma può anche essere utilizzato per contare, ad esempio, il numero di nuove registrazioni di utenti (pseudonimi che appaiono per la prima volta) nell'arco di un giorno o di un mese. Pur essendo per lo più informazioni di natura statistica, le stesse possono essere utilizzate per mettere in atto veri e propri attacchi di discriminazione (in modo da sviluppare effetti differenti per gruppi di utenti differenti): OSS, venendo a conoscenza che un dato pseudonimo di un nuovo utente appare per la prima volta in un determinato giorno, sarà in grado di monitorare le interazioni tra quello specifico utente e SN. Detta informazione, come verrà mostrato in seguito, può facilmente diventare un problema in ordine alla protezione dei dati personali dell'utente.

8.3 IL COLLEGAMENTO DI DATI

Nell'ipotesi appena presa in considerazione, i dati accessibili da parte di OSS forniscono più informazioni dei soli indirizzi IP: ogni log di accesso ne memorizza anche il timestamp. Pertanto, invece di monitorare frequentemente i cambiamenti nella banca dati di SN, OSS può semplicemente fare affidamento sui

riferimenti temporali collegati ad ogni pseudonimo per eseguire lo stesso tipo di discriminazione dell'utente di cui si è detto in precedenza. I riferimenti temporali vengono memorizzati insieme agli indirizzi IP pseudonimizzati, per cui le due informazioni risultano direttamente collegate l'una a l'altra. Sulla base del collegamento di tali dati, OSS può incrementare le proprie informazioni su specifici utenti di SN: ad esempio, un utente specifico accede a SN più spesso al mattino, in pausa pranzo o alla sera? Solo o per lo più di domenica? Solo nelle festività religiose del calendario ortodosso? Solo durante i periodi di vacanze scolastiche in Danimarca?

Ognuna di queste specificazioni aggiuntive consente a OSS di avvicinarsi a una violazione della pseudonimizzazione, basandosi solo sulle marche temporali memorizzate e sulla possibilità di collegare diversi record con pseudonimi identici. Come mostrato, tali tipi di informazioni sono in grado di fornire a OSS alcune caratterizzazioni degli utenti di SN, le quali possono essere considerate informazioni personali.

Tuttavia, il collegamento di dati richiede che agli stessi dataset strutturati vengano accoppiati delle informazioni supplementari, quali – come negli esempi già forniti – il calendario ortodosso o le vacanze scolastiche danesi. Di fatto, gli stessi possono essere considerati come attacchi basati sulle conoscenze di base dell'intruso, come discusso nel capitolo 4, ma con una diversa complessità delle conoscenze di base necessarie. Inoltre, tale tipologia di informazioni estratte è di natura statistica, quindi una buona probabilità di affidabilità, ma comunque non del 100%. In tali ipotesi, quanto maggiore è il numero di dati contenuti nella banca dati, tanto più affidabile (o falsificabile) diventa la possibilità di collegamento. Pertanto, quanto più grande è il social network SN, tanto più semplice diventa per OSS effettuare tali discriminazioni o addirittura attacchi di re-identificazione.

L'esempio di cui sopra riguardava esclusivamente i casi di pseudonimizzazioni di indirizzo IP e timestamp. Ciò pure accadrebbe, e anzi sarebbe ancora più affidabile, nei casi in cui, invece dell'indirizzo IP, si pseudonimizzi l'indirizzo e-mail, poiché quest'ultimo tende a cambiare con minore frequenza e, dunque, rappresenta un identificativo maggiormente univoco per un individuo umano.

8.4 L'ABBINAMENTO DELLA DISTRIBUZIONE DELLE RICORRENZE

Le strutture dei dati di cui all'esempio precedente risultano assai ridotte e semplicistiche, poiché prendono in considerazione esclusivamente la pseudonimizzazione dell'indirizzo IP e del timestamp. Tuttavia, nel caso in cui si disponga di adeguate informazioni di base, le stesse possono essere sufficienti per portare a termine attacchi discriminatori o addirittura attacchi di re-identificazione. Inoltre, gli inserimenti di dati consentono di memorizzare, solitamente, più informazioni rispetto ai due suddetti valori, per cui, in realtà, contengono più dettagli utili ai fini della scoperta degli pseudonimi.

Si consideri ad esempio che SN memorizzi, oltre al timestamp e l'indirizzo IP pseudonimizzati in ogni record di dati, anche il tipo e la versione del browser[31] utilizzato dall'utente, le impostazioni e le preferenze di lingua dell'utente (definite nelle impostazioni del browser), la versione del sistema operativo del computer dell'utente, ecc. Come è stato scoperto dalla Electronic Frontier Foundation nel progetto Panopticlick[32], questa combinazione di impostazioni del browser può già, di per sé, essere sufficiente per identificare in modo univoco un determinato browser – e quindi l'utente – di un sito web online. Dunque, se SN memorizza,

[31] Si deve osservare che questo è il log preimpostato. Per esempio Apache web server.
[32] Si rimanda alla consultazione del sito https://panopticlick.eff.org/

per ogni accesso al proprio sito web, tutte le predette informazioni, anche OSS potrà avere accesso ad esse.

Pesino nell'ipotesi in cui SN eseguisse una sorta di pseudonimizzazione su ciascuna di queste configurazioni (ad es. memorizzando solo un Hash della stringa relativa la versione del browser ricevuta dal browser dell'utente), OSS potrebbe comunque essere in grado di vedere tutte le stringhe della versione del browser pseudonimizzate, calcolare le statistiche relative a quel valore di Hash che appare con una data frequenza nell'intero database di SN, e confrontare la distribuzione di questi diversi valori esistenti con le statistiche reperibili pubblicamente sul sito web di Panopticlickc, così ottenendo, per ogni valore di Hash, la vera stringa relativa alla versione del browser, e ciò nonostante il corretto utilizzo della funzione di pseudonimizzazione. Infatti, il solo fatto che alla distribuzione statistica dei diversi pseudonimi corrisponda la distribuzione statistica dei loro presunti testi in chiaro, può essere sufficiente per risalire agli pseudonimi con un'alta probabilità di successo.

Ovviamente ciò dipenderà, in larga misura, dalla tecnica di pseudonimizzazione utilizzata. Se viene utilizzata una tecnica ingegneristica adeguata, l'aggiunta di metadati all'interno di una funzione di pseudonimizzazione sarà in grado di offrire una migliore protezione da un attacco di ingegneria inversa.

8.5 LE CONOSCENZE SUPPLEMENTARI

Se OSS avesse a disposizione delle conoscenze aggiuntive in merito alle caratteristiche di un determinato utente e cercasse di scoprire, dalla banca dati pseudonimizzata che riceve da SN, i record di dati riferiti a quell'utente, ogni informazione supplementare potrebbe diventare importante. Si prenda in considerazione, ad esempio, il caso in cui OSS sia a conoscenza del fatto che l'utente di riferimento è di genere maschile e utilizza il Browser Chrome su un dispositivo iPad; tale informazione è, di per sé, in grado di restringere notevolmente l'ambito dei profili utente su cui OSS dovrà concentrarsi. Ognuno di questi valori di dati, anche se pseudonimizzato, riduce l'insieme delle possibilità, cioè l'insieme dei profili utente contenuti nella banca dati di SN che può appartenere allo specifico utente di riferimento cercato da OSS. Le informazioni sul browser possono essere trattate attraverso la tecnica di distribuzione delle probabilità di attacco di cui al Capitolo 8.4, escludendo un'ampia cerchia di profili utente con pseudonimi di browser troppo, o troppo poco, corrispondenti alla configurazione "Browser Chrome utilizzato su un dispositivo iPad".

Una volta effettuata questa prima scrematura, sulla base dei profili rimanenti, con un semplice attacco di forza bruta o un attacco di distribuzione statistica, OSS potrà acquisire l'indicazione del genere di appartenenza degli pseudonimi, eliminando circa la metà dei profili utente rimasti. A questo punto, se tutti i restanti profili utente hanno in comune la circostanza per cui il primo accesso a SN è avvenuto tra maggio e luglio 2018, OSS ha già ottenuto un'ulteriore informazione su detto specifico utente: ovvero che si è registrato presso SN in quell'intervallo di tempo. Si tratta di un attacco deduttivo riuscito. Proseguendo nell'analisi dei rimanenti profili utente, OSS potrà scoprire che uno specifico modello di *timestamp* viene utilizzato da SN esclusivamente con due profili utente, così che il cerchio venga ulteriormente circoscritto a due sole ipotesi che corrispondo alle caratteristiche del profilo utente di riferimento (che OSS è stato in grado di osservare in alcune occasioni in passato). Pertanto, l'ambito di ricerca viene ristretto a due soli profili utente.

Le informazioni comuni a entrambi i profili rimasti, essendo sicuramente assai rispondenti al profilo di riferimento cercato da OSS, sono in grado di fornire a OSS un numero consistente di informazioni utili. Per

escludere definitivamente uno dei candidati rimanenti, OSS dovrà semplicemente monitorare in modo specifico l'utilizzo di SN da parte di questi due profili e, al successivo accesso, verificare se lo stesso possa essere stato fatto dall'individuo di riferimento o meno (e ciò basandosi su ulteriori conoscenze di base che OSS ha ottenuto in merito al proprio obiettivo). In definitiva, OSS sarà in grado di collegare uno dei profili utente all'identità dell'obiettivo. In questo modo, OSS sarà anche in grado di individuare tutti gli pseudonimi relativi i valori dei dati di quella persona, consentendo potenzialmente a OSS di identificare o discriminare anche altri profili di utenti.

Va comunque sottolineato che il problema della disponibilità delle informazioni supplementari è "ortogonale" rispetto alla pseudonimizzazione, pur essendo anzitutto una questione di protezione dei dati by design. Pertanto, come già precedente menzionato nel presente trattato, oltre alla pseudonimizzazione, può essere presa in considerazione anche l'introduzione di ostacoli all'interno dei parametri della funzione di pseudonimizzazione, ovvero l'uso della generalizzazione, in modo da rendere meno efficaci gli attacchi di forza bruta (per cui si rimanda alla Sezione 5.6). Questo grado di libertà consente di rafforzare ulteriormente la pseudonimizzazione e proteggere i dati da attacchi rilevanti.

8.6 IL COLLEGAMENTO TRA NUMEROSE FONTI DI DATI

Oltre l'ipotesi appena descritta riguardante SN e OSS, un caso assai più complesso di pseudonimizzazione è rappresentato dall'ipotesi di scambio di dati pseudonimizzati non già tra due sole organizzazioni (SN e OSS), bensì all'interno di un mercato su larga scala. In tali circostanze, più organizzazioni condividono set di dati personali pseudonimizzati, al fine di ottenere una certa funzionalità (ad esempio, la creazione di profili a scopi di marketing), proteggendo al contempo l'identità delle persone interessate. Sul tema, spesso ci si avvale della teoria per cui la pseudonimizzazione sia in grado di impedire la re-identificazione delle persone interessate, così legittimando una siffatta condivisione di dati. A tal proposito, il presente trattato, lungi dall'esprimere giudizi di merito in relazione alla legittimità della condivisione dei set di dati pseudonimizzati, mira ad approfondire la questione della corretta applicazione della pseudonimizzazione in un simile contesto.

Si prenda in considerazione l'ipotesi di un insieme di società composto da A ad E, le quali raccolgono i dati personali dei loro utenti, come i dati raccolti da SN nell'esempio precedente. Il collegamento dei profili degli utenti delle diverse società potrebbe essere effettuato confrontando gli indirizzi e-mail utilizzati dai rispettivi utenti. Ad esempio, se presso le aziende B e D appaiono due profili utente registrati con lo stesso indirizzo e-mail, molto probabilmente apparterranno allo stesso utente interessato. Tuttavia, come già discusso nel Capitolo 7, pure lo stesso indirizzo e-mail rappresenta un dato personale. Perciò, sarà necessario pseudonimizzare gli indirizzi e-mail presenti nei set di dati delle società B e D prima di poterli condividere con A, B, C, D ed E.

In questo caso, la sfida è rappresentata dal fatto che tutte le società coinvolte intendano mantenere la funzionalità dei dati pseudonimizzati ai fini di collegare i profili appartenenti alla stessa persona, senza, tuttavia, ridurre la protezione dell'identità di quell'utente. A tal fine, per poter confrontare e collegare tra loro i record dei dati provenienti dai diversi set di dati, tutte e cinque le aziende dovranno applicare la medesima pseudonimizzazione, utilizzando la stessa funzione di pseudonimizzazione e lo stesso segreto di pseudonimizzazione. In tale ipotesi, tuttavia, vi è un chiaro divario tra la funzionalità (data dalla possibilità di collegare gli indirizzi e-mail pseudonimizzati) e la protezione dell'identità degli utenti di tali indirizzi e-mail. In altre parole, seguendo l'esempio già riportato, B e D dovrebbero poter essere in grado di conoscere che dei

loro determinati record di dati hanno in comune lo stesso indirizzo e-mail, e quindi sono riconducibili allo stesso utente, tuttavia non dovrebbero essere in grado di risalire all'indirizzo e-mail e, quindi, all'identità dell'interessato.

Come discusso nel Capitolo 7, in tali ipotesi, l'utilizzo di funzioni di pseudonimizzazione poco efficaci (quali l'hashing) consente semplici attacchi di forza bruta, congettura o di distribuzione delle probabilità, di cui si è già detto. Con l'aggiunta delle informazioni supplementari (non personali) contenute nei record di dati condivisi, e, probabilmente, con qualche ulteriore conoscenza di base, tali attacchi risultano – nella maggior parte dei casi – agevoli e con buone probabilità di successo. In ulteriore, più le imprese condividono informazioni relative le caratteristiche di un determinato interessato, più un intruso avrà a disposizione informazioni utili a violare la pseudonimizzazione di un utente, con maggiore probabilità che questi attacchi abbiano successo.

I rischi per la privacy permangono, comunque, anche nella generale ipotesi in cui le organizzazioni applichino tecniche di pseudonimizzazione diverse (e anche salde) agli identificativi dei loro utenti (ad es. indirizzi e-mail o indirizzi IP). Si prenda in considerazione l'ipotesi in cui le suddette società da A ed E forniscano tali dati pseudonimizzati a OSS al fine di ottenere, ad esempio, servizi statistici. Se gli pseudonimi forniti sono accompagnati da informazioni sul browser/dispositivo dell'utente come descritto al punto 8.4 (impostazioni del browser, sistema operativo ecc.), e assumendo che tali informazioni sul dispositivo siano uniche per ogni dispositivo[33], allora OSS dovrà semplicemente collegare i diversi pseudonimi forniti dalle diverse società, al medesimo utente corrispondente.

8.7 LE CONTROMISURE

Come discusso nel capitolo 5, le tecniche di pseudonimizzazione casuale (documentale o completa) riducono la possibilità di collegare i diversi pseudonimi presenti nei vari set di dati, così limitando o addirittura eliminando le funzionalità statistiche delle banche dati pseudonimizzate. Allo stesso tempo, limitano la possibilità di collegare diversi record di dati (potenzialmente diffusi in molte organizzazioni) a un medesimo profilo utente. Pertanto, pure nell'ipotesi in cui venga applicata una pseudonimizzazione casuale, se OSS fosse in grado di scoprire se due diversi pseudonimi appartengono allo stesso identificativo, OSS potrebbe comunque essere in grado di eseguire i suddetti attacchi. Analogamente, B e D potrebbero riuscire a re-identificare l'interessato corrispondente ai due profili utente corrispondenti. In questo caso il dilemma tra protezione e funzionalità diventa di nuovo cruciale. Allora, come ci si può difendere in modo affidabile da questa tipologia di attacchi alla pseudonimizzazione?

In base a quanto emerso nel presente trattato, la migliore tecnica di pseudonimizzazione è quella di:

- Considerare l'intero set di dati disponibile.
- Informarsi sulle dimensioni del dominio di input dei singoli valori dei dati.
- Applicare la pseudonimizzazione a tutti i valori dei dati in modo tale che gli attacchi di forza bruta e della ricerca nel dizionario diventino impossibili.
- Eliminare qualsiasi possibilità di attacchi derivanti dalle conoscenze di base o di distribuzione statistica.
- Progettare la funzione di pseudonimizzazione su larga scala in modo che tale set di dati pseudonimizzati mantenga solo il tipo di funzionalità strettamente necessaria ai fini del trattamento, eliminando tutte le altre funzionalità dal set di dati pseudonimizzati.

[33] Il noto termine *device fingerprinting* descrive tale rischio per la privacy.

In relazione all'ipotesi esemplificativa di cui al presente Capitolo, SN potrebbe utilizzare uno schema di pseudonimizzazione tale da pseudonimizzare non solo gli indirizzi IP, ma tutte le possibili combinazioni di indirizzi IP e di timestamp. In tal caso, il collegamento del timestamp a qualsiasi altra fonte di dati esterna diventerebbe impossibile, poiché tali informazioni non sono più disponibili per OSS. Per essere in grado di re-identificare, OSS avrebbe bisogno di conoscere (o indovinare) l'esatta combinazione di indirizzo IP e timestamp. In generale, la pseudonimizzazione di una combinazione di input di dati non può essere ragionevolmente scoperta senza conoscere (o indovinare) tutti i dati inseriti in chiaro. Con questa impostazione, tale pseudonimizzazione impedirebbe in modo assai più rigoroso qualsiasi tentativo da parte di OSS di scoprire uno pseudonimo.

Alcuni esempi di tecniche base utili ai fini di una solida funzione di pseudonimizzazione sono già stati presentati nel Capitolo 5, insieme a una discussione approfondita circa la loro resistenza agli attacchi alla pseudonimizzazione di cui al Capitolo 4. Per estendere tali tecniche a record di dati strutturati è spesso sufficiente considerare l'intero record di dati come input e applicare una combinazione su misura di funzioni Hash e tecniche comuni di anonimizzazione. Tecniche più avanzate di pseudonimizzazione sono state brevemente discusse nel Capitolo 5.6 e in una precedente trattazione dell'ENISA [2].

9. CONCLUSIONI E RACCOMANDAZIONI

Sotto la vigenza del GDPR, la sfida di una corretta applicazione della pseudonimizzazione ai dati personali sta gradualmente diventando un argomento assai dibattuto in diversi settori: a partire dalla ricerca e dal mondo accademico, passando per la giustizia e le forze dell'ordine, fino ad arrivare al settore della compliance di diverse organizzazioni Europee. Nel presente trattato sono state introdotte definizioni e argomenti basilar, unitamente alle varie tecniche, agli attacchi e alle contromisure più rilevanti ai fini di supportare tale prospettato futuro argomento interdisciplinare.

Come illustrato nel presente trattato, la materia della pseudonimizzazione dei dati all'interno di infrastrutture informatiche complesse è un campo impegnativo, variamente influenzato dal contesto, dagli enti coinvolti, dai tipi di dati, dalle informazioni di base e dai dettagli di implementazione. Infatti, nell'ambito della pseudonimizzazione, non esiste una soluzione univoca e semplice, valida per tutti gli approcci e applicabile a tutti i casi pratici possibili. Al contrario, per poter attuare un solido procedimento di pseudonimizzazione, è richiesto un elevato livello di competenza, così da ridurre al minimo la minaccia di attacchi di discriminazione o di re-identificazione, pur mantenendo il grado di funzionalità fondamentale ai fini del trattamento dei dati pseudonimizzati.

A tal scopo, sulla base dell'analisi già fornita nel trattato, qui di seguito vengono riportate alcune conclusioni e raccomandazioni essenziali utili a tutti gli enti interessati circa l'adozione e l'attuazione pratica della pseudonimizzazione.

VERSO LA PSEUDONIMIZZAZIONE ATTRAVERSO UN APPROCCIO BASATO SUL RISCHIO

Nonostante tutte le tecniche di pseudonimizzazione conosciute abbiano le loro proprie, ben chiare, caratteristiche intrinseche, ciò non rende, nella pratica, la scelta del giusto approccio un compito banale. A tal fine è necessario un attento esame del contesto in cui applicare la pseudonimizzazione, considerando tutti gli obiettivi specifici della pseudonimizzazione (da chi proteggere le identità, qual è la funzionalità che si desidera raggiungere dagli pseudonimi ottenuti, e così via), così come la semplicità nella realizzazione. Pertanto, per quanto attiene la scelta della tecnica di pseudonimizzazione più appropriata, occorre adottare un approccio basato sul rischio, in modo da valutare e limitare adeguatamente le relative minacce alla privacy. Infatti, la semplice protezione dei dati aggiuntivi necessari per la re-identificazione, pur essendo un prerequisito, non garantisce necessariamente l'eliminazione di tutti i rischi.

I titolari del trattamento e i responsabili del trattamento dovrebbero considerare attentamente l'attuazione della pseudonimizzazione secondo un approccio basato sul rischio, tenendo conto delle finalità e del contesto generale del trattamento dei dati personali, così come i livelli di funzionalità e di scalabilità che si intendono raggiungere.

I creatori di prodotti, servizi e applicazioni dovrebbero fornire ai titolari del trattamento e ai responsabili del trattamento informazioni adeguate in merito al proprio utilizzo delle tecniche di pseudonimizzazione e ai livelli di sicurezza e di protezione dei dati che forniscono.

Le autorità di regolamentazione (ad esempio le Autorità di controllo e il Comitato Europeo per la Protezione dei Dati) dovrebbero fornire linee guida pratiche ai titolari e ai responsabili del trattamento dei dati personali circa la valutazione del rischio, promuovendo, al contempo, le buone pratiche in materia di pseudonimizzazione.

LA DEFINIZIONE DELLO STATO DELL'ARTE

Per supportare un approccio basato sul rischio nell'ambito della pseudonimizzazione, è essenziale definire lo stato dell'arte nel settore. Infatti, come mostrato in questo rapporto, se da un lato sono disponibili diverse tecniche di pseudonimizzazione, d'altro lato l'applicazione pratica delle stesse può variare, ad esempio, a seconda della tipologia di identificativi o di set di dati. A tal fine, è importante lavorare sui casi d'uso e sugli esempi specifici, ampliando il ventaglio delle opzioni di applicazioni tecniche possibili nel campo della pseudonimizzazione.

La Commissione europea e le Istituzioni europee competenti dovrebbero sostenere la definizione e la diffusione di informazioni sullo stato dell'arte della pseudonimizzazione, in collaborazione con le comunità di ricerca e l'industria del settore.

Le Autorità di regolamentazione (ad esempio le Autorità di controllo e il Comitato Europeo per la Protezione dei Dati) dovrebbero promuovere la pubblicazione delle buone pratiche nel campo della pseudonimizzazione.

L'AVANZAMENTO DELLO STATO DELL'ARTE

Il presente trattato ha focalizzato l'attenzione sulle tecniche basilari di pseudonimizzazione che, al giorno d'oggi, i titolari e i responsabili del trattamento hanno a disposizione. Tuttavia, nelle ipotesi più complesse (che, come visto, sono assai frequenti nella pratica), sarà sempre più indispensabile ricorrere all'uso di tecniche più avanzate (e salde), come quelle provenienti dal settore dell'anonimizzazione. Anzi, la nozione stessa di anonimizzazione dovrebbe essere rivisitata, in quanto i modelli sono in continua evoluzione (e, quindi, l'anonimizzazione diventa, nei casi reali, sempre più impegnativa).

La comunità scientifica dovrebbe lavorare per sviluppare le attuali tecniche di pseudonimizzazione in modo che diventino soluzioni più avanzate in grado di affrontare efficacemente le particolari sfide sorte nell'era dei big data. La Commissione europea e le istituzioni europee competenti dovrebbero sostenere e diffondere tali sforzi.

BIBLIOGRAFIA

[1] ENISA, “Recommendations on shaping technology according to GDPR provisions - An overview on data pseudonymisation”, 2018.

[2] ENISA, “Privacy and data protection by design - from policy to engineering,” 2014.

[3] L. Sweeney, “K-anonymity: A model for protecting privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.

[4] A. Pfitzmann and M. Hansen, “A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management,” 2010.

[5] M. Barbaro, T. Zeller and S. Hansell, A face is exposed for aol searcher no. 4417749, vol. 9, New York Times, 2006, p. 8.

[6] M. Hellman, “A cryptanalytic time-memory trade-off,” IEEE transactions on Information Theory, vol. 26, no. 4, pp. 401-406, 1980.

[7] P. Oechslin, “Making a Faster Cryptanalytic Time-Memory Trade-off,” in CRYPTO 2003, 2003.

[8] J. L. Massey, “Guessing and entropy,” in Proceedings of 1994 IEEE International Symposium on Information Theory, 1994.

[9] Y. Yona and S. Diggavi, “The effect of bias on the guesswork of HASH functions,” in 2017 IEEE International Symposium on Information Theory (ISIT), 2017.

[10] D. G. Malone and W. Sullivan, “Guesswork and entropy,” IEEE Transactions on Information Theory, vol. 50, no. 3, pp. 525--526, 2004.

[11] H. C. Van Tilborg and S. Jajodia, Encyclopedia of cryptography and security, Springer Science & Business Media, 2014.

[12] J. Katz, A. J. Menezes, P. C. Van Oorschot and S. A. Vanstone, Handbook of applied cryptography, CRC press, 1996.

[13] M. J. Dworkin, “SHA-3 Standard: Permutation-Based HASH and Extendable-Output Functions,” 2015.

[14] T. Eastlake and D. Hansen, “US Secure HASH Algorithms (SHA and SHA-based HMAC and HKDF),” 2011.

[15] M. Bellare, R. Canetti and H. Krawczyk, “Keying HASH functions for message authentication,” in Annual international cryptology conference, 1996.

[16] H. Krawczyk, R. Canetti and M. Bellare, “HMAC: Keyed-HASHing for Message Authentication,” RFC, pp. 1-11, 1997.

[17] R. C. Merkle, “A Digital Signature Based on a Conventional Encryption Function,” Advances in Cryptology — CRYPTO '87, pp. 369-378, 1988.

[18] G. Becker, “Merkle Signature Schemes, Merkle Trees and Their Cryptanalysis,” Bochum, 2008.

[19] L. Lamport, “Password authentication with insecure communication,” Communications of the ACM, pp. 770-772, November 1981.

[20] B. H. Bloom, “Space/time trade-offs in HASH coding with allowable errors,” Communications of the ACM, pp. 422-426, July 1970.

[21] S. Weber, “On Transaction Pseudonyms with Implicit Attributes,” Cryptology ePrint Archive: Report 2012/568, https://eprint.iacr.org/2012/568 , 2012.

[22] A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam, “L-diversity: Privacy beyond k-anonymity,” in 22nd International Conference on Data Engineering (ICDE'06), 2006.

[23] N. Li, T. Li and S. Venkatasubramanian, “t-closeness: Privacy beyond k-anonymity and l-diversity,” in 23rd International Conference on Data Engineering, 2007.

[24] C. Dwork and A. Roth, “The Algorithmic Foundations of Differential Privacy,” Foundations and Trends in Theoretical Computer Science, pp. 211-407, August 2014.

[25] L. Sweeney, “Only You, Your Doctor, and Many Others May Know,” Technology Science, vol. 2015092903, no. 9, p. 29, 2015.

[26] H. W and F. Wang, “A Survey of Noninteractive Zero Knowledge Proof System and Its Applications,” The Scientific World Journal, 2014.

[27] IETF, “Internet Engineering Task Force: RFC 791, Internet Protocol DARPA Internet Program Protocol Specification,” 1981.

[28] IETF, “Internet Engineering Task Force: RFC8200, Internet Protocol, Version 6 (IPv6) Specification,” STD 86, 2017.

[29] IETF, “Internet Engineering Task Force: RFC4632, Classless Inter-domain Routing (CIDR): The Internet Address Assignment and Aggregation Plan,” BCP 122, 2006.

[30] IETF, “Internet Engineering Task Force: RFC 5735, Special Use IPv4 Addresses,” 2010.

[31] WP29, “Article 29 Data Protection Working Party: Opinion 4/2007 on the concept of personal data,” 2007.

[32] IETF, “Internet Engineering Task Force: IPFIX Working Group, IP Flow Anonymization Support,”2011

[33] “An Analysis of Google Logs Retention Policies,” Journal of Privacy and Confidentiality, vol. 3, no. 1, 2011.

[34] L. Demir, A. Kumar, M. Cunche and C. Lauradoux, “The pitfalls of HASHing for privacy,” IEEE Communications Surveys & Tutorials, pp. 551-565, 2017.

[35] S. Englehardt, J. Han and A. Narayanan, “I never signed up for this! privacy implications of email tracking,” in Proceedings on Privacy Enhancing Technologies, 2018.

[36] Digital Summit ’s Data Protection Focus Group, “White Paper on Pseudonymization,” 2017.

[37] R. Noumeir, A. Lemay and J.-M. Lina, “Pseudonymization of radiology data for research purposes.,” Journal of digital imaging, vol. 20, no. 3, pp. 284-295, 2007.

[38] I. Polakis, G. Kontaxis, S. Antonatos, E. Gessiou, T. Petsas and E. P. Markatos, “Using social networks to harvest email addresses,” in Proceedings of the 9th Annual ACM Workshop on Privacy in the Electronic Society, 2010.