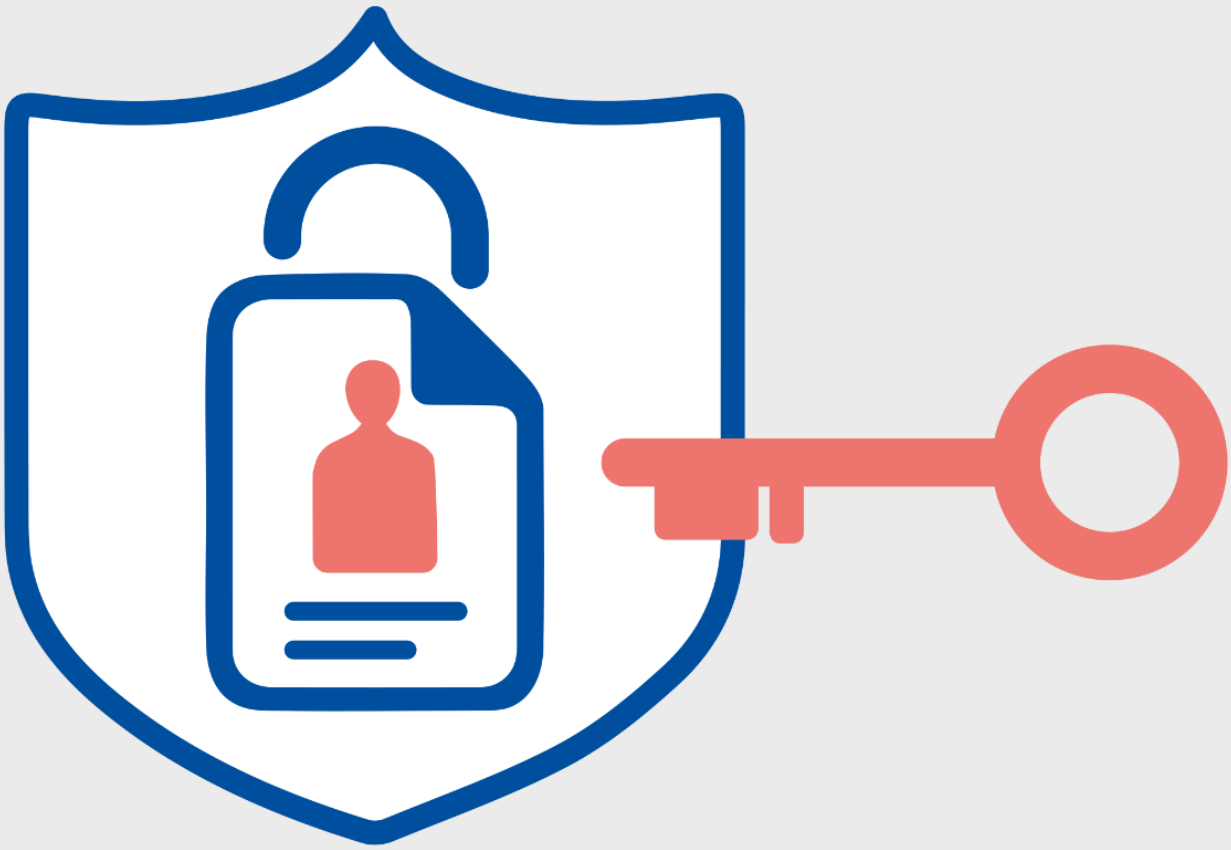
EUROPEAN UNION AGENCY
FOR CYBERSECURITY



TECNICHE DI PSEUDONIMIZZAZIONE E MIGLIORI PRATICHE

Raccomandazioni per sviluppare tecnologie conformi alle disposizioni in materia di protezione dei dati e privacy

NOVEMBRE 2019

INFORMAZIONI SULL'ENISA

L'Agenzia dell'Unione europea per la cibersicurezza (ENISA) è attiva dal 2004 sul fronte della sicurezza informatica in Europa. L'ENISA collabora con l'Unione europea (UE) e i suoi Stati membri, di concerto con il settore privato e i cittadini europei, al fine di formulare consigli e raccomandazioni sulle buone pratiche in materia di sicurezza delle informazioni. Assiste inoltre gli Stati membri dell'UE nell'attuazione della legislazione dell'Unione in materia e lavora per migliorare la resilienza delle infrastrutture critiche informatizzate e di rete in Europa. L'ENISA si adopera per potenziare l'attuale livello di competenza degli Stati membri dell'UE, sostenendo lo sviluppo di comunità transfrontaliere impegnate a migliorare la sicurezza delle reti e delle informazioni in tutta l'UE. Dal 2019 è attiva nel predisporre schemi di certificazione della cibersicurezza. Maggiori informazioni sull'ENISA e sulle sue attività sono disponibili al seguente indirizzo: www.enisa.europa.eu.

CONTATTI

Per contattare gli autori, inviare un'email a isdp@enisa.europa.eu
Per maggiori informazioni sul documento, si prega di contattare il responsabile delle relazioni con i media al seguente indirizzo press@enisa.europa.eu.

COLLABORATORI ESTERNI

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

REDATTORI

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

RINGRAZIAMENTI

I nostri ringraziamenti vanno a Giuseppe D'Acquisto (Garante), Nils Gruschka (Università di Oslo) e Simone Fischer-Hübner (Università di Karlstad) per aver esaminato la presente relazione fornendoci il loro prezioso parere.

NOTE LEGALI

COPYRIGHT

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

INDICE

1. INTRODUZIONE 7

1.1 INFORMAZIONI DI RIFERIMENTO 7

1.2 CAMPO DI APPLICAZIONE E OBIETTIVI 7

1.3 STRUTTURA 8

2. TERMINOLOGIA 9

3. SCENARI DI PSEUDONIMIZZAZIONE 11

3.1 SCENARIO 1: LA PSEUDONIMIZZAZIONE PER USO INTERNO 11

3.2 SCENARIO 2: RESPONSABILE DEL TRATTAMENTO COINVOLTO NELLA PSEUDONIMIZZAZIONE 12

3.3 SCENARIO 3: INVIO DI DATI PSEUDONIMIZZATI A UN RESPONSABILE DEL TRATTAMENTO 12

3.4 SCENARIO 4: IL RESPONSABILE DEL TRATTAMENTO COME ENTITÀ DI PSEUDONIMIZZAZIONE 14

3.5 SCENARIO 5: TERZE PARTI COME ENTITÀ DI PSEUDONIMIZZAZIONE 14

3.6 SCENARIO 6: L'INTERESSATO COME ENTITÀ DI PSEUDONIMIZZAZIONE 15

4. MODELLO DI ATTACCO 17

4.1 ATTACCANTI INTERNI 17

4.2 ATTACCANTI ESTERNI 17

4.3 OBIETTIVI DELL'ATTACCO ALLA PSEUDONIMIZZAZIONE 18

4.3.1 Segreto di pseudonimizzazione 18
4.3.2 Reidentificazione completa 18
4.3.3 Discriminazione 18

4.4 PRINCIPALI TECNICHE DI ATTACCO 19

4.4.1 Attacco a forza bruta 19
4.4.2 Ricerca in un dizionario 20
4.4.3 Inferenze 21

4.5 FUNZIONALITÀ E PROTEZIONE DEI DATI 21

5. TECNICHE DI PSEUDONIMIZZAZIONE 23

5.1 PSEUDONIMIZZAZIONE DI UN SINGOLO IDENTIFICATORE 23

5.1.1 Contatore 23
5.1.2 Generatore di numeri casuali 24
5.1.3 Funzione crittografica di hash 24
5.1.4 Codice di autenticazione del messaggio 24
5.1.5 Crittografia 25

5.2 STRATEGIE DI PSEUDONIMIZZAZIONE 25

5.2.1 Pseudonimizzazione deterministica 25
5.2.2 Pseudonimizzazione randomizzata al documento 26
5.2.3 Pseudonimizzazione completamente randomizzata 26

5.3 SCEGLIERE UNA TECNICA E UNA STRATEGIA DI PSEUDONIMIZZAZIONE 26

5.4 RECUPERO 27

5.5 PROTEZIONE DELLA CHIAVE DI PSEUDONIMIZZAZIONE 28

5.6 TECNICHE AVANZATE DI PSEUDONIMIZZAZIONE 28

6. PSEUDONIMIZZAZIONE DEGLI INDIRIZZI IP 30

6.1 PSEUDONIMIZZAZIONE E LIVELLO DI PROTEZIONE DEI DATI 30

6.2 PSEUDONIMIZZAZIONE E LIVELLO DI FUNZIONALITÀ 31

6.2.1 Livello di pseudonimizzazione 31
6.2.2 Scelta della modalità di pseudonimizzazione 32

7. PSEUDONIMIZZAZIONE DEGLI INDIRIZZI E-MAIL 35

7.1 CONTATORE E GENERATORE DI NUMERI CASUALI (RNG) 35

7.2 FUNZIONE CRITTOGRAFICA DI HASH 37

7.3 CODICE DI AUTENTICAZIONE DEL MESSAGGIO 38

7.4 CRITTOGRAFIA 39

8. PSEUDONIMIZZAZIONE IN PRATICA: UNO SCENARIO PIÙ COMPLESSO 41

8.1 UN ESEMPIO DI SIMULAZIONE 41

8.2 INFORMAZIONI INERENTI AI DATI 42

8.3 DATI COLLEGATI 42

8.4 DISTRIBUZIONE CORRISPONDENTE DELLE OCCORRENZE 43

8.5 CONOSCENZE AGGIUNTIVE 44

8.6 COLLEGAMENTO TRA PIÙ SORGENTI DI DATI 44

8.7 CONTROMISURE 45

9. CONCLUSIONI E RACCOMANDAZIONI 47

SINTESI

Alla luce del regolamento generale sulla protezione dei dati (RGPD) ([1]), il dibattito sulla corretta applicazione della pseudonimizzazione ai dati personali sta divenendo via via più acceso in diverse comunità, dal mondo accademico e della ricerca a quello giudiziario e dell'applicazione delle leggi, fino a toccare il campo della gestione della conformità in varie organizzazioni europee. Sulla base del precedente lavoro sul campo dell'ENISA ([2]), la presente relazione approfondisce le nozioni di base relative alla pseudonimizzazione, vagliando soluzioni tecniche come potenziale sostegno all'attuazione pratica.

In particolare, a partire da una serie di scenari aventi per oggetto la pseudonimizzazione, la presente relazione individua in primo luogo i principali attori che possono essere coinvolti nel processo di pseudonimizzazione e i loro possibili ruoli. Procede quindi ad analizzare i diversi modelli di attacco e le tecniche di attacco adottate contro la pseudonimizzazione, come l'attacco a forza bruta, la ricerca in un dizionario e le inferenze. Presenta inoltre le principali tecniche di pseudonimizzazione (mediante contatore, mediante generatore di numeri casuali, mediante funzione crittografica di hash, mediante codice di autenticazione del messaggio e crittografia) unitamente alle relative politiche (la pseudonimizzazione deterministica, randomizzata in funzione del dato e completamente randomizzata) attualmente disponibili. Si occupa in particolare dei parametri in grado di influenzare la scelta di eventuali tecniche o politiche di pseudonimizzazione sul campo, quali la protezione dei dati, la funzionalità, la scalabilità e la re-identificazione. Si fa un rapido cenno anche alle tecniche di pseudonimizzazione più avanzate. Secondo le sopra enunciate descrizioni, la relazione si basa su due casi d'uso riguardanti la pseudonimizzazione di indirizzi IP e di indirizzi e-mail, analizzando le particolarità connesse a questi specifici identificatori . Esamina inoltre un caso d'uso più complesso relativo alla pseudonimizzazione di più registrazioni di dati, analizzando le possibilità di re-identificazione.

Tra i principali risultati della relazione, è emerso come non esista una soluzione semplice e univoca per la pseudonimizzazione, valida per tutti gli approcci e in tutti gli scenari possibili. Al contrario, per applicare un robusto processo di pseudonimizzazione è necessaria un'elevata competenza, riducendo possibilmente la minaccia di classificazioni o di attacchi di re-identificazione, e mantenendo al contempo il grado di funzionalità necessario per il trattamento dei dati pseudonimizzati.

A tal fine, la relazione trae le seguenti conclusioni e raccomandazioni per tutte le parti interessate riguardo all'adozione e all'attuazione pratiche della pseudonimizzazione dei dati.

UN APPROCCIO ALLA PSEUDONIMIZZAZIONE BASATO SUL RISCHIO

Sebbene tutte le tecniche di pseudonimizzazione ad oggi note presentino caratteristiche intrinseche ormai chiare, non è tuttavia semplice, a livello pratico, individuare la tecnica corretta. È pertanto necessario adottare un approccio basato sul rischio, valutando il livello di protezione richiesto e tenendo conto delle relative esigenze di funzionalità e scalabilità.

([1]) Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio, del 27 aprile 2016, relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati e che abroga la direttiva 95/46/CE (regolamento generale sulla protezione dei dati), https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32016R0679&from=EN

([2]) ENISA, Raccomandazioni per sviluppare tecnologie conformi alle disposizioni del RGPD: una panoramica sulla pseudonimizzazione dei dati, https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-RGPD-provisions

I titolari e i responsabili del trattamento dei dati dovrebbero considerare attentamente l'idea di attuare la pseudonimizzazione secondo un approccio basato sul rischio, tenendo conto dello scopo e del contesto generale del trattamento dei dati personali, nonché dei livelli di funzionalità e scalabilità che intendono raggiungere.

I produttori di beni, servizi e applicazioni dovrebbero fornire informazioni adeguate ai titolari e ai responsabili del trattamento dei dati circa le tecniche di pseudonimizzazione da loro adottate e i livelli di sicurezza e protezione dei dati da queste ultime garantiti.

Le autorità di regolamentazione (quali le autorità preposte alla protezione dei dati e il Comitato europeo per la protezione dei dati) sono chiamate a fornire un orientamento pratico ai titolari e ai responsabili del trattamento dei dati in materia di valutazione del rischio, promuovendo al tempo stesso le migliori pratiche nel campo della pseudonimizzazione.

DEFINIZIONE DELLO STATO DELL'ARTE

Per poter sostenere un approccio alla pseudonimizzazione basato sul rischio, è indispensabile definire lo stato dell'arte del settore. A tal fine, si rivela essenziale lavorare su casi d'uso ed esempi specifici, fornendo più dettagli e le possibili opzioni di attuazione tecnica.

La Commissione europea e le istituzioni dell'UE competenti dovrebbero promuovere la definizione e la diffusione dello stato dell'arte nella pseudonimizzazione, d'intesa con la comunità dei ricercatori e con il settore industriale.

Le autorità di regolamentazione (quali le autorità preposte alla protezione dei dati e il Comitato europeo per la protezione dei dati) dovrebbero promuovere la pubblicazione delle migliori pratiche nel campo della pseudonimizzazione.

PROMUOVERE LO STATO DELL'ARTE

Se, da un lato, la presente relazione è incentrata sulle tecniche di pseudonimizzazione di base attualmente disponibili, è essenziale adottare tecniche più avanzate (e robuste), come quelle derivanti dal campo dell'anonimizzazione, così da affrontare gli scenari sempre più complessi connessi all'applicazione pratica.

La comunità dei ricercatori dovrebbe lavorare per integrare nelle attuali tecniche di pseudonimizzazione soluzioni più avanzate, per fare fronte in maniera efficace alle particolari sfide poste dall'era dei big data. La Commissione europea e le istituzioni dell'UE competenti dovrebbero promuovere e diffondere tali sforzi.

1. INTRODUZIONE

1.1 INFORMAZIONI DI RIFERIMENTO

La pseudonimizzazione è un tipo di trattamento di dati personali che ha attirato maggiormente l'attenzione a seguito dell'adozione del RGPD, in cui viene indicato come un meccanismo appositamente progettato per la sicurezza e la protezione dei dati. Inoltre, nell'ambito del RGPD e subordinatamente al corretto utilizzo, la pseudonimizzazione può, in una certa misura, ridurre gli obblighi legali a carico dei titolari del trattamento dei dati.

Vista la crescente importanza che ha assunto sia per i titolari del trattamento di dati che per gli interessati, l'ENISA ha pubblicato nel 2018 [1] una panoramica sulle nozioni e principali tecniche connesse alla pseudonimizzazione, in base al suo ruolo ai sensi del RGPD. In particolare, a partire dalla definizione di pseudonimizzazione (nonché dalle sue differenze rispetto ad altre tecnologie, come l'anonimizzazione e la crittografia), la relazione va innanzitutto a delineare i principali vantaggi derivanti dalla protezione dei dati mediante pseudonimizzazione. Successivamente a tale analisi, presenta poi alcune potenziali tecniche di pseudonimizzazione, come l'hash, l'hash con chiave o con salt, la crittografia, il dispositivo di autenticazione mediante «token», e altri sistemi pertinenti. Vengono infine prese in esame alcune applicazioni della pseudonimizzazione, con particolare attenzione rivolta all'area delle applicazioni mobili.

Sebbene il suddetto lavoro dell'ENISA tratti alcune questioni chiave della pseudonimizzazione, è necessario condurre ulteriori ricerche e analisi sia per rafforzare il concetto di pseudonimizzazione come misura di sicurezza (articolo 32 del RGPD) sia per delinearne il ruolo, quale strumento appositamente concepito per proteggere i dati (articolo 25 del RGPD). In effetti, come anche indicato nella relazione dell'ENISA, vi è soprattutto la necessità di promuovere le migliori pratiche di pseudonimizzazione e fornire esempi di casi d'uso che potrebbero concorrere a definire lo «stato dell'arte» in tale settore.

In quest'ottica, l'ENISA, nell'ambito del suo programma di attività per il 2019, si è concentrata sull'applicazione pratica della pseudonimizzazione dei dati ([3]).

1.2 CAMPO DI APPLICAZIONE E OBIETTIVI

Il campo di applicazione generale del presente lavoro consiste nel fornire una guida e le migliori pratiche relative all'attuazione tecnica della pseudonimizzazione dei dati.

In particolare, la relazione è intesa a:

- analizzare diversi scenari di pseudonimizzazione e i relativi soggetti coinvolti;
- presentare potenziali tecniche di pseudonimizzazione in rapporto ai relativi modelli di intrusione e di attacco;
- analizzare l'applicazione della pseudonimizzazione su specifici identificatori, in particolare indirizzi IP, indirizzi e-mail e altri insiemi di dati strutturati (casi d'uso);
- trarre le relative conclusioni e formulare raccomandazioni per nuove ricerche nel settore.

([3]) Dato che l'ENISA è chiamata a fornire indicazioni sugli aspetti connessi alla politica in materia di sicurezza delle reti e dell'informazione nell'UE, ne consegue logicamente che affrontare aree di interesse, quali la privacy e la protezione dei dati, rappresenta un'adeguata estensione del suo lavoro, andando a soddisfare le esigenze delle parti interessate. In effetti, l'analisi dell'attuazione pratica della pseudonimizzazione è importante ai fini della sicurezza dei dati personali, come indicato nell'articolo 32 del RGPD.

I casi d'uso sono stati selezionati partendo dal presupposto che gli specifici identificativi (indirizzi IP, indirizzi e-mail, identificativi in insiemi di dati strutturati) rappresentano casi piuttosto comuni in scenari di vita reale differenti. I casi d'uso selezionati sono inoltre espressione di una pluralità di requisiti connessi alla pseudonimizzazione, ossia che vanno dal formato rigido degli indirizzi IP alla struttura più flessibile degli indirizzi e-mail, fino alla natura imprevedibile degli insiemi di dati più ampi.

Il pubblico a cui si rivolge il presente rapporto è composto dai titolari del trattamento dati, i responsabili del trattamento, i produttori di beni, servizi e applicazioni, le Autorità per la protezione dei dati (DPA), nonché da qualsiasi altra parte interessata alla pseudonimizzazione dei dati.

Il documento presuppone una conoscenza basilare dei principi di protezione dei dati personali e del ruolo/processo di pseudonimizzazione. Per una panoramica sulla pseudonimizzazione dei dati ai sensi del RGPD, si rinvia anche ai precedenti lavori svolti sul campo dall'ENISA.

Il dibattito e gli esempi portati nella presente relazione si incentrano unicamente su soluzioni tecniche potenzialmente in grado di favorire la privacy e la protezione dei dati; non devono pertanto essere intesi come un parere giuridico sui casi analizzati.

1.3 STRUTTURA

La relazione è strutturata come segue:

- Il capitolo 2 introduce la terminologia adottata nelle altre sezioni della relazione, con le relative note esplicative, ove necessario.
- Il capitolo 3 fa riferimento ai più comuni scenari di pseudonimizzazione in cui è possibile imbattersi sul campo.
- Il capitolo 4 descrive i possibili modelli di intrusione edi attacco relativi alla pseudonimizzazione (e gli scenari in precedenza illustrati).
- Il capitolo 5 presenta le principali tecniche e politiche di pseudonimizzazione attualmente disponibili.
- I capitoli 6, 7 e 8 analizzano l'applicazione di differenti tecniche di pseudonimizzazione su indirizzi IP, indirizzi e-mail e insiemi di dati più complessi (casi d'uso).
- Il capitolo 8 fa un breve riepilogo delle analisi condotte e formula le principali conclusioni e raccomandazioni per tutte le parti interessate.

La presente relazione rientra nel lavoro svolto dall'ENISA nell'ambito della privacy e della protezione dei dati ([4]), focalizzato sull'analisi di soluzioni tecniche pensate per l'esecuzione del RGPD, della tutela della vita privata fin dalla progettazione e la sicurezza nel trattamento dei dati personali.

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

2. TERMINOLOGIA

Il presente capitolo enumera alcuni termini adottati nella relazione che sono essenziali per favorire la corretta comprensione da parte del lettore. Alcuni di questi termini sono basati sul RGPD, mentre altri si riferiscono a norme tecniche o sono espressamente definiti ai fini della presente pubblicazione.

Vengono in particolare adoperati i seguenti termini:

Per **dato personale** si intende qualsiasi informazione riguardante una persona fisica identificata o identificabile (l'**interessato**); si considera identificabile la persona fisica che può essere identificata, direttamente o indirettamente, con particolare riferimento a un identificativo come il nome, un numero di identificazione, dati relativi all'ubicazione, un identificativo online o a uno o più elementi caratteristici della sua identità fisica, fisiologica, genetica, psichica, economica, culturale o sociale (articolo 4, punto 1), del RGPD).

Il **titolare del trattamento dei dati** o il **titolare** è 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 (articolo 4, punto 7), del RGPD).

Il **responsabile del trattamento dei dati** o il **responsabile** è la persona fisica o giuridica, l'autorità pubblica, il servizio o altro organismo che tratta dati personali per conto del titolare del trattamento (articolo 4, punto 8), del RGPD.

Per **pseudonimizzazione** si intende il trattamento dei dati personali in modo tale che i dati personali non possano più essere attribuiti a un interessato specifico senza l'utilizzo di informazioni aggiuntive, a condizione che tali informazioni aggiuntive siano conservate separatamente e soggette a misure tecniche e organizzative intese a garantire che tali dati personali non siano attribuiti a una persona fisica identificata o identificabile (articolo 4, punto 5), del RGPD ([5]).

L'**anonimizzazione** è un processo mediante il quale i dati personali vengono modificati in modo irreversibile così che il titolare del trattamento, da solo o in collaborazione con altre parti, non possa più identificare direttamente o indirettamente l'interessato (ISO/TS 25237:2017) ([6]).

L'**identificativo** è quel valore che identifica un elemento all'interno di uno schema di identificazione ([7]). Un identificativo univoco è associato a un unico elemento. Nella presente relazione spesso si presuppone che i dati personali vengano associati a identificativi univoci.

Lo **pseudonimo**, anche noto come **nome in codice**, è un'informazione associata all'identificativo di un individuo o ad altri tipi di dati personali (come i dati relativi all'ubicazione).

([5]) Per le relative definizioni tecniche di pseudonimizzazione, cfr. [1].
([6]) Per approfondire l'argomento, inclusa la differenza tra pseudonimizzazione e anonimizzazione, cfr. [1].
([7]) Il Gruppo dell'articolo 29 [31] fa riferimento agli identificativi come a informazioni strettamente collegate a un individuo, in un rapporto privilegiato con quest'ultimo, che ne consentono l'identificazione. La misura in cui un determinato identificatore è in grado di consentire l'identificazione dipende dallo specifico contesto di trattamento dei dati personali. Pertanto, gli identificativi possono corrispondere a singole informazioni (quali nome, indirizzo e-mail, numero di previdenza sociale, ecc.) ma anche a dati più complessi.

Gli pseudonimi possono presentare diversi gradi di associabilità (agli identificativi originali) ([8]). Il grado di associabilità di tipi differenti di pseudonimi è importante per valutare la forza di questi ultimi, ma anche per progettare sistemi di pseudonimizzazione che consentano di raggiungere il grado di associabilità desiderato (per es., nell'analizzare file di registro pseudonimi o in caso di sistemi di reputazione) ([9]).

Una **funzione di pseudonimizzazione**, indicata con P, è una funzione che sostituisce un identificativo ID con uno pseudonimo $pseudo$.

Una chiave **di pseudonimizzazione**, indicata come s è un parametro di una funzione di pseudonimizzazione P. La funzione P non può essere valutata/calcolata nel caso in cui s sia sconosciuto.

La **funzione di recupero**, indicata come R, è una funzione che sostituisce uno pseudonimo $pseudo$ con un identificativo ID, utilizzando la chiave di pseudonimizzazione s. Essa va a invertire la funzione di pseudonimizzazione P.

La **tabella di mappatura di pseudonimizzazione** rappresenta l'azione di una funzione di pseudonimizzazione. Essa associa ciascun identificativo allo pseudonimo corrispondente. A seconda della funzione di pseudonimizzazione P, la relativa tabella di mappatura può corrispondere alla chiave di pseudonimizzazione o a una sua parte.

L'**entità di pseudonimizzazione** è quella che converte gli identificativi in pseudonimi utilizzando un'apposita funzione. Può corrispondere a un titolare del trattamento dei dati, a un responsabile del trattamento dei dati (che esegue la pseudonimizzazione per conto del titolare), a una terza parte fidata o a un interessato, a seconda del contesto di pseudonimizzazione. Va sottolineato che, sulla base di tale definizione, il ruolo dell'entità di pseudonimizzazione è strettamente collegato all'attuazione pratica della pseudonimizzazione in uno specifico contesto ([10]). Tuttavia, nella presente relazione, la responsabilità dell'intero processo di pseudonimizzazione (e del trattamento dei dati in generale) è sempre a carico del titolare del trattamento.

Il **dominio dell'identificativo/dello pseudonimo** si riferisce ai domini, ossia a tutti possibili valori, da cui vengono ricavati l'identificativo e lo pseudonimo. Si può trattare dello stesso dominio o di domini differenti. Possono essere domini finiti o infiniti.

L'attaccante è quell'entità che cerca di decodificare la pseudonimizzazione e di riassociare uno pseudonimo (o un set di dati pseudonimizzato) al suo titolare.

L'**attacco di reidentificazione** è un attacco rivolto alla pseudonimizzazione ad opera di un attaccante, che mira a reidentificare il titolare di uno pseudonimo.

([8]) A tal fine, si può definire uno pseudonimo come un modo per «mascherare» l'identificatore di un individuo, rendendo tale soggetto più o meno identificabile a seconda del contesto.
[9] Per un approfondimento sui gradi di associabilità, cfr. [4].
([10]) Va notato che, in base alla definizione di pseudonimizzazione contenuta nel RGPD (articolo 4, punto 5)), non si fa alcun riferimento a chi detiene le informazioni aggiuntive.

3. SCENARI DI PSEUDONIMIZZAZIONE

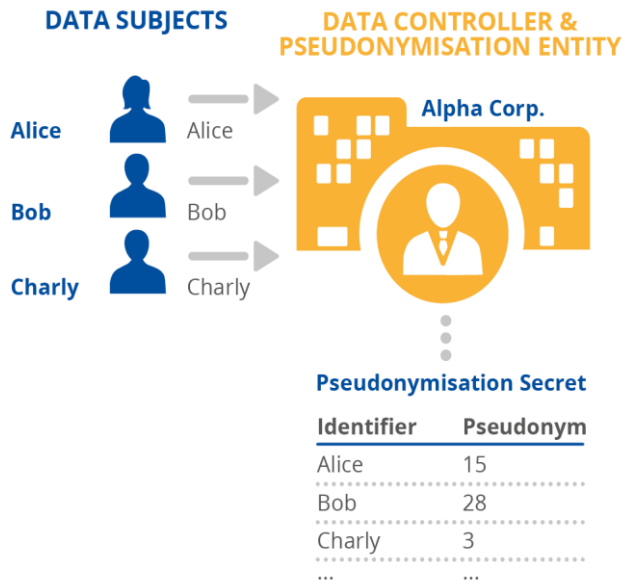
Come già sottolineato in [1], la pseudonimizzazione svolge un ruolo importante nel RGPD come misura di sicurezza (articolo 32 del RGPD), così come nell'ambito della protezione dei dati (articolo 25 del RGPD). Il vantaggio più evidente legato alla pseudonimizzazione consiste nell'occultare l'identità degli interessati a terzi (diversi dall'entità di pseudonimizzazione) nell'ambito di una specifica operazione di trattamento dei dati. La pseudonimizzazione, tuttavia, non si limita a occultare la reale identità, ma concorre all'obiettivo di proteggere i dati anche grazie all'inassociabilità [2], ossia riducendo il rischio che i dati relativi alla privacy vengano collegati tra differenti domini di trattamento dei dati. Inoltre, la pseudonimizzazione (essendo una tecnica di minimizzazione dei dati) può contribuire al principio della minimizzazione dei dati ai sensi del RGPD, come nel caso in cui il titolare del trattamento non debba avere accesso all'identità reale degli interessati, ma solo ai loro pseudonimi. Infine, un altro importante vantaggio connesso alla pseudonimizzazione, da non sottovalutare, è costituito dall'accuratezza dei dati (per un'analisi più dettagliata del ruolo della pseudonimizzazione, cfr. [1]).

Tenendo conto dei vantaggi di cui sopra, il presente capitolo illustra diversi scenari di pseudonimizzazione in cui è possibile imbattersi, elencando per ciascuno i vari attori e gli specifici obiettivi della pseudonimizzazione.

3.1 SCENARIO 1: LA PSEUDONIMIZZAZIONE PER USO INTERNO

Uno scenario comune di pseudonimizzazione dei dati si presenta nel momento in cui i dati vengono raccolti direttamente dai relativi interessati e pseudonimizzati dal titolare del trattamento, per la successiva elaborazione interna.

Figura 1. Esempio di pseudonimizzazione - Scenario 1

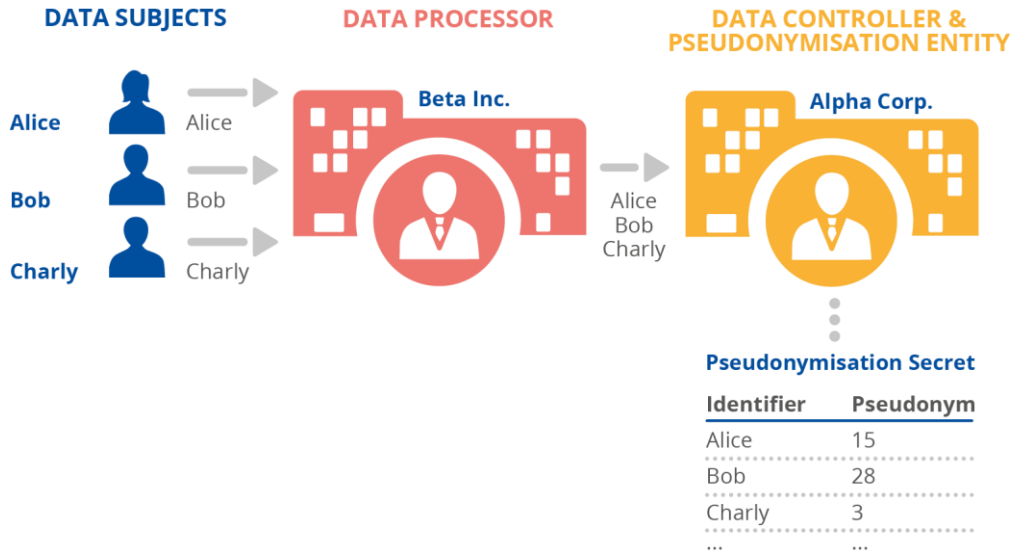


Nella Figura 1, il titolare del trattamento (la Società Alfa) assume il ruolo di entità di pseudonimizzazione, effettuando la selezione e l'assegnazione degli pseudonimi agli identificativi. Occorre sottolineare che gli interessati non necessariamente conoscono o apprendono il loro pseudonimo, poiché la chiave di pseudonimizzazione (in tal caso, la tabella di mappatura della pseudonimizzazione) è noto unicamente alla Società Alfa. Nella fattispecie, il ruolo della pseudonimizzazione consiste nel migliorare la sicurezza dei dati personali in caso di uso interno (ad es. condivisione tra diverse unità del titolare del trattamento) ([11]) o di incidenti di sicurezza.

3.2 SCENARIO 2: RESPONSABILE DEL TRATTAMENTO COINVOLTO NELLA PSEUDONIMIZZAZIONE

Si tratta di una variante dello scenario 1 e vede coinvolto nel trattamento anche un responsabile, che ottiene gli identificativi dagli interessati (per conto del titolare del trattamento). La pseudonimizzazione viene comunque eseguita dal titolare del trattamento.

Figura 2. Esempio di pseudonimizzazione – Scenario 2



Nella Figura 2, a un responsabile del trattamento dei dati (la Società Beta) viene assegnato il compito di raccogliere gli identificativi dagli interessati e di inoltrarli a un titolare del trattamento (la Società Alfa), che esegue infine la pseudonimizzazione. Anche in questo caso il titolare del trattamento è l'entità di pseudonimizzazione. Tale scenario potrebbe includere, per esempio, un fornitore di servizi cloud che mette a disposizione servizi di raccolta dati per conto del titolare del trattamento. A questo punto il titolare del trattamento è nuovamente responsabile di applicare la pseudonimizzazione dei dati prima di un eventuale trattamento successivo. Gli obiettivi della pseudonimizzazione sono gli stessi dello scenario 1 (ma questa volta nel trattamento è coinvolto anche un responsabile).

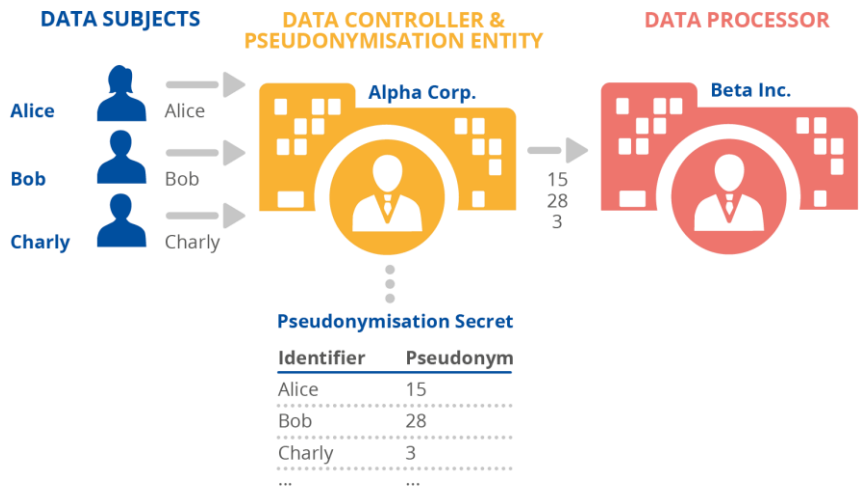
3.3 SCENARIO 3: INVIO DI DATI PSEUDONIMIZZATI A UN RESPONSABILE DEL TRATTAMENTO

Diversamente dal caso precedente, se in questo scenario è sempre il titolare del trattamento a eseguire la pseudonimizzazione, il responsabile del trattamento si limita a ricevere i dati pseudonimizzati da parte del titolare del trattamento.

([11]) Per la nozione di «analisi generale» per uso interno, cfr. il considerando (29) del RGPD.

La figura 3 mostra un titolare del trattamento (la Società Alfa) che raccoglie i dati e ne esegue la pseudonimizzazione (in qualità di entità di pseudonimizzazione). La differenza con i precedenti scenari è data dal fatto che, in questo caso, il titolare del trattamento inoltra i dati pseudonimizzati a un successivo responsabile del trattamento (la Società Beta), come in caso di analisi statistiche o di memorie di dati persistenti. In tale scenario, con la pseudonimizzazione dei dati è possibile raggiungere l'obiettivo di protezione prefissato: La Società Beta non viene a conoscenza degli identificativi degli interessati e non è quindi in grado di reidentificare in maniera diretta le persone fisiche a questi connesse (presupponendo la mancanza di altri attributi che le consentirebbero la reidentificazione). In tal modo, la pseudonimizzazione garantisce la sicurezza dei dati proteggendoli dai responsabili del trattamento.

Figura 3. Esempio di pseudonimizzazione - Scenario 3

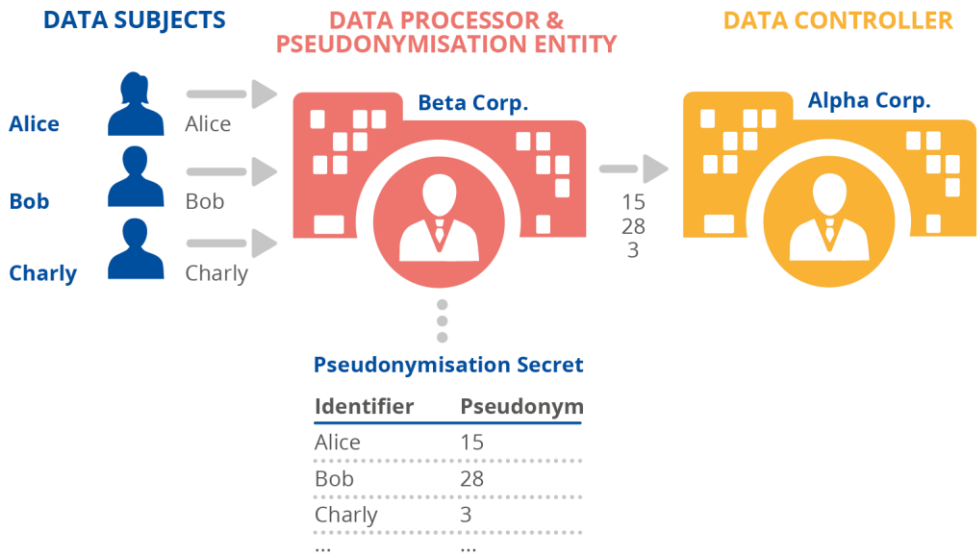


In una variante di tale scenario, i dati pseudonimizzati potrebbero non essere inviati a un responsabile del trattamento, bensì a un altro titolare del trattamento (come nell'ambito di un obbligo legale del titolare del trattamento iniziale o in altri contesti giuridici).

3.4 SCENARIO 4: IL RESPONSABILE DEL TRATTAMENTO COME ENTITÀ DI PSEUDONIMIZZAZIONE

Un altro possibile scenario vede il titolare del trattamento assegnare il processo di pseudonimizzazione a un responsabile del trattamento (per es., un fornitore di servizi cloud che gestisce il segreto di pseudonimizzazione e/o organizza le relative strutture tecniche).

Figura 4: Esempio di pseudonimizzazione - Scenario 4



La Figura 4 mostra il caso in cui i dati personali siano inviati dagli interessati a un responsabile del trattamento (la Società Beta), che va poi a eseguire la pseudonimizzazione, agendo quindi in qualità di entità di pseudonimizzazione per conto del titolare del trattamento (la Società Alfa). I dati pseudonimizzati vengono poi inoltrati al titolare del trattamento. In questo specifico scenario, il titolare del trattamento memorizza solamente i dati pseudonimizzati. Ne consegue una maggiore sicurezza a livello del titolare del trattamento, grazie alla de-identificazione dei dati (per es. nel caso di una violazione presso quest'ultimo). Tuttavia, il titolare del trattamento è in tutti i casi in grado di reidentificare l'interessato tramite il responsabile del trattamento. A tale proposito, acquisisce una notevole importanza anche la sicurezza a livello del responsabile del trattamento.

In una variante di questo scenario, vari responsabili del trattamento potrebbero essere coinvolti nel processo di pseudonimizzazione, come una sequenza di entità di pseudonimizzazione (catena di responsabili del trattamento).

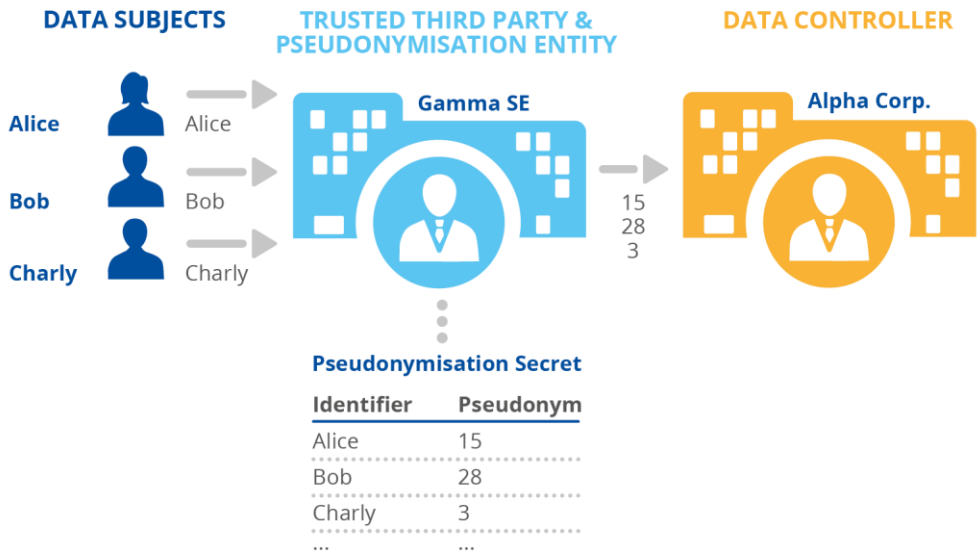
3.5 SCENARIO 5: TERZE PARTI COME ENTITÀ DI PSEUDONIMIZZAZIONE

In questo scenario, la pseudonimizzazione viene eseguita da una terza parte (non dal responsabile del trattamento), che va poi a inoltrare i dati al titolare del trattamento. Nel caso presente, contrariamente allo scenario 4, il titolare del trattamento non può accedere agli identificativi degli interessati (in quanto la terza parte non ricade sotto il suo controllo).

La Figura 5 mostra il caso in cui i dati personali siano inviati a una terza parte (la Società Gamma) che va poi a eseguire la pseudonimizzazione, in qualità di entità di pseudonimizzazione. I dati pseudonimizzati vengono in seguito inoltrati al titolare del trattamento (la Società Alfa). In tale scenario, il titolare del trattamento non è in grado di associare direttamente o indirettamente i record di dati individuali ai rispettivi interessati. La sicurezza e la protezione dei dati risultano in tal modo potenziate a livello del titolare del

trattamento conformemente al principio della minimizzazione dei dati. Il presente scenario può verificarsi laddove il titolare del trattamento non necessiti di accedere all'identità degli interessati (ma unicamente ai loro pseudonimi).

Figura 5. Esempio di pseudonimizzazione - Scenario 5



Tale scenario potrebbe assumere particolare rilievo in caso di titolari del trattamento congiunti (co-titolari), in cui uno esegue la pseudonimizzazione (operando in qualità di terza parte fidata, cfr. la figura 5), mentre l'altro si limita a ricevere i dati pseudonimizzati per un ulteriore trattamento.

Un'interessante variante di questo scenario (che richiede ulteriori analisi) è rappresentata dal caso in cui la terza parte fidata corrisponda a più di un'entità, in grado di creare e recuperare pseudonimi solo in maniera congiunta (o basandosi eventualmente su uno schema di condivisione segreta), così che non si debba fare affidamento su un'unica entità.

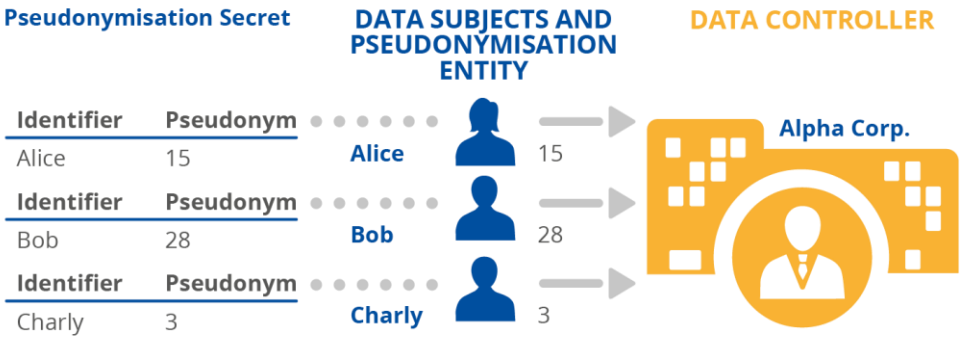
3.6 SCENARIO 6: L'INTERESSATO COME ENTITÀ DI PSEUDONIMIZZAZIONE

Si tratta di uno speciale caso di pseudonimizzazione in cui gli pseudonimi vengono creati dagli stessi interessati, nel quadro del processo complessivo di pseudonimizzazione.

Come si osserva nell'esempio della figura 6, ciascun individuo genera il proprio pseudonimo, per poi inoltrare da quel momento i dati pertinenti servendosi di quest'ultimo ([12]).

([12]) Si rammenta che lo pseudonimo può essere unico o differire a seconda dei vari servizi e applicazioni (cfr. il capitolo 5).

Figura 6: Esempio di pseudonimizzazione - Scenario 6



Un esempio di tale sistema di pseudonimizzazione dei dati sarebbe l'utilizzo della chiave pubblica di una coppia di chiavi nei sistemi blockchain (quale il Bitcoin) per ottenere lo pseudonimo. In questo caso, la pseudonimizzazione mira a impedire al titolare del trattamento di venire a conoscenza ([13]) degli identificativi degli interessati, consentendo a questi ultimi di avere il controllo del processo di pseudonimizzazione; ovviamente, la responsabilità dello schema complessivo di pseudonimizzazione ricade nuovamente sul titolare del trattamento ([14]). Anche questo caso è in linea con il principio della minimizzazione dei dati e si può applicare alle situazioni in cui il titolare del trattamento non abbia bisogno di avere accesso agli identificativi originali (ovvero, quando gli pseudonimi sono sufficienti per una specifica operazione di trattamento dei dati).

[13] Vale a dire che il titolare del trattamento non acquisisce alcun segreto di pseudonimizzazione che consenta una reidentificazione diretta.
[14] Per questo caso, cfr. l'articolo 11 del RGPD.

4. MODELLO DI ATTACCO

Come indicato nel capitolo 3, il principale obiettivo della pseudonimizzazione è di limitare l'associabilità tra un set di dati pseudonimizzato e i titolari degli pseudonimi, così da proteggere l'identità degli interessati. Questo tipo di protezione è in genere finalizzato a contrastare le azioni compiute da un attaccante per eseguire una reidentificazione.

Il presente capitolo prende in esame i possibili modelli antagonisti e i diversi tipi di attacchi di reidentificazione particolarmente rilevanti per la pseudonimizzazione. A tal fine, vengono affrontate le nozioni di antagonisti interni ed esterni, esaminando i ruoli che potrebbero assumere negli scenari di pseudonimizzazione precedentemente evocati. Comprendere tali tematiche è essenziale per approfondire l'utilizzo delle tecniche di pseudonimizzazione, che verrà trattato nei capitoli a seguire.

4.1 ATTACCANTI INTERNI

In base all'accezione comune del termine, nell'ambito della sicurezza IT, un attaccante interno dispone di specifiche conoscenze, capacità o autorizzazioni (rapportate al suo obiettivo)[15]. Nel contesto della pseudonimizzazione, ciò implica che l'attaccante sia in grado di ottenere informazioni sulla chiave di pseudonimizzazione e/o relative informazioni di rilievo.

Per es., riprendendo gli scenari 1, 2, 3 e 4 descritti nel capitolo 3, un intruso interno potrebbe trovarsi presso il titolare del trattamento (come nel caso di un dipendente di tale titolare). Tuttavia, in base agli scenari 2 e 4, potrebbe trovarsi anche presso il responsabile del trattamento (per es., il dipendente malintenzionato di un contraente). In ultimo, prendendo in esame lo scenario 5, l'attaccante interno potrebbe trovarsi presso una terza parte fidata (la quale funge da entità di pseudonimizzazione). Le terze parti che potrebbero legittimamente avere accesso ai dati personali (quali un'autorità di controllo o di contrasto) non sono per convenzione considerate antagoniste ([16]).

4.2 ATTACCANTI ESTERNI

Contrariamente agli attaccanti interni, un attaccante esterno non ha accesso diretto alla chiave di pseudonimizzazione o ad altre informazioni pertinenti. Tuttavia, questo tipo di attaccante può avere accesso a un set di dati pseudonimizzato, oltre a essere in grado di eseguire l'operazione di pseudonimizzazione su valori arbitrari dei dati di ingresso da lui selezionati (per es., potrebbe avere accesso a un'implementazione black box della funzione di pseudonimizzazione o potrebbe costringere l'entità di pseudonimizzazione a pseudonimizzare input arbitrari). L'obiettivo di un attaccante esterno è di accrescere le sue informazioni riguardo al set di dati pseudonimizzato, venendo per es. a conoscenza dell'identità associata a un determinato pseudonimo (ricavando ulteriori informazioni su tale identità dai dati aggiuntivi contenuti nel set di dati connesso a tale pseudonimo).

[15] Secondo il Centro minacce interne del CERT presso il Software Engineering Institute (SEI) della Carnegie Mellon University, si definisce «minaccia interna» la possibilità da parte di un individuo che ha o ha autorizzato l'accesso alle risorse di un'organizzazione, di utilizzare il suo accesso, in modo intenzionale o non intenzionale, al fine di influire negativamente sull'organizzazione, https://insights.sei.cmu.edu/insider-threat/2017/03/cert-definition-of-insider-threat---updated.html

([16]) Occorre tuttavia notare che la legittimità di tale accesso potrebbe essere messa in discussione qualora non fosse rispettato il principio di minimizzazione dei dati (come nel caso in cui un'autorità di controllo ottenesse l'accesso al segreto di pseudonimizzazione anziché ricevere in forma esplicita soltanto i dati personali cui avrebbe diritto). Tali scenari rientrerebbero nel modello degli antagonisti interni, in quanto la terza parte ha un accesso interno legittimo, analogamente all'entità di pseudonimizzazione.

Considerati gli scenari presentati nel capitolo 3, si deve per definizione considerare un attaccante esterno qualsiasi soggetto che agisca dolosamente in tutti gli scenari indicati e che non faccia parte né lavori per conto dell'entità di pseudonimizzazione. Un titolare del trattamento dei dati (con finalità dolose) può assumere il ruolo di attaccante esterno nell'ambito dello scenario 5 o dello scenario 6. Anche un responsabile del trattamento dei dati (con finalità dolose) può assumere tale ruolo nell'ambito dello scenario 3.

4.3 OBIETTIVI DELL'ATTACCO ALLA PSEUDONIMIZZAZIONE

A seconda del contesto e del metodo di pseudonimizzazione, l'attaccente può avere obiettivi diversi contro i dati pseudonimizzati, quali il recupero del segreto di pseudonimizzazione, la reidentificazione completa o la discriminazione. Se la maggior parte degli esempi illustrati nei paragrafi seguenti si concentra sull'individuazione dell'identità «reale» degli interessati, va tuttavia notato che un attacco andato a buon fine non è (solamente) questione di retroingegnerizzazione, ma deriva soprattutto dalla capacità di individuare uno specifico soggetto all'interno di un gruppo (anche se l'identità «reale» non viene rivelata).

4.3.1 Segreto di pseudonimizzazione

In questo caso, l'attaccante si focalizza sull'individuazione della chiave di pseudonimizzazione (ovvero, quando essa viene utilizzato). Si tratta della tipologia di attacco più grave, in quanto l'uso della chiave di pseudonimizzazione consente all'attaccante di reidentificare qualsiasi pseudonimo nel set di dati (reidentificazione completa o discriminazione), nonché di eseguire ulteriori processi di pseudonimizzazione sul set di dati.

4.3.2 Reidentificazione completa

Quando l'obiettivo dell'attacco è la reidentificazione completa, l'attaccante intende pervenire all'associazione sussistente tra uno o più pseudonimi e l'identità dei loro titolari. Questa tipologia di attaccante è stata ampiamente esaminata in letteratura (cfr. per es. [3], [4], [5]).

L'attacco di reidentificazione più grave consiste nella reidentificazione di tutti gli pseudonimi. L'attaccante può adoperare due strategie per conseguire tale obiettivo: recuperare ogni identificativo a partire dallo pseudonimo corrispondente in modo indipendente oppure risalire al segreto di pseudonimizzazione (cfr. 4.3.1). La forma meno grave degli attacchi di reidentificazione completa implica invece un attaccante che può unicamente reidentificare un sottoinsieme di pseudonimi nel set di dati. Si consideri, per es., un set di dati pseudonimizzato dei voti degli studenti di un corso universitario. Ogni voce all'interno del set di dati contiene uno pseudonimo corrispondente all'identità dello studente (nome e cognome) e un secondo pseudonimo relativo al genere dello studente (associando per es. le studentesse a numeri pari e gli studenti a numeri dispari). Un attaccante porterà completamente a segno l'attacco di reidentificazione nel caso in cui riesca a recuperare il nome, il cognome e il genere di uno studente.

4.3.3 Discriminazione

Un attacco di discriminazione è finalizzato a identificare le proprietà del titolare di uno pseudonimo (almeno una). Tali proprietà potrebbero non portare direttamente a individuare l'identità del titolare dello pseudonimo, ma potrebbero essere sufficienti a differenziarlo in qualche modo.

Considerando l'esempio dei voti degli studenti illustrato poc'anzi, il relativo set di dati può contenere, come pseudonimi, due numeri pari tra molti numeri dispari. I numeri pari corrispondono alle studentesse mentre i numeri dispari sono associati agli studenti (l'attaccante è a conoscenza di tale elemento). Entrambi i numeri pari hanno ottenuto il 100 % come risultato all'esame finale. Supponiamo inoltre che nel set di dati pseudonimizzato non vi siano altri studenti che abbiano ottenuto il 100 %. Nel caso in cui l'attaccante apprenda poi che un

determinato studente ha ottenuto il 100 % in questo corso, verrà immediatamente a sapere che tale studente era di sesso femminile. Viceversa, se l'attaccante apprende che uno studente di quel corso era di sesso femminile, verrà immediatamente a sapere che tale studentessa aveva ottenuto il 100 %. Occorre sottolineare che, in questo caso, l'attaccante non apprende l'identità del titolare di uno pseudonimo, ma viene esclusivamente a conoscenza di alcune sue proprietà (ovvero il valore del genere o del voto). Dato che diversi studenti condividono la medesima combinazione di proprietà, l'attaccante non è in grado di individuare l'esatto record di dati di un particolare titolare di pseudonimo. L'acquisizione di tali informazioni aggiuntive potrebbe già bastare all'attaccante per eseguire la discriminazione o potrebbe favorire un successivo attacco mirato ad acquisire conoscenze generali per risalire all'identità celata dietro a uno pseudonimo.

4.4 PRINCIPALI TECNICHE DI ATTACCO

Esistono tre principali tecniche per decodificare una funzione di pseudonimizzazione: attacchi a forza bruta (ricerca esaustiva), ricerca in un dizionario e inferenze[17]. L'efficacia di tali attacchi dipende da diversi parametri, tra cui:

- la quantità di informazioni sul titolare dello pseudonimo (l'interessato) contenute nello pseudonimo;
- le conoscenze generali dell'attaccante;
- la dimensione del dominio dell'identificativo;
- la dimensione del dominio dello pseudonimo;
- la scelta e la configurazione della funzione di pseudonimizzazione utilizzata (che include la dimensione della chiave di pseudonimizzazione).

Vengono di seguito descritte le sopraindicate tecniche di attacco.

4.4.1 Attacco a forza bruta

La funzionalità di tale tecnica di attacco è subordinata alla capacità dell'attaccante di calcolare la funzione di pseudonimizzazione (ovvero, non vi è alcuna chiave di pseudonimizzazione) o dal suo accesso a un'implementazione black box della funzione di pseudonimizzazione. A seconda dell'obiettivo dell'attacco, potrebbero essere applicabili ulteriori condizioni. Se l'attacco a forza bruta serve a ottenere una reidentificazione completa (ovvero l'individuazione dell'identità originale), il dominio dell'identificativo deve essere finito e relativamente piccolo. L'attaccante, per ogni pseudonimo incontrato, può tentare di recuperare l'identificativo originale applicando la funzione di pseudonimizzazione su ciascun valore del dominio dell'identificativo, fino a quando non trova una corrispondenza.

Tabella 1: Pseudonimizzazione del mese di nascita

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

[17] Va notato che, come indicato in precedenza, possono essere utilizzati anche altri attributi (oltre allo pseudonimo e ai dati pseudonimizzati) per identificare un individuo. Per approfondimenti, cfr. il capitolo 8.

Consideriamo la pseudonimizzazione di un mese di nascita in un set di dati. La dimensione del dominio dell'identificativo corrisponde a 12, pertanto un attaccante può rapidamente elencare tutte le possibilità. In questo caso, gli pseudonimi associati a ogni mese vengono calcolati come la somma del codice ASCII delle prime tre lettere del mese di nascita (con l'iniziale maiuscola). Consideriamo che un attaccante si sia imbattuto nello pseudonimo 301. Può applicare su ogni mese di nascita la funzione di pseudonimizzazione fino a quando non trova il mese che corrisponde al valore 301. La Tabella 1 mostra i calcoli effettuati dall'attaccante per reidentificare lo pseudonimo 301, dando luogo alla tabella di mappatura della funzione di pseudonimizzazione.

Ovviamente, affinché l'attacco vada a buon fine, la dimensione del dominio degli identificativi risulta essere decisiva. In caso di domini di identificativi di piccole dimensioni, come nell'esempio sopraindicato, è assai semplice eseguire un attacco a forza bruta. In caso di domini di identificativi di grandezza infinita, gli attacchi a forza bruta diventano generalmente impraticabili. Se la dimensione del dominio degli identificativi è troppo grande, la reidentificazione completa risulta estremamente difficile e l'attaccante può solamente eseguire un attacco di discriminazione.

In questo caso, l'attaccante può infatti considerare un sottodominio del dominio degli identificativi per il quale calcolare tutti gli pseudonimi. Ritorniamo all'esempio della tabella 1, supponendo che il dominio sia di piccole dimensioni. Supponiamo che l'attaccantevoglia differenziare le persone con un mese di nascita che comincia con la lettera G da quelle che hanno un mese di nascita che inizia con una lettera diversa. Tale sottodominio contiene gennaio e giugno. L'attaccante può effettuare una ricerca esaustiva su questo sottodominio calcolando gli pseudonimi corrispondenti a gennaio e a giugno. Se trova uno pseudonimo diverso da 281 o 301, allora viene a sapere che il mese di nascita non inizia con la lettera G.

Nel caso in cui si utilizzasse una chiave di pseudonimizzazione, basterebbe un dominio dell'identificativo di piccole dimensioni per impedire l'attacco (poiché l'attaccante non sarebbe in grado di calcolare la funzione di pseudonimizzazione e purché non vi sia accesso a un'implementazione «black box» di tale funzione). In tal caso, è possibile eseguire un attacco a forza bruta sull'intero spazio delle chiavi di pseudonimizzazione: in altre parole l'attaccante verifica in modo esaustivo tutti i possibili segreti, andando a calcolare per ciascuno di essi la funzione di recupero. Tale attacco avrà buon esito qualora l'attaccante indovinasse la chiave di pseudonimizzazione, indipendentemente dalle dimensioni del dominio dell'identificativo. Pertanto, per contrastare un simile attacco, il numero delle potenziali chiavi di pseudonimizzazione deve essere grande a sufficienza, così da renderlo praticamente impossibile.

4.4.2 Ricerca in un dizionario

La ricerca nel dizionario è un'ottimizzazione dell'attacco a forza bruta, in quanto può consentire di risparmiare sui costi computazionali. L'attaccante infatti, per effettuare una completa reidentificazione o discriminazione, deve confrontarsi con una grande quantità di pseudonimi. Di conseguenza, calcola preliminarmente un (enorme) insieme di pseudonimi e salva il risultato in un dizionario. Ogni voce nel dizionario contiene uno pseudonimo associato all'identificativo o alle informazioni corrispondenti. Ogni volta che l'attaccante deve reidentificare uno pseudonimo, andrà a ricercarlo nel dizionario. Tale ricerca ha un costo precedente al calcolo pari a quello di una ricerca esaustiva e memorizza il risultato in una memoria capiente. La reidentificazione di uno pseudonimo presenta unicamente il costo di una ricerca nel dizionario. La ricerca nel dizionario è fondamentalmente il calcolo e l'archiviazione della tabella di mappatura. Utilizzando le tabelle di Hellman [6] o le rainbow table [7] è perfino possibile raggiungere compromessi

tempo/memoria, al fine di ampliare ulteriormente la gamma. Tuttavia, tale attacco presenta specifiche varianti, in cui vengono utilizzate informazioni aggiuntive sul modo in cui opera la funzione di pseudonimizzazione. Tali attacchi possono funzionare anche in caso di domini di input infiniti.

4.4.3 Inferenze

Questo tipo di attacco si basa su alcune conoscenze generali (quali la distribuzione di probabilità o qualsiasi altra informazione aggiuntiva) di cui l'attaccante può disporre relativamente ad alcuni (o tutti) i titolari di pseudonimo, alla funzione di pseudonimizzazione o al set di dati. La ricerca esaustiva e la ricerca in un dizionario presuppongono implicitamente che tutti gli identificativi presentino la stessa probabilità o frequenza di occorrenze. Alcuni identificativi possono tuttavia essere più frequenti di altri. Servirsi delle caratteristiche statistiche degli identificativi è un processo noto come «inferenza» [8], [9], [10] ed è ampiamente adoperato da chi viola le password. Si noti che è possibile applicare le inferenze anche in caso di domini degli identificativi di enormi dimensioni. L'attaccante non deve necessariamente avere accesso alla funzione di pseudonimizzazione (poiché è possibile eseguire la discriminazione anche tramite una semplice analisi della frequenza degli pseudonimi osservati).

Consideriamo un caso riguardante pseudonimi corrispondenti a «nomi di persona». È difficile analizzare integralmente il dominio dei «nomi di persona». Tuttavia, l'attaccante sa quali sono i «nomi di persona» più diffusi (Tabella 2) e può effettuare una ricerca esaustiva o una ricerca nel dizionario sul dominio dei «nomi di persona» più utilizzati, così da ottenere la discriminazione.

Tabella 2. Elenco di nomi di persona diffusi

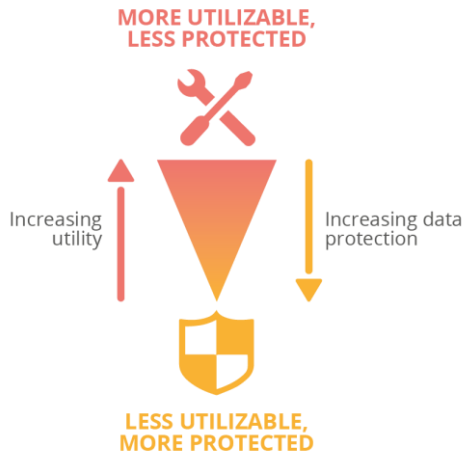
Nomi di persona più diffusi					
Leonardo	Sofia	Francesco	Giulia	Alessandro	Aurora

Supponiamo un caso simile, ma con un dominio di identificativi infinito. È possibile definire un sottodominio finito di identificativi inclusi nel set di dati. Se l'attaccante riesce a indovinare tale sottodominio, può eseguire una ricerca esaustiva (cfr. il capitolo 6 per il relativo caso d'uso sulla pseudonimizzazione dell'indirizzo e-mail). A seconda della quantità di informazioni generali o di metadati detenuti dall'attaccante e della quantità di informazioni associabili trovate nel set di dati pseudonimizzato, questo tipo di attacco può consentire di scoprire l'identità di un singolo possessore di pseudonimo, di una parte di essi o dell'intero set di dati. Soprattutto per set di dati di piccole dimensioni, tali attacchi possono presentare alte percentuali di successo.

4.5 FUNZIONALITÀ E PROTEZIONE DEI DATI

A seconda della funzione di pseudonimizzazione prescelta, uno pseudonimo può contenere alcune informazioni sull'identificativo originale. Ogni tipo di pseudonimo può pertanto essere soggetto a un attacco di reidentificazione come quelli sopra descritti. Per es., un attaccante con sufficienti conoscenze generali potrebbe essere in grado di riassociare lo pseudonimo al suo identificativo effettuando un'inferenza.

Figura 7. Funzionalità e protezione dei dati



In molti casi, tuttavia, le informazioni aggiuntive sull'identificativo originale contenute nello pseudonimo vengono conservate per effettuare un'associazione tra gli stessi pseudonimi, che deve essere eseguita da un titolare del trattamento valido. Per es., uno pseudonimo può mantenere l'anno della data di nascita di una persona come parte dello pseudonimo (per es. «AAAA-1999»). È così possibile classificare gli pseudonimi sulla base dell'anno di nascita, per es. rispetto all'età, allo stato giuridico (bambino o adulto), alle condizioni di vita (in età scolastica/lavoratore/pensionato) o simili. Potrebbe trattarsi di una caratteristica intenzionale della funzione di pseudonimizzazione impiegata, così da consentire ai titolari del trattamento di effettuare tale classificazione anche su dati pseudonimizzati.

Chiaramente, la scelta della funzione di pseudonimizzazione, se da un lato può rafforzare l'funzionalità degli pseudonimi creati, può dall'altro indebolire la protezione, a causa di tale approccio. Si potrebbe dunque considerare un compromesso tra funzionalità e protezione dei dati (cfr. la figura 7). Nel momento in cui si applica la pseudonimizzazione a scenari del mondo reale, occorrerebbe analizzare questo compromesso con la debita attenzione, così da ottimizzare l'funzionalità per gli scopi prefissati e al contempo proteggere quanto più possibile i possesssori di uno pseudonimo (gli interessati).

5. TECNICHE DI PSEUDONIMIZZAZIONE

In base ai modelli di attaccante e ai tipi di attacchi illustrati nel capitolo 4, il presente capitolo passa rapidamente in rassegna le tecniche e le politiche di pseudonimizzazione ad oggi più comuni. Per un'analisi più dettagliata delle primitive crittografiche, cfr. [1].

In linea di principio, una funzione di pseudonimizzazione associa gli identificativi agli pseudonimi. Una funzione di pseudonimizzazione deve presentare un requisito fondamentale. Consideriamo due diversi identificativi $ID1$ e $ID2$ e i loro corrispondenti pseudonimi $pseudo1$ e $pseudo2$. Una funzione di pseudonimizzazione deve verificare che $pseudo1$ sia differente da $pseudo2$. In caso contrario, il recupero dell'identificativo risulterebbe ambiguo: l'entità di pseudonimizzazione non potrebbe infatti determinare se $pseudo1$ corrisponde a $ID1$ o a $ID2$. Tuttavia, un singolo identificativo ID può essere associato a più pseudonimi ($pseudo1$, $pseudo2\ldots$) purché l'entità di pseudonimizzazione sia in grado di invertire tale operazione. In tutti i casi, secondo la definizione di pseudonimizzazione (cfr. il capitolo 2), esistono alcune informazioni aggiuntive che consentono l'associazione degli pseudonimi con gli identificativi originali; si tratta della cosiddetta chiave di pseudonimizzazione. Il caso più semplice di chiave di pseudonimizzazione consiste nella tabella di mappatura della pseudonimizzazione.

Nelle seguenti sezioni vengono innanzitutto definite le principali opzioni per pseudonimizzare un singolo identificativo. Si passa quindi a descrivere le diverse politiche di pseudonimizzazione, mettendo a confronto le rispettive caratteristiche di implementazione. Si fa inoltre riferimento ai principali criteri su cui può basarsi un titolare del trattamento per scegliere una tecnica di pseudonimizzazione. Vengono infine esaminate le possibilità di recupero della pseudonimizzazione da parte della relativa entità.

5.1 PSEUDONIMIZZAZIONE DI UN SINGOLO IDENTIFICATORE

Partendo dalla pseudonimizzazione di un singolo identificativo, vengono di seguito elencati alcuni possibili approcci, con relativi vantaggi e limiti.

5.1.1 Contatore

Il contatore è la più semplice forma di pseudonimizzazione. Gli identificativi sono sostituiti da un numero scelto da un contatore monotono. In primo luogo, viene impostato un seme s su 0 (a titolo esemplificativo), che viene poi incrementato. Onde evitare ambiguità, è fondamentale che i valori prodotti dal contatore non si ripetano mai.

I vantaggi del contatore derivano dalla sua semplicità, che lo rende un buon candidato per set di dati non complessi e di piccole dimensioni. In termini di protezione dei dati, il contatore fornisce pseudonimi che non sono associabili agli identificativi iniziali (sebbene il carattere sequenziale del contatore possa comunque fornire informazioni sull'ordine dei dati all'interno di un set). Tale soluzione può tuttavia presentare problemi di implementazione e scalabilità in caso di set di dati più sofisticati e di grandi dimensioni, poiché occorrerebbe in tal caso archiviare la tabella completa di mappatura della pseudonimizzazione.

5.1.2 Generatore di numeri casuali

Il generatore di numeri casuali è un meccanismo che produce, all'interno di un set, valori che presentano tutti la stessa probabilità di essere selezionati, risultando pertanto imprevedibili ([18]). Questo approccio è simile al contatore, con la differenza che all'identificativo viene assegnato un numero casuale. Per creare tale mappatura, vi sono due opzioni: un generatore di numeri casuali vero e proprio o un generatore di numeri pseudo casuali crittograficamente sicuro (cfr. [11] per le definizioni esatte). Occorre notare che, in entrambi i casi, è possibile incorrere in collisioni, se non si presta la dovuta attenzione ([19]). Si verifica una collisione nel momento in cui due identificativi vengono associati al medesimo pseudonimo. La probabilità di incorrere in una collisione dipende dal noto paradosso del compleanno [12].

Il generatore di numeri casuali fornisce una solida protezione dei dati (poiché, a differenza del contatore, per creare ogni pseudonimo si va a utilizzare un numero casuale, rendendo difficile l'estrazione di informazioni riguardanti l'identificativo iniziale, a meno che la tabella di mappatura non sia stata compromessa). Come menzionato in precedenza, le collisioni possono rappresentare un problema, unitamente alla scalabilità (occorre memorizzare la completa tabella di mappatura della pseudonimizzazione), a seconda dello scenario di implementazione.

5.1.3 Funzione crittografica di hash

Una funzione crittografica di hash prende stringhe di input di lunghezza arbitraria e le associa ad output di lunghezza fissa [13] [14]. Essa presenta le proprietà riportate di seguito.

- Unidirezionale: è computazionalmente impraticabile trovare input che si associno a output specificati in precedenza.
- Senza collisioni: è computazionalmente impraticabile trovare due input distinti che si associno al medesimo output.

Si applica una funzione crittografica di hash direttamente all'identificativo, così da ottenere lo pseudonimo corrispondente: $Pseudo = H(ID)$. Il dominio dello pseudonimo dipende dalla lunghezza del digest (output) prodotto dalla funzione.

Come indicato in [1], se da un lato una funzione di hash può contribuire in modo significativo all'integrità dei dati, dall'altro viene in genere considerata una tecnica di pseudonimizzazione debole, in quanto soggetta ad attacchi a forza bruta e di ricerca nel dizionario. I capitoli 6, 7 e seguenti riportano specifici esempi di tale vulnerabilità.

5.1.4 Codice di autenticazione del messaggio

Questa primitiva può essere considerata come una funzione di hash con chiave. È molto simile alla soluzione precedente, salvo per l'introduzione di una chiave segreta atta a generare lo pseudonimo. Se non si è a conoscenza di tale chiave, non è possibile associare gli identificativi agli pseudonimi. HMAC [15] [16] è di gran lunga la più diffusa modalità di codice di autenticazione del messaggio impiegata nei protocolli Internet.

Come indicato in [1], il Codice di autenticazione del messaggio è generalmente considerato una tecnica di pseudonimizzazione solida dal punto di vista della protezione dei dati poiché, a meno che la chiave non sia stata compromessa, è impossibile decodificare lo pseudonimo. Possono essere applicabili diverse varianti di tale metodo, che presentano differenti requisiti di funzionalità e scalabilità dell'entità di pseudonimizzazione (per esempi più specifici, cfr. i capitoli 6 e 7 di seguito).

([18]) Si noti che, invece dei numeri, è possibile utilizzare una sequenza casuale di caratteri.
([19]) Il rischio di collisioni si fa minimo nel momento in cui vengono generati numeri pseudo casuali molto grandi (per es. di 100 cifre).

5.1.5 Crittografia

Il presente rapporto prende in esame soprattutto la crittografia simmetrica (deterministica) e, in particolare, le cifrature a blocchi come l'AES, insieme alle loro modalità operative [11]. Si utilizza una cifratura a blocchi per crittografare un identificativo servendosi di una chiave segreta, che è sia chiave di pseudonimizzazione sia la chiave da impiegare per il recupero. L'uso di cifrature a blocchi ai fini della pseudonimizzazione deve misurarsi con la dimensione del blocco. Gli identificativi possono avere dimensioni minori o maggiori rispetto alla dimensione del blocco di input della cifratura a blocchi. In caso di identificativi di dimensioni inferiori, occorre considerare la possibilità del riempimento [11]. Nel caso in cui la dimensione degli identificativi sia invece maggiore della dimensione del blocco, per risolvere il problema si può scegliere tra due opzioni: comprimere gli identificativi, così che la loro dimensione diventi inferiore a quella del blocco, oppure, se non si può procedere alla compressione, utilizzare una modalità operativa (come la modalità Counter, CTR). Quest'ultima opzione, tuttavia, presuppone la gestione di un parametro extra: il vettore di inizializzazione.

Come indicato in [1], la crittografia può essere impiegata anche come una solida tecnica di pseudonimizzazione, che presenta alcune proprietà simili al Codice di autenticazione del messaggio. I capitoli 6 e 7 offrono a tal riguardo alcuni esempi specifici.

Sebbene il presente rapporto si incentri principalmente su schemi di crittografia deterministica, vi è l'alternativa della crittografia probabilistica, valida soprattutto nei casi in cui occorra ricavare pseudonimi diversi per lo stesso identificativo (cfr. la politica di pseudonimizzazione completamente randomizzata illustrata di seguito). Per un approfondimento, cfr. [1].

5.2 STRATEGIE DI PSEUDONIMIZZAZIONE

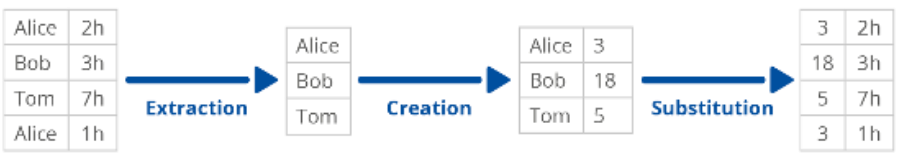
Se la scelta della tecnica di pseudonimizzazione risulta essenziale, è di pari importanza, ai fini della sua applicazione pratica, la relativa strategia (o modalità) di implementazione.

La presente sezione prende in esame la questione più generale della pseudonimizzazione di un database o di qualsiasi documento contenente k identificativi. Consideriamo un identificativo ID che appare più volte in due set di dati A e B. Successivamente alla pseudonimizzazione, l'identificativo ID viene sostituito in base a una delle seguenti stategie: pseudonimizzazione deterministica, randomizzata al documento e completamente randomizzata.

5.2.1 Pseudonimizzazione deterministica

In tutti i database e ogniqualvolta appare, ID viene sempre sostituito con lo stesso pseudonimo $pseudo$. Esso è uniforme all'interno di un database e tra database differenti. Per implementare tale modalità, occorre anzitutto estrarre l'elenco degli identificativi univoci contenuti nel database. In secondo luogo, l'elenco viene associato agli pseudonimi e gli identificativi sono infine sostituiti agli pseudonimi nel database (cfr. la Figura 8).

Figura 8. Pseudonimizzazione deterministica

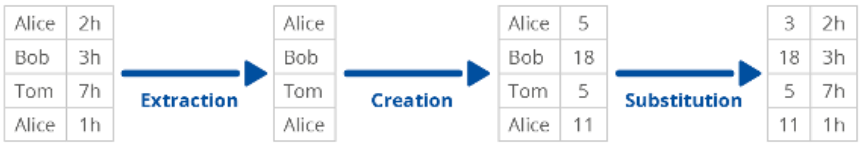


Per implementare la pseudonimizzazione deterministica possono essere impiegate tutte le tecniche indicate nel capitolo 5.1.

5.2.2 Pseudonimizzazione randomizzata al documento

Ogniqualvolta ID appare in un database, viene sostituito con un differente pseudonimo ($pseudo_1$, $pseudo_2$, e così via). Tuttavia, ID viene sempre associato alla medesima raccolta di ($pseudo_1$, $pseudo_2$) nel set di dati A e B.

Figura 9. Pseudonimizzazione randomizzata al documento



In questo caso, la pseudonimizzazione risulta uniforme solo tra database differenti. Si crea la tabella di mappatura adoperando tutti gli identificativi presenti nel database. Ogni occorrenza di un determinato identificativo (ossia Sofia nella Figura 9) viene trattata in maniera indipendente.

5.2.3 Pseudonimizzazione completamente randomizzata

Infine, per ogni occorrenza di ID all'interno di un database A o B, ID viene sostituito con uno pseudonimo differente ($pseudo_1$, $pseudo_2$). Questo caso è un esempio di pseudonimizzazione completamente randomizzata. È possibile considerare tale strategia come un'ulteriore estensione della pseudonimizzazione randomizzata al documento. Effettivamente, qualora applicate a un singolo documento, queste due strategie si comportano allo stesso modo. Tuttavia, se lo stesso documento viene pseudonimizzato due volte mediante una pseudonimizzazione completamente randomizzata, si ottengono due output diversi. Con la pseudonimizzazione randomizzata al documento, si ottiene invece due volte lo stesso output. In altre parole, nella pseudonimizzazione randomizzata al documento la casualità è selettiva (per es., solo per Sofia), mentre nella pseudonimizzazione completamente randomizzata la casualità è globale (si applica a qualsiasi record di dati).

5.3 SCEGLIERE UNA TECNICA E UNA STRATEGIA DI PSEUDONIMIZZAZIONE

La scelta di una tecnica e di una strategia di pseudonimizzazione dipende da diversi parametri, in primo luogo dal livello di protezione dei dati e dall'funzionalità del set di dati pseudonimizzato (che l'entità di pseudonimizzazione intende raggiungere). In termini di protezione, come esaminato nelle sezioni precedenti, l'RNG (random number generator), i codici di autenticazione del messaggio e la crittografia rappresentano le tecniche più efficaci, appositamente mirate a contrastare ricerche esaustive, ricerche nel dizionario e congetture. Tuttavia, a seconda dei requisiti di funzionalità, l'entità di pseudonimizzazione potrebbe optare per una combinazione di differenti approcci o per delle varianti dell'approccio selezionato. Allo stesso modo, relativamente alle strategie di pseudonimizzazione, la pseudonimizzazione completamente randomizzata offre il miglior livello di protezione, ma non consente alcun confronto tra database. Le funzioni di pseudonimizzazione randomizzata al documento e deterministica offrono invece funzionalità, ma consentono anche l'associabilità tra record di dati. È possibile applicare soluzioni specifiche in base agli identificativi da pseudonimizzare (per esempi più specifici, cfr. i capitoli 6 e 7).

Inoltre, la complessità di un determinato schema in termini di implementazione e scalabilità potrebbe destare preoccupazione all'entità di pseudonimizzazione, portandola a interrogarsi sul grado di difficoltà della pseudonimizzazione applicata agli identificativi e sulla possibilità che quest'ultima influisca sulla dimensione del database.

Tabella 3. Confronto tra differenti tecniche in termini di flessibilità (formato dell' identificativo) e dimensione dello pseudonimo

Metodo	Dimensione dell' identificativo	Dimensione m dello pseudonimo in bit
Contatore	Qualsiasi	$m = log_2$k
Generatore di numeri casuali	Qualsiasi	$m \gg 2log_2k$
Funzione di hash	Qualsiasi	Fissa o $m \gg 2log_2k$
Codici di autent. del messaggio	Qualsiasi	Fissa o $m \gg 2log_2k$
Crittografia	Fissa ([20])	Fissa o uguale all'identificativo

Per identificativi di dimensione variabile è possibile adottare la maggior parte delle soluzioni, eccetto per determinate scelte in caso di crittografia. La dimensione di uno pseudonimo dipende da k, il numero degli identificativi presenti nel database. Per generatori di numeri casuali, funzioni di hash e codice di autenticazione del messaggio, vi è il rischio di collisioni: occorre dunque scegliere con cura la dimensione dello pseudonimo (cfr. il paradosso del compleanno). Le funzioni di hash e il codice di autenticazione del messaggio sono appositamente concepiti per assicurare che la dimensione del digest prevenga qualunque rischio di collisioni. Infine, la dimensione dello pseudonimo generato da uno schema di crittografia può essere fissa o corrispondere a quella dell'identificativo originale. La Tabella 3 mostra la scalabilità degli approcci sopra enunciati rispetto alla funzione di recupero.

5.4 RECUPERO

Poiché, per definizione, l'uso di informazioni aggiuntive è fondamentale per la pseudonimizzazione, l'entità di pseudonimizzazione è chiamata a implementare un meccanismo di recupero. Questo meccanismo può essere più o meno complesso, a seconda della funzione di pseudonimizzazione. Esso consiste in genere nell'utilizzare uno pseudonimo $pseudo$ e un segreto di pseudonimizzazione S, per recuperare l'identificativo corrispondente ID. Ciò può per es. verificarsi quando l'entità di pseudonimizzazione ha rilevato un'anomalia nel suo sistema e deve contattare le entità designate. Tale «anomalia» può corrispondere, per es., a una violazione dei dati, e l'entità pseudonimizzazione è tenuta a informarne gli interessati, ai sensi del RGPD. Questo meccanismo di recupero potrebbe rivelarsi necessario anche per consentire agli interessati di esercitare i propri diritti (ai sensi degli articoli 12-21 del RGPD).

Tabella 4. Confronto tra differenti tecniche in base al meccanismo di recupero

Metodo	Recupero basato su pseudonimo
Contatore	Tabella di mappatura
Generatore di numeri casuali	Tabella di mappatura
Funzione di hash	Tabella di mappatura
Codici di autent. del messaggio	Tabella di mappatura
Crittografia	De-crittografia

([20]) La crittografia che utilizza una cifratura a blocchi lavora con input di dimensione fissa. Alcune modalità di funzionamento (come il CTR) consentono tuttavia di lavorare su input di qualsiasi dimensione.

La maggior parte dei metodi in precedenza illustrati richiede che l'entità di pseudonimizzazione mantenga la tabella di mappatura tra gli identificativi e gli pseudonimi, per eseguire il recupero dell'identificativo, ad eccezione della crittografia (Tabella 4). In effetti, è possibile applicare la crittografia direttamente all'identificativo.

5.5 PROTEZIONE DELLA CHIAVE DI PSEUDONIMIZZAZIONE

Affinché la pseudonimizzazione sia efficiente, l'entità di pseudonimizzazione deve sempre proteggere la chiave di pseudonimizzazione mediante adeguate misure tecniche e organizzative. Ciò dipende chiaramente dallo specifico scenario di pseudonimizzazione (cfr. il capitolo 3).

In primo luogo, occorre isolare la chiave di pseudonimizzazione dal set di dati, non dovendo mai, per es., essere gestiti in uno stesso file (in caso contrario, un attaccante potrà facilmente recuperare gli identificativi). In secondo luogo, occorre eliminare in modo sicuro la chiave di pseudonimizzazione da qualsiasi supporto non sicuro (memoria e sistemi). Terzo, bisogna far sì che solide politiche di controllo dell'accesso assicurino che solo le entità autorizzate possano accedere a tale chiave. Occorre inoltre un sistema di accesso sicuro, che tenga traccia di tutte le richieste di accesso alla chiave. Infine, la chiave di pseudonimizzazione, se memorizzata in un computer, deve essere crittografata. Il computer, per garantire tale crittografia, necessita a sua volta di una corretta gestione e archiviazione delle chiavi.

5.6 TECNICHE AVANZATE DI PSEUDONIMIZZAZIONE

Oltre alle sopra elencate tecniche di pseudonimizzazione, ve ne sono altre più avanzate, adatte a diversi contesti. Illustrarle in dettaglio una ad una eccederebbe l'ambito della presente relazione, pertanto si procederà a una rapida rassegna di tali tecniche, per quei lettori che fossero interessati.

Oltre al semplice hashing di dati, strutture più avanzate quali gli alberi Merkle [17, 18] utilizzano hash di set di hash, per es. $h_3=\text{hash}(h_1,h_2)$, per ottenere pseudonimi ben articolati, che possono essere solo parzialmente identificati. Analogamente, le catene di hash [19] si basano sull'hashing ripetuto dei valori di hash, per es. $h_4=h_3(h_2(h_1(x)))$, per produrre un valore che, per reidentificare i dati originali di un determinato pseudonimo, richiede più operazioni di hashing inverso. Un esempio di tale tecnica di hashing è dato dalla catena di pseudonimizzazione, che coinvolge diverse entità di pseudonimizzazione, che vanno a prendere gli pseudonimi creati dalla precedente entità come input per crearne di nuovi (per es., applicando un ulteriore livello di hash). Una catena simile rimarrà valida anche qualora un attaccante riuscisse a scoprire tutte le pseudonimizzazioni, tranne una, che sono state applicate nella catena totale, il che la rende una tecnica di pseudonimizzazione particolarmente robusta. Si tratta inoltre di una pratica diffusa, adottata tra le altre cose negli studi clinici.

Se il dominio di input si estende su più dimensioni (cfr. il capitolo 8, per un esempio), i filtri di Bloom [20], oltre a essere utilizzati come tecnica di anonimizzazione, possono essere impiegati anche per eseguire in modo efficiente una pseudonimizzazione computazionalmente praticabile su tutte le possibili combinazioni di valori di input su domini diversi, nonostante il problema dell'esplosione combinatoria.

Rappresentano ulteriori approcci degni di interesse gli pseudonimi di transazione collegabili e/o l'associabilità di pseudonimi controllati con l'opzione di re-identificazione progressiva [21].

Per concludere, tutte le tecniche atte ad aumentare efficacemente l'anonimizzazione possono risultare utili anche per la pseudonimizzazione, come nel caso delle comuni tecniche per il k-anonimato [3, 22, 23], per la privacy differenziale, ecc. [25] Per le relative descrizioni, cfr. [2].

Tra le altre soluzioni degne di interesse, figurano la dimostrazione a conoscenza zero [26] e il più vasto ambito delle credenziali basate su attributi [2].

6. PSEUDONIMIZZAZIONE DEGLI INDIRIZZI IP

Servendoci delle tecniche e delle informazioni precedentemente illustrate, in questo capitolo viene presentato uno specifico caso d'uso relativo alla pseudonimizzazione degli indirizzi IP.

Un indirizzo IP serve a identificare in modo univoco un dispositivo all'interno di una rete IP. Si distinguono due tipi di indirizzi IP: IPv4 [27] e IPv6 [28]. Nel presente caso d'uso, il rapporto si concentra su IPv4, che è il più comune; applicare i concetti sopra enunciati a IPv6 si rivelerebbe infatti piuttosto complesso, eccedendo l'ambito di questo documento. Un indirizzo IPv4 consiste in 32 bit (128 bit nel caso di IPv6) suddivisi in un prefisso di rete (byte più significativi) e in un identificativo host (byte meno significativi) con l'aiuto di una maschera di sottorete. Vengono spesso rappresentati tramite notazione decimale puntata, composta da 4 numeri decimali tra 0 e 255 separati da punti, come 127.0.0.1. La dimensione del prefisso di rete e dell'identificativo di host dipende da quella del blocco CIDR (Classless Inter-Domain Routing [29]). Vi sono inoltre indirizzi IP speciali, come 127.0.0.1 (localhost) o 224.0.01. (multicast). Questi indirizzi speciali sono definiti in [30] e vengono suddivisi in 15 classi.

L'Autorità per l'assegnazione dei numeri per Internet (IANA) gestisce l'intero spazio degli indirizzi IP con l'ausilio di cinque registri regionali di internet (Regional Internet Registry, RIR). Questi assegnano sottoinsiemi di indirizzi IP a organizzazioni locali quali fornitori di servizi Internet (ISP), che a loro volta assegnano gli indirizzi ai dispositivi degli utenti finali. L'assegnazione di ogni indirizzo IP è documentata dal RIR corrispondente nel cosiddetto database WHOIS ([21]). L'assegnazione può essere statica oppure dinamica (come nel caso in cui venga impiegato un DHCP (protocollo di configurazione IP dinamica)).

Lo status giuridico degli indirizzi IP è stato dibattuto dalla Corte di giustizia dell'Unione europea nell'ambito della causa C-582/14 Breyer contro la Repubblica Federale di Germania ([22]). Che siano statici o dinamici, gli indirizzi IP vengono comunque considerati dati personali. Ciò ha trovato peraltro conferma nel parere 4/2007 sul concetto di dati personali espresso dal Gruppo dell'articolo 29 per la tutela dei dati [31]. Le tracce di database o di rete contenenti indirizzi IP devono essere pertanto protette, e la pseudonimizzazione rappresenta ovviamente una funzione di protezione, che da un lato consente l'uso di indirizzi IP, e dall'altro impedisce la loro associabilità a individui specifici. Ciò detto, per scegliere la tecnica di pseudonimizzazione per gli indirizzi IP più opportuna, occorre trovare un buon compromesso tra funzionalità e protezione dei dati. In effetti, anche in questo caso il titolare del trattamento potrebbe aver bisogno di calcolare statistiche o rilevare modelli (in caso di errata configurazione di un dispositivo o per la qualità dei servizi) nel database pseudonimizzato. Se in ambito applicativo funzionalità e protezione dei dati non possono essere considerate separatamente, esse vengono tuttavia separate per favorire una migliore comprensione.

6.1 PSEUDONIMIZZAZIONE E LIVELLO DI PROTEZIONE DEI DATI

Il principale problema connesso alla pseudonimizzazione di indirizzi IP è rappresentato dalla dimensione dello spazio di input (dominio dell'identificativo), dal momento che ci sono solo 2^{32} indirizzi IP possibili. Ciò può consentire a un eventuale attaccante di eseguire una ricerca

([21]) Per maggiori informazioni, cfr. https://whois.icann.org
[22] Per maggiori informazioni, cfr.: https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX%3A62014CJ0582

esaustiva e una ricerca nel dizionario, così da muovere attacchi di reidentificazione completa o di discriminazione, qualora la funzione di pseudonimizzazione non sia stata scelta in modo adeguato.

Considerate le caratteristiche sopra enunciate, in questo caso d'uso le funzioni crittografiche di hash si rivelano particolarmente vulnerabili. Si è considerato l'esempio di un indirizzo IP pseudonimizzato con la funzione di hash SHA-256. Un attaccante provvisto di uno pseudonimo/digest può servirsi di strumenti già esistenti ([23]) per effettuare una ricerca esaustiva. La Tabella 5 mostra la durata di tale ricerca su un semplice computer portatile che monta un processore Intel(R) Core(TM) i7-8650U, CPU a 1.90GHz (8 core), nonché la dimensione del dizionario. Anche nel peggiore dei casi, ci vogliono solamente due minuti circa per recuperare l'indirizzo IP associato a un determinato pseudonimo.

Tabella 5. Costi pratici di un attacco contro la pseudonimizzazione con funzione di hash

Classe IP	Numero di possibili IP	Tempo di ricerca esaustiva	Dimensione del 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

Supponiamo inoltre che l'intruso intenda determinare se uno pseudonimo corrisponde a uno specifico indirizzo [30]. Il suo attacco di discriminazione non dovrà essere effettuato su tutti i 2^{32} indirizzi IP possibili, ma solo sugli specifici indirizzi IP 588.518.401.

Il caso summenzionato mostra come la pseudonimizzazione degli indirizzi IP che impiega unicamente la funzione crittografica di hash non sia efficace. Per la protezione dei dati vanno dunque preferite altre funzioni di pseudonimizzazione, come il codice di autenticazione del messaggio, la crittografia con chiave segreta ad hoc o il generatore di numeri casuali. Come in precedenza affermato, un attaccante non può eseguire gli stessi attacchi, in quanto questi metodi utilizzano una chiave segreta (Codice di autenticazione del messaggio e crittografia) oppure una fonte di casualità (per l'RNG). Può essere impiegato anche un contatore, ma occorre prestare attenzione alle possibili previsioni (dovute alla sua natura sequenziale).

6.2 PSEUDONIMIZZAZIONE E LIVELLO DI FUNZIONALITÀ

Come già ribadito, in caso di indirizzi IP l'funzionalità può rappresentare un requisito essenziale per l'entità di pseudonimizzazione, come per il calcolo di statistiche o per la sicurezza della rete. Di conseguenza, l'approccio adottato (indipendentemente dalla tecnica prescelta) deve favorire un livello adeguato di protezione pur preservando alcune informazioni utili di base (ricavabili dagli indirizzi IP). Nella presente sezione, verranno prese in esame due dimensioni connesse a questo tema: la possibilità di minimizzare il livello/campo applicativo della pseudonimizzazione degli indirizzi IP e la scelta della stategia di pseudonimizzazione (modalità).

6.2.1 Livello di pseudonimizzazione

Nella sezione precedente, si è presupposto che la pseudonimizzazione fosse applicata all'intero indirizzo IP (32 bit). Tuttavia, per accrescere l'funzionalità, è possibile applicarla solo sui bit meno significativi dell'indirizzo (identificativo di host), conservando il prefisso di rete. Tale tecnica è chiamata pseudonimizzazione «con conservazione del prefisso» [32] e consente di

[23] Per es., dei software di violazione password, come «John the Ripper» o altri.

identificare l'origine globale di un pacchetto (rete) senza sapere quale dispositivo all'interno della rete lo abbia inviato. È fondamentale conoscere il numero di dispositivi per un determinato prefisso. La Tabella 5 mostra prefissi di varie dimensioni. Questa tecnica viene già impiegata da diversi fornitori di servizi per pseudonimizzare gli indirizzi IP (cfr. per es. [33]).

6.2.2 Scelta della modalità di pseudonimizzazione

La scelta della modalità di pseudonimizzazione ha forti ripercussioni sull'funzionalità e sul livello di protezione dei dati, indipendentemente dalla scelta di una determinata tecnica di pseudonimizzazione. Nella presente sezione, tale relazione viene ulteriormente analizzata con l'ausilio di un esempio specifico.

Consideriamo la pseudonimizzazione di indirizzi IP di partenza e di destinazione in una traccia di rete. La Tabella 6 fornisce gli indirizzi di origine e di destinazione dei primi pacchetti di una richiesta HTTP tra un client (145.254.160.237) e un server (65.208.228.223).

Tabella 6. Origine e destinazione di una richiesta HTTP

Numero del pacchetto	Origine	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

Applichiamo al caso sopraindicato la pseudonimizzazione deterministica utilizzando, per es., un RNG. Ogni indirizzo IP è associato a uno pseudonimo univoco. La tabella di mappatura ottenuta è illustrata nella Tabella 7. Dopo aver applicato la pseudonimizzazione deterministica, si ottiene la Tabella 8.

Tabella 7. Tabella di mappatura per la pseudonimizzazione deterministica

Indirizzo IP	Pseudonimo
145.254.160.237	238
65.208.228.223	47

Tabella 8. Indirizzi di origine e destinazione trasformati tramite la pseudonimizzazione deterministica

Numero del pacchetto	Origine	Destinazione
Pacchetto 1	238	47
Pacchetto 2	47	238
Pacchetto 3	238	47

Pacchetto 4	238	47
Pacchetto 5	47	238

Confrontiamo le informazioni ricavabili dalla traccia di rete originale (Tabella 6) e Tabella 8. Come emerge dal confronto, da ambedue le tracce (originale e pseudonimizzata) è possibile ricavare il numero totale di indirizzi IP coinvolti e la quantità dei pacchetti inviati da ciascun indirizzo durante la comunicazione. Pertanto, con la pseudonimizzazione degli indirizzi IP nella tabella 8, è possibile ottenere lo stesso livello di analisi statistica (e, quindi, di funzionalità) degli indirizzi IP.

Consideriamo adesso il caso di una pseudonimizzazione randomizzata al documento con un RNG. Ogniqualvolta si incontra un indirizzo IP, questo viene trasformato in uno pseudonimo differente. Per es., l'indirizzo IP 145.254.160.237 viene associato a 5 pseudonimi: 39, 71, 48, 136 e 120 (Tabella 9). Dopo aver applicato la pseudonimizzazione randomizzata al documento, si ottiene la Tabella 10.

Tabella 9. Tabella di mappatura per la pseudonimizzazione randomizzata al documento

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 trasformati tramite la pseudonimizzazione randomizzata al documento

Numero del pacchetto	Origine	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, se nella Tabella 6 e nella Tabella 8 era possibile contare due indirizzi IP, ciò non è più possibile nella Tabella 10, in cui vengono virtualmente coinvolti 10 indirizzi IP. È pertanto diminuito il livello di funzionalità (ma è accresciuto il livello di protezione). Ovviamente, applicare una pseudonimizzazione completamente randomizzata presenta un impatto ancora più forte sull'funzionalità. La Tabella 11 confronta a tal fine le differenti modalità di pseudonimizzazione.

Tabella 11. Modalità di pseudonimizzazione e funzionalità

Modalità di pseudonimizzazione			
Funzionalità	**Deterministica**	**Randomizzata al documento**	**Completamente randomizzata**
Statistiche (conteggio, ecc.)	SÌ	NO	NO
Semantica del protocollo	SÌ	NO	NO
Confronto tra differenti tracce	SÌ	SÌ	NO

Ovviamente, non esiste una soluzione univoca per questo problema, e la scelta finale dipende dai requisiti di funzionalità e protezione dell'entità di pseudonimizzazione.

7. PSEUDONIMIZZAZIONE DEGLI INDIRIZZI E-MAIL

In questo capitolo, la pseudonimizzazione degli indirizzi e-mail è considerata come un caso d'uso più specifico delle tecniche presentate in precedenza nel documento.

Un indirizzo di posta elettronica (e-mail) costituisce un identificativo tipico di un individuo. Un indirizzo e-mail si presenta con il formato locale@dominio, in cui la parte «locale» corrisponde all'utente che possiede l'indirizzo e la parte «dominio» al fornitore del servizio di posta. Gli indirizzi e-mail sono generalmente utilizzati in diverse applicazioni; per es., possono costituire l'identificativo principale di un individuo che si registra a un servizio elettronico. Inoltre, gli indirizzi e-mail sono generalmente presenti in molti database, dove possono essere presenti anche altri identificativi, quali i nomi delle persone.

Gli utenti tendono a utilizzare lo stesso indirizzo e-mail per diverse applicazioni, condividendolo con varie organizzazioni, per es. quando si registrano a un account online. Inoltre, gli indirizzi e-mail sono spesso resi pubblici online, nonostante sia stato dimostrato che possono essere facilmente reperiti o indovinati ([24]). Per via di queste caratteristiche speciali, quando gli indirizzi e-mail vengono utilizzati come identificativi, è particolarmente importante proteggerli.

In questo caso d'uso, gli indirizzi e-mail sono considerati identificativi (per es. in un database o servizio online), durante l'analisi dell'applicazione a questi di diverse tecniche di pseudonimizzazione. Generalmente il processo di pseudonimizzazione è eseguito da un'entità di pseudonimizzazione (per es. il titolare del trattamento dei dati) nell'ambito dell'operazione/ fornitura di un servizio.

7.1 CONTATORE E GENERATORE DI NUMERI CASUALI (RNG)

Tenendo presente le descrizioni del capitolo 5, è possibile utilizzare sia il contatore che l'RNG per la pseudonimizzazione delle e-mail con l'uso di una tabella di mappatura, come quella mostrata nell'esempio della Tabella 12. Chiaramente, la pseudonimizzazione è forte a condizione che la tabella di mappatura sia protetta e memorizzata separatamente dai dati pseudonimizzati.

Tabella 12. Esempio di pseudonimizzazione degli indirizzi e-mail con RNG o contatore (pseudonimizzazione completa)

Indirizzo di posta elettronica	Pseudonimo (generatore di numeri casuali)	Pseudonimo (generatore a contatore)
alice@abc.eu	328	10
bob@wxyz.com	105	11
eve@abc.eu	209	12
john@qed.edu	83	13

([24]) In effetti, è stato dimostrato che anche il recupero di una semplice informazione di base, per es. i nomi degli utenti di un social network, consente di raccogliere in modo efficiente milioni di indirizzi e-mail [38].

alice@wxyz.com	512	14
mary@clm.eu	289	15

Nell'esempio della tabella 12, sia il contatore che l'RNG producono pseudonimi che non rivelano alcuna informazione sugli identificativi iniziali (indirizzi e-mail) e non consentono ulteriori analisi (per es. analisi statistiche) sugli pseudonimi. Per aumentare l'funzionalità, è possibile applicare la pseudonimizzazione solo a una parte dell'indirizzo e-mail, per es. la parte locale (senza modificare la parte del dominio, cfr. Tabella 13).

Tabella 13. Esempio di pseudonimizzazione degli indirizzi e-mail con RNG o contatore (pseudonimizzazione solo della parte locale)

Indirizzo di posta elettronica	Pseudonimo (generatore di numeri casuali)	Pseudonimo (generatore a 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**

Come mostrato nella Tabella 13, quando le e-mail sono pseudonimizzate, è ancora possibile riconoscere il dominio e, pertanto, condurre analisi pertinenti (per es. il numero di utenti di posta elettronica provenienti dallo stesso dominio). Come illustrato in precedenza nel documento, il contatore può essere più debole in termini di protezione in quanto consente previsioni in virtù della sua natura sequenziale (per es. nei 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 nel database).

A partire da questo semplice caso, a seconda del livello di protezione dei dati e di funzionalità che l'entità di pseudonimizzazione deve raggiungere, è possibile ottenere diverse varianti mantenendo diversi livelli di informazione negli pseudonimi (per es. su domini e parti locali identici, ecc.).

Tabella 14. Esempio di pseudonimizzazione degli indirizzi e-mail con RNG: vari livelli di funzionalità

Indirizzo di posta elettronica	Pseudonimo (RNG) che conserva le informazioni su domini identici	Pseudonimo (RNG) che conserva anche le informazioni su paesi/estensioni identici	Pseudonimo (RNG) che conserva le informazioni su parti locali e domini identici	Pseudonimo (RNG) che conserva le informazioni su paesi/ estensioni, 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 nella scalabilità della tecnica nei casi di set di dati di grandi dimensioni, soprattutto se è necessario che lo stesso pseudonimo sia sempre assegnato allo stesso indirizzo (cioè in uno scenario pseudonimico deterministico come nella Tabella 12). In effetti, in tal caso, l'entità di pseudonimizzazione deve eseguire un controllo incrociato in tutta la tabella di pseudonimizzazione ogni volta che deve essere pseudonimizzato un nuovo elemento. La complessità aumenta in casi di implementazione più sofisticati come quelli mostrati nella tabella 14 (per es. quando l'entità di pseudonimizzazione deve classificare gli indirizzi e-mail con lo stesso dominio o lo stesso paese senza rivelare tale dominio/paese).

7.2 FUNZIONE CRITTOGRAFICA DI HASH

Come indicato in [34], il numero totale di account e-mail globali è stimato approssimativamente a 4,7 miliardi ≈ 2^{32} (poiché, nonostante le dimensioni che nella pratica sono teoricamente infinite dello spazio degli indirizzi e-mail validi, gli indirizzi esistenti si trovano in uno spazio molto più piccolo). Tale fatto, come menzionato anche in precedenza nel capitolo, rende gli indirizzi e-mail facilmente individuabili o indovinabili ([25]), il che comporta che le funzioni crittografiche di hash risultano essere una tecnica debole per la pseudonimizzazione [34]. In effetti, qualsiasi persona interna all'organizzazione o intruso esterno, che abbia accesso a un elenco pseudonimizzato di indirizzi e-mail, può facilmente eseguire un attacco di ricerca nel dizionario (Figura 10). Questa osservazione è rilevante per tutti gli scenari di pseudonimizzazione presentati nel capitolo 3 (indipendentemente dal fatto che l'entità di pseudonimizzazione sia il titolare del trattamento dati, il responsabile o una terza parte fidata).

Figura 10. Ricavare un indirizzo e-mail dal suo valore hash

Nonostante le sopracitate insidie delle funzioni crittografiche di hash, è opportuno sottolineare che, come indicato in [35], i fornitori di servizi spesso condividono indirizzi e-mail con terze parti, semplicemente eseguendo l'hashing. Un esempio concreto è il funzionamento dei cosiddetti elenchi di pubblici personalizzati, che offre alle aziende la possibilità di confrontare i valori hash degli indirizzi e-mail dei clienti per definire elenchi comuni di clienti ([26]).

Nonostante i significativi rischi di protezione dei dati illustrati sopra, i valori degli hash crittografici potrebbero ancora essere utili in determinate condizioni, per es. per la codifica interna degli indirizzi e-mail (come nel caso di attività di ricerca) e come meccanismo di convalida/integrità per un titolare del trattamento dei dati (cfr. anche in [1]). Le funzioni di hash possono anche essere utilizzate per pseudonimizzare parti di un indirizzo e-mail (per es. solo la

[25] Teoricamente, se tutti gli indirizzi possibili fossero disponibili per un instruso, persino un attacco a forza bruta sarebbe praticabile; in ogni caso, tuttavia, lo spazio (relativamente) ridotto degli indirizzi e-mail indica che effettivamente potrebbe avere esito positivo anche una congettura casuale sugli indirizzi e-mail. Ancora peggio, nell'era dei big data, le congetture casuali potrebbero non essere nemmeno necessarie poiché gli indirizzi e-mail validi sono spesso pubblicamente disponibili o possono essere facilmente derivati in contesti specifici (per es. se il dominio e il formato di una specifica organizzazione sono noti).
[26] Per es., cfr: https://www.facebook.com/business/help/112061095610075?helpref=faq_content

parte del dominio), consentendo in tal modo di ottenere una certa funzionalità per gli pseudonimi derivati; se la parte rimanente è pseudonimizzata con un metodo più forte (per es. Codice di autenticazione del messaggio), il rischio di ricavare l'intero indirizzo e-mail iniziale è significativamente ridotto.

7.3 CODICE DI AUTENTICAZIONE DEL MESSAGGIO

Rispetto al semplice hashing, un codice di autenticazione del messaggio offre notevoli vantaggi in termini di protezione dei dati anche per la pseudonimizzazione dell'indirizzo e-mail, purché la chiave segreta sia archiviata in modo sicuro. Inoltre, l'entità di pseudonimizzazione può utilizzare chiavi segrete diverse, per settori diversi, per generare per es. pseudonimi di settore diversi per lo stesso indirizzo e-mail. È possibile utilizzare un Codice di autenticazione del messaggio per impedire al titolare del trattamento di accedere agli indirizzi e-mail nei casi in cui l'accesso agli pseudonimi sia sufficiente per lo scopo specifico del trattamento (per es. negli scenari 5 e 6 del capitolo 3). Si pensi per es. ai messaggi pubblicitari basati sugli interessi, in cui gli inserzionisti devono associare uno pseudonimo univoco per ciascun individuo, senza però che questo riveli l'identità originale dell'utente [36].

Come nelle tecniche precedenti, per aumentare l'funzionalità degli pseudonimi, è possibile valutare diversi scenari di attuazione nella pratica. Per es., un possibile approccio potrebbe consistere nell'applicare il Codice di autenticazione del messaggio separatamente a diverse parti dell'indirizzo e-mail (per es. alle parti locale e di dominio), utilizzando la stessa chiave segreta. La Figura 11 mostra un esempio caratteristico: l'impiego della stessa chiave per ciascun MAC determina la generazione degli stessi sotto-pseudonimi per le parti di dominio corrispondenti (in colore verde) ogni volta che i domini dell'indirizzo e-mail sono identici. Tuttavia, poiché l'output di un Codice di autenticazione del messaggio ha una dimensione fissa, che è generalmente molto più grande di quella dell'indirizzo e-mail iniziale ([27]), gli pseudonimi risultanti possono essere di dimensioni piuttosto grandi (con un ulteriore incremento nel caso in cui le parti diverse siano pseudonimizzate separatamente).

Figura 11. Usare un MAC per generare indirizzi e-mail pseudonimizzati con alcune funzionalità

Un aspetto importante per quanto riguarda l'implementazione pratica del Codice di autenticazione del messaggio è il recupero. Va sottolineato che anche l'entità di pseudonimizzazione dei dati, che ha accesso alla chiave segreta, non è in grado di ricavare direttamente gli pseudonimi; tale operazione è possibile solo in forma indiretta, riproducendo gli pseudonimi per ciascun indirizzo e-mail noto al fine di vedere le corrispondenze con l'elenco pseudonimizzato. Chiaramente, se è disponibile una tabella di mappatura di pseudonimizzazione, ricavare gli pseudonimi è un'operazione semplice, ma in tal caso aumentano anche i requisiti di archiviazione. Per questi motivi, il Codice di autenticazione del messaggio probabilmente non è la tecnica di pseudonimizzazione più pratica nei casi in cui il

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

titolare del trattamento dei dati debba essere in grado di associare facilmente gli pseudonimi agli indirizzi e-mail (per es. in alcuni casi illustrati nei capitoli 3.1 e 3.2).

7.4 CRITTOGRAFIA

Un'alternativa al Codice di autenticazione del messaggio è la crittografia, applicata soprattutto in modo deterministico, ovvero utilizzando una chiave segreta per produrre uno pseudonimo per ciascun indirizzo e-mail (crittografia simmetrica). La distribuzione è più pratica in questo caso, poiché non è necessario prevedere una tabella di mappatura della pseudonimizzazione: il recupero è possibile direttamente attraverso il processo di de-crittografia [37].

Vale la pena sottolineare che, sebbene alcuni algoritmi crittografici asimmetrici (chiave pubblica) possano essere implementati in modo deterministico ([28]), non sono raccomandati per la pseudonimizzazione di indirizzi e-mail (o per altri tipi di dati, cfr. anche in [1]). Per es., supponiamo che l'entità di pseudonimizzazione debba generare, per ciascun indirizzo e-mail, pseudonimi diversi per utenti/destinatari diversi, interni o esterni (con il presupposto che ciascun destinatario sarà in grado di identificare nuovamente i propri dati, ma non i dati pseudonimizzati di altri destinatari). Una possibilità per raggiungere questo obiettivo sarebbe quella di crittografare le e-mail con la chiave pubblica di ciascun destinatario, consentendo così solo al destinatario specifico di eseguire la decrittografia. Tuttavia, supponendo che le chiavi pubbliche siano in linea di principio disponibili a chiunque, qualsiasi attaccante può lanciare un attacco di ricerca nel dizionario basato su indirizzi e-mail noti (o indovinati) (come quello mostrato nella Figura 10, in cui viene utilizzata la crittografia a chiave pubblica con una chiave pubblica nota anziché una funzione di hash).

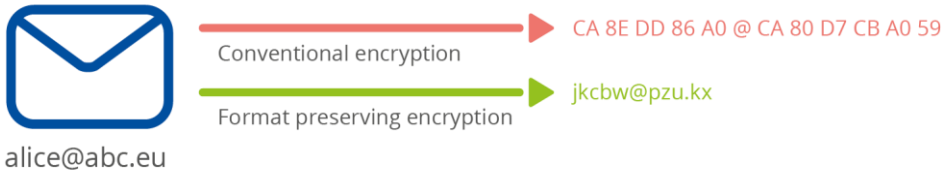
La natura della crittografia non consente di default l'utilizzo dei dati pseudonimizzati. La crittografia separata delle parti di un indirizzo e-mail può essere sufficiente per ridurre questo problema, analogamente ai codici di autenticazione del messaggio (cfr. Figura 11), in cui il Codice di autenticazione del messaggio può essere sostituito da un algoritmo di crittografia. In generale, per consentire agli pseudonimi di contenere informazioni utili, è possibile utilizzare specifiche tecniche crittografiche; di seguito viene fornito un esempio che illustra la crittografia con protezione del formato.

CRITTOGRAFIA CON CONSERVAZIONE DEL FORMATO (FPE)

Uno schema di database può prevedere un determinato tipo di dati per campi specifici. Per es., un indirizzo e-mail contiene di norma una parte locale (info), seguita da un simbolo @, che a sua volta è seguito da un dominio. Se il titolare del trattamento dei dati non ha bisogno di conservare gli indirizzi e-mail iniziali, ma è tenuto comunque a conservare un elenco pseudonimizzato mantenendo la struttura del database, la crittografia con protezione del formato si rivela una procedura adeguata. Esistono diverse implementazioni note relative alla crittografia con conservazione del formato, basate su schemi crittografici noti ([29]). Ad ogni modo, 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 nella parte locale degli indirizzi e-mail – è sufficiente per garantire che lo pseudonimo derivato abbia la forma desiderata. La differenza tra FPE e crittografia convenzionale è illustrata nella Figura 12.

([28]) Nonostante il fatto che, per motivi di sicurezza, un algoritmo a chiave pubblica debba essere in linea di principio probabilistico [1].
[29] Cfr. per es. https://csrc.nist.gov/publications/detail/sp/800-38g/rev-1/draft, una bozza attuale redatta dal NIST dei metodi adeguati di crittografia con conservazione del formato nei casi in cui la dimensione del dominio sia troppo piccola.
([30]) La sostituzione di un carattere è un caso speciale di crittografia (sebbene possano sorgere problemi di sicurezza se tale sostituzione non viene implementata correttamente).

Figura 12. Crittografia con conservazione del formato vs crittografia convenzionale per ricavare pseudonimi dagli indirizzi e-mail



Nella Figura 12 è stato utilizzato un cifrario a flusso simmetrico per la crittografia convenzionale, per garantire che lo pseudonimo derivato abbia la stessa lunghezza dell'indirizzo iniziale (i caratteri dello pseudonimo derivato sono non alfanumerici e vengono quindi indicati in forma esadecimale).

Si noti che, a seconda del caso, può essere necessario progettare adeguatamente le implementazioni della FPE, affinché non emergano schemi che potrebbero far trapelare informazioni sull'identità degli individui.

8. PSEUDONIMIZZAZIONE IN PRATICA: UNO SCENARIO PIÙ COMPLESSO

Come si può dedurre dai due precedenti casi d'uso illustrati nei capitoli 6 e 7, la pseudonimizzazione anche di tipi di dati più semplici, come indirizzi IP o indirizzi e-mail, è un compito impegnativo e soggetto a errori. Quando si tratta di sistemi del mondo reale, tuttavia, spesso non è la scelta della tecnica di pseudonimizzazione utilizzata per uno o due identificativi specifici a causare la maggior parte dei problemi; è l'implicita associabilità tra un insieme di pseudonimi e altri valori di dati che sono integrati in una struttura di dati più complessa. L'esempio più comune è quello di un servizio online che crea profili utente al momento della registrazione e li arricchisce con informazioni personali ogni volta che diventano disponibili nuovi dati. In questo caso, anche se l'indirizzo e-mail dell'utente e tutti gli indirizzi IP trovati nei registri di accesso di tale utente vengono rigorosamente pseudonimizzati come indicato in precedenza, sussiste ancora una forte minaccia di reidentificazione o discriminazione anche per la struttura stessa dei dati pseudonimizzati. Nella presente sezione ci apprestiamo ad analizzare questi casi più complessi di pseudonimizzazione dei dati.

8.1 UN ESEMPIO DI SIMULAZIONE

Ai fini della presente trattazione, ipotizziamo un esempio di scenario molto simile a quelli più comunemente diffusi nel mondo reale: un social network. L'operatore immaginario, SocialNetwork Inc. (di seguito SN), agisce in qualità di titolare del trattamento dei dati e consente ai suoi utenti (che si presumono essere esclusivamente umani) di registrare un account che viene archiviato nel datacenter di SN. Con tale account, gli utenti possono utilizzare una serie di funzioni che consentono loro, per es., di collegarsi ad altri utenti, organizzazioni o argomenti di interesse. Al momento della registrazione, gli utenti di SN devono fornire il loro nome reale, sotto forma di nome e cognome, il nickname, la data di nascita e il genere, insieme a una serie di informazioni personali facoltative (località, interessi, dati biometrici, ecc.) e a 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, inclusi una marcatura temporale (timestamp) e l'indirizzo IP di accesso.

Per migliorare la conformità con il RGPD, la dirigenza di SN ha deciso di pseudonimizzare gli indirizzi IP nei registri di accesso secondo le tecniche illustrate nel capitolo 6. Le altre informazioni sono mantenute in chiaro, in quanto occorre presentarle all'utente sui siti web di SN laddove sia necessario o per eseguire controlli e convalide (per es. la data di nascita è necessaria per calcolare l'età e verificare che l'utente abbia più di 16 anni nel momento in cui accede a servizi speciali). In questo caso, la pseudonimizzazione dell'indirizzo e-mail non è possibile, poiché SN deve essere in grado di inviare agli utenti e-mail con notifiche (e altri contenuti).

Ipotizziamo una seconda organizzazione immaginaria, Online Security Services Corp. (di seguito OSS), che agisce in qualità di responsabile del trattamento dei dati per conto di SN, con il compito di effettuare la manutenzione dei servizi di archiviazione e sicurezza per alcune parti del database utenti di SN. In questa posizione, OSS ha accesso ai file di registro pseudonimizzati di SN, ovvero agli indirizzi IP e alle marcature temporali pseudonimizzate di

tutti gli accessi al sito web, ma non agli indirizzi IP originali. In tale contesto, OSS non può reidentificare gli utenti appartenenti a un indirizzo IP in quanto i dati sono archiviati in un database diverso presso SN che non è accessibile a OSS. Pertanto, per quanto riguarda la pseudonimizzazione, ci troviamo nello scenario del capitolo 3.3, in cui SN è il titolare del trattamento dei dati e OSS il successivo responsabile del trattamento dei dati.

8.2 INFORMAZIONI INERENTI AI DATI

A prima vista, OSS non è in grado di decodificare la pseudonimizzazione degli indirizzi IP eseguita da SN, presupponendo che quest'ultima abbia utilizzato una funzione di pseudonimizzazione sufficientemente forte. A seconda della funzione di pseudonimizzazione, e in particolare della politica di pseudonimizzazione (cfr. il capitolo 5.2), OSS potrebbe ancora essere in grado di dedurre se un determinato pseudonimo occorre frequentemente, raramente, solo una volta o mai nel database. Ciò potrebbe non essere sufficiente di per sé per scoprire un'identità, ma può comunque essere utilizzato per identificare gli utenti che accedono frequentemente. Se un registro di accesso contiene uno pseudonimo con un'alta frequenza di occorrenza, OSS può dedurre che si tratta probabilmente di un utente abituale di SN. Viceversa, se uno pseudonimo occorre per la prima volta in un set di dati, molto probabilmente questo utente si è appena registrato su SN e ha effettuato l'accesso al suo account utente per la prima volta, oppure l'indirizzo IP di un utente registrato è cambiato (il che può accadere frequentemente, rendendo le osservazioni finora formulate di natura puramente probabilistica).

Questo tipo di informazioni inerenti ai dati può già essere utile per OSS, per es. per sapere quanti utenti di SN sono utenti continui e quanti si registrano una volta senza tornare una seconda (con un certo margine di errore dovuto al cambio di indirizzi IP). Queste informazioni possono già risultare fondamentali nella relazione commerciale tra SN e OSS.

Oltre a tali informazioni inerenti ai dati, il fatto che OSS abbia accesso continuo al database di SN consente un altro tipo di raccolta di informazioni per OSS: monitorando continuamente il set di dati archiviato per SN, OSS apprende il cambiamento di tale set di dati. Ciò include banalmente il numero totale di accessi al sito web di SN, ma può anche essere utilizzato, per es., per contare il numero di registrazioni di nuovi utenti (pseudonimi che occorrono per la prima volta) al giorno o al mese. Pur essendo per lo più di natura statistica, queste informazioni possono già essere utilizzate per organizzare veri e propri attacchi di discriminazione (così da ideare impatti diversi su diversi gruppi di utenti): OSS apprende quale pseudonimo del nuovo utente viene visualizzato per primo in un determinato giorno, consentendogli di monitorare la quantità di interazioni che questo specifico utente ha con SN. Queste informazioni possono facilmente diventare un problema di protezione dei dati dell'interessato, come verrà mostrato in seguito.

8.3 DATI COLLEGATI

Nello scenario di simulazione, i dati accessibili a OSS forniscono più informazioni rispetto ai soli indirizzi IP: ogni voce di registro memorizza infatti anche la marcatura temporale dell'accesso. Quindi, invece di monitorare frequentemente le modifiche nel database su SN, OSS può semplicemente fare affidamento sulle marcature temporali collegate a ogni pseudonimo, per eseguire lo stesso tipo di discriminazione di prima. Le marcature temporali sono memorizzate insieme agli indirizzi IP pseudonimizzati, sono quindi direttamente collegate una a una a tali informazioni. Sulla base di questi dati collegati, OSS può aumentare di molto le proprie conoscenze su specifici utenti di SN: un utente specifico accede a SN di più al mattino, a pranzo o la sera? Solo o principalmente la domenica? Solo durante le festività religiose del calendario ortodosso? Solo durante i periodi di vacanze scolastiche in Danimarca?

Ciascuna di queste ulteriori caratterizzazioni consentono a OSS di avvicinarsi a una violazione della pseudonimizzazione, basandosi solo sulle marcature temporali memorizzate e sulla

possibilità di associare record di dati diversi con pseudonimi identici. Come si può vedere, questo tipo di informazioni inizia a fornire a OSS alcune caratterizzazioni degli utenti di SN che possono essere considerate informazioni personali. L'associazione richiede tuttavia che vengano associate ulteriori informazioni agli stessi set di dati strutturati, come per es. il calendario ortodosso o le vacanze scolastiche danesi. Pertanto, questi possono essere considerati come attacchi basati su conoscenze generali, come discusso nel capitolo 4, ma con una complessità variabile delle conoscenze necessarie. Inoltre, questo tipo di informazioni estratte è di natura statistica, quindi non affidabile al 100 %, ma con un certo grado di probabilità. In questo caso, più sono le voci contenute nel database, più diventa affidabile (o falsificabile) un'ipotesi di collegamento. Pertanto, più è grande il social network SN, più diventa facile per OSS eseguire tale discriminazione o persino attacchi di re-identificazione.

Questo esempio contemplava solo un indirizzo IP e una marcatura temporale pseudonimizzati. Ciò vale anche, con un livello ancora superiore di affidabilità, per un indirizzo e-mail pseudonimizzato al posto di un indirizzo IP pseudonimizzato, poiché quest'ultimo tende a cambiare più frequentemente, e quindi rappresenta più di un identificativo univoco per un individuo.

8.4 DISTRIBUZIONE CORRISPONDENTE DELLE OCCORRENZE

Le strutture di dati dell'esempio di cui sopra sono piuttosto piccole e semplicistiche: solo indirizzo IP e marcatura temporale. Eppure, risultano sufficienti per attacchi di discriminazione o addirittura attacchi di re-identificazione, purché vi siano sufficienti conoscenze generali. Inoltre, i dati immessi del mondo reale contengono in genere più informazioni rispetto a questi due semplici valori, pertanto forniscono più dettagli da utilizzare per scoprire gli pseudonimi.

Si consideri che SN memorizza più di una semplice marcatura temporale e di un indirizzo IP pseudonimizzato in ogni record di dati, per es. memorizza anche il tipo e la versione del browser ([31]) utilizzato da quell'utente, l'insieme e le preferenze dei linguaggi naturali parlati dall'utente (come definito nelle impostazioni del browser), la versione del sistema operativo del computer dell'utente, ecc. Come è stato mostrato dalla Electronic Frontier Foundation nel progetto Panopticlick ([32]), questa combinazione di impostazioni del browser da sola può già essere sufficiente per identificare in modo univoco un determinato browser – e quindi l'utente – di un sito web online. Se in un dato momento SN memorizza tutte queste informazioni per ciascun accesso al proprio sito web, OSS potrebbe avervi accesso.

Anche se SN esegue una sorta di pseudonimizzazione su ciascuna di queste configurazioni (per es. memorizzando solo un hash con chiave della stringa della versione del browser ricevuta dal browser dell'utente), OSS può ancora vedere tutte quelle stringhe pseudonimizzate della versione del browser, calcolare le statistiche sulla frequenza di apparizione del valore di hash nel database complessivo di SN e confrontare tale distribuzione di diversi valori esistenti con le statistiche disponibili pubblicamente raccolte sul sito web Panopticlick, per scoprire la vera stringa della versione del browser che si trova dietro ogni valore di hash, nonostante il corretto utilizzo della funzione di pseudonimizzazione. Il semplice fatto che la distribuzione statistica di diversi pseudonimi corrisponda alla distribuzione statistica dei loro presunti testi in chiaro può essere sufficiente per svelare quegli pseudonimi, con un'alta probabilità di successo.

Questo ovviamente dipende molto dall'approccio di pseudonimizzazione prescelto. Se viene applicato un approccio di ingegnerizzazione adeguato, l'aggiunta di metadati all'argomento

([31]) Occorre notare che questo è il comportamento di default del registro, per es. del server web Apache.
([32]) https://panopticlick.eff.org/

della funzione di pseudonimizzazione può offrire una maggiore protezione contro la retroingegnerizzazione.

8.5 CONOSCENZE AGGIUNTIVE

Se OSS dispone di ulteriori conoscenze sulle caratteristiche di un determinato utente e sta cercando di ricavare i record di dati dell'utente dal database pseudonimizzato che riceve da SN, ogni informazione aggiuntiva può rivelarsi fondamentale. Se OSS sa che l'utente target specifico è maschio e utilizza il browser Chrome su un iPad, questa informazione da sola restringe in modo significativo l'insieme di possibilità dei profili utente visti da OSS. Ognuno di questi valori di dati, anche se pseudonimizzati, riduce l'insieme di possibilità, ovvero l'insieme dei profili utente contenuti nel database SN che possono appartenere all'utente target specifico cercato da OSS. Le informazioni del browser possono essere affrontate con un attacco di distribuzione di probabilità illustrato nella sezione 8.4, rimuovendo gran parte dei profili utente con pseudonimi del browser con troppe o troppe poche occorrenze che corrispondono alla specifica probabilità di configurazione «Chrome su un iPad».

Dai profili rimanenti, un banale attacco a forza bruta o di distribuzione statistica rivela a OSS quale pseudonimo corrisponde a un determinato genere, eliminando circa la metà dei restanti profili utente. Se ora tutti i restanti profili utente hanno in comune il fatto che il primo accesso a SN sia avvenuto tra maggio e luglio 2018, OSS ha già imparato qualcosa su quel determinato utente: lui o lei si è registrato/a a SN in quel periodo di tempo. Questo è un attacco inferenziale riuscito. Analizzando ulteriormente i profili utente rimanenti, OSS può apprendere un modello specifico di marcatura temporale dell'utilizzo di SN relativo a due di tali profili utente, in modo tale che corrispondano al presunto modello di utilizzo dell'individuo target (che in passato OSS è stato in grado di osservare in alcune occasioni). Pertanto, il set target della ricerca viene ridotto a solo due profili utente.

Ogni informazione che entrambi i profili hanno in comune deve quindi valere anche per l'individuo target specifico, il che fornisce a OSS già molte informazioni sul loro target di ricerca. Per eliminare il falso candidato restante, OSS può semplicemente monitorare in modo specifico l'utilizzo di SN da parte di questi due profili, e all'accesso successivo verificare se tale accesso possa aver avuto origine dal loro individuo target o meno (sulla base di ulteriori conoscenze generali ottenute da quei fatti che OSS aveva già appreso sul proprio target). Alla fine, OSS è in grado di associare il profilo utente all'identità di destinazione. In tal modo, OSS è anche in grado di scoprire tutte le pseudonimizzazioni eseguite sui valori dei dati di quell'individuo, dando inoltre a OSS la possibilità di individuare o discriminare rispetto ad altri profili utente.

Tuttavia, va notato che il problema delle informazioni aggiuntive disponibili è «traversale» alla pseudonimizzazione, pur trattandosi principalmente di una questione di protezione dei dati. Pertanto, come già accennato in precedenza nel rapporto, oltre alla pseudonimizzazione, si può prendere in considerazione l'aggiunta di rumore (injection of noise) ai parametri della funzione di pseudonimizzazione o l'uso della generalizzazione, al fine di rendere meno efficaci gli attacchi a forza bruta (cfr. anche il capitolo 5.6). Tale grado di libertà è un modo per rafforzare ulteriormente la pseudonimizzazione e proteggersi dai relativi attacchi.

8.6 COLLEGAMENTO TRA PIÙ SORGENTI DI DATI

Oltre allo scenario sopra enunciato di SN e OSS, emerge uno scenario di pseudonimizzazione ancora più interessante, laddove si considera non solo la partecipazione di due organizzazioni (SN e OSS), ma si ipotizza un mercato su larga scala di dati pseudonimizzati. In tali scenari, un gran numero di diverse organizzazioni condividono set di dati personali pseudonimizzati, con l'intento di consentire alcune funzionalità(per es. la creazione di profili a fini di marketing), proteggendo al contempo l'identità degli interessati. L'argomentazione spesso sollevata in tali scenari è che la pseudonimizzazione impedisce la re-identificazione degli interessati,

legittimando così tale condivisione dei dati. Il presente rapporto non prende posizione sulla legittimità della condivisione dei set di dati pseudonimizzati, ma affronta le questioni relative alla corretta applicazione della pseudonimizzazione in tale contesto.

Prendiamo un insieme di società dalla A alla E, che raccolgono tutti i dati personali sui propri utenti, come i dati raccolti da SN nell'esempio precedente. Si potrebbe procedere all'associazione dei profili utente di diverse società confrontando gli indirizzi e-mail da questi utilizzati. Se due profili utente si trovano, per es., presso le società B e D, registrati esattamente con lo stesso indirizzo e-mail, molto probabilmente appartengono al medesimo interessato. Tuttavia, lo stesso indirizzo e-mail è ovviamente costituito da dati personali, come è stato illustrato nel capitolo 7. Diventa quindi necessario applicare la pseudonimizzazione agli indirizzi e-mail nei set di dati di B e D, prima di condividerli tra A, B, C, D ed E.

La difficoltà in questo caso consiste nel fatto che tutti i partecipanti vogliono mantenere la funzionalità dei dati pseudonimizzati per associare i profili appartenenti alla stessa persona, senza ridurre la protezione dell'identità di quell'utente. Pertanto, tutte e cinque le società devono applicare la stessa pseudonimizzazione, utilizzando la stessa funzione e lo stesso segreto di pseudonimizzazione, al fine di poter confrontare e associare record di dati provenienti da set di dati tra loro differenti. In questo caso, c'è una chiara discrepanza tra la funzionalità (di associare gli indirizzi e-mail pseudonimizzati) e la protezione (degli utenti di tali indirizzi e-mail). In altre parole, B e D devono essere in grado, ed essere autorizzate, di apprendere che i loro record di dati specifici condividono lo stesso indirizzo e-mail e appartengono quindi allo stesso utente, ma non devono essere in grado di sapere di quale indirizzo e-mail – e quindi di quale soggetto – si tratti.

Come illustrato nel capitolo 7, in tali scenari l'uso di funzioni di pseudonimizzazione deboli (come il semplice hashing) consente banali attacchi a forza bruta, di inferenza o di distribuzione di probabilità, come già evidenziato sopra. Grazie a dati aggiuntivi (non personali) contenuti nei record di dati condivisi e ad altre eventuali conoscenze generali, questi attacchi si rivelano attuabili e ampiamente efficaci in molti scenari. E non è tutto: maggiore è il numero di aziende che condividono informazioni sugli attributi di una determinata persona interessata, maggiori sono le informazioni a disposizione di un attaccante che intenda invertire la funzione di pseudonimizzazione utilizzata, andando così ad aumentare la probabilità di successo di tali attacchi.

Possono verificarsi rischi per la privacy anche nello scenario più generale in cui le organizzazioni applichino tecniche di pseudonimizzazione differenti (e persino forti) agli identificativi dei propri utenti (per es. e-mail o indirizzo IP). Ipotizziamo che il suddetto gruppo di aziende da A a E fornisca tali dati pseudonimi a OSS, al fine di ottenere, per es., servizi statistici. Se gli pseudonimi forniti sono accompagnati da informazioni sul browser/dispositivo dell'utente come descritto nella sezione 8.4 (impostazioni del browser, sistema operativo, ecc.), rammentando che tali informazioni devono essere univoche per ciascun dispositivo ([33]), OSS può semplicemente associare tra loro pseudonimi differenti, forniti da società diverse, che corrispondono allo stesso utente.

8.7 CONTROMISURE

Come illustrato nel capitolo 5, le tecniche di pseudonimizzazione randomizzata al documento e completamente randomizzata riducono l'associazione tra pseudonimi diversi da set di dati differenti, andando così a mitigare o addirittura eliminare le caratteristiche statistiche dei database pseudonimizzati. Allo stesso tempo, esse limitano la possibilità di associare diversi record di dati (potenzialmente distribuiti su molte organizzazioni) a un unico profilo utente.

([33]) Il noto termine «device fingerprinting» descrive questo rischio per la privacy.

Pertanto, pur applicando la pseudonimizzazione randomizzata, OSS potrebbe comunque essere in grado di eseguire gli attacchi descritti sopra nel caso in cui possa scoprire se due diversi pseudonimi appartengono allo stesso identificativo. Allo stesso modo, B e D possono re-identificare l'interessato che si cela dietro i profili utente condivisi. In questo caso, risulta di nuovo evidente il compromesso tra protezione e utilità.

Come ci si può difendere, dunque, da questi attacchi alla pseudonimizzazione in modo affidabile?

Secondo l'analisi svolta in questo rapporto, l'approccio migliore alla pseudonimizzazione è quello di:

- Considerare l'intero set di dati disponibile.
- Apprendere quali sono le dimensioni del dominio di input dei valori dei dati di un individuo.
- Applicare la pseudonimizzazione su tutti i valori dei dati in modo tale che gli attacchi a forza bruta e di ricerca nel dizionario diventino impraticabili.
- Eliminare qualsiasi possibilità di attacchi basati su conoscenze generali o di distribuzione statistica.
- Progettare la funzione di pseudonimizzazione su larga scala, in modo tale che il set di dati pseudonimizzato mantenga solo il tipo di utilità necessaria ai fini del trattamento, rimuovendo invece tutte le altre utilità dal set di dati pseudonimizzato.

Ai fini dello scenario descritto in questo capitolo, SN può utilizzare uno schema di pseudonimizzazione che pseudonimizza non solo gli indirizzi IP stessi, ma anche tutte le possibili combinazioni di indirizzi IP e marcature temporali. Pertanto, diventa impossibile associare la marcatura temporale a qualsiasi sorgente di dati esterna, poiché tali informazioni non sono più disponibili per OSS. Per una corretta re-identificazione, OSS dovrebbe conoscere (o indovinare) l'esatta combinazione di indirizzo IP e marcatura temporale. In generale, non è ragionevolmente possibile scoprire la pseudonimizzazione di una combinazione di input di dati nel caso in cui non si conoscano (o non si indovinino) tutti i dati di input in chiaro. In questo contesto, tale pseudonimizzazione bloccherebbe qualsiasi tentativo di OSS di scoprire un determinato pseudonimo in modo di gran lunga più solido.

Nel capitolo 5 abbiamo già illustrato alcuni esempi di tecniche di base per ottenere funzioni di pseudonimizzazione efficaci, approfondendone la resilienza contro gli attacchi alla pseudonimizzazione descritti nel capitolo 4. Affinché valgano anche per record di dati strutturati, è spesso sufficiente considerare l'intero record di dati come input e applicare alla pseudonimizzazione in generale una combinazione personalizzata di funzioni hash con chiave e tecniche comuni. Nel capitolo 5.6 e in un precedente rapporto dell'ENISA [2] sono state brevemente illustrate le tecniche più avanzate di pseudonimizzazione.

9. CONCLUSIONI E RACCOMANDAZIONI

Alla luce del RGPD, il dibattito sulla corretta applicazione della pseudonimizzazione ai dati personali sta divenendo via via più acceso in diverse comunità, dal mondo accademico e della ricerca a quello giudiziario e dell'applicazione delle leggi, fino a toccare il campo della gestione della conformità in varie organizzazioni europee. In questa relazione sono state introdotte alcune nozioni di base, insieme alle relative definizioni, tecniche, attacchi e contromisure a sostegno dei futuri lavori interdisciplinari.

Come mostrato nella presente relazione, il campo della pseudonimizzazione dei dati in infrastrutture informatiche complesse risulta impegnativo, poiché dipende strettamente da questioni di contesto, entità coinvolte, tipologie di dati, conoscenze generali e dettagli di implementazione. È in effetti emerso come non esista una soluzione semplice e univoca per la pseudonimizzazione, in grado di funzionare per tutti gli approcci e in tutti gli scenari possibili. Al contrario, per applicare un processo di pseudonimizzazione solido è necessaria un'elevata competenza, riducendo possibilmente la minaccia di discriminazioni o di attacchi di re-identificazione, e mantenendo al contempo il grado di funzionalità necessario per il trattamento dei dati pseudonimizzati.

A tal fine, sulla base delle analisi condotte nel presente rapporto, è possibile trarre le seguenti conclusioni e raccomandazioni per tutte le parti interessate, riguardo all'adozione e all'attuazione della pseudonimizzazione dei dati.

UN APPROCCIO ALLA PSEUDONIMIZZAZIONE BASATO SUL RISCHIO

Sebbene tutte le tecniche di pseudonimizzazione ad oggi note presentino caratteristiche intrinseche ormai chiare, non è tuttavia semplice, a livello pratico, individuare l'approccio corretto. È necessario un attento esame del contesto in cui occorre applicare la pseudonimizzazione, tenendo conto di tutti gli scopi prefissati per il caso specifico (da chi devono essere nascoste le identità, qual è l'funzionalità desiderata per gli pseudonimi derivati, ecc.), nonché della facilità di attuazione. Per scegliere la corretta tecnica di pseudonimizzazione risulta pertanto necessario adottare un approccio basato sul rischio, così da valutare in maniera adeguata e ridurre le relative minacce alla privacy. In effetti, il semplice fatto di proteggere i dati aggiuntivi necessari per la re-identificazione, sebbene si tratti di un prerequisito, non garantisce necessariamente l'eliminazione di tutti i rischi.

I titolari e i responsabili del trattamento dei dati dovrebbero considerare attentamente l'idea di attuare la pseudonimizzazione secondo un approccio basato sul rischio, tenendo conto dello scopo e del contesto generale del trattamento dei dati personali, nonché dei livelli di funzionalità e scalabilità che intendono raggiungere.

I produttori di beni, servizi e applicazioni dovrebbero fornire informazioni adeguate ai titolari e ai responsabili del trattamento dei dati circa le tecniche di pseudonimizzazione da loro adottate e i livelli di sicurezza e protezione dei dati da queste ultime garantiti.

Le autorità di regolamentazione (quali le autorità preposte alla protezione dei dati personali e il comitato europeo per la protezione dei dati) sono chiamate a fornire un orientamento pratico ai titolari e ai responsabili del trattamento dei dati in materia di valutazione del

rischio, promuovendo al tempo stesso le migliori pratiche nel campo della pseudonimizzazione.

DEFINIZIONE DELLO STATO DELL'ARTE

Per poter sostenere un approccio alla pseudonimizzazione basato sul rischio, è indispensabile definire lo stato dell'arte del settore. Se infatti, come mostrato nella presente relazione, sono disponibili diverse tecniche di pseudonimizzazione, può variare la loro applicazione pratica, per es. tra diversi tipi di identificativi o set di dati. A tal fine, si rivela essenziale lavorare su casi d'uso ed esempi specifici, fornendo maggiori dettagli e possibili opzioni di implementazione tecnica di pseudonimizzazione.

La Commissione europea e le istituzioni UE competenti dovrebbero promuovere la definizione e la diffusione dello stato dell'arte nella pseudonimizzazione, d'intesa con la comunità scientifica e con il settore industriale.

Le autorità di regolamentazione (quali le autorità preposte alla protezione dei dati personali e il Comitato europeo per la protezione dei dati) dovrebbero promuovere la pubblicazione delle migliori pratiche nel campo della pseudonimizzazione.

PROMUOVERE EVOLUZIONI DELLO STATO DELL'ARTE

La presente relazione è soprattutto incentrata sulle tecniche di pseudonimizzazione di base ad oggi disponibili per i titolari e i responsabili del trattamento dei dati. Tuttavia, per scenari più complessi (che, come mostrato nel rapporto, si presentano abbastanza spesso nell'applicazione pratica), si rende sempre più necessario l'uso di tecniche più avanzate (e solide), come quelle derivate dall'area dell'anonimizzazione. Non solo: occorre rivedere la nozione stessa di anonimizzazione, poiché i modelli di attacco si stanno evolvendo (rendendola sempre più impegnativa in scenari di vita reale).

La comunità scientifica dovrebbe lavorare per integrare tra le attuali tecniche di pseudonimizzazione soluzioni più avanzate, per far fronte in modo efficace alle particolari sfide poste dall'era dei big data. La Commissione europea e istituzioni UE competenti sono chiamate a promuovere e diffondere tali sforzi.

Citazioni

[1] M.J. Dworkin, *SHA-3 Standard: Permutation-Based Hash and Extendable-Output Functions*, 2015.

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

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

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

[5] M. Hellman, *A cryptanalytic time-memory trade-off,* «*IEEE transactions on Information Theory*»*,* vol. 26 (1980), n. 4, pp. 401-406.

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

[7] D.G. Malone e W. Sullivan, *Guesswork and entropy*, «IEEE Transactions on Information Theory*»,* vol. 50 (2004), n. 3, pp. 525-526.

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

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

[10] M. Bellare, R. Canetti e H. Krawczyk, *Keying hash functions for message authentication* in *Annual international cryptology conference*, 1996.

[11] L. Demir, A. Kumar, M. Cunche e C. Lauradoux, *The pitfalls of hashing for privacy*, «IEEE Communications Surveys and Tutorials*»,* 2017, pp. 551-565.

[12] H. Krawczyk, R. Canetti e M. Bellare, *HMAC: Keyed-Hashing for Message Authentication*, RFC, 1997*,* pp. 1-11.

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

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

[15] I. Polakis, G. Kontaxis, S. Antonatos, E. Gessiou, T. Petsas e 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.

[16] T. Eastlake e D. Hansen, *US Secure Hash Algorithms (SHA and SHA-based HMAC and HKDF)*, 2011.

[17] A. Narayanan e V. Shmatikov, *Robust De-anonymization of Large Sparse Datasets* in *IEEE Symposium on Security and Privacy*, 2008.

[18] ENISA, *Recommendations on shaping technology according to GDPR provisions - An overview on data pseudonymisation*, Atene, 2018.

[19] WP29, Gruppo di lavoro articolo 29 per la protezione dei dati: *Parere 4/2007 sul concetto di dati personali*, 2007.

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

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

[22] IETF, *Internet Engineering Task Force: RFC 5735, Special Use IPv4 Addresses*, 2010.

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

[24] G. Becker, *Merkle Signature Schemes, Merkle Trees and Their Cryptanalysis*, Bochum, 2008.

[25] L. Lamport, *Password authentication with insecure communication*, «Communications of the ACM», novembre 1981, pp. 770-772.

[26] B. H. Bloom, *Space/time trade-offs in hash coding with allowable errors*, «Communications of the ACM», luglio 1970, pp. 422-426.

[27] L. Sweeney, *K-anonymity: A model for protecting privacy*, «International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems», vol. 10 (2002), n. 5, pp. 557-570.

[28] L. Sweeney, *Only You, Your Doctor, and Many Others May Know*, «Technology Science», vol. 2015092903 (2015), n. 9, p. 29.

[29] C. Dwork e A. Roth, *The Algorithmic Foundations of Differential Privacy*, «Foundations and Trends in Theoretical Computer Science», agosto 2014, pp. 211-407.

[30] R. Noumeir, A. Lemay e J.M. Lina, *Pseudonymization of radiology data for research purposes*, «Journal of digital imaging», vol. 20 (2007), n. 3, pp. 284-295.

[31] Y. Yona e S. Diggavi, *The effect of bias on the guesswork of hash functions* in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017.

[32] ENISA, *Privacy and data protection by design - from policy to engineering,* 2014.

[33] Digital Summit Data Protection Focus Group, *White Paper on Pseudonymization*, 2017.

[34] P. Oechslin, *Making a Faster Cryptanalytic Time-Memory Trade-off* in *CRYPTO 2003*, 2003.

[35] IETF, *Internet Engineering Task Force: RFC 791, Internet Protocol DARPA Internet Program Protocol Specification*, 1981.

[36] IETF, *Internet Engineering Task Force: IPFIX Working Group, IP Flow Anonymization Support*, 2011.

[37] *An Analysis of Google Logs Retention Policies*, «Journal of Privacy and Confidentiality», vol. 3 (2011), n. 1.

[38] S. Weber, *On Transaction Pseudonyms with Implicit Attributes*, *Cryptology ePrint Archive: Report 2012/568, <https://eprint.iacr.org/2012/568>,* 2012.

[39] W.H. e F.W., *A Survey of Noninteractive Zero Knowledge Proof System and Its Applications*, «The Scientific World Journal», 2014.

TP-01-19-810-EN-N

INFORMAZIONI SULL'ENISA

L'Agenzia europea per la sicurezza delle reti e dell'informazione (ENISA) è attiva dal 2004 sul fronte della sicurezza informatica in Europa. L'ENISA collabora con l'Unione europea e i suoi Stati membri, di concerto con il settore privato e i cittadini europei, al fine di delineare consigli e raccomandazioni sulle buone pratiche in materia di sicurezza informatica. Assiste inoltre gli Stati membri UE nell'attuazione della legislazione europea in materia e lavora per migliorare la resilienza delle infrastrutture critiche informatizzate e di rete. L'ENISA si adopera per potenziare l'attuale livello di competenza degli Stati membri UE, sostenendo lo sviluppo di comunità transnazionali impegnate a migliorare la sicurezza delle reti e informazioni in tutta l'Unione europea. Dal 2019 è attiva nell'elaborare schemi di certificazione per la sicurezza informatica. Maggiori informazioni sull'ENISA e le sue attività sono disponibili al seguente indirizzo: www.enisa.europa.eu.

ENISA
European Union Agency for Cybersecurity

Athens Office
1 Vasilissis Sofias Str
151 24 Marousi, Attiki, Greece

Heraklion office
95 Nikolaou Plastira
700 13 Vassilika Vouton, Heraklion, Greece

enisa.europa.eu

ISBN 978-92-9204-307-0
doi: 10,2824/247,711