GRUPPO DI LAVORO ARTICOLO 29 PER LA PROTEZIONE DEI DATI

0829/14/IT
WP216

Parere 05/2014 sulle tecniche di anonimizzazione

adottato il 10 aprile 2014

[NdT] Ai fini del presente parere, con "responsabile del trattamento" e con "incaricato del trattamento" si intendono rispettivamente il "titolare" e il "responsabile" di cui all'articolo 4, lettera f) e lettera g) del decreto legislativo 30 giugno 2003, n. 196 (codice in materia di protezione dei dati personali).

IL GRUPPO PER LA TUTELA DELLE PERSONE CON RIGUARDO AL TRATTAMENTO DEI DATI PERSONALI

istituito ai sensi della direttiva 95/46/CE del Parlamento europeo e del Consiglio del 24 ottobre 1995,

visti gli articoli 29 e 30 di detta direttiva,

visto il proprio regolamento,

HA ADOTTATO IL PRESENTE PARERE:

SINTESI

Nel presente parere, il Gruppo di lavoro esamina l'efficacia e i limiti delle tecniche di anonimizzazione esistenti rispetto al quadro giuridico dell'UE in materia di protezione dei dati e fornisce raccomandazioni per l'impiego di tali tecniche, tenendo conto del rischio residuo di identificazione insito in ciascuna di esse.

Il Gruppo di lavoro riconosce il valore potenziale dell'anonimizzazione, in particolare come strategia per consentire alle persone e alla società in senso lato di fruire dei vantaggi dei "dati aperti", attenuando al contempo i rischi per le persone interessate. Tuttavia, dagli studi di casi e dalle pubblicazioni di ricerca è emerso quanto sia difficile creare insiemi di dati effettivamente anonimi mantenendo al contempo tutte le informazioni sottostanti necessarie per espletare l'attività richiesta.

Alla luce della direttiva 95/46/CE e di altri strumenti giuridici rilevanti dell'UE, l'anonimizzazione è il risultato del trattamento di dati personali volto a impedire irreversibilmente l'identificazione. Nel procedere in tal senso, i responsabili del trattamento devono tener conto di svariati elementi e prendere in considerazione tutti i mezzi che "possono ragionevolmente" essere utilizzati per l'identificazione (dal responsabile del trattamento o da altri).

L'anonimizzazione costituisce un trattamento successivo dei dati personali; in quanto tale, deve soddisfare il requisito di compatibilità tenendo conto delle motivazioni giuridiche e delle circostanze del trattamento successivo. Inoltre, i dati resi anonimi non rientrano nell'ambito di applicazione della legislazione in materia di protezione dei dati, tuttavia le persone interessate potrebbero comunque avere diritto a tale tutela in base ad altre disposizioni (ad esempio, quelle che proteggono la riservatezza delle comunicazioni).

Il presente parere illustra le principali tecniche di anonimizzazione, ossia la randomizzazione e la generalizzazione. In particolare, il parere esamina l'aggiunta del rumore statistico, le permutazioni, la privacy differenziale, l'aggregazione, il *k*-anonimato, la *l*-diversità e la *t*-vicinanza. Ne illustra i principi, i punti di forza e di debolezza, nonché gli errori e gli insuccessi comuni connessi all'impiego di ciascuna tecnica.

Il parere esamina l'affidabilità di ogni tecnica sulla base di tre criteri:

i) è ancora possibile individuare una persona,

ii) è ancora possibile collegare i dati relativi a una persona, e

iii) è possibile dedurre informazioni riguardanti una persona?

Conoscere i principali punti di forza e debolezza di ciascuna tecnica è utile per decidere come progettare un processo di anonimizzazione adeguato in un determinato contesto.

Viene presa in esame anche la pseudonimizzazione al fine di chiarire alcune insidie e convinzioni erronee: la pseudonimizzazione non è un metodo di anonimizzazione. Si limita a ridurre la correlabilità di un insieme di dati all'identità originaria di una persona interessata, e rappresenta pertanto una misura di sicurezza utile.

Il parere giunge alla conclusione che le tecniche di anonimizzazione possono fornire garanzie di protezione della sfera privata e possono essere utilizzate per creare efficaci procedure di

anonimizzazione, ma soltanto se la loro applicazione viene progettata in maniera adeguata – nel senso che i requisiti preliminari (contesto) e l'obiettivo o gli obiettivi della procedura di anonimizzazione devono essere definiti in modo chiaro per ottenere l'anonimizzazione desiderata e produrre al contempo dati utili. La soluzione ottimale dovrebbe essere decisa caso per caso, se possibile utilizzando una combinazione di tecniche diverse e tenendo conto delle raccomandazioni pratiche formulate nel presente parere.

Infine, i responsabili del trattamento devono essere consapevoli che un insieme di dati resi anonimi può comunque presentare rischi residui per le persone interessate. Di fatto, da un lato anonimizzazione e reidentificazione sono argomenti attivi di ricerca e vengono regolarmente pubblicate nuove scoperte in materia e, dall'altro lato, persino i dati resi anonimi, come le statistiche, possono essere utilizzati per arricchire i profili esistenti delle persone, determinando quindi nuovi problemi di protezione dei dati. L'anonimizzazione non va pertanto considerata un'operazione una tantum e i relativi rischi dovrebbero essere oggetto di un riesame periodico da parte dei responsabili del trattamento.

1 Introduzione

Sebbene dispositivi, sensori e reti creino ingenti volumi e nuove tipologie di dati, e i costi di archiviazione dei dati stiano diventando trascurabili, crescono l'interesse e la domanda del pubblico riguardo al riutilizzo di tali dati. I "dati aperti" possono offrire vantaggi innegabili alla società, alle persone e alle organizzazioni, ma solo se viene rispettato il diritto di ognuno alla protezione dei propri dati personali e della propria sfera privata.

L'anonimizzazione potrebbe costituire una strategia valida per preservare i vantaggi e attenuare i rischi. Una volta che un insieme di dati viene reso effettivamente anonimo e le persone non sono più identificabili, le norme dell'Unione in materia di protezione dei dati non sono più applicabili. Tuttavia, dagli studi di casi e dalle pubblicazioni di ricerca emerge con chiarezza che non è così semplice creare un insieme di dati effettivamente anonimo a partire da un ampio insieme di dati personali, mantenendo al contempo tutte le informazioni sottostanti necessarie per espletare l'attività richiesta. Ad esempio, un insieme di dati considerato anonimo potrebbe essere combinato con un altro insieme di dati in maniera tale da consentire l'identificazione di una o più persone.

Nel presente parere, il Gruppo di lavoro esamina l'efficacia e i limiti delle tecniche di anonimizzazione esistenti rispetto al quadro giuridico dell'UE in materia di protezione dei dati e fornisce raccomandazioni per un uso oculato e responsabile di tali tecniche al fine di realizzare un processo di anonimizzazione.

2 Definizioni e analisi giuridica

2.1. Definizioni nel contesto giuridico dell'UE

La direttiva 95/46/CE fa riferimento all'anonimizzazione nel considerando 26 per escludere i dati resi anonimi dal campo di applicazione della legislazione in materia di protezione dei dati:

> *"considerando che i principi della tutela si devono applicare a ogni informazione concernente una persona identificata o identificabile; che, per determinare se una persona è identificabile, è opportuno prendere in considerazione l'insieme dei mezzi che possono essere ragionevolmente utilizzati dal responsabile del trattamento o da altri per identificare detta persona; che i principi della tutela non si applicano a dati resi anonimi in modo tale che la persona interessata non è più identificabile; che i codici di condotta ai sensi dell'articolo 27 possono costituire uno strumento utile di orientamento sui mezzi grazie ai quali dati possano essere resi anonimi e registrati in modo da rendere impossibile l'identificazione della persona interessata;"*[1].

Un'attenta lettura del considerando 26 fornisce una definizione concettuale di anonimizzazione. Dal considerando 26 si desume che per rendere anonimi determinati dati, gli stessi devono essere privati di elementi sufficienti per impedire l'identificazione della

[1] Va osservato inoltre che questo è lo stesso approccio seguito nel progetto di regolamento dell'Unione in materia di protezione dei dati, al considerando 23, "per determinare se una persona è identificabile, è opportuno prendere in considerazione l'insieme dei mezzi che possono essere ragionevolmente utilizzati dal responsabile del trattamento o da altri per identificare detta persona".

persona interessata. Più precisamente, i dati devono essere trattati in maniera tale da non poter più essere utilizzati per identificare una persona fisica utilizzando "l'insieme dei mezzi che possono essere ragionevolmente utilizzati" dal responsabile del trattamento o da altri. Un fattore importante è che il trattamento deve essere irreversibile. La direttiva non specifica come si debba o si possa effettuare il processo di anonimizzazione[2]. L'accento è posto sul risultato: i dati devono essere tali da non consentire l'identificazione della persona interessata mediante "l'insieme" dei mezzi che "possono" essere "ragionevolmente" utilizzati. Si fa riferimento ai codici di condotta come strumento per stabilire possibili meccanismi di anonimizzazione e alla conservazione in una forma tale da "rendere impossibile" l'identificazione della persona interessata. La direttiva stabilisce pertanto requisiti molto rigorosi.

Anche la direttiva relativa alla vita privata e alle comunicazioni elettroniche (direttiva 2002/58/CE) fa riferimento ad "anonimizzazione" e "dati anonimi" in un contesto molto simile. Il considerando 26 sancisce che:

> *"I dati relativi al traffico utilizzati per la commercializzazione dei servizi di comunicazione o per la fornitura di servizi a valore aggiunto dovrebbero inoltre essere cancellati o resi anonimi dopo che il servizio è stato fornito".*

Di conseguenza, l'articolo 6, paragrafo 1, sancisce che:

> *"I dati sul traffico relativi agli abbonati ed agli utenti, trattati e memorizzati dal fornitore di una rete pubblica o di un servizio pubblico di comunicazione elettronica devono essere cancellati o resi anonimi quando non sono più necessari ai fini della trasmissione di una comunicazione, fatti salvi i paragrafi 2, 3 e 5 del presente articolo e l'articolo 15, paragrafo 1."*

All'articolo 9, paragrafo 1, si legge inoltre:

> *"Se i dati relativi all'ubicazione diversi dai dati relativi al traffico, relativi agli utenti o abbonati di reti pubbliche di comunicazione o servizi di comunicazione elettronica accessibili al pubblico possono essere sottoposti a trattamento, essi possono esserlo soltanto a condizione che siano stati resi anonimi o che l'utente o l'abbonato abbiano dato il loro consenso, e sempre nella misura e per la durata necessaria per la fornitura di un servizio a valore aggiunto."*

Il fondamento logico è che il risultato dell'anonimizzazione quale tecnica applicata ai dati personali dovrebbe essere, allo stato attuale della tecnologia, permanente come una cancellazione, vale a dire dovrebbe rendere impossibile il trattamento dei dati personali.[3]

---

[2] Tale concetto viene ripreso e approfondito a pagina 8 del presente parere.

[3] Si rammenta che l'anonimizzazione viene definita anche in norme internazionali quali ISO 29100 come processo nel quale le informazioni personali identificabili (IPI) sono modificate irreversibilmente in modo tale che un titolare di IPI non possa più essere identificato direttamente o indirettamente, né dal singolo responsabile del trattamento di IPI né dallo stesso in collaborazione con altri (ISO 29100:2011). Anche per l'ISO l'elemento fondamentale è l'irreversibilità delle modifiche subite dai dati personali per consentirne l'identificazione diretta o indiretta. Da questo punto di vista, esiste un considerevole livello di convergenza con i principi e concetti alla base della direttiva 95/46/CE. Ciò vale anche per le definizioni che compaiono in alcune leggi nazionali (ad esempio, in Italia, Germania e Slovenia), dove l'accento è posto sulla non identificabilità e si fa riferimento allo "sforzo sproporzionato" per la reidentificazione (D, SI). Tuttavia, la legge francese in materia di protezione dei dati prevede che i dati rimangano dati personali anche se è estremamente difficile e improbabile reidentificare la persona interessata – vale a dire, non vi sono disposizioni che fanno riferimento al test di "ragionevolezza".

2.2. Analisi giuridica

L'analisi della formulazione del testo relativo all'anonimizzazione negli strumenti più significativi dell'UE in materia di protezione dei dati consente di porre in evidenza quattro caratteristiche chiave:

- l'anonimizzazione può essere il risultato del trattamento di dati personali allo scopo di impedire irreversibilmente l'identificazione della persona interessata;

- possono essere previste diverse tecniche di anonimizzazione, non esiste alcuna norma prescrittiva nella legislazione dell'Unione;

- deve essere attribuita importanza agli elementi contestuali: è opportuno prendere in considerazione "l'insieme" dei mezzi che "possono essere ragionevolmente" utilizzati per l'identificazione da parte del responsabile del trattamento o di altri, prestando particolare attenzione a ciò che ultimamente, allo stato attuale della tecnologia, è diventato "ragionevolmente utilizzabile" (dato l'incremento della potenza di calcolo e degli strumenti disponibili);

- l'anonimizzazione presenta un fattore di rischio intrinseco: occorre tenerne conto nel valutare la validità di qualsiasi tecnica di anonimizzazione – compresi gli impieghi possibili dei dati "resi anonimi" mediante tale tecnica – e vanno soppesate la gravità e la probabilità di tale rischio.

Nel presente parere, si utilizza l'espressione "tecnica di anonimizzazione", invece di "anonimato" o "dati anonimi", per precisare il rischio intrinseco residuo di reidentificazione correlato a qualsiasi misura tecnico-organizzativa tesa a rendere "anonimi" i dati.

2.2.1. Legittimità del processo di anonimizzazione

In primo luogo, l'anonimizzazione è una tecnica che si applica ai dati personali al fine di ottenere una deidentificazione irreversibile. Pertanto, l'assunto di partenza è che i dati personali devono essere stati raccolti e trattati in conformità alla legislazione applicabile in materia di conservazione dei dati in un formato identificabile.

In tale contesto, il processo di anonimizzazione, inteso come trattamento di dati personali per ottenerne l'anonimizzazione, rappresenta un "trattamento successivo". In quanto tale, il trattamento deve superare la prova di compatibilità conformemente agli orientamenti forniti dal Gruppo di lavoro nel suo parere 03/2013 sulla limitazione della finalità[4].

Ne consegue che, in linea di principio, la base giuridica per l'anonimizzazione può essere individuata in ciascuna delle motivazioni citate all'articolo 7 (compreso l'interesse legittimo del responsabile del trattamento), a condizione che siano soddisfatti anche i requisiti di qualità dei dati di cui all'articolo 6 della direttiva e tenuto debitamente conto delle circostanze specifiche e di tutti i fattori citati nel parere del Gruppo di lavoro sulla limitazione della finalità[5].

[4] Parere 03/2013 del Gruppo di lavoro articolo 29 disponibile all'indirizzo: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

[5] Ciò significa in particolare che occorre condurre una valutazione sostanziale alla luce di tutte le circostanze rilevanti, prestando particolare attenzione ai seguenti fattori chiave:

D'altro canto, vanno citate anche le disposizioni contenute nell'articolo 6, paragrafo 1, lettera e), della direttiva 95/46/CE (ma anche nell'articolo 6, paragrafo 1 e nell'articolo 9, paragrafo 1, della direttiva relativa alla vita privata e alle comunicazioni elettroniche), in quanto dimostrano la necessità di conservare i dati personali "in modo da consentire l'identificazione" per un arco di tempo non superiore a quello necessario al conseguimento delle finalità per le quali sono stati rilevati o successivamente trattati.

Di per sé, la disposizione insiste sul fatto che i dati personali andrebbero almeno resi anonimi "per impostazione predefinita" (in conformità a diversi requisiti giuridici, come quelli menzionati nella direttiva relativa alla vita privata e alle comunicazioni elettroniche riguardo ai dati relativi al traffico). Se il responsabile del trattamento desidera conservare tali dati personali una volta conseguite le finalità del trattamento originario o successivo, dovrebbero essere utilizzate tecniche di anonimizzazione in modo da impedire irreversibilmente l'identificazione.

Di conseguenza, il Gruppo di lavoro ritiene che l'anonimizzazione, quale trattamento successivo di dati personali, possa essere considerata compatibile con le finalità originarie del trattamento, ma solo a condizione che il processo di anonimizzazione sia tale da produrre informazioni rese anonime nel senso descritto nel presente documento.

Va inoltre sottolineato che l'anonimizzazione deve essere conforme ai vincoli giuridici richiamati dalla Corte di giustizia europea nella sua decisione in merito alla causa C-553/07 (*College van burgemeester en wethouders van Rotterdam/M.E.E. Rijkeboer*), in relazione alla necessità di conservare i dati in forma identificabile in modo da consentire, ad esempio, l'esercizio dei diritti di accesso da parte delle persone interessate. La Corte di giustizia ha decretato che "*L'art. 12, lett. a), della direttiva [95/46/CE] impone agli Stati membri di prevedere il diritto di accesso alle informazioni sui destinatari o sulle categorie di destinatari dei dati nonché sul contenuto delle informazioni comunicate non solo per il presente, ma anche per il passato. Spetta agli Stati membri fissare il termine per la conservazione di tali informazioni nonché il corrispondente accesso alle stesse che costituiscano un giusto equilibrio tra, da una parte, l'interesse della persona di cui trattasi a tutelare la propria sfera privata, in particolare, tramite i mezzi di intervento e le possibilità di agire in giudizio previste dalla direttiva e, dall'altra, l'onere che l'obbligo di conservare tali informazioni comporta per il responsabile del trattamento*."

Ciò assume particolare rilevanza nel caso in cui un responsabile del trattamento si basi sull'articolo 7, lettera f), della direttiva 95/46/CE per quanto riguarda l'anonimizzazione: occorre sempre trovare un equilibrio tra l'interesse legittimo del responsabile del trattamento e i diritti e le libertà fondamentali delle persone interessate.

Ad esempio, un'inchiesta condotta nel 2012-2013 dal garante olandese per la protezione dei dati riguardo all'utilizzo delle tecnologie di *deep packet inspection* (ispezione approfondita dei pacchetti) da parte di quattro gestori di telefonia mobile ha individuato l'esistenza di un motivo giuridico basato sull'articolo 7, lettera f), della direttiva 95/46/CE per rendere anonimi i dati relativi al traffico quanto prima possibile dopo la loro rilevazione. In effetti, l'articolo 6

---

a) il rapporto tra le finalità per le quali sono stati raccolti i dati personali e le finalità del loro trattamento successivo;
b) il contesto in cui sono stati raccolti i dati personali e le ragionevoli aspettative delle persone interessate circa il loro ulteriore impiego;
c) la natura dei dati personali e l'impatto del trattamento successivo sulle persone interessate;
d) le misure di salvaguardia adottate dal responsabile del trattamento per garantire un trattamento equo e per prevenire ripercussioni indesiderate sulle persone interessate.

della direttiva relativa alla vita privata e alle comunicazioni elettroniche stabilisce che i dati sul traffico relativi agli abbonati ed agli utenti, trattati e memorizzati dal fornitore di una rete pubblica o di un servizio pubblico di comunicazione elettronica devono essere cancellati o resi anonimi quanto prima possibile. In questo caso, poiché è consentito ai sensi dell'articolo 6 della direttiva relativa alla vita privata e alle comunicazioni elettroniche, esiste un motivo giuridico corrispondente nell'articolo 7 della direttiva concernente la protezione dei dati. Si potrebbe anche presentare l'argomentazione al contrario: se una tipologia di trattamento dei dati non è consentita ai sensi dell'articolo 6 della direttiva relativa alla vita privata e alle comunicazioni elettroniche, non può esservi un motivo giuridico nell'articolo 7 della direttiva concernente la protezione dei dati.

2.2.2. Potenziale identificabilità dei dati resi anonimi

Il Gruppo di lavoro ha approfondito il concetto di dati personali nel parere 4/2007 concernente i dati personali, concentrandosi sugli elementi essenziali della definizione contenuta nell'articolo 2, lettera a), della direttiva 95/46/CE, compresa la parte "identificata o identificabile" di tale definizione. In questo contesto, il Gruppo di lavoro ha nuovamente concluso che "i dati anonimizzati sarebbero quindi dati anonimi già corrispondenti a una persona identificabile ma che non ne permettono più l'identificazione".

Il Gruppo di lavoro ha pertanto già chiarito che la direttiva suggerisce l'esame dei "mezzi … che possono essere ragionevolmente utilizzati" quale criterio da applicare per valutare se il processo di anonimizzazione sia sufficientemente affidabile, vale a dire se l'identificazione sia diventata "ragionevolmente" impossibile. Il contesto e le circostanze particolari di un caso specifico incidono direttamente sull'identificabilità. Nell'allegato tecnico al presente parere, si fornisce un'analisi dell'impatto della scelta della tecnica più appropriata.

Com'è stato già sottolineato, la ricerca, gli strumenti e la potenza di calcolo si evolvono continuamente. Pertanto, non è possibile né utile fornire un elenco esauriente delle circostanze in cui l'identificazione non risulta più possibile. Alcuni fattori chiave meritano comunque di essere considerati e illustrati.

In primo luogo, si può sostenere che i responsabili del trattamento debbano concentrarsi sui mezzi concreti necessari per invertire il processo di anonimizzazione, in particolare per quanto riguarda i costi e le competenze necessarie a mettere in atto tali sistemi e la valutazione della loro probabilità e gravità. Ad esempio, gli sforzi compiuti e i costi sostenuti ai fini dell'anonimizzazione (in termini di tempi e risorse richiesti) dovrebbero essere misurati in rapporto alla crescente disponibilità di mezzi tecnici a basso costo per identificare le persone nelle banche dati, all'accessibilità pubblica sempre maggiore di altre banche dati (come quelle rese accessibili in relazione alle politiche in materia di "dati aperti"), e ai numerosi esempi di anonimizzazione incompleta che comportano effetti conseguenti avversi e a volte irreparabili per le persone interessate[6]. Va sottolineato che il rischio di identificazione può aumentare col tempo e dipende anche dallo sviluppo della tecnologia dell'informazione e della comunicazione. Le disposizioni giuridiche, se del caso, devono essere pertanto

[6] È interessante notare che gli emendamenti del Parlamento europeo al progetto di regolamento generale sulla protezione dei dati, nella versione presentata di recente (21 ottobre 2013), sanciscono specificamente nel considerando 23 che "Per accertare la ragionevole probabilità che i mezzi siano utilizzati per identificare la persona, è opportuno prendere in considerazione tutti i fattori obiettivi, tra cui i costi e il tempo necessario per l'identificazione, tenendo conto sia delle tecnologie disponibili al momento del trattamento sia dello sviluppo tecnologico".

formulate in maniera neutrale sotto il profilo tecnologico e idealmente dovrebbero tener conto dei cambiamenti del potenziale di sviluppo della tecnologia dell'informazione[7].

In secondo luogo, "l'insieme dei mezzi che possono essere ragionevolmente utilizzati per determinare se una persona è identificabile" sono quelli che devono essere utilizzati "dal responsabile del trattamento o da altri". Pertanto, è essenziale comprendere che quando un responsabile del trattamento non cancella i dati originali (identificabili) a livello di evento, e trasmette poi parte di questo insieme di dati (ad esempio, dopo l'eliminazione o il mascheramento dei dati identificabili), l'insieme di dati risultante contiene ancora dati personali. Soltanto se il responsabile del trattamento aggrega i dati a un livello in cui i singoli eventi non sono più identificabili si può definire anonimo l'insieme di dati risultante. Ad esempio, se un'organizzazione raccoglie dati sugli spostamenti delle persone, i tipi di spostamenti individuali a livello di evento rientrano ancora tra i dati personali per tutte le parti coinvolte, fintantoché il responsabile del trattamento (o altri) ha ancora accesso ai dati non trattati originali, anche se gli identificatori diretti sono stati espunti dall'insieme dei dati forniti a terzi. Tuttavia, se il responsabile del trattamento cancella i dati non trattati e fornisce a terzi solamente statistiche aggregate ad alto livello, ad esempio "il lunedì sulla rotta X i passeggeri sono più numerosi del 160% rispetto al martedì", i dati possono essere definiti anonimi.

Un'efficace soluzione di anonimizzazione impedisce a tutte le parti di identificare una persona in un insieme di dati, di collegare due dati all'interno di un insieme di dati (o tra due insiemi distinti di dati) e di dedurre informazioni da tale insieme di dati. In generale, eliminare elementi direttamente identificanti non è pertanto di per sé sufficiente a garantire che l'identificazione della persona interessata non sia più possibile. Spesso è necessario adottare misure supplementari per prevenire l'identificazione, ancora una volta a seconda del contesto e degli scopi del trattamento cui sono destinati i dati resi anonimi.

ESEMPIO
I profili di dati genetici costituiscono un esempio di dati personali che possono essere a rischio di identificazione se l'unica tecnica utilizzata è l'eliminazione dell'identità del donatore, data la natura unica di determinati profili. È già stato dimostrato in letteratura[8] che la combinazione di risorse genetiche pubblicamente accessibili (ad esempio, registri genealogici, necrologie, risultati delle interrogazioni dei motori di ricerca) e di metadati concernenti i donatori di DNA (data della donazione, età, luogo di residenza) possono rivelare l'identità di determinate persone, anche se il DNA è stato donato "in forma anonima".

Entrambe le famiglie di tecniche di anonimizzazione, ossia randomizzazione e generalizzazione dei dati[9], presentano carenze; tuttavia, ognuna di esse può rivelarsi adeguata, in circostanze e contesti specifici, per conseguire lo scopo desiderato senza compromettere la sfera privata delle persone interessate. Va chiarito che per "identificazione" non si intende solo la possibilità di recuperare il nome e/o l'indirizzo di una persona, ma anche la potenziale identificabilità mediante individuazione, correlabilità e deduzione. Inoltre, l'applicabilità della normativa in materia di protezione dei dati non dipende dalle intenzioni del responsabile del trattamento o del destinatario. Nella misura in cui i dati sono identificabili, si applicano le norme in materia di protezione dei dati.

[7] Cfr. il parere 4/2007 del Gruppo di lavoro articolo 29, pag. 15.
[8] Cfr. John Bohannon, *Genealogy Databases Enable Naming of Anonymous DNA Donors*, Science, Vol. 339, n. 6117 (18 gennaio 2013), pag. 262.
[9] Le principali caratteristiche e differenze di queste due tecniche di anonimizzazione sono descritte alla sezione 3 del presente documento ("Analisi tecnica").

Nei casi in cui un insieme di dati sottoposto a una tecnica di anonimizzazione (anonimizzato e reso pubblico dal responsabile del trattamento originario) sia oggetto di trattamento da parte di terzi, questi ultimi possono procedere in modo legittimo senza necessariamente tener conto dei requisiti in materia di protezione dei dati, a condizione che non possano (direttamente o indirettamente) identificare le persone interessate nell'insieme di dati originario. Tuttavia, tali terzi sono tenuti a prendere in considerazione tutti i fattori contestuali e circostanziali summenzionati (tra cui le caratteristiche specifiche delle tecniche di anonimizzazione applicate dal responsabile del trattamento originario) al momento di decidere come utilizzare e, soprattutto, mettere insieme tali dati anonimizzati per le loro finalità – in quanto le conseguenze che ne derivano possono comportare tipologie diverse di responsabilità a loro carico. Nei casi in cui i fattori e le caratteristiche in questione siano tali da comportare un rischio inaccettabile di identificazione delle persone interessate, il trattamento rientra nuovamente nell'ambito di applicazione della normativa in materia di protezione dei dati.

Quanto precede non va in alcun modo considerato un elenco esaustivo, ma fornisce un orientamento generale sull'approccio da adottare per valutare la potenziale identificabilità di un insieme di dati specifico sottoposto ad anonimizzazione in base alle diverse tecniche disponibili. Tutti i fattori suddetti possono essere considerati alla stregua di altrettanti fattori di rischio che devono essere valutati sia dai responsabili del trattamento quando anonimizzano gli insiemi di dati, sia dai terzi quando utilizzano tali insiemi di dati "resi anonimi" per le loro finalità.

2.2.3. Rischi connessi all'utilizzo di dati resi anonimi

All'atto di valutare il ricorso alle tecniche di anonimizzazione, i responsabili del trattamento devono tener conto dei rischi di seguito descritti.

- Un rischio specifico consiste nel considerare i dati pseudonimizzati equivalenti ai dati resi anonimi. La sezione relativa all'analisi tecnica precisa che i dati pseudonimizzati non possono essere equiparati a informazioni rese anonime, in quanto continuano a permettere l'identificazione delle singole persone e le rendono trasversalmente collegabili a diversi insiemi di dati. Gli pseudonimi consentono potenzialmente l'identificabilità e rientrano pertanto nel campo di applicazione del regime giuridico della protezione dei dati. Tale aspetto assume una particolare rilevanza nel contesto della ricerca scientifica, statistica o storica[10].

ESEMPIO
Un esempio tipico di convinzione erronea concernente la pseudonimizzazione è costituito dal famoso "caso AOL (America On Line)". Nel 2006 è stata resa pubblica una banca dati contenente venti milioni di parole chiave di ricerca relative a più di 650 000 utenti riferite a un arco di tempo di tre mesi, e l'unica misura a tutela della sfera privata applicata consisteva nel sostituire l'identificativo degli utenti AOL con un attributo numerico. La conseguenza è stata l'identificazione pubblica e la localizzazione di alcuni utenti. Le stringhe di risultati pseudonimizzate dei motori di ricerca, soprattutto se unite ad altri attributi quali indirizzi IP o altri parametri di configurazione dei clienti, possiedono un potere molto elevato di identificazione.

- Un secondo errore consiste nel ritenere che dati adeguatamente anonimizzati (che soddisfano tutte le condizioni e criteri summenzionati e non rientrano per definizione nel campo di applicazione della direttiva sulla protezione dei dati) privino le persone di qualsivoglia salvaguardia, anzitutto per il motivo che all'utilizzo di tali dati potrebbero essere applicabili altri atti legislativi. Ad esempio, l'articolo 5, paragrafo 3, della direttiva relativa alla vita privata e alle comunicazioni elettroniche vieta l'archiviazione di "informazioni" di qualsiasi tipo (comprese informazioni non personali) e l'accesso alle stesse su apparecchi

[10] Cfr. anche il parere 4/2007 del Gruppo di lavoro articolo 29, pagg. 18-20.

terminali in assenza del consenso dell'abbonato/utente, in quanto ciò rientra nel principio più generale di riservatezza delle comunicazioni.

- Un terzo caso di negligenza potrebbe scaturire dal non considerare l'impatto sulle persone, in determinate circostanze, di dati adeguatamente anonimizzati, soprattutto nel caso della definizione di profili. La sfera della vita privata di una persona è tutelata dall'articolo 8 della Convenzione europea dei diritti dell'uomo e dall'articolo 7 della Carta dei diritti fondamentali dell'UE; di conseguenza, benché le norme in materia di protezione dei dati possano non essere più applicabili a tale tipologia di dati, l'utilizzo degli insiemi di dati resi anonimi e pubblicati ad uso di terzi potrebbe determinare una perdita di riservatezza. Occorre prestare particolare attenzione quando si gestiscono informazioni rese anonime, soprattutto ogniqualvolta tali informazioni vengono utilizzate (spesso in combinazione con altri dati) per prendere decisioni che producono effetti (sebbene indirettamente) sulle persone. Come già precisato nel presente parere e chiarito dal Gruppo di lavoro in particolare nel parere sul concetto di "limitazione della finalità" (parere 03/2013)[11], le aspettative legittime delle persone interessate in merito a trattamenti successivi dei loro dati vanno valutate alla luce dei fattori contestuali rilevanti, quali la natura del rapporto tra le persone interessate e i responsabili del trattamento, gli obblighi normativi applicabili e la trasparenza dei trattamenti di dati.

3 Analisi tecnica, affidabilità delle tecnologie ed errori ricorrenti

Esistono diverse pratiche e tecniche di anonimizzazione che presentano gradi variabili di affidabilità. La presente sezione affronta i punti principali che i responsabili del trattamento devono prendere in considerazione nell'applicarle e verte in particolare sul livello di garanzia che una data tecnica consente di ottenere tenendo conto dello stato attuale della tecnologia e di tre rischi essenziali per l'anonimizzazione:

- *individuazione*, che corrisponde alla possibilità di isolare alcuni o tutti i dati che identificano una persona all'interno dell'insieme di dati;
- *correlabilità*, vale a dire la possibilità di correlare almeno due dati concernenti la medesima persona interessata o un gruppo di persone interessate (nella medesima banca dati o in due diverse banche dati). Se un intruso riesce a determinare (ad esempio mediante un'analisi della correlazione) che due dati sono assegnati allo stesso gruppo di persone, ma non è in grado di identificare alcuna persona del gruppo, la tecnica fornisce una protezione contro l'individuazione, ma non contro la correlabilità;
- *deduzione*, vale a dire la possibilità di desumere, con un alto grado di probabilità, il valore di un attributo dai valori di un insieme di altri attributi.

Pertanto, una soluzione che elimini i tre rischi suddetti sarebbe utile per impedire la reidentificazione effettuata mediante i mezzi più probabili e ragionevoli che potrebbero essere utilizzati dal responsabile del trattamento e da altri. A tale riguardo, il Gruppo di lavoro sottolinea che le tecniche di deidentificazione e di anonimizzazione sono oggetto di ricerca continua e che tale ricerca ha ripetutamente dimostrato che nessuna tecnica è di per sé esente da carenze. In generale, esistono due diversi approcci all'anonimizzazione: il primo si basa

[11] Disponibile all'indirizzo http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

sulla ***randomizzazione***, mentre il secondo si basa sulla ***generalizzazione***. Il parere esamina inoltre altri concetti quali *pseudonimizzazione, privacy differenziale, l-diversità, t-vicinanza*.

Il presente parere utilizza la seguente terminologia in questa sezione: un insieme di dati è composto da diversi dati relativi a persone (le persone interessate). Ogni dato si riferisce a una persona interessata ed è composto da un insieme di valori (o "immissioni", ad esempio: 2013) per ciascun attributo (ad esempio, anno). Un insieme di dati è una raccolta di dati che possono essere configurati alternativamente come tabella (o insieme di tabelle) o come grafico annotato/ponderato, come attualmente avviene sempre più spesso. Gli esempi riportati nel parere fanno riferimento a tabelle, ma si possono applicare anche ad altre rappresentazioni grafiche di dati. Combinazioni di attributi relativi a una persona interessata o a un gruppo di persone interessate possono essere definite quasi-identificatori. In alcuni casi, un insieme di dati può contenere dati molteplici relativi a un'unica persona. Un "intruso" è un soggetto terzo (vale a dire, né il responsabile del trattamento né l'incaricato del trattamento) che accede ai dati originali in maniera accidentale o intenzionale.

3.1. Randomizzazione

La randomizzazione è una famiglia di tecniche che modifica la veridicità dei dati al fine di eliminare la forte correlazione che esiste tra i dati e la persona. Se i dati sono sufficientemente incerti non possono più essere riferiti a una persona specifica. Di per sé la randomizzazione non riduce l'unicità di ogni dato, in quanto ciascun dato può comunque essere ancora estrapolato da un'unica persona interessata, ma può rappresentare una tutela dagli attacchi/rischi di deduzione e può essere affiancata da tecniche di generalizzazione per fornire maggiori garanzie di tutela della sfera privata. Potrebbe essere necessario applicare tecniche supplementari per garantire che un dato non possa identificare una singola persona.

3.1.1. Aggiunta del rumore statistico

La tecnica dell'aggiunta del rumore statistico può rivelarsi utile soprattutto nel caso in cui gli attributi possano avere un effetto avverso importante sulle persone e consiste nel modificare gli attributi contenuti nell'insieme di dati in modo tale da renderli meno accurati mantenendo nel contempo la distribuzione generale. All'atto di trattare un insieme di dati, un osservatore parte dal presupposto che i valori siano accurati, ma ciò corrisponde solo limitatamente al vero. Ad esempio, se l'altezza di una persona è stata originariamente misurata approssimandola al centimetro più vicino, l'insieme di dati anonimizzati potrebbe contenere un'altezza accurata solo con un'approssimazione di +-10cm. Se la tecnica viene applicata in maniera efficace, eventuali terzi non riescono a identificare una persona né possono riparare i dati o altrimenti desumere in che modo gli stessi sono stati modificati.

Di solito, l'aggiunta del rumore statistico deve essere affiancata da altre tecniche di anonimizzazione, quali l'eliminazione degli attributi ovvi e dei quasi-identificatori. Il livello di rumore statistico dovrebbe dipendere dal livello di informazioni richieste e dall'impatto sulla sfera privata delle persone in seguito alla divulgazione degli attributi protetti.

3.1.1.1. Garanzie

- *Individuazione:* è ancora possibile individuare i dati riferiti a una persona (magari in modo non identificabile) anche se i dati sono meno affidabili.
- *Correlabilità:* è ancora possibile correlare i dati della stessa persona, ma i dati sono meno affidabili e pertanto un dato reale può essere correlato a un altro che è stato aggiunto artificialmente (ad esempio, per creare rumore statistico). In alcuni casi, un'attribuzione errata potrebbe esporre una persona interessata a un livello di rischio significativo e persino maggiore di una corretta.
- *Deduzione:* gli attacchi tramite deduzione sono possibili, ma la probabilità di successo è minore e potrebbero comparire alcuni falsi positivi (e falsi negativi).

3.1.1.2. Errori comuni

- Aggiungere rumore statistico incoerente: se il rumore statistico non è semanticamente plausibile (vale a dire, è "fuori scala" e non rispetta la logica tra gli attributi in un dato insieme), un intruso che acceda alla banca dati potrebbe filtrare il rumore statistico e, in alcuni casi, rigenerare le voci mancanti. Inoltre, se l'insieme di dati è troppo scarso[12], permane la possibilità di correlare a una fonte esterna le immissioni di dati inseriti per creare rumore.
- Presumere che l'aggiunta di rumore statistico sia sufficiente: l'aggiunta di rumore statistico è una misura complementare che ostacola il recupero dei dati personali da parte di un eventuale intruso. A meno che il rumore statistico non copra le informazioni contenute nell'insieme di dati, non si dovrebbe presumere che l'aggiunta di rumore statistico rappresenti una soluzione a sé stante per l'anonimizzazione.

3.1.1.3. Casi di insuccesso dell'aggiunta di rumore statistico

Un esperimento di reidentificazione molto famoso è quello condotto sulla banca dati dei clienti del fornitore di contenuti video Netflix. I ricercatori hanno analizzato le proprietà geometriche della banca dati costituita da più di 100 milioni di giudizi su una scala da 1 a 5 espressi da quasi 500 000 utenti su oltre 18 000 film; tali dati erano stati resi pubblici dalla società dopo che erano stati "resi anonimi" in base a una politica interna in materia di tutela della sfera privata che prevedeva l'eliminazione di tutte le informazioni che identificavano i clienti tranne i giudizi e le date. Era stato anche aggiunto il rumore statistico quando i giudizi erano leggermente migliorati o peggiorati.

Nonostante ciò, è stato appurato che il 99% dei dati degli utenti poteva essere identificato in maniera univoca all'interno dell'insieme di dati utilizzando 8 giudizi e date con errori di 14 giorni come criteri di selezione, mentre l'abbassamento dei criteri di selezione (2 giudizi e un errore di 3 giorni) consentiva comunque di identificare il 68% degli utenti[13].

3.1.2. Permutazione

La tecnica, che consiste nel mescolare i valori degli attributi all'interno di una tabella in modo tale che alcuni di essi risultino artificialmente collegati a diverse persone interessate, è utile quando è importante mantenere l'esatta distribuzione di ciascun attributo all'interno dell'insieme di dati.

[12] Tale concetto viene ulteriormente approfondito nell'allegato, pag. 30.
[13] Narayanan, A., & Shmatikov, V. (2008, maggio). *Robust de-anonymization of large sparse datasets*. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on* (pagg. 111-125). IEEE.

La permutazione può essere considerata una forma speciale di aggiunta di rumore statistico. Nella tecnica classica di aggiunta del rumore, gli attributi vengono modificati mediante valori randomizzati. La generazione di rumore statistico coerente può rappresentare un'operazione difficile da effettuare, mentre modificare solo marginalmente i valori degli attributi potrebbe non tutelare adeguatamente la sfera privata. In alternativa, le tecniche di permutazione modificano i valori contenuti nell'insieme di dati semplicemente permutandoli da un dato all'altro. Tali scambi garantiscono che gamma e distribuzione dei valori rimangano invariate, a differenza delle correlazioni tra valori e persone. Se tra due o più attributi sussiste un legame logico o una correlazione statistica e gli stessi vengono permutati in maniera indipendente, tale legame verrà meno. Può pertanto essere importante permutare un insieme di attributi correlati in modo da non spezzare il legame logico, altrimenti un intruso potrebbe individuare gli attributi permutati e invertire la permutazione.

Ad esempio, se consideriamo un sottogruppo di attributi in un insieme di dati medici quali "motivi del ricovero/sintomi/reparto responsabile", nella maggior parte dei casi tra i valori sussisteranno legami logici forti e la permutazione di un unico valore verrebbe pertanto individuata e potrebbe persino essere invertita.

Analogamente all'aggiunta di rumore statistico, la permutazione di per sé potrebbe non garantire l'anonimizzazione e dovrebbe pertanto essere sempre accompagnata dall'eliminazione degli attributi ovvi/quasi-identificatori.

3.1.2.1. Garanzie

- *Individuazione:* analogamente all'aggiunta di rumore statistico, permane la possibilità di individuare i dati di una persona, ma gli stessi sono meno affidabili.
- *Correlabilità:* se la permutazione riguarda attributi e quasi-identificatori, potrebbe impedire una correlazione "corretta" di attributi a un insieme di dati sia internamente sia esternamente, ma consentirebbe comunque una correlabilità "non corretta", in quanto un'immissione reale potrebbe essere associata a una persona interessata diversa.
- *Deduzione:* permane la possibilità di trarre deduzioni dall'insieme di dati, specialmente se gli attributi sono correlati o uniti da un legame causale logico forte; tuttavia, non sapendo quali attributi sono stati permutati, l'intruso deve tener conto del fatto che la propria deduzione si basa su un'ipotesi errata e quindi può ricorrere soltanto alla deduzione probabilistica.

3.1.2.2. Errori comuni

- Selezione dell'attributo sbagliato: la permutazione di attributi non sensibili o non rischiosi non offrirebbe grandi vantaggi in termini di protezione dei dati personali. Se gli attributi sensibili/rischiosi fossero ancora associati all'attributo originario, un intruso potrebbe ancora estrapolare informazioni sensibili sulle persone.
- Permutazione casuale degli attributi: se tra due attributi sussiste una forte correlazione, la loro permutazione casuale non fornisce garanzie degne di nota. L'errore comune in questione è illustrato nella tabella 1.
- Ritenere che la permutazione sia sufficiente: analogamente all'aggiunta di rumore statistico, la permutazione di per sé non garantisce l'anonimato e dovrebbe essere affiancata da altre tecniche, come l'eliminazione degli attributi ovvi.

3.1.2.3. Casi di insuccesso della permutazione

L'esempio dimostra come permutare gli attributi in maniera casuale produca garanzie di tutela della sfera privata inconsistenti se tra diversi attributi sussistono legami logici. In seguito al tentativo di anonimizzazione, è molto semplice dedurre il reddito di ogni persona in base alla professione (e all'anno di nascita). Ad esempio, da un esame diretto dei dati è possibile sostenere che l'amministratore delegato presente nella tabella è nato molto presumibilmente nel 1957 e percepisce la retribuzione più elevata, mentre il disoccupato è nato nel 1964 e ha il reddito più basso.

Anno	Sesso	Professione	Reddito (permutato)
1957	M	Ingegnere	70k
1957	M	Amministratore delegato	5k
1957	M	Disoccupato	43k
1964	M	Ingegnere	100k
1964	M	Dirigente	45k

Tabella 1. Esempio di anonimizzazione inefficace mediante la permutazione di attributi correlati

3.1.3. Privacy differenziale

La privacy differenziale[14] appartiene alla famiglia delle tecniche di randomizzazione, ma adotta un approccio diverso: mentre l'inserimento del rumore statistico interviene prima, al momento dell'eventuale pubblicazione dell'insieme di dati, la privacy differenziale può essere utilizzata quando il responsabile del trattamento genera opinioni anonimizzate di un insieme di dati e conserva al contempo una copia dei dati originali. Le opinioni anonimizzate sono solitamente generate attraverso un sottogruppo di interrogazioni per terzi specifici. Il sottogruppo presenta una certa dose di rumore statistico casuale aggiunto appositamente a posteriori. La privacy differenziale suggerisce al responsabile del trattamento la quantità e la forma di rumore statistico che va aggiunto per ottenere le garanzie di tutela della sfera privata richieste[15]. In tale contesto, è particolarmente importante continuare a controllare (almeno per ogni nuova interrogazione) che non sussista la possibilità di identificare una persona nell'insieme dei risultati dell'interrogazione. Occorre tuttavia chiarire che le tecniche di privacy differenziale non modificano i dati originari e pertanto, finché questi permangono, il responsabile del trattamento è in grado di identificare le persone all'interno dei risultati delle interrogazioni di privacy differenziale tenendo conto dell'insieme dei mezzi che possono essere ragionevolmente utilizzati. Tali risultati vanno trattati alla stregua di dati personali.

Un vantaggio offerto da un approccio basato sulla privacy differenziale consiste nel fatto che gli insiemi di dati sono forniti a terzi autorizzati in risposta a una richiesta specifica e non tramite la pubblicazione di un unico insieme di dati. Per agevolare la revisione, il responsabile del trattamento può acquisire un elenco di tutte le interrogazioni e richieste e accertarsi che terzi non accedano a dati per i quali non possiedono l'autorizzazione. Un'interrogazione può anche essere sottoposta a tecniche di anonimizzazione, tra cui l'aggiunta di rumore statistico o la sostituzione, per offrire un'ulteriore tutela della sfera privata. L'individuazione di un meccanismo interattivo interrogazione-risposta valido che possa rispondere alle domande in

[14] Dwork, C. (2006). *Differential privacy*. In *Automata, languages and programming* (pagg. 1-12). Springer Berlin Heidelberg.
[15] Cfr. Ed Felten (2012) *Protecting privacy by adding noise*. URL: https://techatftc.wordpress.com/2012/06/21/protecting-privacy-by-adding-noise/.

maniera sufficientemente accurata (vale a dire, con meno rumore statistico possibile) e tutelare al contempo la sfera privata costituisce un tema di ricerca ancora aperto.

Per contenere gli attacchi tramite deduzione e correlabilità occorre tenere traccia delle interrogazioni formulate da un soggetto e osservare le informazioni acquisite sulle persone interessate; allo stesso modo, le banche dati di "privacy differenziale" non dovrebbero essere utilizzate su motori di ricerca aperti che non offrono alcuna rintracciabilità dei soggetti che formulano le interrogazioni.

3.1.3.1 Garanzie

- *Individuazione:* se vengono prodotte solo statistiche e le norme applicate all'insieme sono scelte in maniera oculata, non dovrebbe essere possibile utilizzare le risposte per individuare una persona.
- *Correlabilità:* utilizzando richieste multiple potrebbe essere possibile correlare le informazioni relative a una persona specifica tra due risposte.
- *Deduzione:* è possibile dedurre informazioni su persone o gruppi ricorrendo a richieste multiple.

3.1.3.2. Errori comuni

- Non aggiungere una quantità sufficiente di rumore statistico: per impedire un collegamento con le conoscenze di base, la sfida consiste nel fornire elementi di prova minimi sull'eventuale contributo all'insieme di dati da parte di una specifica persona interessata o un gruppo di persone interessate. La difficoltà maggiore, dal punto di vista della protezione dei dati, consiste nel riuscire a generare una quantità adeguata di rumore da aggiungere alle risposte vere in modo da tutelare la sfera privata delle persone e mantenere contemporaneamente l'utilità delle risposte pubblicate.

3.1.3.3 Casi di insuccesso della privacy differenziale

Trattare ciascuna interrogazione in modo indipendente: una combinazione di risultati di interrogazioni può consentire la divulgazione di informazioni che avrebbero dovuto essere riservate. In assenza di una cronologia delle interrogazioni, un intruso potrebbe formulare domande multiple a una banca dati di "privacy differenziale" in modo da restringere progressivamente il campo del campione prodotto fino a far emergere, in maniera deterministica o con una probabilità molto alta, un carattere specifico di un'unica persona interessata o di un gruppo di persone interessate. Inoltre, come ulteriore raccomandazione si dovrebbe evitare di pensare erroneamente che i dati siano anonimi per altri, anche se il responsabile del trattamento può ancora identificare la persona interessata nella banca dati originaria prendendo in considerazione l'insieme dei mezzi che possono essere ragionevolmente utilizzati.

3.2. Generalizzazione

La generalizzazione rappresenta la seconda famiglia di tecniche di anonimizzazione e consiste nel generalizzare, o diluire, gli attributi delle persone interessate modificando la rispettiva scala o ordine di grandezza (vale a dire, una regione anziché una città, un mese anziché una settimana). Sebbene possa essere efficace per impedire l'individuazione, la generalizzazione non consente un'anonimizzazione che risulti efficace in tutti i casi; in particolare, presuppone approcci quantitativi specifici e sofisticati per impedire la correlabilità e la deduzione.

3.2.1. Aggregazione e *k*-anonimato

Le tecniche di aggregazione e *k*-anonimato sono volte a impedire l'individuazione di persone interessate mediante il loro raggruppamento con almeno k altre persone. A tale scopo, i valori degli attributi sono sottoposti a una generalizzazione tale da attribuire a ciascuna persona il medesimo valore. Ad esempio, riducendo il grado di dettaglio di una località da città a Stato, si include un numero più elevato di persone interessate. Le date di nascita individuali possono essere generalizzate in una serie di date o raggruppate per mese o anno. Altri attributi numerici (ad esempio, retribuzioni, peso, altezza o il dosaggio di un farmaco) possono essere generalizzati mediante il ricorso a intervalli di valori (ad esempio, retribuzione 20 000 EUR – 30 000 EUR). Tali metodi possono essere utilizzati nei casi in cui la correlazione di valori puntuali di attributi possa creare quasi-identificatori.

3.2.1.1. Garanzie

- *Individuazione:* poiché i medesimi attributi sono condivisi da k utenti, non dovrebbe più essere possibile individuare una persona all'interno di un gruppo di k utenti.
- *Correlabilità:* benché la correlabilità sia limitata, permane la possibilità di collegare i dati per gruppi di k utenti. All'interno di tale gruppo, la probabilità che due dati corrispondano agli stessi pseudoidentificatori è pari a 1/k (che potrebbe essere significativamente più elevata della probabilità che tali informazioni siano non correlabili).
- *Deduzione:* il difetto principale del modello di *k*-anonimato consiste nel fatto che non protegge da alcun tipo di attacco tramite deduzione. In effetti, se tutte le k persone rientrano in uno stesso gruppo e se è noto a quale gruppo appartiene una persona, è piuttosto semplice recuperare il valore di tale proprietà.

3.2.1.2. Errori comuni

- Trascurare alcuni quasi-identificatori: per quanto riguarda il *k*-anonimato, un parametro cruciale è la soglia di k. Più alto è il valore di k, maggiori sono le garanzie di tutela della sfera privata. Un errore comune consiste nell'aumentare artificialmente il valore k riducendo l'insieme di quasi-identificatori considerato. La riduzione dei quasi-identificatori agevola la creazione di raggruppamenti di *k*-utenti grazie al potere intrinseco di identificazione associato agli altri attributi (specialmente se alcuni di essi sono sensibili o possiedono un livello molto elevato di entropia, come accade nel caso di attributi molto rari). Non prendere in considerazione tutti i quasi-identificatori al momento della selezione dell'attributo da generalizzare è un errore cruciale; se alcuni attributi possono essere utilizzati per individuare una persona in un raggruppamento di k, la generalizzazione non tutela alcune persone (cfr. l'esempio riportato nella tabella 2).
- Valore di k troppo basso: optare per un valore di k basso è altrettanto problematico. Se k è troppo basso, il peso di una persona all'interno di un raggruppamento è troppo significativo e gli attacchi tramite deduzione vanno più spesso a segno. Ad esempio, se k=2, la probabilità che le due persone abbiano in comune la medesima proprietà è più elevata che per k>10.
- Non raggruppare le persone con lo stesso fattore ponderale: anche raggruppare un insieme di persone con una distribuzione disomogenea degli attributi può essere problematico. L'incidenza del dato di una persona su un insieme di dati finirà per variare: alcune rappresenteranno una quota significativa delle informazioni inserite,

mentre i contributi di altre rimarranno alquanto trascurabili. Pertanto, è importante accertarsi che k sia sufficientemente elevato da impedire che singole persone rappresentino una quota eccessivamente ampia degli elementi di un raggruppamento.

3.1.3.3. Casi di insuccesso del *k*-anonimato

Il problema principale del *k*-anonimato è che non protegge dagli attacchi tramite deduzione. Nell'esempio che segue, se l'intruso sa che una persona specifica è presente nell'insieme di dati ed è nata nel 1964, sa anche che la stessa ha avuto un attacco cardiaco. Inoltre, se è noto che l'insieme di dati in questione proviene da un'organizzazione francese, è possibile dedurre che tutte le persone del gruppo risiedono a Parigi, in quanto le prime tre cifre dei codici postali parigini sono 750*).

Anno	Sesso	Codice postale	Diagnosi
1957	M	750*	Attacco cardiaco
1957	M	750*	Colesterolo
1957	M	750*	Colesterolo
1964	M	750*	Attacco cardiaco
1964	M	750*	Attacco cardiaco

Tabella 2. Esempio di *k*-anonimizzazione progettata in maniera inadeguata

3.2.2. *L*-*L*-diversità/*T*-vicinanza

La *l*-*l*-diversità amplia il *k*-anonimato per impedire gli attacchi tramite deduzione deterministica facendo sì che in ciascuna classe di equivalenza ogni attributo abbia almeno *l* valori diversi.

Un obiettivo fondamentale da conseguire è limitare la presenza di classi di equivalenza con una scarsa variabilità degli attributi, in modo tale che un eventuale intruso che disponga di una conoscenza di base di una persona interessata specifica rimanga sempre con un grado di incertezza significativo.

La *l*-*l*-diversità è utile per proteggere i dati dagli attacchi tramite deduzione quando i valori degli attributi sono ben distribuiti. Va tuttavia precisato che la tecnica in questione non è in grado di impedire la fuga di informazioni se gli attributi all'interno di una partizione sono distribuiti in maniera disomogenea o rientrano in un intervallo ridotto di valori o significati semantici. In definitiva, la *l*-*l*-diversità è soggetta ad attacchi tramite deduzione probabilistica.

La *t*-vicinanza rappresenta un affinamento della *l*-*l*-diversità nel senso che mira a creare classi equivalenti che assomigliano alla distribuzione iniziale di attributi nella tabella. La tecnica in oggetto è utile quando è importante mantenere i dati quanto più possibile prossimi a quelli originali; a tale scopo, alla classe di equivalenza viene imposto un ulteriore vincolo, vale a dire che non solo devono esistere almeno *l* valori diversi all'interno di ogni classe di equivalenza, ma anche che ogni valore è rappresentato tante volte quante sono necessarie per rispecchiare la distribuzione iniziale di ciascun attributo.

3.2.2.1. Garanzie

- *Individuazione:* come il *k*-anonimato, la *l*-*l*-diversità e la *t*-vicinanza garantiscono che i dati relativi a una persona non possano essere individuati all'interno della banca dati.

- *Correlabilità:* la *l*-diversità e la *t*-vicinanza non rappresentano un miglioramento rispetto al *k*-*k*-anonimato per quanto riguarda la non correlabilità. Il problema è analogo a quello di ogni raggruppamento: la probabilità che le stesse informazioni appartengano alla medesima persona interessata è più elevata di 1/N (dove N rappresenta il numero di persone interessate nella banca dati).
- *Deduzione:* il principale vantaggio offerto dalla *l*-diversità e dalla *t*-vicinanza rispetto al *k*-anonimato consiste nel fatto che viene eliminata la possibilità di attaccare tramite deduzione una banca dati "*l*-diversa" o "*t*-*t*-vicina" con una sicurezza del 100%.

3.2.2.2. Errori comuni

- Proteggere valori di attributi sensibili mescolandoli con altri attributi sensibili: non è sufficiente avere due valori di un attributo in un raggruppamento per garantire la tutela della sfera privata. Di fatto, la distribuzione dei valori sensibili in ogni raggruppamento dovrebbe assomigliare alla distribuzione di tali valori nella popolazione totale, o perlomeno dovrebbe essere uniforme in tutto il raggruppamento.

3.2.2.3. Casi di insuccesso della *l*-diversità

Nella tabella sottostante la *l*-diversità riguarda l'attributo "diagnosi"; tuttavia, se si sa che una persona nata nel 1964 è presente nella tabella, è ancora possibile ipotizzare con una probabilità molto elevata che la stessa abbia avuto un attacco cardiaco.

Anno	Sesso	Codice postale	Diagnosi
1957	M	750*	**Attacco cardiaco**
1957	M	750*	**Colesterolo**
1957	M	750*	**Colesterolo**
1957	M	750*	**Colesterolo**
1964	M	750*	**Attacco cardiaco**
1964	M	750*	**Attacco cardiaco**
1964	M	750*	**Attacco cardiaco**
1964	M	750*	**Colesterolo**
1964	M	750*	**Attacco cardiaco**
1964	M	750*	**Attacco cardiaco**
1964	M	750*	**Attacco cardiaco**
1964	M	750*	**Attacco cardiaco**
1964	M	750*	**Attacco cardiaco**
1964	M	750*	**Attacco cardiaco**
1964	M	750*	**Attacco cardiaco**

Tabella 3. Una tabella *l*-diversa in cui i valori "diagnosi" non sono uniformemente distribuiti

Nome	Data di nascita	Sesso
Smith	1964	M
Rossi	1964	M
Dupont	1964	M
Jansen	1964	M
Garcia	1964	M

Tabella 4. Se un intruso sapesse che tali persone sono presenti nella tabella 3, potrebbe dedurre che le stesse hanno avuto un attacco cardiaco

4. Pseudonimizzazione

La pseudonimizzazione consiste nel sostituire un attributo (solitamente un attributo univoco) di un dato con un altro. La persona fisica potrebbe pertanto essere ancora identificata in maniera indiretta; di conseguenza, la pseudonimizzazione, se utilizzata da sola, non consente di ottenere un insieme di dati anonimo. Viene comunque trattata nel presente parere tenuto conto delle numerose convinzioni fallaci e degli errori che ne accompagnano l'utilizzo.

La pseudonimizzazione riduce la correlabilità di un insieme di dati all'identità originale di una persona interessata; in quanto tale, rappresenta una misura di sicurezza utile, ma non un metodo di anonimizzazione.

Il risultato della pseudonimizzazione può essere indipendente dal valore iniziale (come accade nel caso di un numero casuale generato dal responsabile del trattamento o di un cognome scelto dalla persona interessata) o può essere estrapolato dai valori originali di un attributo o insieme di attributi, ad esempio una funzione di hash o un sistema di crittografia.

Le tecniche di pseudonimizzazione più usate sono le seguenti:

- crittografia con chiave segreta: in questo caso, chi conosce la chiave può facilmente risalire all'identificazione di ogni persona interessata decrittando l'insieme di dati, in quanto i dati personali sono ancora contenuti all'interno dell'insieme di dati, pur se in forma crittografata. Ipotizzando di applicare un sistema di crittografia avanzato, la decrittazione può avvenire solamente se si è a conoscenza della chiave;
- funzione di hash: corrisponde a una funzione che, a partire da un'immissione di dati di qualsiasi dimensione (l'immissione potrebbe essere costituita da un unico attributo o da un insieme di attributi), restituisce comunque un'emissione di dimensione fissa; tale funzione non può essere invertita, vale a dire che non esiste più il rischio di inversione associato alla crittografia. Tuttavia, se l'intervallo di valori di immissione relativi alla funzione di hash è noto, la funzione stessa consente di riprodurli al fine di desumere il valore corretto associato a un dato specifico. Ad esempio, se un insieme di dati è stato pseudonimizzato effettuando l'hashing del numero nazionale di identificazione, lo stesso può essere estrapolato semplicemente effettuando l'hashing di tutti i possibili valori di immissione e raffrontando il risultato con i valori contenuti nell'insieme di dati. Le funzioni di hash sono solitamente progettate per essere

relativamente rapide da calcolare e sono soggette ad attacchi brutali[16]. Si possono inoltre creare tabelle precalcolate per consentire l'inversione in blocco di un insieme consistente di valori hash.

Il ricorso a una funzione di hash con salt (che prevede l'aggiunta di un valore casuale, noto come "salt", all'attributo oggetto di hashing) può ridurre la probabilità di estrapolare il valore di immissione, tuttavia permane la possibilità di calcolare con mezzi ragionevolmente utilizzabili il valore dell'attributo originario che si cela dietro al risultato di una funzione di hash con salt[17];

- funzione di hash cifrato con chiave memorizzata: corrisponde a una funzione di hash particolare che utilizza una chiave segreta quale immissione aggiuntiva (la differenza rispetto alla funzione di hash con salt è che il salt abitualmente non è segreto). Un responsabile del trattamento può riprodurre la funzione sull'attributo utilizzando la chiave segreta, ma un intruso avrebbe molte più difficoltà a riprodurre la funzione senza conoscere la chiave, in quanto il numero di possibilità da vagliare è sufficientemente elevato da risultare impraticabile;
- crittografia deterministica o funzione di hash cifrato con cancellazione della chiave: questa tecnica può essere equiparata alla selezione di un numero casuale quale pseudonimo di ciascun attributo contenuto nell'insieme di dati seguita dalla cancellazione della tabella delle corrispondenze. Tale soluzione consente[18] di ridurre il rischio di correlabilità tra i dati personali contenuti nell'insieme di dati e quelli relativi alla medesima persona presenti in un altro insieme di dati in cui viene utilizzato uno pseudonimo diverso. Se si ricorre a un algoritmo particolarmente avanzato, un intruso ha notevoli difficoltà computazionali a cercare di decriptare o riprodurre la funzione, in quanto dovrebbe provare tutte le chiavi possibili, visto che la chiave non è disponibile;
- tokenizzazione: questa tecnica si applica solitamente (anche se non unicamente) nel settore finanziario per sostituire i numeri delle carte d'identità con valori che presentano un'utilità ridotta per un eventuale intruso. Si tratta di una tecnica derivata dalle precedenti in quanto si basa tipicamente sull'applicazione di un meccanismo di crittografia univoca o sull'assegnazione, tramite una funzione indicizzata, di un numero sequenziale o di un numero generato casualmente che non deriva matematicamente dai dati originali.

4.1. Garanzie

- *Individuazione:* permane la possibilità di individuare i dati delle persone, in quanto queste ultime sono ancora identificate da un attributo unico che è il risultato della funzione di pseudonimizzazione (= l'attributo pseudonimizzato).
- *Correlabilità:* la correlabilità rimane un'operazione di semplice effettuazione tra dati che utilizzano lo stesso attributo pseudonimizzato per fare riferimento alla stessa persona. Anche se per la stessa persona interessata vengono utilizzati diversi attributi pseudonimizzati, la correlabilità potrebbe essere comunque effettuata mediante altri attributi. Solamente nel caso in cui nessun altro attributo contenuto nell'insieme di dati

[16] Attacchi del genere consistono nel provare tutte le immissioni plausibili al fine di costruire tabelle di corrispondenza.
[17] Soprattutto se è noto il tipo di attributo (nome, codice fiscale, data di nascita, ecc.). Per aumentare il numero di calcoli, si potrebbe ricorrere a una funzione di hash derivata da una chiave, in cui il valore calcolato viene sottoposto a diversi hashing con un salt breve.
[18] A seconda degli altri attributi contenuti nell'insieme di dati e della cancellazione dei dati originali.

possa essere utilizzato per identificare la persona interessata e se è stato eliminato ogni legame tra l'attributo originario e quello pseudonimizzato (compresa la cancellazione dei dati originali) non sussiste alcun riferimento incrociato ovvio tra due insiemi di dati che utilizzano attributi pseudonimizzati diversi.

- *Deduzione:* gli attacchi all'identità reale di una persona interessata tramite deduzione sono possibili all'interno dell'insieme di dati o tra diversi insiemi di dati che utilizzano lo stesso attributo pseudonimizzato per una persona, oppure se gli pseudonimi sono molto evidenti e non mascherano adeguatamente l'identità originale della persona interessata.

4.2. Errori comuni

- Ritenere che un insieme di dati pseudonimizzati sia anonimizzato: spesso i responsabili del trattamento presumono che eliminare o sostituire uno o più attributi sia sufficiente per rendere anonimo un insieme di dati. Molti esempi hanno dimostrato l'erroneità di tale convinzione; la semplice modifica dell'identità non impedisce l'identificazione di una persona interessata se l'insieme di dati continua a contenere quasi-identificatori o se i valori di altri attributi consentono comunque di identificare una persona. In molti casi identificare una persona all'interno di un insieme di dati pseudonimizzato può essere facile come con i dati originali. Occorre adottare misure supplementari per poter considerare l'insieme di dati effettivamente anonimizzato, tra cui l'eliminazione e la generalizzazione degli attributi o la cancellazione dei dati originali o almeno la loro estrema aggregazione.
- Errori comuni quando si utilizza la pseudonimizzazione quale tecnica per ridurre la correlabilità:

    - utilizzare la stessa chiave in banche dati diverse: l'eliminazione della correlabilità di diversi insiemi di dati dipende in larga misura dall'utilizzo di un algoritmo cifrato e dal fatto che a un'unica persona corrispondono diversi attributi pseudonimizzati in diversi contesti. Pertanto, è importante evitare di utilizzare la stessa chiave in banche dati diverse per poter ridurre la correlabilità;
    - utilizzare chiavi diverse ("chiavi a rotazione") per diversi utenti: si potrebbe essere tentati di utilizzare chiavi diverse per gruppi di utenti diversi e di cambiare la chiave in funzione dell'utilizzo (ad esempio, utilizzare la stessa chiave per registrare 10 immissioni relative allo stesso utente). Tuttavia, tale operazione, se non progettata in maniera adeguata, potrebbe determinare la comparsa di modelli che potrebbero parzialmente ridurre i vantaggi desiderati. Ad esempio, introdurre una rotazione della chiave secondo regole specifiche per persone specifiche agevolerebbe la correlabilità delle informazioni corrispondenti a una determinata persona. Inoltre, la scomparsa di un dato pseudonimizzato ricorrente nella banca dati nel momento in cui ne compare uno nuovo potrebbe lasciar intuire che entrambi i dati si riferiscono alla medesima persona fisica;
    - conservare la chiave: se la chiave segreta viene conservata insieme ai dati pseudonimizzati e gli stessi sono danneggiati, l'intruso potrebbe riuscire a collegare con facilità i dati pseudonimizzati ai loro attributi originali. La stessa eventualità si verifica se la chiave viene conservata separatamente dai dati ma non in maniera sicura.

4.3. Carenze della pseudonimizzazione

- *Sanità*

1. Nome, indirizzo, data di nascita	2. Durata della prestazione assistenziale speciale	3. Indice di massa corporea	6. Numero di riferimento della coorte di ricerca
[redacted]	< 2 anni	15	QA5FRD4
	> 5 anni	14	2B48HFG
	< 2 anni	16	RC3URPQ
	> 5 anni	18	SD289K9
	< 2 anni	20	5E1FL7Q

Tabella 5. Esempio di pseudonimizzazione tramite hashing (nome, indirizzo, data di nascita) che può essere facilmente invertita

È stato creato un insieme di dati per esaminare il rapporto tra il peso di una persona e il diritto al pagamento di una prestazione assistenziale speciale. L'insieme di dati originario conteneva il nome, l'indirizzo e la data di nascita delle persone interessate, ma tali informazioni sono state cancellate. Il numero di riferimento della coorte di ricerca è stato generato dai dati cancellati mediante una funzione di hash. Benché il nome, l'indirizzo e la data di nascita siano stati cancellati dalla tabella, se si conoscono il nome, l'indirizzo e la data di nascita di una persona interessata oltre alla funzione di hash utilizzata, è facile calcolare i numeri di riferimento della coorte di ricerca.

- *Social Network*

  È stato dimostrato[19] che è possibile estrapolare informazioni sensibili su persone specifiche dai grafici di social-network malgrado le tecniche di "pseudonimizzazione" applicate a tali dati. Un provider di un social network ha erroneamente ritenuto che la pseudonimizzazione fosse efficace per impedire l'identificazione dopo aver venduto i dati ad altre aziende a fini di marketing e di pubblicità. Il provider aveva sostituito i nomi reali con soprannomi, ma evidentemente ciò non è stato sufficiente a rendere anonimi i profili degli utenti, in quanto i rapporti tra le diverse persone sono unici e possono essere utilizzati come identificatori.

- *Localizzazioni*

  I ricercatori del MIT[20] hanno recentemente analizzato un insieme di dati pseudonimizzato contenente 15 mesi di coordinate di mobilità spaziotemporale di 1,5 milioni di persone in un territorio compreso in un raggio di 100 km. Hanno dimostrato che il 95% delle persone poteva essere identificato mediante quattro luoghi, e che bastavano due luoghi per identificare più del 50% delle persone interessate (uno di tali luoghi è noto, essendo molto probabilmente "casa" o "ufficio") con un margine molto ridotto di protezione della sfera privata, benché le identità delle persone fossero state pseudonimizzate sostituendo i loro attributi reali […] con altre etichette.

[19] A. Narayanan e V. Shmatikov, "*De-anonymizing social networks*", nel trentesimo simposio dell'IEEE sulla sicurezza e la sfera privata, 2009.
[20] Y.-A. de Montjoye, C. Hidalgo, M. Verleysen e V. Blondel, "*Unique in the Crowd: The privacy bounds of human mobility*," Nature, num. 1376, 2013.

5. Conclusioni e raccomandazioni

5.1. Conclusioni

Le tecniche di deidentificazione e anonimizzazione sono oggetto di intense ricerche e il presente parere ha coerentemente dimostrato che ciascuna tecnica presenta vantaggi e svantaggi. Nella maggior parte dei casi non è possibile fornire raccomandazioni minime circa i parametri da utilizzare, in quanto ogni insieme di dati va studiato caso per caso.

In molti casi, un insieme di dati resi anonimi può continuare a presentare un rischio residuo per le persone interessate. In effetti, anche quando non è più possibile recuperare con precisione il dato di una persona, potrebbe permanere la possibilità di raccogliere informazioni su tale persona con l'ausilio di altre fonti di informazione accessibili (pubblicamente o meno). Va precisato che oltre all'impatto diretto sulle persone interessate prodotto dalle conseguenze di un processo di anonimizzazione carente (fastidio, spreco di tempo e sensazione di perdita di controllo per il fatto di essere stati inclusi in un raggruppamento inconsapevolmente o senza previo consenso), un'anonimizzazione carente potrebbe determinare altri effetti collaterali indiretti nel caso in cui una persona interessata sia erroneamente inclusa nell'obiettivo di un intruso come conseguenza del trattamento di dati resi anonimi, soprattutto se le finalità dell'intruso sono dolose. Il Gruppo di lavoro sottolinea pertanto che le tecniche di anonimizzazione possono fornire garanzie di tutela della sfera privata, ma solamente se la loro applicazione viene progettata in maniera adeguata, nel senso che i requisiti preliminari (contesto) e l'obiettivo o gli obiettivi del processo di anonimizzazione devono essere specificati in maniera chiara per conseguire il livello di anonimizzazione desiderato.

5.2. Raccomandazioni

- Alcune tecniche di anonimizzazione presentano limiti intrinseci. I responsabili del trattamento devono esaminare con attenzione tali limiti prima di utilizzare una determinata tecnica per effettuare un processo di anonimizzazione. Devono prendere in considerazione le finalità perseguite tramite l'anonimizzazione, come la tutela della sfera privata delle persone all'atto della pubblicazione di un insieme di dati o l'autorizzazione del recupero di un'informazione da un insieme di dati.
- Nessuna tecnica descritta nel presente documento soddisfa con certezza i criteri di un'effettiva anonimizzazione (vale a dire, impossibilità di individuare una persona, nessuna correlabilità tra i dati relativi a una persona e nessuna deduzione in merito alle persone). Tuttavia, poiché alcuni di questi rischi possono essere evitati in tutto o in parte applicando una determinata tecnica, occorre prestare particolare attenzione quando si decide di applicare una determinata tecnica alla situazione specifica o di ricorrere a un insieme di tali tecniche al fine di accrescere l'affidabilità dell'esito.

La tabella sottostante fornisce una panoramica dei punti di forza e debolezza delle tecniche considerate sulla base di tre requisiti di base.

	Sussiste ancora il rischio di individuazione?	Sussiste ancora il rischio di correlabilità?	Sussiste ancora il rischio di deduzione?
Pseudonimizzazione	Sì	Sì	Sì
Aggiunta di rumore statistico	Sì	Forse no	Forse no
Sostituzione	Sì	Sì	Forse no
Aggregazione o *k*-anonimato	No	Sì	Sì
L-diversità	No	Sì	Forse no
Privacy differenziale	Forse no	Forse no	Forse no
Hashing/tokenizzazione	Sì	Sì	Forse no

Tabella 6. Punti di forza e debolezza delle tecniche considerate

- La soluzione ottimale va decisa caso per caso. Una soluzione (vale a dire, un processo di anonimizzazione completo) che soddisfi tutti e tre i criteri rappresenterebbe una protezione valida dall'identificazione effettuata mediante l'insieme dei mezzi che possono essere ragionevolmente utilizzati dal responsabile del trattamento o da altri.
- Ogniqualvolta una proposta non soddisfa uno dei criteri, occorre effettuare una valutazione approfondita dei rischi di identificazione. Tale valutazione dovrebbe essere trasmessa all'autorità competente, se la legge nazionale impone che tale autorità valuti o autorizzi il processo di anonimizzazione.

Per attenuare i rischi di identificazione, occorre prendere in considerazione le buone pratiche di seguito descritte.

Buone pratiche di anonimizzazione

In generale

- Non affidarsi all'approccio "pubblica e dimentica". Dato il rischio residuo di identificazione, i responsabili del trattamento sono tenuti a:
    - 1. individuare periodicamente i nuovi rischi e rivalutare il rischio o i rischi residui,
    - 2. valutare se i controlli presenti per i rischi individuati siano sufficienti e adeguarli di conseguenza; E
    - 3. monitorare e controllare i rischi.
- Nell'ambito di tali rischi residui, tenere conto del potenziale di identificazione della parte non anonimizzata di un insieme di dati (se presente), soprattutto se la stessa è unita alla parte anonimizzata, oltre che di possibili correlazioni tra attributi (ad esempio tra dati su località geografiche e livello di ricchezza).

Elementi contestuali

- Le finalità da conseguire con l'insieme di dati anonimizzato dovrebbero essere specificate con chiarezza, in quanto svolgono un ruolo fondamentale nel determinare il rischio di identificazione.
- A ciò si accompagna l'esame di tutti gli elementi contestuali rilevanti, come ad esempio la natura dei dati originali, i meccanismi di controllo esistenti (comprese misure di sicurezza per limitare l'accesso agli insiemi di dati), la dimensione del campione (caratteristiche quantitative), la disponibilità di fonti pubbliche di informazione (a cui si affidino i

destinatari), la trasmissione prevista di dati a terzi (limitata, illimitata, ad esempio su Internet, ecc.).

- Occorre prestare attenzione a possibili intrusi esaminando la vulnerabilità dei dati a determinati attacchi mirati (anche in questo caso, la sensibilità delle informazioni e la natura dei dati rappresentano fattori chiave).

Elementi tecnici

- I responsabili del trattamento sono tenuti a comunicare la tecnica di anonimizzazione/la combinazione di tecniche che intendono utilizzare, soprattutto se prevedono di pubblicare l'insieme di dati anonimizzati.
- Occorre eliminare dall'insieme di dati gli attributi ovvi (ad esempio rari)/i quasi-identificatori.
- Se si ricorre alle tecniche di aggiunta del rumore statistico (nella randomizzazione), il livello di rumore aggiunto ai dati va determinato in funzione del valore di un attributo (vale a dire, non dovrebbe essere aggiunto alcun rumore fuori scala), dell'impatto sulle persone interessate degli attributi da proteggere e/o del diradamento dell'insieme di dati.
- Se si ricorre alla privacy differenziale (nella randomizzazione), occorre considerare la necessità di tenere traccia delle interrogazioni in modo da individuare eventuali interrogazioni che violano la sfera privata, in quanto l'intrusività delle interrogazioni è cumulativa.
- Se si attuano tecniche di generalizzazione, è fondamentale che il responsabile del trattamento non si limiti a un solo criterio di generalizzazione anche per il medesimo attributo, vale a dire che occorre selezionare diversi gradi di dettaglio delle località o diversi intervalli temporali. La selezione del criterio da applicare dev'essere determinata dalla distribuzione dei valori degli attributi nella popolazione interessata. Non tutte le distribuzioni si prestano a essere generalizzate, ossia non è possibile adottare un approccio indifferenziato alla generalizzazione. Occorre garantire la variabilità all'interno delle classi di equivalenza; ad esempio, dovrebbe essere scelta una soglia specifica a seconda degli "elementi contestuali" summenzionati (dimensione del campione, ecc.) e se tale soglia non viene raggiunta, il campione specifico dovrebbe essere scartato (oppure dovrebbe essere stabilito un criterio di generalizzazione diverso).

ALLEGATO
Introduzione alle tecniche di anonimizzazione

A.1. Introduzione

All'anonimato sono associate interpretazioni diverse nell'Unione europea – in alcuni paesi corrisponde all'anonimato computazionale (vale a dire, dovrebbe essere computazionalmente difficile, persino per il responsabile del trattamento che opera in collaborazione con altri, identificare direttamente o indirettamente una delle persone interessate) e in altri paesi all'anonimato perfetto (vale a dire, dovrebbe essere impossibile, persino per il responsabile del trattamento che opera in collaborazione con altri, identificare direttamente o indirettamente una delle persone interessate). Ciononostante, in entrambi i casi la "anonimizzazione" corrisponde al processo mediante il quale i dati sono resi anonimi. La differenza consiste in ciò che viene considerato un livello accettabile dal punto di vista del rischio di reidentificazione.

Si possono prevedere vari casi di utilizzo per i dati resi anonimi, dalle indagini sociali, alle analisi statistiche, allo sviluppo di nuovi servizi/prodotti. Talvolta, persino tali attività con finalità generiche possono avere un impatto su specifiche persone interessate e vanificare la presunta natura anonima dei dati trattati. Si possono fornire numerosi esempi, dall'avvio di iniziative di marketing mirate, all'attuazione di misure pubbliche sulla base della definizione del profilo degli utenti, dei loro comportamenti o dei loro schemi di mobilità[21].

Purtroppo, al di là delle affermazioni generiche non esiste una metrica sperimentata che consenta di valutare in anticipo il tempo o gli sforzi necessari per effettuare una reidentificazione dopo il trattamento o, in alternativa, di selezionare la procedura più adeguata da attuare se si desidera abbassare la probabilità che un insieme di dati pubblicati riconduca a un insieme identificato di persone interessate.

"L'arte dell'anonimizzazione", come sono talvolta definite tali pratiche nella letteratura scientifica[22], è una nuova disciplina scientifica ancora agli albori ed esistono molte pratiche che consentono di ridurre il potere di identificazione degli insiemi di dati; occorre tuttavia ribadire con chiarezza che la maggioranza di tali pratiche non impedisce la correlazione tra i dati trattati e le persone interessate. In talune circostanze, l'identificazione di insiemi di dati ritenuti anonimi si è rivelata del tutto possibile, in altre situazioni si è constatata la comparsa di falsi positivi.

In generale, esistono due diversi approcci: uno è basato sulla generalizzazione degli attributi, l'altro sulla randomizzazione. L'esame dei dettagli e delle sottigliezze di tali tecniche consente di approfondire la conoscenza del potere di identificazione dei dati e di fare nuova luce sul concetto stesso di dati personali.

A.2. "Anonimizzazione" per randomizzazione

Un'opzione di anonimizzazione consiste nel modificare i valori effettivi per impedire correlazioni tra i dati resi anonimi e i valori originali. Tale obiettivo può essere conseguito mediante un ampio spettro di metodologie che vanno dall'aggiunta di rumore statistico alla sostituzione dei dati (permutazione). Va sottolineato che l'eliminazione di un attributo

[21] Ad esempio il caso di TomTom nei Paesi Bassi (cfr. l'esempio riportato al paragrafo 2.2.3).
[22] Jun Gu, Yuexian Chen, Junning Fu, Huanchun Peng, Xiaojun Ye, *Synthesizing: Art of Anonymization, Database and Expert Systems Applications Lecture Notes in Computer Science* – Springer – Volume 6261, 2010, pagg. 385-399.

rappresenta una forma estrema di randomizzazione dell'attributo stesso (essendo l'attributo totalmente mascherato dal rumore statistico).

In talune circostanze, la finalità dell'intero trattamento non consiste tanto nella pubblicazione di un insieme di dati randomizzati, bensì nel consentire l'accesso a tali dati mediante interrogazioni. In questo caso il rischio per la persona interessata consiste nella probabilità che un intruso riesca a estrapolare informazioni da una serie di interrogazioni diverse all'insaputa del responsabile del trattamento. Per garantire l'anonimato delle persone presenti nell'insieme di dati non dovrebbe essere possibile desumere che una determinata persona interessata ha contribuito all'insieme di dati, spezzando così il legame con qualsiasi tipo di informazione di base di cui un intruso potrebbe disporre.

L'aggiunta di rumore statistico in linea con la risposta all'interrogazione può ridurre ulteriormente il rischio di reidentificazione. Tale approccio, noto in letteratura anche col nome di privacy differenziale[23], si distingue da quelli descritti in precedenza in quanto conferisce ai responsabili della pubblicazione dei dati un maggiore controllo sull'accesso ai dati rispetto alla divulgazione al pubblico. L'aggiunta di rumore statistico ha due obiettivi principali: da un lato, proteggere la sfera privata delle persone interessate che compaiono nell'insieme di dati e, dall'altro lato, preservare l'utilità delle informazioni pubblicate. In particolare, l'intensità del rumore deve essere proporzionata al livello dell'interrogazione (un numero eccessivo di interrogazioni su persone che richiedono una risposta troppo accurata finiscono per aumentare la probabilità di identificazione). Oggigiorno l'applicazione efficace della randomizzazione deve essere valutata caso per caso, in quanto nessuna tecnica offre una metodologia a prova di errore, poiché esistono esempi di fuga di informazioni sugli attributi di una persona interessata (che fosse o meno inclusa nell'insieme di dati) persino nei casi in cui il responsabile del trattamento riteneva che l'insieme di dati fosse anonimizzato.

Potrebbe essere utile esaminare esempi specifici per chiarire le potenziali carenze della randomizzazione quale mezzo per garantire l'anonimizzazione. Ad esempio, nel contesto dell'accesso interattivo, interrogazioni considerate innocue per la sfera privata potrebbero rappresentare un rischio per le persone interessate. Di fatto, se l'intruso sa che un sottogruppo S di persone compare nell'insieme di dati che contiene informazioni sull'incidenza dell'attributo A nella popolazione P, effettuando una semplice interrogazione composta dalle seguenti due domande "Quante persone nella popolazione P possiedono l'attributo A?" e "Quante persone nella popolazione P, eccetto quelle appartenenti al sottogruppo S, possiedono l'attributo A?" potrebbe riuscire a determinare (mediante sottrazione) il numero di persone nel sottogruppo S che possiedono effettivamente l'attributo A, in modo deterministico o mediante deduzione di probabilità. In ogni caso, la sfera privata delle persone del sottogruppo S potrebbe essere gravemente compromessa, soprattutto in base alla natura dell'attributo A.

Si può inoltre considerare che se una persona interessata non compare nell'insieme di dati, ma il suo rapporto con i dati dell'insieme in questione è noto, la pubblicazione dell'insieme di dati può tradursi in un rischio per la sua sfera privata. Ad esempio, se è noto che "il valore dell'attributo A della persona in questione si scosta dal valore medio della popolazione di una quantità X", chiedendo semplicemente al responsabile della banca dati di effettuare l'operazione non lesiva della sfera privata di estrapolare il valore medio dell'attributo A l'intruso può dedurre con esattezza un dato personale relativo a una persona interessata specifica.

[23] Cynthia Dwork, *Differential Privacy*, *International Colloquium on Automata, Languages and Programming* (ICALP) 2006, pagg. 1–12.

L'inserimento di alcune inesattezze relative nei valori effettivi di una banca dati è un'operazione che deve essere progettata in maniera adeguata. Occorre aggiungere una quantità di rumore statistico sufficiente a proteggere la sfera privata, ma anche abbastanza limitata da preservare l'utilità dei dati. Ad esempio, se il numero di persone interessate con un particolare attributo è molto esiguo o la sensibilità dell'attributo è elevata, potrebbe essere preferibile ricorrere a un intervallo o a una frase generica quale "un numero ridotto di casi, possibilmente pari addirittura a zero" invece di specificare il numero effettivo. In tal modo, anche se il meccanismo di aggiunta del rumore è noto in anticipo, viene preservata la sfera privata della persona interessata in quanto permane un grado di incertezza. Dal punto di vista dell'utilità, se l'inesattezza è progettata in maniera adeguata, i risultati sono comunque utili per finalità statistiche o decisionali.

La randomizzazione delle banche dati e la privacy differenziale richiedono ulteriori riflessioni. In primo luogo, il grado esatto di distorsione può variare significativamente a seconda del contesto (il tipo di interrogazione, la dimensione della popolazione nella banca dati, la natura dell'attributo e il suo potere intrinseco di identificazione) e non è possibile prevedere una soluzione indifferenziata. Inoltre, il contesto può cambiare nel tempo e il meccanismo interattivo dovrebbe essere modificato di conseguenza. Calibrare il rumore implica tenere traccia dei rischi cumulativi per la sfera privata che ogni meccanismo interattivo comporta per le persone interessate. Il meccanismo di accesso ai dati dovrebbe pertanto essere dotato di segnalazioni che si attivano quando viene raggiunto un bilancio dei "costi per la sfera privata" e le persone interessate potrebbero essere esposte a rischi specifici se viene effettuata una nuova interrogazione, al fine di aiutare il responsabile del trattamento a determinare il grado adeguato di distorsione da immettere volta per volta nei dati personali.

D'altro canto, occorre anche considerare il caso in cui i valori degli attributi vengono cancellati (o modificati). Una soluzione comunemente utilizzata per gestire alcuni valori atipici per gli attributi è cancellare l'insieme di dati correlati alle persone atipiche o i valori atipici. In quest'ultimo caso, è importante accertarsi che l'assenza di valori non diventi di per sé un elemento di identificazione di una persona interessata.

Esaminiamo ora la randomizzazione mediante sostituzione di attributi. Un errore ricorrente nel caso dell'anonimizzazione è la sua equiparazione alla crittografia o alla codifica mediante chiavi. Tale errore si basa su due convinzioni, vale a dire, a) che una volta che alcuni attributi di un dato in una banca dati (ad esempio nome, indirizzo, data di nascita) vengono crittografati o sostituiti da una stringa apparentemente randomizzata in seguito a un'operazione di codifica tramite chiave quale la funzione di hash con chiave, tale dato sia "anonimizzato", e b) che l'anonimizzazione sia più efficace se la lunghezza della chiave è adeguata e l'algoritmo di crittografia è avanzato. Tale convinzione erronea è diffusa tra i responsabili del trattamento e merita una spiegazione, anche relativamente alla pseudonimizzazione e ai rischi presumibilmente più bassi ad essa associati.

In primo luogo, le tecniche in questione perseguono finalità radicalmente diverse: la crittografia come pratica di sicurezza si propone di garantire la riservatezza di un canale di comunicazione tra parti identificate (esseri umani, dispositivi o parti di software/hardware) per evitare intercettazioni o divulgazione non intenzionale. La codifica tramite chiave corrisponde a una traduzione semantica dei dati che dipende da una chiave segreta. D'altro canto, l'obiettivo dell'anonimizzazione consiste nell'impedire l'identificazione delle persone evitando collegamenti nascosti tra attributi e persone.

Di per sé, né la crittografia né la codifica tramite chiave permettono di conseguire l'obiettivo di rendere non identificabile una persona interessata, in quanto i dati originari sono ancora

disponibili o deducibili, per lo meno da parte del responsabile del trattamento. Attuare semplicemente una traduzione semantica dei dati personali, come accade nel caso della codifica con chiave, non esclude la possibilità di ripristinare i dati nella loro struttura originaria, applicando l'algoritmo al contrario o mediante attacchi brutali, a seconda della natura degli schemi, o in seguito a una violazione dei dati. La crittografia più avanzata può garantire un livello più elevato di protezione dei dati, vale a dire che gli stessi risultano incomprensibili in mancanza della chiave di decodifica, ma non sono necessariamente anonimizzati. Fintantoché la chiave o i dati originali rimangono accessibili (anche in caso di terzi fidati, contrattualmente vincolati a fornire un servizio di deposito sicuro della chiave), permane la possibilità di identificare una persona interessata.

Affidarsi semplicemente alla solidità del meccanismo di crittografia quale misura del grado di "anonimizzazione" di un insieme di dati è fuorviante, in quanto molti altri fattori tecnici e organizzativi incidono sulla sicurezza generale di un meccanismo di crittografia o di una funzione di hash. In letteratura sono stati segnalati molti attacchi andati a buon fine che eludevano completamente l'algoritmo, in quanto facevano leva su una custodia carente delle chiavi (ad esempio, l'esistenza di una modalità predefinita meno sicura) o su altri fattori umani (ad esempio, password deboli per il recupero della chiave). Infine, un sistema di crittografia scelto con una dimensione della chiave prestabilita dovrebbe garantire la riservatezza per un determinato periodo di tempo (la maggior parte delle chiavi attuali dovranno essere ridimensionate intorno al 2020), sebbene un processo di anonimizzazione non dovrebbe avere limiti temporali.

È ora opportuno soffermarsi sui limiti della randomizzazione (o sostituzione e rimozione) degli attributi esaminando vari esempi di insuccesso del processo di anonimizzazione per randomizzazione verificatisi negli ultimi anni, nonché le ragioni del loro fallimento.

Un caso noto che riguarda la pubblicazione di dati anonimizzati in maniera carente è quello del premio Netflix[24]. Se si prende un dato generico all'interno di una banca dati nella quale sono stati randomizzati diversi attributi relativi a una persona interessata, si nota che ogni dato può essere ulteriormente suddiviso in due sottodati come segue: {attributi randomizzati, attributi "in chiaro"}, in cui gli attributi in chiaro possono essere una qualsiasi combinazione di dati ipoteticamente non personali. Un'osservazione specifica che può essere fatta sull'insieme di dati del premio Netflix scaturisce dalla constatazione che ogni dato può essere rappresentato da un punto in uno spazio pluridimensionale in cui ogni attributo in chiaro rappresenta una coordinata. Utilizzando questa tecnica, ogni insieme di dati può essere considerato una costellazione di punti in uno spazio pluridimensionale che può presentare un grado elevato di diradamento, vale a dire che i punti sono distanti gli uni dagli altri. Di fatto, la distanza tra i punti può essere tale che, dopo aver diviso lo spazio in regioni ampie, ogni regione contiene solamente un dato. Nemmeno l'aggiunta di rumore statistico riesce ad avvicinare i dati in modo tale da farli rientrare nella stessa regione pluridimensionale. Ad esempio, nell'esperimento Netflix, i dati erano sufficientemente unici, con soltanto 8 giudizi sui film raccolti in un arco di 14 giorni. Dopo l'aggiunta di rumore sia ai giudizi sia alle date, non vi era alcuna sovrapposizione di regioni. In altre parole, la selezione stessa di soli 8 film oggetto di giudizio costituiva una sorta di impronta digitale dei giudizi espressi, non condivisa tra due persone interessate all'interno della banca dati. Sulla base di tale osservazione geometrica, i ricercatori hanno confrontato l'insieme di dati ritenuti anonimi di Netflix con un'altra banca dati pubblica contenente giudizi sui film (l'IMBD), individuando così gli utenti che avevano espresso un giudizio sugli stessi film nei medesimi intervalli di tempo. Poiché la

[24] Arvind Narayanan, Vitaly Shmatikov: *Robust De-anonymization of Large Sparse Datasets*. Simposio dell'IEEE sulla sicurezza e la sfera privata 2008: 111-125.

maggioranza degli utenti presentava una corrispondenza di uno a uno, è stato possibile importare le informazioni ausiliarie recuperate nella banca dati IMDB nell'insieme di dati pubblicati da Netflix, conferendo pertanto un'identità a tutti i dati reputati anonimi.

È importante sottolineare che si tratta di una proprietà generale: la parte residua di una qualsiasi banca dati "randomizzata" continua a possedere un potere di identificazione molto alto a seconda della rarità della combinazione degli attributi residui. Si tratta di una raccomandazione che i responsabili del trattamento dovrebbero sempre tenere presente quando selezionano la randomizzazione quale tecnica per conseguire l'anonimizzazione desiderata.

Molti esperimenti di reidentificazione di questo tipo hanno seguito un approccio simile, proiettando due banche dati sullo stesso sottospazio. È una metodologia di reidentificazione molto potente, che di recente ha trovato applicazione in molte aree diverse. Ad esempio, un esperimento di identificazione effettuato su un social network[25] ha sfruttato il grafico sociale degli utenti pseudonimizzati tramite etichette. In questo caso, l'attributo utilizzato per l'identificazione è stato l'elenco dei contatti di ciascun utente, in quanto è stato dimostrato che la probabilità che due persone abbiano il medesimo elenco dei contatti è molto bassa. Sulla base di tale ipotesi intuitiva, si è constatato che un sottografico delle connessioni interne di un numero di nodi molto limitato costituisce un'impronta digitale topologica per il recupero dei dati nascosti nel network, e che, una volta identificata tale sottorete, è possibile identificare una parte consistente di tutto il social network. Per fornire qualche cifra sulla potenza di un attacco del genere, è stato dimostrato che utilizzando meno di 10 nodi (che possono dare luogo a milioni di configurazioni di sottorete diverse, ognuna delle quali costituisce potenzialmente un'impronta digitale topologica) è possibile esporre ad attacchi di reidentificazione un social network di più di 4 milioni di nodi pseudonimizzati e 70 milioni di collegamenti, nonché compromettere la privacy di un numero elevato di connessioni. Va precisato che tale approccio di reidentificazione non è mirato al contesto specifico dei social network, bensì è sufficientemente generico da poter essere potenzialmente adattato ad altre banche dati in cui sono registrati i rapporti tra gli utenti (ad esempio rubrica telefonica, corrispondenza tramite posta elettronica, luoghi di incontro, ecc.).

Un altro modo per identificare un dato presumibilmente anonimo si basa sull'analisi dello stile di scrittura (stilometria)[26]. Sono stati già sviluppati vari algoritmi volti a estrarre la metrica da un testo oggetto di analisi, compresa la frequenza dell'utilizzo di una particolare terminologia, la comparsa di schemi grammaticali specifici e il tipo di punteggiatura. Tutte queste proprietà possono essere utilizzate per collegare un testo apparentemente anonimo allo stile di scrittura di un autore identificato. I ricercatori hanno reperito lo stile di scrittura di oltre 100 000 blog e sono attualmente in grado di identificare automaticamente l'autore di un post con un grado di precisione ormai prossimo all'80%; l'accuratezza di tale tecnica è destinata ad aumentare anche con l'utilizzo di altri segnali, quali la localizzazione o altri metadati contenuti nel testo.

Il potere di identificazione che utilizza la semantica di un dato (vale a dire, la parte residua non randomizzata di un dato) è una tematica che merita maggiore attenzione da parte della comunità di ricerca e dell'industria. La recente rivelazione dell'identità dei donatori di DNA

[25] L. Backstrom, C. Dwork, e J. M. Kleinberg. *Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography*, Atti della sedicesime conferenza internazionale sul World Wide Web WWW'07, pagg. 181-190 (2007).

[26] http://33bits.org/2012/02/20/is-writing-style-sufficient-to-deanonymize-material-posted-online/.

(2013)[27] dimostra che sono stati compiuti ben pochi progressi dal noto caso di AOL (2006) – quando venne resa nota pubblicamente una banca dati contenente venti milioni di parole chiave di ricerca relative a più di 650 000 utenti nell'arco di 3 mesi. Il risultato è stato che alcuni utenti di AOL sono stati identificati e localizzati.

Un'altra famiglia di dati che raramente viene anonimizzata semplicemente eliminando l'identità delle persone interessate o crittografando in parte alcuni attributi sono i dati di localizzazione. Gli schemi di mobilità degli esseri umani sono sufficientemente unici da consentire alla parte semantica dei dati di localizzazione (i luoghi in cui la persona interessata si trovava in un momento specifico), persino in assenza di altri attributi, di rivelare molti tratti di una persona interessata[28]. Studi accademici rappresentativi l'hanno dimostrato diverse volte[29].

A questo proposito, occorre mettere in guardia dal ricorso a pseudonimi quale modo per garantire alle persone interessate una protezione adeguata da fughe di identità o attributi. Se la pseudonimizzazione si basa sulla sostituzione di un'identità con un altro codice unico, presumere che ciò costituisca un meccanismo di deidentificazione efficace è ingenuo e non tiene conto della complessità delle metodologie di identificazione e dei contesti molteplici in cui le stesse possono essere applicate.

A.3. "Anonimizzazione" per generalizzazione

Un esempio semplice potrebbe essere utile a chiarire l'approccio che si basa sulla generalizzazione degli attributi.

Esaminiamo il caso in cui un responsabile del trattamento decida di pubblicare una tabella semplice che contiene tre informazioni o attributi: un numero di identificazione, unico per ogni dato, un'identificazione di localizzazione, che collega la persona interessata al luogo in cui vive, e un'identificazione di proprietà, che indica la proprietà posseduta dalla persona interessata; ipotizziamo inoltre che tale proprietà sia uno di due valori distinti, indicati generalmente da {P1, P2}:

[27] I dati genetici sono un esempio particolarmente significativo di dati sensibili a rischio di reidentificazione nel caso in cui l'unico meccanismo utilizzato per "anonimizzarli" sia l'eliminazione dell'identità dei donatori. Cfr. l'esempio citato al precedente paragrafo 2.2.2. Cfr. anche John Bohannon, *Genealogy Databases Enable Naming of Anonymous DNA Donors*, Science, Vol. 339, n. 6117 (18 gennaio 2013), pag. 262.

[28] Tale questione è stata affrontata in alcune legislazioni nazionali. Ad esempio, in Francia le statistiche di localizzazione pubblicate sono anonimizzate mediante generalizzazione e permutazione. Di conseguenza, INSEE pubblica statistiche che sono generalizzate mediante l'aggregazione di tutti i dati in una superficie di 40 000 metri quadrari. Il grado di dettaglio dell'insieme di dati è tale da preservare l'utilità dei dati e le permutazioni impediscono gli attacchi di deanonimizzazione nelle aree diradate. Più in generale, l'aggregazione di questa famiglia di dati e la loro permutazione fornisce garanzie forti di tutela da attacchi di deduzione e deanonimizzazione (http://www.insee.fr/en/).

[29] de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M. & Blondel, V.D. *Unique in the Crowd: The privacy bounds of human mobility*. Nature. 3, 1376 (2013).

Identità seriale	Identità di localizzazione	Proprietà
#1	Roma	P1
#2	Madrid	P1
#3	Londra	P2
#4	Parigi	P1
#5	Barcellona	P1
#6	Milano	P2
#7	New York	P2
#8	Berlino	P1

Tabella A1. Campione di persone interessate raggruppate per localizzazione e patrimonio P1 e P2

Se qualcuno, denominato intruso, sa in anticipo che una persona interessata specifica (il bersaglio) che abita a Milano è inclusa nella tabella, dopo aver esaminato la tabella può giungere alla conclusione che #6, essendo l'unica persona interessata con quell'identità di localizzazione, possiede anche la proprietà P2.

Questo esempio essenziale mette in evidenza gli elementi principali di qualsiasi procedura di identificazione applicata a un insieme di dati che è stato sottoposto a un presunto processo di anonimizzazione. In particolare, un intruso (accidentalmente o intenzionalmente) dispone di informazioni di base su alcune o tutte le persone interessate di un insieme di dati e si propone di mettere in relazione tali conoscenze di base con i dati presenti nell'insieme di dati pubblicato per ottenere un quadro più chiaro delle caratteristiche delle persone interessate.

Per rendere meno efficace o meno immediato il collegamento dei dati con qualsiasi tipo di conoscenza di base, il responsabile del trattamento potrebbe concentrarsi sull'identità di localizzazione, sostituendo la città in cui abitano le persone interessate con una regione più ampia, ad esempio lo Stato. La tabella risultante sarebbe la seguente.

Identità seriale	Identità di localizzazione	Proprietà
#1	Italia	P1
#2	Spagna	P1
#3	Regno Unito	P2
#4	Francia	P1
#5	Spagna	P1
#6	Italia	P2
#7	Stati Uniti	P2
#8	Germania	P1

Tabella A2. Generalizzazione della tabella A1 per nazionalità

Con questa nuova aggregazione dei dati, la conoscenza di base di un intruso circa una persona interessata identificata (diciamo, "il bersaglio abita a Roma e compare nella tabella") non consente di trarre alcuna conclusione chiara sulla proprietà, in quanto i due italiani che compaiono nella tabella hanno proprietà distinte, rispettivamente P1 e P2. L'intruso si ritroverebbe con un'incertezza del 50% sulla proprietà del soggetto bersaglio. Questo semplice esempio mostra gli effetti della generalizzazione sulla pratica dell'anonimizzazione. Di fatto, sebbene questo accorgimento della generalizzazione possa essere efficace per dimezzare la probabilità di identificare un bersaglio italiano, non sarebbe altrettanto utile nel caso di un bersaglio proveniente da altri luoghi (ad esempio, gli Stati Uniti).

Inoltre, l'intruso avrebbe ancora la possibilità di acquisire informazioni su un bersaglio spagnolo. Se la conoscenza di base è del tipo "il bersaglio abita a Madrid e compare nella tabella" oppure "il bersaglio abita a Barcellona e compare nella tabella", l'intruso può dedurre con una certezza del 100% che il bersaglio possiede la proprietà P1. Ne consegue che la generalizzazione non garantisce lo stesso grado di tutela della sfera privata o la stessa resistenza ad attacchi tramite deduzione che colpiscano tutta la popolazione presente nell'insieme di dati.

Seguendo tale ragionamento, si potrebbe essere tentati di concludere che una generalizzazione maggiore potrebbe essere utile per impedire qualsiasi collegamento, ad esempio una generalizzazione per continente. La tabella risultante sarebbe la seguente.

Identità seriale	Identità di localizzazione	Proprietà
#1	Europa	P1
#2	Europa	P1
#3	Europa	P2
#4	Europa	P1
#5	Europa	P1
#6	Europa	P2
#7	Nord America	P2
#8	Europa	P1

Tabella A3. Generalizzazione della tabella A1 per continente

Con questo tipo di aggregazione, tutte le persone interessate presenti nella tabella, ad eccezione di quella che abita negli Stati Uniti, sarebbero tutelate dagli attacchi mediante collegamento e identificazione, e un'eventuale conoscenza di base del tipo "il bersaglio abita a Madrid e compare nella tabella" oppure "il bersaglio abita a Milano e compare nella tabella" comporterebbe un certo grado di probabilità riguardo alla proprietà relativa a una data persona interessata (P1 con una probabilità del 71,4% e P2 con una probabilità del 28,6%), anziché un collegamento diretto. Inoltre, tale generalizzazione supplementare si traduce in una perdita evidente e radicale di informazioni: la tabella non consente di individuare eventuali correlazioni tra le proprietà e il luogo, vale a dire se un luogo specifico sia in grado di determinare con maggiore probabilità l'una o l'altra delle due proprietà, in quanto rivela solamente le cosiddette distribuzioni "marginali", vale a dire la probabilità assoluta di riscontrare la proprietà P1 e P2 a livello di popolazione (pari rispettivamente al 62,5% e al 37,5% nel nostro esempio) e in ogni continente (rispettivamente, come precisato, 71,4% e 28,6% in Europa e 100% e 0% nell'America settentrionale).

L'esempio dimostra inoltre che la pratica della generalizzazione incide sull'utilità pratica dei dati. Attualmente sono disponibili alcuni strumenti di ingegneria che consentono di stabilire in anticipo (vale a dire, prima della pubblicazione di un insieme di dati) qual è il livello più appropriato di generalizzazione degli attributi, in modo da ridurre il rischio di identificazione delle persone interessate presenti nella tabella senza compromettere eccessivamente l'utilità dei dati pubblicati.

K-anonimato

Un modo per prevenire gli attacchi tramite collegamenti sulla base della generalizzazione degli attributi è il noto sistema denominato *k*-anonimato. Tale pratica è nata in seguito a un esperimento di reidentificazione condotto alla fine degli anni '90 in cui una società privata

statunitense, attiva nel settore sanitario, aveva reso pubblico un insieme di dati ritenuti anonimizzati. L'anonimizzazione consisteva nell'eliminare i nomi delle persone interessate, tuttavia l'insieme dei dati conteneva ancora dati di carattere sanitario e altri attributi quali il codice di avviamento postale (l'identità del luogo in cui abitavano i soggetti), il sesso e la data di nascita completa. La stessa tripletta di attributi {codice di avviamento postale, sesso e data di nascita completa} era stata inoltre inserita in altri registri pubblicamente accessibili (ad esempio, la lista elettorale) e pertanto aveva potuto essere utilizzata da un ricercatore universitario per collegare l'identità delle persone interessate specifiche agli attributi contenuti nell'insieme di dati pubblicato. La conoscenza di base di cui l'intruso (il ricercatore) disponeva poteva essere la seguente: "so che la persona interessata presente nella lista elettorale con una tripletta specifica {codice di avviamento postale, sesso e data di nascita completa} è unica. Nell'insieme di dati pubblicato esiste un dato con tale tripletta". È stato osservato empiricamente[30] che la grande maggioranza (più dell'80%) delle persone interessate presenti nel registro pubblico utilizzato nell'esperimento di ricerca in oggetto è stata associata in maniera univoca a una tripletta specifica, che ne ha reso possibile l'identificazione. Di conseguenza, in questo caso i dati non erano stati adeguatamente anonimizzati.

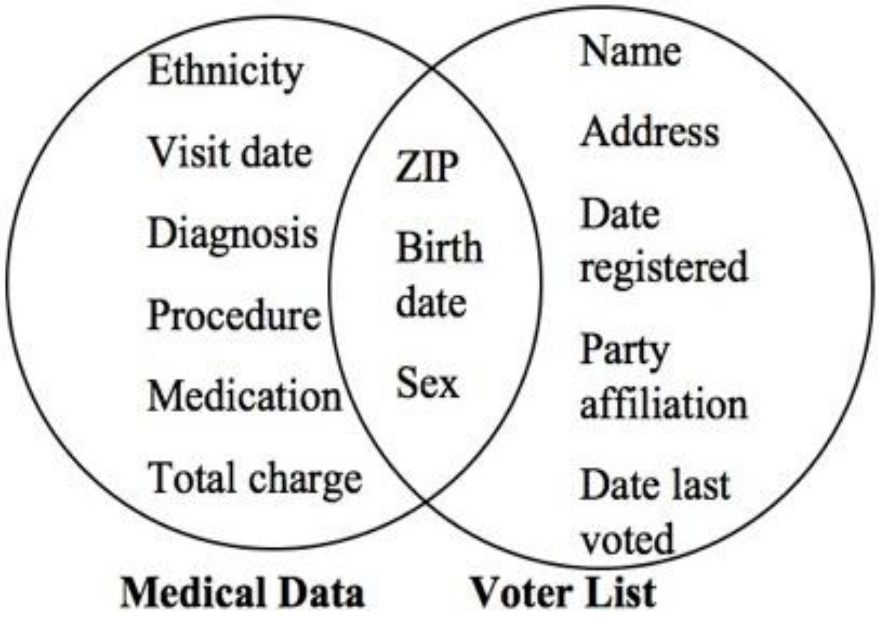

Figura A1. Reidentificazione tramite collegamento dei dati

Al fine di ridurre l'efficacia di attacchi simili tramite collegamenti, si è sostenuto che i responsabili del trattamento dovrebbero in primo luogo esaminare l'insieme di dati e raggruppare gli attributi che potrebbero ragionevolmente essere utilizzati da un intruso per collegare la tabella pubblicata a un'altra fonte ausiliaria; ogni gruppo dovrebbe comprendere almeno *k* combinazioni identiche di attributi generalizzati (vale a dire, dovrebbe rappresentare una classe di equivalenza di attributi). Gli insiemi di dati dovrebbero essere poi pubblicati solo dopo essere stati ripartiti in tali gruppi omogenei. Gli attributi selezionati per la generalizzazione sono noti in letteratura come quasi-identificatori, in quanto la loro conoscenza "in chiaro" comporterebbe l'immediata identificazione delle persone interessate.

Molti esperimenti di identificazione hanno dimostrato la debolezza di tabelle *k*-anonimizzate progettate in maniera carente. Ad esempio, tale eventualità potrebbe verificarsi perché gli altri

[30] L. Sweeney. *Weaving Technology and Policy Together to Maintain Confidentiality*. *Journal of Law, Medicine &Ethics*, 25, nn. 2 e 3 (1997): 98-110.

attributi in una classe di equivalenza sono identici (come accade per la classe di equivalenza di persone interessate spagnole nell'esempio della tabella A2) oppure la loro distribuzione è molto sbilanciata con una prevalenza elevata di un attributo specifico, o ancora perché il numero di dati in una classe di equivalenza è molto basso e consente in entrambi i casi la deduzione per probabilità, oppure perché non esiste alcuna differenza "semantica" significativa tra gli attributi in chiaro delle classi di equivalenza (ad esempio, la misura quantitativa di tali attributi potrebbe essere effettivamente diversa ma numericamente molto vicina, oppure potrebbero appartenere a una gamma di attributi semanticamente simili, ad esempio, la medesima categoria di rischio di credito, o la stessa famiglia di patologie), per cui l'insieme di dati potrebbe ancora rivelare molte informazioni sulle persone interessate consentendo attacchi tramite collegamenti[31]. Un punto importante da sollevare a questo proposito è che ogniqualvolta i dati sono diradati (ad esempio, una proprietà specifica compare molto di rado in un'area geografica), e una prima aggregazione non è in grado di raggruppare i dati con un numero sufficiente di ricorrenze di diverse proprietà (ad esempio, in un'area geografica è ancora possibile localizzare un numero basso di ricorrenze di poche proprietà), occorre un'ulteriore aggregazione degli attributi per conseguire l'anonimizzazione desiderata.

L-diversità

Partendo da tali osservazioni, nel corso degli anni sono state proposte varianti del *k*-anonimato e sono stati sviluppati alcuni criteri di progettazione per potenziare la pratica di anonimizzazione mediante generalizzazione, al fine di attenuare i rischi di attacchi tramite collegamenti. Tali varianti e criteri si basano sulle proprietà probabilistiche degli insiemi di dati. Nello specifico, si aggiunge un ulteriore vincolo, e cioè che ogni attributo in una classe di equivalenza ricorra almeno *l* volte, in modo che un eventuale intruso si ritrovi sempre con un grado di incertezza notevole sugli attributi pur possedendo una conoscenza di base su una persona interessata specifica. Ciò equivale a dire che un insieme di dati (o partizione) dovrebbe possedere un numero minimo di ricorrenze di una proprietà selezionata: tale accorgimento dovrebbe attenuare il rischio di reidentificazione. Questo è l'obiettivo che si propone di conseguire la pratica di anonimizzazione della *l*-diversità. Un esempio di tale pratica compare nelle tabelle A4 (i dati originali) e A5 (il risultato del trattamento). Come si evince dalle tabelle, progettando adeguatamente l'identità della localizzazione e le età delle persone nella tabella A4, la generalizzazione degli attributi comporta un incremento ragguardevole dell'incertezza riguardante gli attributi effettivi di ogni persona interessata coinvolta nell'indagine. Ad esempio, anche se un intruso sa che una persona interessata appartiene alla prima classe di equivalenza, non può accertare se tale persona possieda le proprietà X, Y o Z, in quanto in tale classe (e in tutte le altre classi di equivalenza) esiste almeno un dato che presenta tali proprietà.

[31] Va precisato che è possibile stabilire correlazioni anche una volta che i dati sono stati raggruppati per attributi. Quando il responsabile del trattamento conosce i tipi di correlazione che desidera verificare, può selezionare gli attributi maggiormente rilevanti. Ad esempio, i risultati delle indagini PEW non sono soggetti ad attacchi di deduzione con un grado elevato di dettaglio e sono comunque molto utili per individuare correlazioni tra demografia e interessi (http://www.pewinternet.org/Reports/2013/Anonimato-online.aspx).

Numero seriale	Identità di localizzazione	Età	Proprietà
1	111	38	X
2	122	39	X
3	122	31	Y
4	111	33	Y
5	231	60	Z
6	231	65	X
7	233	57	Y
8	233	59	Y
9	111	41	Z
10	111	47	Z
11	122	46	Z
12	122	45	Z

Tabella A4. Tabella con le persone raggruppare in base a localizzazione, età e tre proprietà X, Y e Z

Numero seriale	Identità di localizzazione	Età	Proprietà
1	11*	<50	X
4	11*	<50	Y
9	11*	<50	Z
10	11*	<50	Z
5	23*	>50	Z
6	23*	>50	X
7	23*	>50	Y
8	23*	>50	Y
2	12*	<50	X
3	12*	<50	Y
11	12*	<50	Z
12	12*	<50	Z

Tabella A5. Esempio di versione 1-diversa della tabella A4

T-T-vicinanza

L'approccio noto come *t-t*-vicinanza si applica al caso specifico di attributi all'interno di una partizione che sono distribuiti in maniera disomogenea o appartengono a una gamma ridotta di valori o significati semantici. Si tratta di un ulteriore miglioramento dell'anonimizzazione per generalizzazione e consiste nel disporre i dati in modo da ottenere classi di equivalenza che riproducano per quanto possibile la distribuzione iniziale degli attributi nell'insieme di dati originario. A tale scopo si ricorre a una procedura in due fasi illustrata qui di seguito. La tabella A6 rappresenta la banca dati originaria che comprende i dati in chiaro delle persone interessate, raggruppati in base a localizzazione, età, retribuzione e due famiglie di proprietà semanticamente simili, rispettivamente (X1, X2, X3) e (Y1, Y2, Y3) (ad esempio, categorie simili di rischio di credito, patologie simili). In primo luogo, la tabella è *l-diversificata* con l=1 (tabella A7), mediante il raggruppamento dei dati in classi di equivalenza semanticamente simili e con una scarsa anonimizzazione mirata; la stessa viene successivamente trattata in modo da ottenere la *t*-vicinanza (tabella A8) e una variabilità più alta all'interno di ogni partizione. Di fatto, con la seconda fase, ogni classe di equivalenza contiene dati di entrambe le famiglie di proprietà. Va osservato che l'identità di localizzazione e l'età presentano gradi di dettaglio diversi nelle varie fasi del processo, e pertanto ogni attributo potrebbe richiedere criteri di generalizzazione diversi per ottenere l'anonimizzazione desiderata, e ciò presuppone

a sua volta una progettazione specifica e un carico computazionale adeguato da parte dei responsabili del trattamento.

Numero seriale	Identità di localizzazione	Età	Retribuzione	Proprietà
1	1127	29	30k	X1
2	1112	22	32k	X2
3	1128	27	35k	X3
4	1215	43	50k	X2
5	1219	52	120k	Y1
6	1216	47	60k	Y2
7	1115	30	55k	Y2
8	1123	36	100k	Y3
9	1117	32	110k	X3

Tabella A6. Tabella con persone raggruppate in base a localizzazione, età, retribuzione e due famiglie di proprietà

Numero seriale	Identità di localizzazione	Età	Retribuzione	Proprietà
1	11**	2*	30k	X1
2	11**	2*	32k	X2
3	11**	2*	35k	X3
4	121*	>40	50k	X2
5	121*	>40	120k	Y1
6	121*	>40	60k	Y2
7	11**	3*	55k	Y2
8	11**	3*	100k	Y3
9	11**	3*	110k	X3

Tabella A7. Versione *l*-diversa della tabella A6

Numero seriale	Identità di localizzazione	Età	Retribuzione	Proprietà
1	112*	<40	30k	X1
3	112*	<40	35k	X3
8	112*	<40	100k	Y3
4	121*	>40	50k	X2
5	121*	>40	120k	Y1
6	121*	>40	60k	Y2
2	111*	<40	32k	X2
7	111*	<40	55k	Y2
9	111*	<40	110k	X3

Tabella A8. Versione *t*-vicina della tabella A6

Va ribadito con chiarezza che l'obiettivo di generalizzare gli attributi delle persone interessate secondo modalità così raffinate talvolta può essere conseguito solamente per un numero esiguo di dati e non per tutti. Le buone pratiche dovrebbero garantire che ciascuna classe di equivalenza contenga molte persone e che non sia più esposta ad attacchi tramite deduzione. In ogni caso, tale approccio presuppone una valutazione approfondita dei dati disponibili da parte dei responsabili del trattamento oltre che la valutazione combinatoria di varie alternative

(ad esempio, diverse ampiezze di gamma, diversi gradi di dettaglio della localizzazione o dell'età, ecc.). In altre parole, l'anonimizzazione per generalizzazione non può essere il risultato di un primo tentativo approssimativo da parte dei responsabili del trattamento di sostituire i valori analitici degli attributi in un dato in base alla gamma di valori, in quanto occorrono approcci quantitativi più specifici, come la valutazione dell'entropia degli attributi all'interno di ciascuna partizione o la misurazione della distanza tra le distribuzioni degli attributi originari e la distribuzione all'interno di ogni classe di equivalenza.