Università di Catania

Dipartimento di Ingegneria Elettrica, Elettronica e Informatica
Corso di Laurea in Ingegneria Informatica

Federico Nicotra

Implementazione di un attacco all'anonimizzazione

Progetto di Internet Security

Anno Accademico 2021/22

Indice

1. Stato dell'arte

Definizione di anonimizzazione

Secondo lo standard di Health Informatics ISO/TS 25237:20 "*L'anonimizzazione è un processo mediante il quale i dati personali vengono modificati in modo irreversibile così che il titolare del trattamento, da solo o in collaborazione con altre parti, non possa più identificare direttamente o indirettamente l'interessato*". Più sinteticamente, seguendo un passaggio della Opinion 05/2014 on Anonymisation Techniques del Working Party art.29: "*... l'anonimizzazione è una tecnica che si applica ai dati personali al fine di ottenere una deidentificazione irreversibile*". I dati personali possono dunque essere considerarsi effettivamente e sufficientemente anonimizzati se non si riferiscono a una persona fisica identificabile o se sono stati resi anonimi in modo tale da rendere l'interessato non più identificabile.

Per identificazione (o de-identificazione, o re-identificazione) si intende la possibilità di associare univocamente ai dati, con o senza ulteriori analisi, i proprietari degli stessi. Sono tutt'oggi in corso degli studi sui meccanismi di anonimizzazione; pertanto, la conoscenza dell'efficacia delle varie tecniche di anonimizzazione è in continua evoluzione. Non è detto che dei dati anonimizzati lo rimangano per sempre o che una tecnica rimarrà efficace per lungo tempo.

Differenze con la pseudonimizzazione

La pseudonimizzazione è un altro modo per rendere anonimi dei dati. Tuttavia, questa tecnica prevede che i dati, tramite informazioni aggiuntive, possano essere deanonimizzati, al contrario dell'anonimizzazione. Il GDPR definisce che i "*dati pseudonimi sono quei dati personali nei quali gli elementi identificativi sono stati sostituiti da elementi diversi, quali stringhe di caratteri o numeri (hash), oppure sostituendo al nome un nickname, purché sia tale da rendere estremamente difficoltosa l'identificazione dell'interessato. Ovviamente il soggetto che detiene la chiave per decifrare i dati (cioè collegare l'elemento pseudonimo al dato personale) deve garantire adeguate misure contro possibili abusi*". Nell'anonimizzazione il processo inverso è fortemente scoraggiato anche dal fatto che chi crea dati anonimizzati non deve avere la possibilità di tornare indietro, come diversamente espresso nella citazione precedente riguardante la pseudonimizzazione.

GDPR su anonimizzazione e pseudonimizzazione

Il Regolamento Generale per la Protezione dei Dati (GDPR) deve essere applicato su dati sottoposti a pseudonimizzazione poiché ancora riconducibili a persone fisiche. I dati, tuttavia, sottoposti ad anonimizzazione non subiscono alcun trattamento dal regolamento. Nel Recital 26, infatti, si legge che:

"È auspicabile applicare i principi di protezione dei dati a tutte le informazioni relative a una persona fisica identificata o identificabile. I dati personali sottoposti a pseudonimizzazione, i quali potrebbero essere attribuiti a una persona fisica mediante l'utilizzo di ulteriori informazioni, dovrebbero essere considerati informazioni su una persona fisica identificabile. Per stabilire l'identificabilità di una persona è opportuno considerare tutti i mezzi, come l'individuazione, di cui il titolare del trattamento o un terzo può ragionevolmente avvalersi per identificare detta persona fisica direttamente o indirettamente. Per accertare la ragionevole probabilità di utilizzo dei mezzi per identificare la persona fisica, si dovrebbe prendere in considerazione l'insieme dei fattori obiettivi, tra cui i costi e il tempo necessario per l'identificazione, tenendo conto sia delle tecnologie disponibili al momento del trattamento, sia degli sviluppi tecnologici. I principi di protezione dei dati non dovrebbero pertanto applicarsi a informazioni anonime, vale a dire informazioni che non si riferiscono a una persona fisica identificata o identificabile o a dati personali resi sufficientemente anonimi da impedire o da non consentire più l'identificazione dell'interessato. Il presente regolamento non si applica pertanto al trattamento di tali informazioni anonime, anche per finalità statistiche o di ricerca."

Utilità dell'anonimizzazione

Il rilascio di dati anonimizzati può rendersi utile per scopi di ricerca o statistici come iscrizioni a corsi di laurea negli atenei italiani, andamento epidemia covid per ospedale o log accesso a siti web mantenendo solo la data di accesso e la pagina visitata. Togliendo informazione al dataset se ne limiteranno le modalità di utilizzo quindi chi crea dataset anonimizzati ne dovrà tenere conto.

Principali rischi e soluzioni

I maggiori rischi cui un garante della privacy deve far fronte quando rilascia dei dati anonimizzati sono i seguenti:

- Individuazione, che corrisponde alla possibilità di isolare uno o un gruppo di tuple che individuano una sola persona all'interno del dataset;
- Collegabilità, che corrisponde alla possibilità di collegare almeno due tuple associate allo stesso soggetto o gruppo di individui;

- Inferenza, che consiste nella possibilità di dedurre, con alta probabilità, il valore di un attributo basandosi sui valori di altri attributi.

Un meccanismo di anonimizzazione che tiene conto di ognuno dei precedenti punti può definirsi sufficientemente robusto contro la re-identificazione eseguita con i mezzi più probabili e ragionevoli che il titolare del trattamento e qualsiasi terza parte possano utilizzare.

Gli studi eseguiti negli anni hanno portato a due famiglie di approcci per l'anonimizzazione, ognuna avendo la propria robustezza contro i tre rischi precedenti. Una si basa sulla randomizzazione e l'altra sulla generalizzazione dei valori dei campi. La *randomizzazione* è una famiglia di tecniche che comportano l'alterazione dei dati di partenza, modificando la veridicità e l'esattezza dei dati: se i dati sono sufficientemente incerti e meno accurati, non possono più essere riferiti a una persona specifica. Nello specifico troviamo:

- Noise addition, che consiste nel modificare di una certa quantità un attributo mantenendo la stessa distribuzione generale;
- Permutazione, che consiste nello scambiare il valore di un attributo con quello di un'altra tupla;
- Privacy differenziale, che consiste nella noise addition applicata a delle views del dataset generate dal garante per terzi.

L'idea alla base delle *generalizzazioni* è quella di sostituire un valore specifico con un sinonimo generico, che però mantenga una certa informazione che non renda inutile il dataset. Questo procedimento tende a omogeneizzare le tuple così da ottenere un certo grado di incertezza quando si tenta di de-anonimizzare. Gli approcci in letteratura riportano:

- Aggregazione e K-Anonimato, che garantisce che ogni individuo di cui sono presenti dati non sia distinguibile da almeno altri K-1 soggetti all'interno della collezione di dati. Per far sì che avvenga, gli attributi vengono generalizzati creando o intervalli di valori (es. Età: 18-25, 26-35 ...) o generalizzando il concetto espresso dal valore (es. Italia → Europa Occidentale);
- L-Diversità garantisce che per ogni classe di equivalenza ogni attributo ha almeno L valori diversi. Esso limita che delle classi di equivalenza contengano elementi, i quali attributi hanno poca variabilità;
- T-Vicinanza, è presentata come un perfezionamento della L-Diversità, in quanto mira a creare classi di equivalenza che assomiglino alla distribuzione iniziale degli attributi nella tabella. Questa tecnica è utile quando è importante mantenere i dati il più vicino possibile a quelli originali; a tal fine,

viene posto un ulteriore vincolo alla classe di equivalenza, ovvero che non solo debbano esistere almeno L valori diversi all'interno di ciascuna classe di equivalenza, ma anche che ogni valore sia rappresentato tante volte quanto necessario per rispecchiare la distribuzione iniziale di ciascun attributo.

Al di là delle definizioni e delle tecniche di anonimizzazione è importante sottolineare che l'anonimizzazione non è per sempre. Come, teoricamente, non esiste un algoritmo crittografico a prova di bruteforce, non esiste una tecnica di anonimizzazione che renda completamente anonimi i dati. Si tratta sempre di capire quanto è difficoltoso, o oneroso, rompere l'anonimizzazione. Con l'aumentare della potenza computazionale e della disponibilità dei dati agli attaccanti, dei dati che adesso si considerano anonimi in futuro potrebbero non esserlo più. Il garante della privacy deve valutare il rischio di de-anonimizzazione dei dati che rilascia.

La tabella in basso riporta i punti di forza e i punti deboli delle tecniche citate considerando i tre rischi principali:

Tecnica	Rischio identificazione	Rischio collegabilità	Rischio inferenza
Pseudonimizzazione	Sì	Sì	Sì
Noise addition	Sì	Probabilmente no	Probabilmente no
Permutazione	Sì	Sì	Probabilmente no
Aggregazione e K-Anonimato	No	Sì	Sì
L-Diversità	No	Sì	Probabilmente no
Privacy differenziale	Probabilmente no	Probabilmente no	Probabilmente no

Tabella 1 – Tecniche e rischi principali

2. Implementazione attacco

Scenario iniziale

Alex e Bruno sono due compagni di classe dell'Istituto di Istruzione Superiore Enrico Mobili di Anonilandia. I due sono molto amici e non ci sono segreti tra loro. Una mattina Alex non si presentò a scuola. Bruno dunque decise di scrivergli un messaggio chiedendogli la ragione per cui non si fosse presentato ma non ricevette alcuna risposta. Tentò di chiamarlo ma il telefono squillava a vuoto. Due giorni dopo Alex tornò a scuola e alla richiesta di spiegazioni da parte di Bruno, egli rispose che avrebbe dovuto fare dei controlli di salute di routine da un po' di tempo e per non procrastinare ancora ha deciso di farlo in quei giorni. Alex rassicurò Bruno sulla sua salute dicendogli che stava bene. Bruno però non riusciva a credergli del tutto. Guardandolo sembrava stare bene ma qualcosa gli faceva intuire che non era più come prima. Per tutto il mese tentò di chiedere se ci fosse qualcosa di diverso ma Alex continuava a rispondere che andava tutto bene. Bruno, dopo tutto questo tempo, non si era ancora convinto e decise di andare in fondo a questa storia.

I tre centri di diagnostica di Anonilandia (*Centro Diagnostico Rita Levi-Montalcini*, *Centro Diagnostico Umberto Veronesi*, *Centro Diagnostico Camillo Golgi*) rilasciano mensilmente dei report pubblici per fini statistici contenenti una lista di degenze. Sono inclusi solo i pazienti che sono stati dimessi in quel mese e che hanno dato il consenso a partecipare alla statistica.

L'attaccante

Bruno vuole consultare l'ultimo report mensile dei centri diagnostici della zona in cerca di qualcosa che possa essere riconducibile al suo amico Alex, potendo in questo modo venire a conoscenza della malattia diagnosticatagli. Ciò che tenterà dunque di eseguire sarà un attacco di $individuazione$.

Bruno è a conoscenza dell'*età* e del *genere* di Alex, nonché del *periodo di assenza* a scuola (Alex si è assentato dal 04/10/2021 al 06/10/2021). Tuttavia, Bruno non sa i giorni di ricovero e dimissione, tantomeno il centro diagnostico in cui Alex è stato. Bruno, dunque, *suppone* che Alex sia stato in uno dei tre centri diagnostici presenti in zona e *suppone* che abbia dato il consenso di inclusione nella statistica.

Dettagli sul dataset

Il database, completo di tutti i dati, di proprietà del centro diagnostico, contiene uno storico di tutti i dati personali della persona ricoverata nel medesimo. Si ipotizzi che contenga nome, cognome, data di nascita, codice fiscale, luogo di nascita, residenza, sesso, data di ricovero, data di dimissione e malattia diagnosticata. Per poter costruire il dataset da pubblicare bisognerà selezionare solo gli attributi che sono strettamente necessari ai fini di ricerca. Attributi come il nome, il cognome, la data di nascita, il codice fiscale, il genere, il luogo di nascita e la residenza sono attributi che individuano una persona in mezzo ad un gruppo molto vasto di persone, specialmente il codice fiscale, unico per ogni persona. È opportuno quindi selezionare solo gli attributi veramente rilevanti ai fini di ricerca e statistica. Si ipotizzi che, di tutti questi attributi personali, si includa solamente il genere e l'età della persona poiché, ai fini medici, si possa distinguere come certe malattie si comportino con l'avanzare dell'età, nei rispettivi generi. Infine, si includano le date di ricovero e dimissioni per un maggiore dettaglio delle durate delle degenze. Si include anche la malattia diagnosticata, oggetto di studio per la ricerca.

Selezionati gli attributi da rilasciare è prudente valutare la possibilità di reidentificazione da parte di attaccanti. Per mitigare il rischio è opportuno applicare dei metodi di generalizzazione o randomizzazione ai dati rilasciati. Nel dataset in esame si ipotizza che si sia applicata la generalizzazione dell'età e della malattia diagnosticata. Per esempio, nel dataset è inclusa la voce "diabete". Essa è volutamente generica perché esistono diverse tipologie di diabete e dunque, così facendo, si introduce un certo grado di incertezza che aiuta il processo di anonimizzazione: si sta utilizzando *aggregazione*. Per l'età si esegue lo stesso tipo di generalizzazione, creando delle fasce di età basate sulla cifra delle decine. Se l'età da generalizzare è minore di 10 allora verrà scritto "0*". Si sarebbe potuto scrivere anche "00-09"; si avrebbe avuto lo stesso tipo di generalizzazione.

Corpo dell'attacco

L'attaccante scarica i dataset dell'ultimo mese dei centri diagnostici e ottiene tre file con la stessa struttura, descritta in Figura 1, contenenti in totale 450 righe.

Egli osserva i dataset e comprende che con dei filtri adeguati può ricercare le tuple che rispetterebbero il profilo di Alex. Gli basterà filtrare per:

- *Age*: 1*

- *Gender*: M
- *Admission*: compresa tra la prima e l'ultima data di assenza
- *Dimission*: compresa tra la prima e l'ultima data di assenza

Bruno ha così ottenuto tre tabelle contenenti delle tuple da cui è sicuro di poter trarre delle informazioni, descritte in Figura 2, Figura 3 e Figura 4.

Age	Gender	Admission	Dimission	Disease
6*	F	2021-09-30	2021-10-02	varicella
0*	M	2021-10-21	2021-10-22	varicella
5*	M	2021-10-10	2021-10-20	vaiolo
4*	M	2021-10-12	2021-10-14	morbillo
1*	F	2021-10-15	2021-10-18	morbillo
3*	F	2021-10-03	2021-10-04	morbillo
1*	F	2021-10-11	2021-10-12	varicella

Figura 1 – Struttura dataset

Può escludere facilmente la "varicella" (Figura 2) poiché è risaputo che il periodo di guarigione può essere molto lungo. Resterebbe con due possibilità in mano: la prima è che ad Alex sia stato diagnosticato un tipo di "diabete" (Figura 3), la seconda è che abbia contratto una malattia che sarebbe riconducibile ad un tipo di "influenza" (Figura 2). Dalla tabella in Figura 4 non si può dedurre nulla di certo. Si potrebbe però ipotizzare che Alex non sia solito andare al Centro Diagnostico Umberto Veronesi, il che potrebbe risultare utile per ricerche future.

Bruno, nonostante la sua profonda conoscenza di Alex, non ha ottenuto alcuna certezza poiché, ammesso che le due supposizioni siano rispettate, avrebbe un'incertezza sul risultato. Se solo una di queste supposizioni non fosse rispettata, quindi o non è andato in questi centri diagnostici, o non ha dato il consenso per la statistica, o entrambi, l'attacco perde di significato.

Age	Gender	Admission	Dimission	Disease
1*	M	2021-10-05	2021-10-05	varicella
1*	M	2021-10-06	2021-10-06	influenza

Figura 2 – Centro D. Camillo Golgi

Age	Gender	Admission	Dimission	Disease
1*	M	2021-10-04	2021-10-06	diabete

Figura 3 – Centro D. Rita Levi-Montalcini

Age	Gender	Admission	Dimission	Disease

Figura 4 – Centro D. Umberto Veronesi (dataset vuoto)

3. Conclusioni

Si può dedurre che per attacchi del genere la conoscenza sia di vitale importanza perché senza di essa un attacco all'anonimizzazione sarebbe molto difficile, se non impossibile. Se per esempio un attaccante esterno alla classe, che non conosce il periodo di assenza di Alex, avesse fatto lo stesso attacco probabilmente non sarebbe arrivato alle stesse conclusioni di Bruno perché avrebbe avuto decine e decine di tuple e possibilità in più. Il modo migliore per evitare che questi attacchi possano accadere è cercare di implementare quante più misure di anonimizzazione possibili, bilanciando sempre utilità del dataset e rischio residuo. Una misura possibile sarebbe quella di chiedere al paziente esplicitamente se essere o meno inserito nei report statistici (come nello scenario). Questo informerà il paziente delle attività del centro di diagnostica, dello scopo delle stesse e dei rischi a cui egli si espone. Un'altra misura da non sottovalutare sarebbe quella di restringere la visibilità del dataset a terzi autorizzati. Questo ne limiterebbe l'uso improprio da utenti maligni (non autorizzati) e inoltre darebbe la possibilità al garante della privacy di tracciare le richieste e decidere caso per caso a chi rilasciare il dataset. Quando possibile, inoltre, bisogna implementare la privacy differenziale (*randomizzazione*) cosicché il dataset originale rimanga alla sola disponibilità e visibilità del proprietario. Quest'ultima soluzione, basandosi su un sistema di viste anonimizzate, permette al proprietario di valutare quali e quante siano le query che un terzo è abilitato ad eseguire prima che il dataset possa rilevare delle informazioni sensibili.

4. Bibliografia / Sitografia

[1] The Anonymisation Problem – Computerphile – https://www.youtube.com/watch?v=puQvpyf0W-M

[2] Utilizzo di dati social per la deanonimizzazione di tracce GPS – https://it.readkong.com/page/utilizzo-di-dati-social-per-la-deanonimizzazione-di-tracce-6998445

[3] Guidance on Anonymisation and Pseudonymisation – https://www.dataprotection.ie/sites/default/files/uploads/2022-04/Anonymisation%20and%20Pseudonymisation%20-%20latest%20April%202022.pdf

[4] Le principali tecniche di anonimizzazione – https://privacygdpr.it/news-privacy-sanita/le-principali-tecniche-di-anonimizzazione/

[5] Anonimizzazione dei dati personali: significato, benefici e dubbi in ottica GDPR – https://www.privacy.it/2021/05/11/anonimizzazione-gpdr-massimini/

[6] ARTICLE 29 DATA PROTECTION WORKING PARTY – Opinion 05/2014 on Anonymisation Techniques – https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[7] Dati Personali: anonimizzazione e pseudonimizzazione – https://www.altalex.com/documents/news/2021/06/08/dati-personali-anonimizzazione-e-pseudonimizzazione

[8] Dato Personale – https://protezionedatipersonali.it/dato-personale

[9] Data re-identification – https://en.wikipedia.org/wiki/Data_re-identification

Source Code at – https://github.com/fedenicotra/AnonymizationAttack