ANONIMIZZAZIONE E PSEUDONIMIZZAZIONE

PERCHÉ TOGLIERE GLI IDENTIFICATORI NON BASTA?

Francesca Pratesi

Università di Pisa

Email: pratesi@di.unipi.it

Knowledge Discovery and Delivery Lab
(ISTI-CNR & Università di Pisa)
www-kdd.isti.cnr.it

GDPR e identificativi online

- **Le persone fisiche possono essere associate a identificativi online prodotti dai dispositivi, dalle applicazioni, dagli strumenti e dai protocolli utilizzati**, quali gli indirizzi IP, a marcatori temporanei (cookies) o a identificativi di altro tipo, come i tag di identificazione a radiofrequenza. Tali identificativi possono lasciare tracce che, in particolare se combinate con identificativi univoci e altre informazioni ricevute dai server, possono essere utilizzate per creare profili delle persone fisiche e identificarle. (considerando 30)

GDPR e identificazione

- **È auspicabile applicare i principi di protezione dei dati a tutte le informazioni relative a una persona fisica identificata o identificabile.** (considerando 26)
- L'applicazione della **pseudonimizzazione** ai dati personali può ridurre i rischi per gli interessati e aiutare i titolari del trattamento e i responsabili del trattamento a rispettare i loro obblighi di protezione dei dati. [. . .] (considerando 28)
- [. . .] Al fine di poter dimostrare la conformità con il presente regolamento, il titolare del trattamento dovrebbe adottare politiche interne e attuare misure che [. . .] potrebbero consistere, tra l'altro, nel ridurre al minimo il trattamento dei dati personali, **pseudonimizzare i dati personali il più presto possibile**, offrire trasparenza per quanto riguarda le funzioni e il trattamento di dati personali, consentire all'interessato di controllare il trattamento dei dati e consentire al titolare del trattamento di creare e migliorare caratteristiche di sicurezza. (considerando 78)
- Il trattamento di dati personali a fini di archiviazione nel pubblico interesse, di ricerca scientifica o storica o a fini statistici dovrebbe essere soggetto a garanzie adeguate per i diritti e le libertà dell'interessato, in conformità del presente regolamento. [. . .] il titolare del trattamento ha valutato la fattibilità di conseguire tali finalità trattando dati personali che non consentono o non consentono più di identificare l'interessato, purché esistano garanzie adeguate (come ad esempio la **pseudonimizzazione** dei dati personali). (considerando156)

GDPR e identificazione (2)

- Laddove il trattamento per una finalità diversa da quella per la quale i dati personali sono stati raccolti non sia basato sul consenso dell'interessato o su un atto legislativo dell'Unione o degli Stati membri [. . .] il titolare del trattamento tiene conto, tra l'altro: [. . .] e) dell'esistenza di garanzie adeguate, che possono comprendere la cifratura o la **pseudonimizzazione**. (Articolo 6 comma 4)

- [..] sia al momento di determinare i mezzi del trattamento sia all'atto del trattamento stesso il titolare del trattamento mette in atto misure tecniche e organizzative adeguate, quali la **pseudonimizzazione**, volte ad attuare in modo efficace i principi di protezione dei dati [. . .] (Articolo 25 comma 1)

Pseudonimizzazione

Sostituire un dato identificativo (es. nomi, codice fiscale) con **un valore surrogato** che spesso è chiamato **token**

Il valore surrogato deve essere:

- irreversibile senza informazione aggiuntiva
- distinguibile dal valore originale

Obiettivo della Pseudonimizzazione

- Ridurre il rischio di rendere pubblici dei dati che permettono la re-identificazione diretta, cioè possibile senza informazione aggiuntiva
- A tal fine si deve mantenere la corrispondenza tra il valore originale e il relativo token in una locazione differente dalla tabella pseudonima

Locazione A

Nomi	Valore Surrogato
Anna Verdi	11779
Luisa Rossi	12121
Giorgio Giallo	21177
Luca Nero	41898
Elisa Bianchi	56789
Enrico Rosa	65656

Locazione B

ID	Sesso	Data Nascita	CAP	DIAGNOSI
11779	F	1962	300122	Cancro
12121	F	1960	300133	Gastrite
21177	M	1950	300111	Infarto
41898	M	1955	300112	Emicrania
56789	F	1965	300200	Lussazione
65656	M	1953	300115	Frattura

Come generare i valori surrogati?

Tecniche di crittografia basate su chiave segreta

- Queste tecniche criptano i dati identificativi usando una chiave segreta. La decifratura può essere fatta solo usando questa chiave che è conosciuta solo dal controllore dei dati

Tecniche basate su funzioni di hash

- Queste tecniche usano una funzione che, dato un identificatore (composto da uno o più attributi), restituisce un valore di dimensione fissa. La funzione non deve essere invertibile.

Tecniche basate su funzioni di tipo keyed-hash

- Queste tecniche sono molto simili alle precedenti, solo che la funzione hash per calcolare il valore surrogato a partire da un identificatore ha bisogno anche di una chiave segreta, rendendo in questo modo il lavoro di un possibile attaccante più difficile

Come generare i valori surrogati? (2)

Tecniche basate su funzioni di tipo keyed-hash con cancellazione della chiave

- Queste tecniche sono uguali alle precedenti, solo che dopo la generazione del valore surrogato, che sostituirà l'identificatore originale, la tabella di corrispondenza tra il valore originale e quello generato verrà cancellata

Tecniche di tokenizzazione

- Queste tecniche sostituiscono l'identificatore con un token generato con un meccanismo di crittografia, con una funzione che genera un numero sequenziale o con un numero casuale.

Tecniche basate su funzioni di tipo salted-hash

- La chiave qua è nota come salt; la differenza principale è che il salt non è necessariamente segreto (un buon salt deve essere il più possible unico e variare spesso) rendendo molto difficili attacchi basati su dizionario.

Anonimizzazione vs Pseudonimizzazione

- Sono due termini distinti spesso confusi
- Dati anonimi e pseudoanonimizzati sono considerati come due categorie molto diverse dal punto di vista legale
- Pseudonimizzazione sostituisce l'identità di una persona in modo tale che la re-identificazione di una persona richiede la conoscenza e l'uso di informazione aggiuntiva
- Anonimizzazione garantisce la protezione dei dati contro la re-identificazione diretta o indiretta di una soggetto

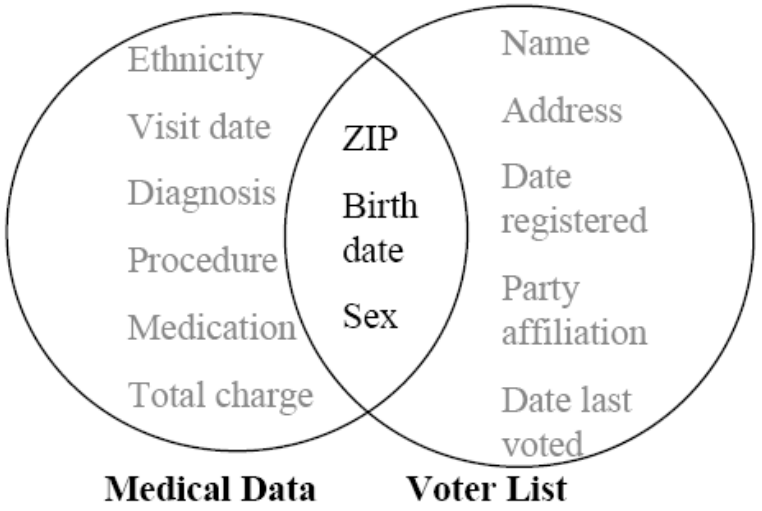
La Pseudonimizzazione è sufficiente per la protezione dei dati?

I dati Pseudoanonimizzati sono ancora Dati Personali!

Il caso del Governatore del Massachussetts

- Un attacco eseguito su una tabella contenente informazioni mediche su alcuni pazienti ha permesso la re-identificazione del governatore del MA.
  - MA ha pubblicato i dati medici di alcuni pazienti (sinistra)
  - Dati sugli elettori del Massachussetts; dataset pubblico (destra)

- Incrociando I dati abbiamo:
  - 6 persone con la stessa data di nascita
  - 3 erano uomini
  - 1 solo con lo stesso CAP



Latanya Sweeney: k-Anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5): 557-570 (2002)

GDPR e identificazione (3)

- L'introduzione esplicita della pseudonimizzazione nel presente regolamento non è quindi intesa a precludere **altre misure di protezione dei dati**. (considerando 28)

- **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. (considerando 26)

GDPR e soluzioni proposte

- La tutela dei diritti e delle libertà delle persone fisiche relativamente al trattamento dei dati personali richiede l'adozione di misure tecniche e organizzative adeguate per garantire il rispetto delle disposizioni del presente regolamento. Al fine di poter dimostrare la conformità con il presente regolamento, il titolare del trattamento dovrebbe adottare politiche interne e attuare misure che soddisfino in particolare i principi della protezione dei dati fin dalla progettazione e della **protezione dei dati di default**. Tali misure potrebbero consistere, tra l'altro, nel **ridurre al minimo il trattamento dei dati personali**, pseudonimizzare i dati personali il più presto possibile, offrire **trasparenza** per quanto riguarda le funzioni e il trattamento di dati personali, consentire all'interessato di **controllare il trattamento** dei dati e consentire al titolare del trattamento di creare e migliorare caratteristiche di **sicurezza**. (considerando 78)

- [. . .]Tali garanzie dovrebbero assicurare che siano state predisposte misure tecniche e organizzative al fine di garantire, in particolare, il principio della **minimizzazione** dei dati. (considerando 156)

- **Valutazione d'impatto sulla protezione dei dati** (Articolo 35)

Privacy-by-Design

- Approccio proattivo
- Sono necessarie delle analisi preliminari su:
  1) tipo di dato
  2) modello di attacco
  3) servizi da fornire

Tecniche di Anonimizzazione

- Tecniche basate su generalizzazione e soppressione
  - K-anonymity
  - l-diversity
  - t-closeness
- Tecniche basate su Randomizzazione
  - Differential-privacy
- Tecniche che garantiscono la privacy in sistemi distribuiti
  - Secure Multi-party computation
- Tecniche di anonimato per il cloud computing

K-Anonymity

- k-anonymity nasconde ogni persona tra altre k-1
  - All'interno della tabella ogni combinazione di attributi quasi identificatori dovrebbe apparire almeno **k** volte
  - La re-identificazione di un individuo incrociando i dati della tabella con informazione aggiuntiva non dovrebbe essere possibile con una confidenza > 1/k
- Come ottenere la k-anonymity?
  - Generalizzazione: pubblicare attributi con valori più generali (data una gerarchia)
  - Soppressione: rimuove record, cioè non si pubblicano outliers. Chiaramente il numero di record da rimuovere deve essere limitato
- Bilanciamento tra privacy e utilità dei dati
  - È necessario non anonimizzare più del necessario
  - Minimizzare la distorsione dei dai

Rendere i dati anonimi: generalizzazione

Governatore: Anno Nascita = 1950, CAP = 300111

ID	Sesso	Anno Nascita	CAP	DIAGNOSI
1	F	1/3/1962	300122	Cancro
3	F	12/2/1960	300133	Gastrite
2	M	4/8/1950	300111	Infarto
4	M	3/11/1955	300112	Emicrania
5	F	22/1/1965	300200	Lussazione
6	M	31/11/1953	300115	Frattura

Rendere i dati anonimi: generalizzazione (2)

Governatore: Anno Nascita = 1950, CAP = 300111

ID	Sesso	Data Nascita	CAP	DIAGNOSI
1	F	[1960-1956]	300***	Cancro
3	F	[1960-1956]	300***	Gastrite
2	M	[1950-1955]	30011*	Infarto
4	M	[1950-1955]	30011*	Emicrania
5	F	[1960-1956]	300***	Lussazione
6	M	[1950-1955]	30011*	Frattura

Vulnerabilità della K-anonymity

ID	Sesso	Data Nascita	CAP	DIAGNOSI
1	F	1962	300122	Cancro
3	F	1960	300133	Gastrite
2	M	1950	300111	Infarto
4	M	1950	300111	Infarto
5	M	1950	300111	Infarto
6	M	1953	300115	Frattura

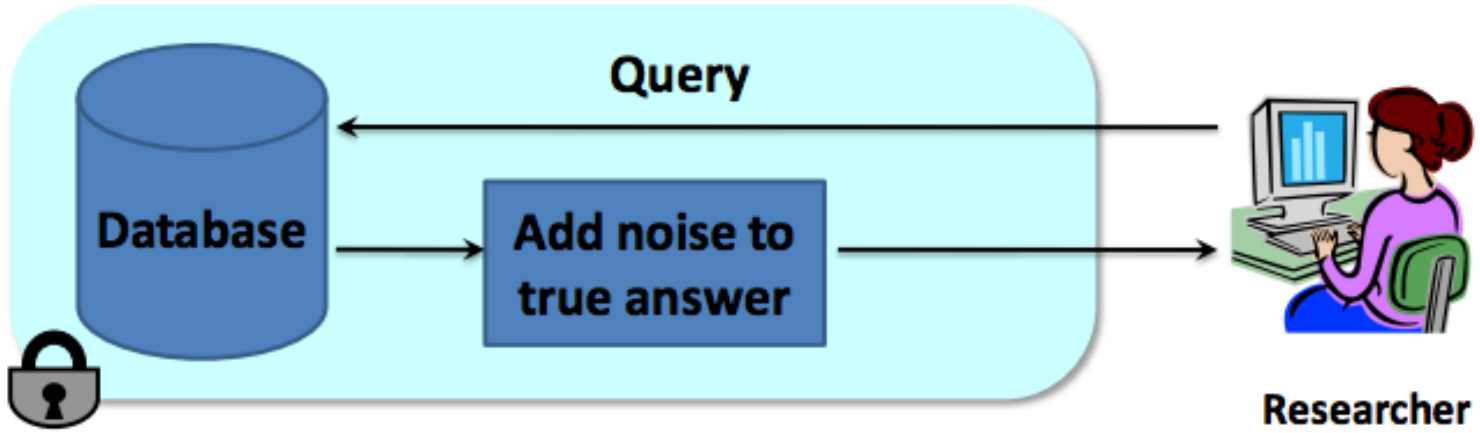
l-Diversity

- *l*-diversity
  - Ogni classe di equivalenza ha almeno L valori sensibili distinti

ID	Sesso	Data Nascita	CAP	DIAGNOSI
1	F	1962	300122	Cancro
3	F	1960	300133	Gastrite
2	M	1950	300111	Infarto
4	M	1950	300111	Emicrania
5	M	1950	300111	Lussazione
6	M	1953	300115	Frattura

Differential Privacy

- Il rischio di privacy di una persona non dovrebbe aumentare con la sua presenza in un database



- Aggiungere rumore alla risposta in modo tale che :
  - Ogni risposta non riveli troppa informazione sul database
  - Le risposte perturbate siano molto simili a quelle originali

Cynthia Dwork: Differential Privacy. ICALP (2) 2006: 1-12

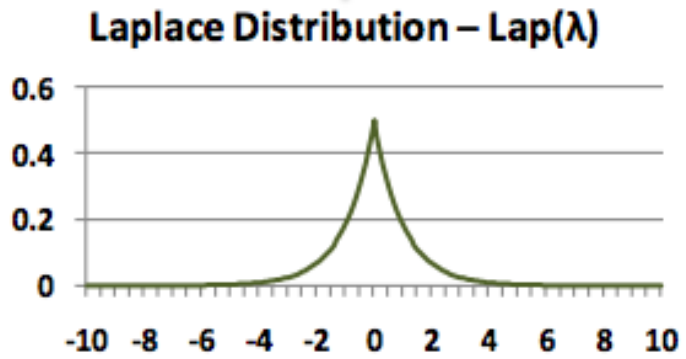
Attacco

Name	Has Diabetes
Alice	yes
Bob	no
Mark	yes
John	yes
Sally	no
Jack	yes

1) Quante persone hanno il diabete? **4**
2) Quante persone, esclusa Alice, hanno il Diabete? **3**

- **Quindi l'attaccante può inferire che Alice ha il diabete**
- **Soluzione**: rendere le due risposte simili

1) La prima risposta diventa 4+1 = 5
2) La seconda risposta diventa 3+2.5=5.5

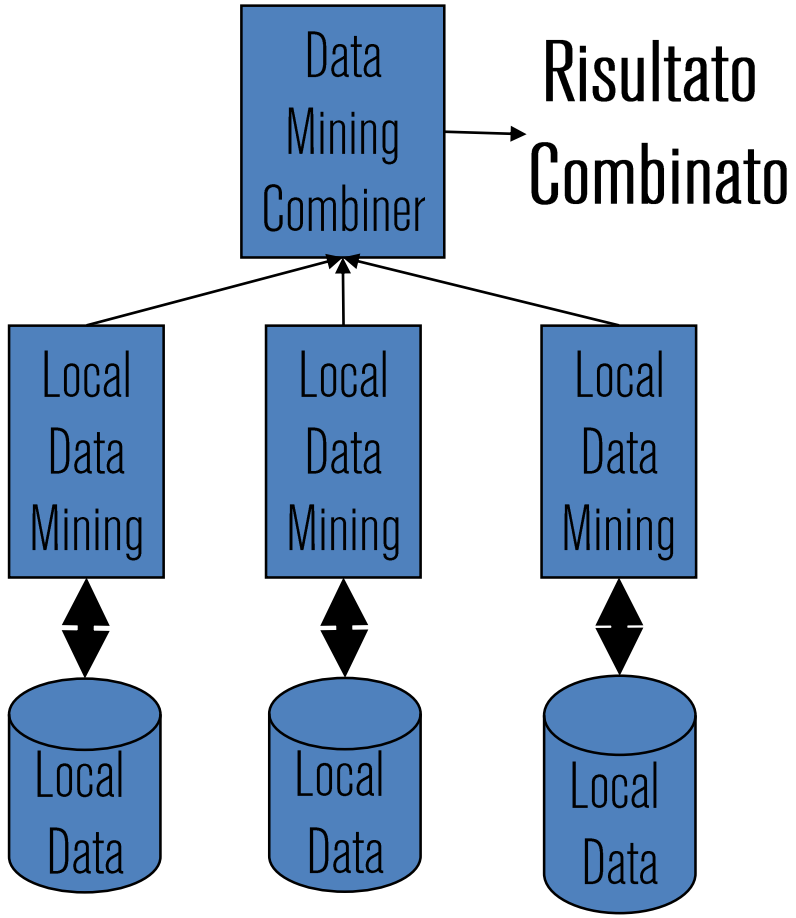


Secure Multyparty Computation

Come calcolare delle funzioni a partire dai dati distribuiti nelle diverse locazioni senza condividere tali dati

Diversi protocolli per il calcolo sicuro

- della somma
- dell'unione tra insieme
- della dimensione dell'intersezione
- del prodotto scalare



La protezione dei dati non riguarda solo dati standard “tabulari”!

Modelli di Privacy per qualsiasi contesto

- Dati di Mobility
  - GPS, Telefonia . . . .
- Dati di acquisto
  - Carte di credito, negozi, ...
- Log delle ricerche
  - Motori di ricerca come Google, Tiscali, ...
- Dati di reti sociali
  - Facebook, Google Plus, Instagram, ...

FAIR

First AId for Responsible data scientist

TRY IT ON:
fair.sobigdata.eu/moodle

The SoBigData online course developed to ensure that people are familiar with the basic elements about ethics, data protection, and intellectual property law

SoBigData

- **Francesca Pratesi**
- Università di Pisa
- Email: pratesi@di.unipi.it