Pseudonymisation

https://bit.ly/2OyWD2u

Cédric Lauradoux

November 22, 2019

Personal data

'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;

How does the data identify the person?

- An identified person can be distinguished from a group of persons.
- **Direct identification** provides the true identity of a person: his/her real name and any additional information that can remove any ambiguity (possible namesake)
- **Indirect identification** can qualify a content or who is performing the identification.

Indirect Identification

- Indirect identification by content is related to the concept of identifiers.
- An identifier is a value that identifies an element within an identification scheme. A unique identifier is associated to only one element or person.
- A quasi-identifier is not by itself a unique identifier but is sufficiently well correlated with an individual. Combine with other quasi-identifiers, they can create a profil (unique identifier)!

Example: quasi-identifiers

- Is your birthday (day+month) an identifier ?
  This is **not a unique identifier** if you consider a group of size greater than 23 (birthday paradox).

- Same question but now for (day+month+year)? This is **not a unique identifier** if you consider the overall population.

- In both cases, it becomes a unique identifier if you consider a small group!

Data

- Personal data $\rightarrow$ GDPR
- Pseudonymised data $\rightarrow$ GDPR recitals
- Anonymous data $\rightarrow$ GDPR recitals
- Anonymised data $\rightarrow$ not in GDPR!
- Encrypted (personal) data $\rightarrow$ not in GDPR!

Why is it like that?

- Pseudonymised and encrypted data are personal data! You MUST apply the GDPR on those data.

- Anonymous and anonymised data are not personal data! You do not need to apply the GDPR on those data.

Pseudonymised data

- ‘pseudonymisation‘ means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;

Pseudonymised data

- The data controller can recover the identity of any subjects using additional information.
- Any third parties can not recover the identity of any subjects because they do not have the additional information.
- Therefore, indirect identification is still possible. **Pseudonymised data are still personal data.**

Anonymous data

‘anonymous data‘ means any information not relating to any identified or identifiable natural person (‘data subject’);

- **They are out of the scope of the GDPR!**

Anonymised data

- **Anonymised data** were personal data which have been processed into anonymous data using an anonymisation function.

- Anonymised data are out of the scope of the GDPR but not anonymisation function because it is a processing of personal data.

Encrypted (personal) data

- Encrypted (personal) data are personal data that have processed by an encryption function with a secret key held by the data controller.

- Indirect identification is still possible if you have the encryption key. **Therefore, encrypted data are still personal data.**

Pseudonymisation

Computer science

- Pseudonymisation is a processing of personal data in which **identifiers** are replaced by **pseudonyms**.

- Recovery is a processing of personal data in which pseudonyms are replaced by the original identifiers. Recovery can only be executed by a **legitimate party** and cannot be executed otherwise.

Example

Identifier	Disease	Date
Alice	Flu	08/02/2019
Bob	Tonsillitis	10/02/2019
Charlie	Flu	11/20/2019
Alice	Gastroenteritis	12/30/2019
Bob	Cholesterol	02/07/2020
Charlie	Allergy	04/17/2020
David	Diabetes	05/26/2020
Bob	Hypertension	05/11/2020

Example

Pseudonym	Disease	Date
13	Flu	08/02/2019
2	Tonsillitis	10/02/2019
25	Flu	11/20/2019
13	Gastroenteritis	12/30/2019
2	Cholesterol	02/07/2020
25	Allergy	04/17/2020
42	Diabetes	05/26/2020
2	Hypertension	05/11/2020

Pseudonymisation

Mathematics

- Pseudonymisation is a **binary relation** $\mathcal{P}$. It is a triplet (A, B, G), with A the set of identifiers, B the set of pseudonyms and G a subset of the Cartesian product $A \times B$ defined as $\{(x, y) | x \in A \text{ and } y \in B\}$. G is called the graph of $\mathcal{P}$.

- Let consider $A = \{\text{Alice}, \text{Bob}, \text{Charlie}\}$ (identifier) and $B = \{1, 2, 3, 4, 5\}$ (pseudonym).

Example

- A pseudonymisation relation $\mathcal{P}$ is defined by:

$$G = \{(Alice, 3), (Alice, 5), (Bob, 2), (Charlie, 1)\}.$$

The graph G of the pseudonymisation relation $\mathcal{P}$ can also be represented by its binary transition matrix $\mathbf{M}$:

$$\mathbf{M} = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{bmatrix} 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{matrix} \text{Alice} \\ \text{Bob} \\ \text{Charlie} \end{matrix}$$

Recovery

- Recovery is the converse binary relation $\mathcal{R} = \mathcal{P}^{-1}$. It is the triplet (B, A, G^{-1}). It is also an **injective function** because:
  - each $b \in B$ is related to at most one element of A.
  - $\forall\ y,\ z \in B$ and $x \in A$ such that $y\mathcal{R}x$ and $z\mathcal{R}x$ $\Rightarrow y = z$.
- The corresponding recovery function $\mathcal{R}$ is defined by:

$$G^{-1} = \{(3, Alice), (5, Alice), (2, Bob), (1, Charlie)\}.$$

Conditions

- **Condition 1.** We must have $|A| \leq |B|$.
- If $|A| \geq |B|$, $x \neq z, y \in B,$ such that $x\mathcal{P}y$ and $z\mathcal{P}y$. This is not pseudonymisation but anonymisation.
- **Condition 2.** A binary relation $\mathcal{P}$ is a pseudonymisation relation if and only if G and $\mathbf{M}$ are secret.
- If you know G, you know G^{-1}. . .

Privacy provisions

- We consider only the pseudonyms!
  We discard any other information.

Pseudonym	Disease	Date
13		
2		
25		
13		
2		
25		
42		
2		

Set reversal

Goal 1

Given B, the adversary can recover A.

- Example: $B = \{2, 13, 25, 42\}$ if the adversary succeeds a set reversal attack, he/she knows:

$$A = \{Alice, Bob, Charlie, David\}.$$

But does not know G!

He/she has reduced the space of possible candidates.

Existential pseudonym reversal

Goal 2

Given a pseudonym $b \in B$, the adversary find $a \in A$ such that $b\mathcal{R}a$.

- The adversary finds that $(42, \mathrm{David})$.
  But he/she has no clue on the other pseudonyms.

Universal pseudonym reversal

Goal 3

$\forall b \in B$, the adversary can find $a \in A$ such that $b \mathcal{R} a$.

- The adversary knows G (or G^{-1}) or $\mathbf{M}$ (or $\mathbf{M}^t$)

$$\mathbf{M} = \begin{array}{c} \begin{matrix} 2 & 13 & 25 & 42 \end{matrix} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \begin{matrix} \\ \text{Alice} \\ \text{Bob} \\ \text{Charlie} \\ \text{David} \end{matrix}$$

Discrimination

Goal 4

Let consider a subset $C \subset A$. Given C and a pseudonym $b \in B$, the adversary can determine if the identifier $a \in A$ such $b\mathcal{R}a$ belongs to C or not.

- $C = \{Alice\}$ and $\bar{C} = \{Bob, Charlie, David\}$. Discrimination

Anonymisation

vs pseudonymisation

- Different techniques than pseudonymisation.
- **Evaluation:** We consider the full database!
  We must be unable to recover the subjects identity!
- Let have a look at a few anonymisation techniques

Anonymisation

Identifier	Disease	Date
13	Flu	08/02/2019
2	Tonsillitis	10/02/2019
25	Flu	11/20/2019
13	Gastroenteritis	12/30/2019
2	Cholesterol	02/07/2020
25	Allergy	04/17/2020
42	Diabetes	05/26/2020
2	Hypertension	05/11/2020

Permutation

Identifier	Disease	Date
13	Tonsillitis	08/02/2019
2	Flu	10/02/2019
25	Hypertension	11/20/2019
13	Gastroenteritis	12/30/2019
2	Cholesterol	02/07/2020
25	Allergy	04/17/2020
42	Diabetes	05/26/2020
2	Flu	05/11/2020

Generalisation and minimisation

Identifier	Disease	Date
13	Short Term	2019
2	Short Term	2019
25	Short Term	2019
13	Short Term	2019
2	Long Term	2020
25	Long Term	2020
42	Long Term	2020
2	Long Term	2020

Adding noise

Identifier	Disease	Date
13	Flu	08/02/2019
2	Tonsillitis	10/02/2019
25	Flu	11/20/2019
13	Gastroenteritis	12/30/2019
2	Cholesterol	02/07/2020
25	Flu	04/17/2020
42	Diabetes	05/20/2020
2	Hypertension	05/11/2020

Systematisation

- Anonymity set, k-anonymity, differential privacy. . .

- **Evaluation (attacks):**
  - Singling-out: extract the records of an individual.
  - Linkability: link the records of a group
  - Inference: deduce new attributes from records

Example

- During WWII, the IJN used the following scheme to protect any messages:
  - ◇ name/locations pseudonymisation,
  - ◇ encryption (using JN-25).
- In 1939, JN-25 was broken by the US Navy. . .
- . . . but they struggle to break the pseudonyms!

Intercepted communication (1943)

ON 18 APRIL CINC COMBINED FLEET VISIT RYZ, R__ AND RXP FOLLOWING SCHEDULE:

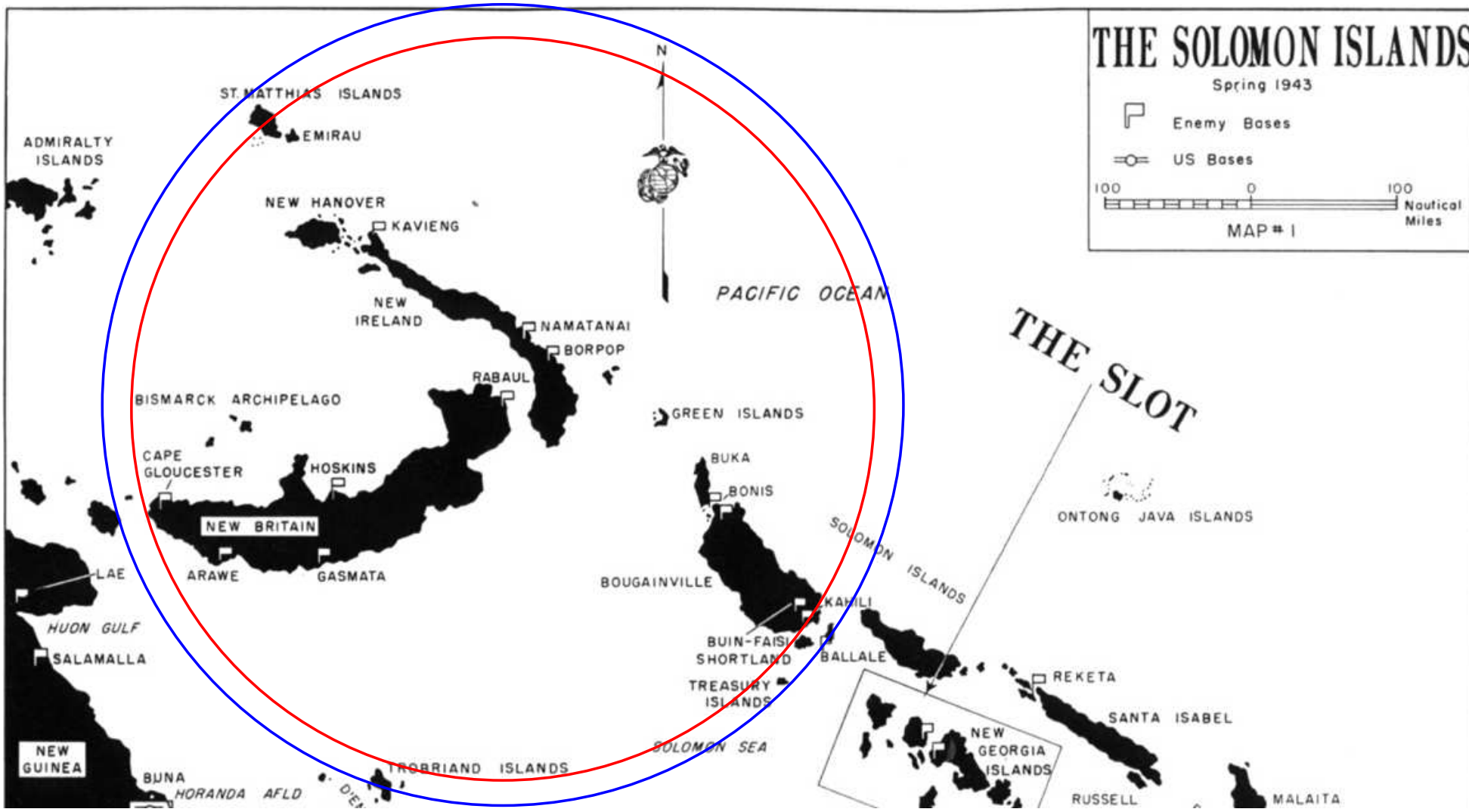
1. DEPART RR AT 0600 IN A MEDIUM ATTACK PLANE ESCORTED BY 6 FIGHTERS. ARRIVE AT RYZ AT 0800. PROCEED BY MINESWEEPERS TO R__ ARRIVING AT 0840. (HAVE MINESWEEPER READY AT #1 BASE.)...

2. AT EACH OF THE ABOVE PLACES THE CINC WILL MAKE SHORT TOUR OF INSPECTION...

Inference attack

- `CINC` = Amiral Isoroku Yamamoto.
- By crossing data, USN analysts get convinced that RR = Rabaul.
- `MEDIUM ATTACK PLANE` = Mitsubishi G4M
  speed: 170 MN/h
- **Duration + Speed → Distance**

Death of Admiral Yamamoto



How pseudonymisation is operated?

- **One-time pseudonymisation:** the adversary can access only one pseudonymised database

- **Many-time pseudonymisation:** the adversary can access multiples pseudonymised databases

Attacks

- **Pseudonym only attack**: the default situation.
  All 4 goals applies.
- **Know identifier attack**:
  Set reversal does not apply.
- **Chosen identifier attack**: the most complicated!
  Set reversal does not apply.

Practical implementation

- **We need to define what is A.**
- **We need to define what is B.**
- **We need to define how we choose $\mathcal{P}$ and $\mathcal{R}$.**

Example

Identifier	Disease	Date
Alice	Flu	08/02/2019
Bob	Tonsillitis	10/02/2019
Charlie	Flu	11/20/2019
Alice	Gastroenteritis	12/30/2019
Bob	Cholesterol	02/07/2020
Charlie	Allergy	04/17/2020
David	Diabetes	05/26/2020
Bob	Hypertension	05/11/2020

Defining the identifier set A

- **Deterministic pseudonymisation**
  $A = \{\text{Alice, Bob, Charlie, David}\}$

- **Randomized pseudonymisation**
  $A = \{\text{Alice, Bob, Charlie, Alice, Bob,}$
  $\text{Charlie, David, Bob}\}$

- How to handle repetitions?

Defining the pseudonym set B

- $B = A$: **set-preserving pseudonymisation**
  Set reversal does not apply

- $B = Id$ with $A \subset Id$:
  format-preserving pseudonymisation

- Otherwise **format-transforming pseudonymisation**

Deterministic pseudonymisation

Implementation

- **Implementation 1:** extract the unique identifiers.
  Complexity: sorting $(n \log_2(n))$

- **Implementation 2:** apply a deterministic function.
  It can be applied on the fly: no complexity!

Arbitrary numbers: counter

Identifier	Pseudonym
Alice	0
Bob	1
Charlie	2
David	3

- Monotonic counter (no repetition)
- Often used by university. . .
- . . . predictable!

Random numbers

Identifier	Pseudonym
Alice	34
Bob	629
Charlie	5
David	17

- Be careful collision can occur!
  Birthday paradox.
- **Unpredictable!**

Cryptographic hash functions

- A cryptographic hash function is defined by:

$$H : \{0,1\}^\star \to \{0,1\}^d$$

- **Properties:**
  - Resistant to collision;
  - Resistant to pre-image.
- **Example:** MD5, SHA1, SHA256, SHA3.
- Anybody can compute a pseudonym from an identifier.

Authentication codes

- It can be viewed as a keyed hash function:

$$H : \{0,1\}^k \times \{0,1\}^\star \to \{0,1\}^d$$

  There is now a secret key K!

- **Example:** HMAC-SHA256, AES-CBC-MAC, SHA3.

- You need to know K and the identifier to compute a pseudonym.

Deterministic encryption

- It can be viewed as a keyed hash function:

$$E : \{0,1\}^k \times \{0,1\}^m \to \{0,1\}^m$$

$$D : \{0,1\}^k \times \{0,1\}^m \to \{0,1\}^m$$

$$\mathrm{Id} = D \circ E$$

  There is now a secret key K!

- **Example:** AES-ECB-128, AES-ECB-256, RSA.

- You need to know K and the identifier to compute a pseudonym.

Evaluation

One-time pseudonymisation

- Pseudonym-only attack

	Goal 1	Goal 2	Goal 3	Goal 4	G^{-1}
Counter	✓	✓	✓	✗	✗
Random numb.	✓	✓	✓	✓	✗
Hash function	✗	✗	✗	✗	✗
Auth. codes	✓	✓	✓	✓	✗
Det. encrypt.	✓	✓	✓	✓	✓

Evaluation

One-time pseudonymisation

- Known identifier attack

	Goal 1	Goal 2	Goal 3	Goal 4	G^{-1}
Counter	n/a	✗	✗	✗	✗
Random numb.	n/a	✓	✓	✓	✗
Hash function	n/a	✗	✗	✗	✗
Auth. codes	n/a	✓	✓	✓	✗
Det. encrypt.	n/a	✓	✓	✓	✓

Evaluation

One-time pseudonymisation

- Chosen identifier attack

	Goal 1	Goal 2	Goal 3	Goal 4	G^{-1}
Counter	n/a	✗	✗	✗	✗
Random numb.	n/a	✓	✓	✓	✗
Hash function	n/a	✗	✗	✗	✗
Auth. codes	n/a	✓	✓	✓	✗
Det. encrypt.	n/a	✓	✓	✓	✓

Evaluation

Many-time pseudonymisation

- Known and chosen identifier attack

	Goal 1	Goal 2	Goal 3	Goal 4	G^{-1}
Counter	n/a	✗	✗	✗	✗
Random numb.	n/a	✗	✓	✗	✗
Hash function	n/a	✗	✗	✗	✗
Auth. codes	n/a	✗	✓	✗	✗
Det. encrypt.	n/a	✗	✓	✗	✓

Randomized pseudonymisation

- It is possible to use randomized encryption:
  - AES-CTR-128, AES-CBC-128. . .
  - Elgamal, Paillier. . .

- How to transform deterministic pseudonymisation into randomized pseudonymisation ?
  - change the key (encryption+ auth. codes only),
  - cascading,
  - salting.

Salting

Identifier	Disease	Date
Alice	Flu	08/02/2019
Bob	Tonsillitis	10/02/2019
Charlie	Flu	11/20/2019
Alice	Gastroenteritis	12/30/2019
Bob	Cholesterol	02/07/2020
Charlie	Allergy	04/17/2020
David	Diabetes	05/26/2020
Bob	Hypertension	05/11/2020

Salting

Identifier	Disease	Date
1,Alice	Flu	08/02/2019
2,Bob	Tonsillitis	10/02/2019
3,Charlie	Flu	11/20/2019
4,Alice	Gastroenteritis	12/30/2019
5,Bob	Cholesterol	02/07/2020
6,Charlie	Allergy	04/17/2020
7,David	Diabetes	05/26/2020
8,Bob	Hypertension	05/11/2020

- Deterministic pseudonymisation on salt,identifier.

Salting

Identifier	Disease	Date
3	Flu	08/02/2019
7	Tonsillitis	10/02/2019
56	Flu	11/20/2019
19	Gastroenteritis	12/30/2019
67	Cholesterol	02/07/2020
42	Allergy	04/17/2020
12	Diabetes	05/26/2020
99	Hypertension	05/11/2020

Evaluation

Many-time pseudonymisation

- Known and chosen identifier attack

	Goal 1	Goal 2	Goal 3	Goal 4	G^{-1}
Counter	n/a	✗	✗	✗	✗
Random numb.	n/a	✓	✓	✓	✗
Hash function	n/a	✗	✗	✗	✗
Auth. codes	n/a	✓	✓	✓	✗
Det. encrypt.	n/a	✓	✓	✓	✓

Conclusion

- Best solutions are encryption or authentication codes.
- **Nothing is better or better than nothing?**
- It is only one step toward protecting privacy!