Security of Personal Data Processing Event

Data Protection by design for SMEs

Pseudonymisation and encryption

Dr. Konstantinos Limniotis

Hellenic Data Protection Authority

klimniotis at dpa.gr

Athens, October 8th , 2018

Overview

- Pseudonymisation and encryption as data protection mechanisms
  - The notion of (non)-anonymous data
  - Pseudonymous vs. anonymous data
  - Pseudonymous vs. encrypted data
  - Examples - Discussion

Introduction

- Data minimization issues
  - Are the actual identities of the individuals needed for the processing?
    - E.g. data anonymisation for research purposes
    - Note that the notion of users identifiers is quite broad
  - Even if no (direct) identification is needed, is there any possibility that in some cases re-identification will be necessary?
  - Even if no (direct) identification is needed, is tracking of the individuals needed?
    - E.g. gps anonymous services

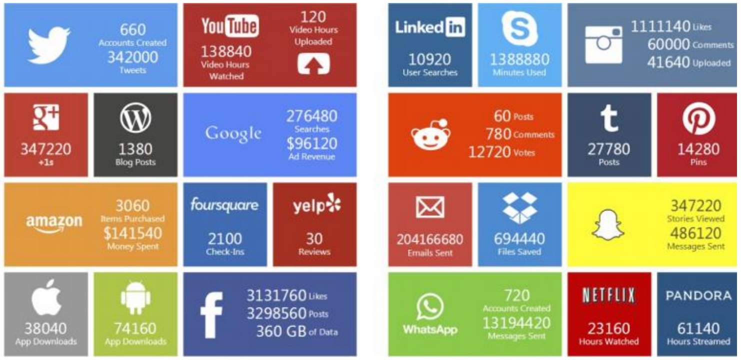
- Data Security issues
  - Is the confidentiality of personal data ensured?
  - Are the identities of the users properly “hidden”?

- Which are the roles of pseudonymisation and encryption?

Fallacies on anonymity

- Wrong assumption: If the users identities are not “obvious”, then the data are anonymous

- The right assumption: Even if the identities are not obvious, one could possibly reveal (some of) them via appropriately utilizing other background knowledge
- Especially the “big data era” raises several privacy concerns



The “wide” notion of identifiers

- Not only direct identifiers, but also quasi-identifiers could possibly allow (depending on the context) identification

~~Identifier~~	Quasi-identifiers			Sensitive attribute
~~Name~~	DOB	Gender	Zipcode	Disease
~~Andre~~	11/1/76	Male	53715	Heart Disease
~~Beth~~	13/4/86	Female	53715	Hepatitis
~~Carol~~	28/2/76	Male	53703	Brochitis
~~Dan~~	21/1/76	Male	53703	Broken Arm
~~Ellen~~	13/4/86	Female	53706	Flu
~~Eric~~	28/2/76	Female	53706	Hang Nail

- There is a unique triplet {5-digit ZIP, gender, date of birth} for the 87% of the citizens in U.S.A [Sweeney, 2002].
- Anonymisation techniques do exist, to alleviate such privacy concerns

Personal and anonymous data

Definitions (General Data Protection Regulation - GDPR)

- The term "personal data" refers to any information relating to an identified or identifiable natural person
- The data protection principles do not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person
  - However, to determine whether a natural person is identifiable, account should be taken **of all the means reasonably likely to be used**, such as singling out, by any person to identify – directly or indirectly – the natural person
  - To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.
- *In simple words, we should be very careful when charactering data as anonymous data*
  - *Have we thoroughly examined whether identification is practically fully impossible?*

Pseudonymous data

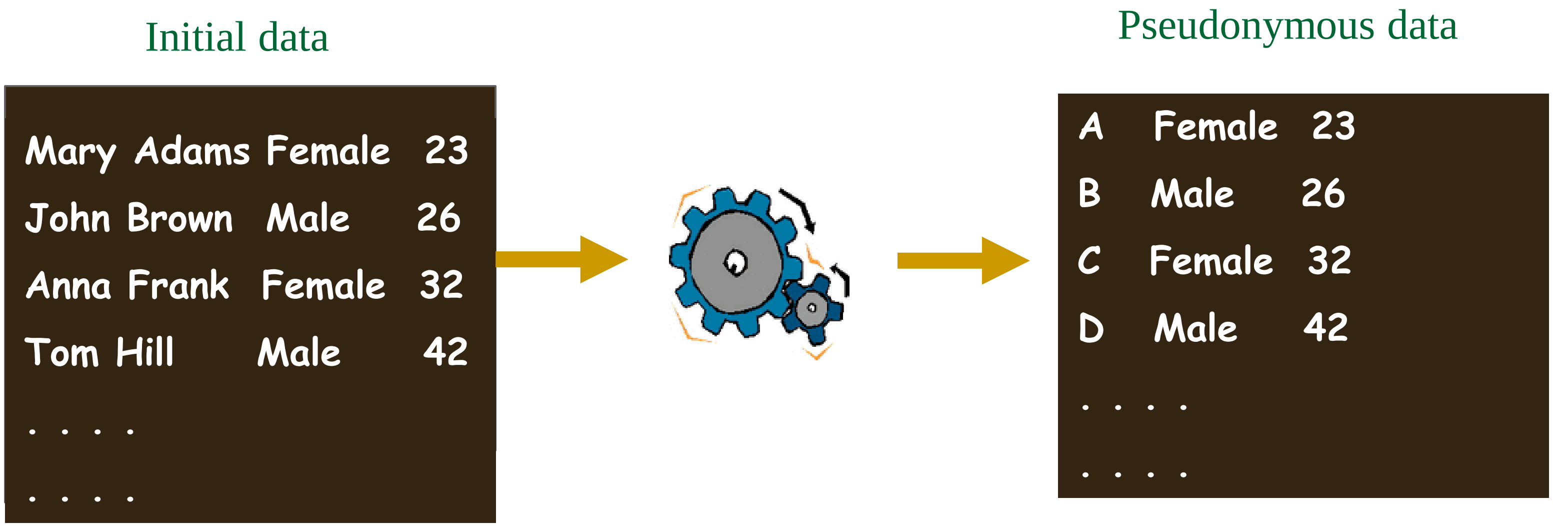
Definitions (General Data Protection Regulation - GDPR)

- "Pseudonymisation" means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific person 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

- Personal data which have undergone pseudonymisation should be considered to be information on an identifiable natural person.
  - That is pseudonymization does not result in anonymous data and, thus, personal data protection principles apply to pseudonymized data

- Pseudonymization should be considered as a potential safeguard to reduce the data protection risks and not as anonymization
  - **Probably necessary, in some cases, to ensure proportionality of the processing**

Pseudonymisation in a simplified form



Initial data

Mary Adams	Female	23
John Brown	Male	26
Anna Frank	Female	32
Tom Hill	Male	42
. . . .		
. . . .		

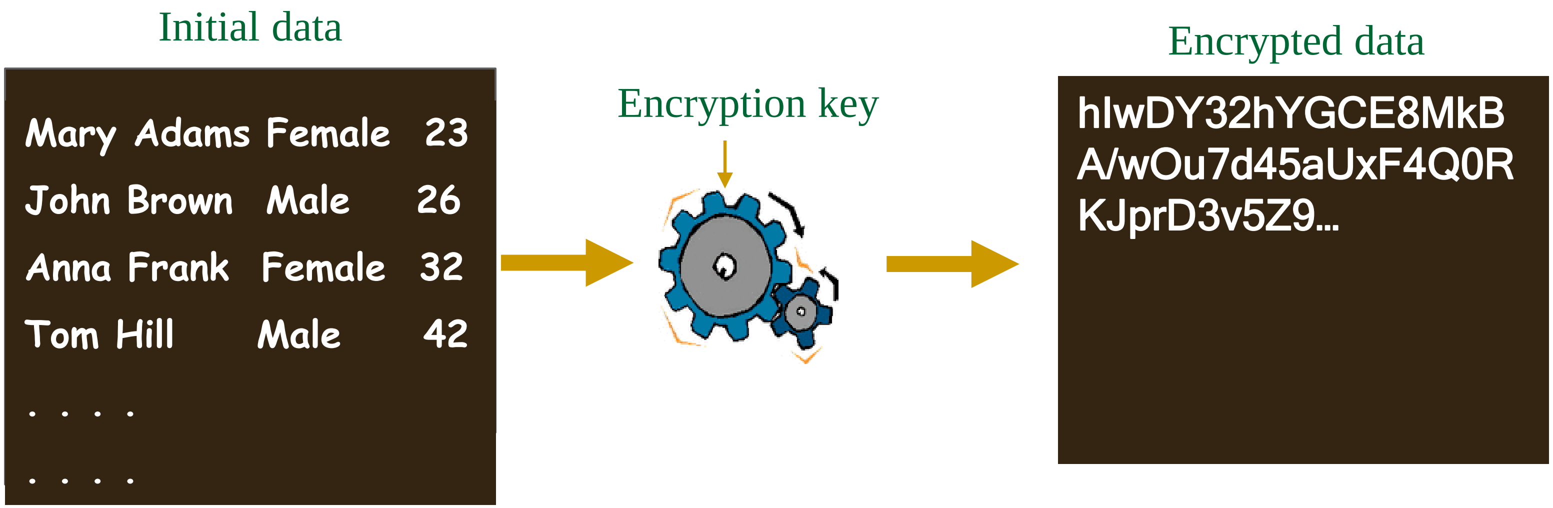
Pseudonymous data

A	Female	23
B	Male	26
C	Female	32
D	Male	42
. . . .		
. . . .		

- The associations between identifiers (Mary Adams, John Brown, …) with their pseudonyms (A, B, …) should be somehow “protected”
  - The pseudonyms can be generated by taking into account more than one users attributes (identifiers / quasi-identifiers)
- Depending on the scope, the purpose of the processing and the relevant risks, pseudonymisation may also necessitate the implementation of “anonymization techniques”

Pseudonymisation vs. encryption

A typical encryption scheme



Initial data

Mary Adams	Female	23
John Brown	Male	26
Anna Frank	Female	32
Tom Hill	Male	42
. . . .		
. . . .		

Encryption key

Encrypted data

hIwDY32hYGCE8MkB
A/wOu7d45aUxF4Q0R
KJprD3v5Z9…

- Encryption is the main instrument to achieve confidentiality
  - Other security services are also addressed (data integrity, entity authentication etc.)
- Only those acquiring the decryption key are able to recover the initial data from the encrypted
- So encryption is clearly different from pseudonymisation
  - Pseudonymous data can be also encrypted
- However, cryptographic techniques can be used in deriving pseudonyms

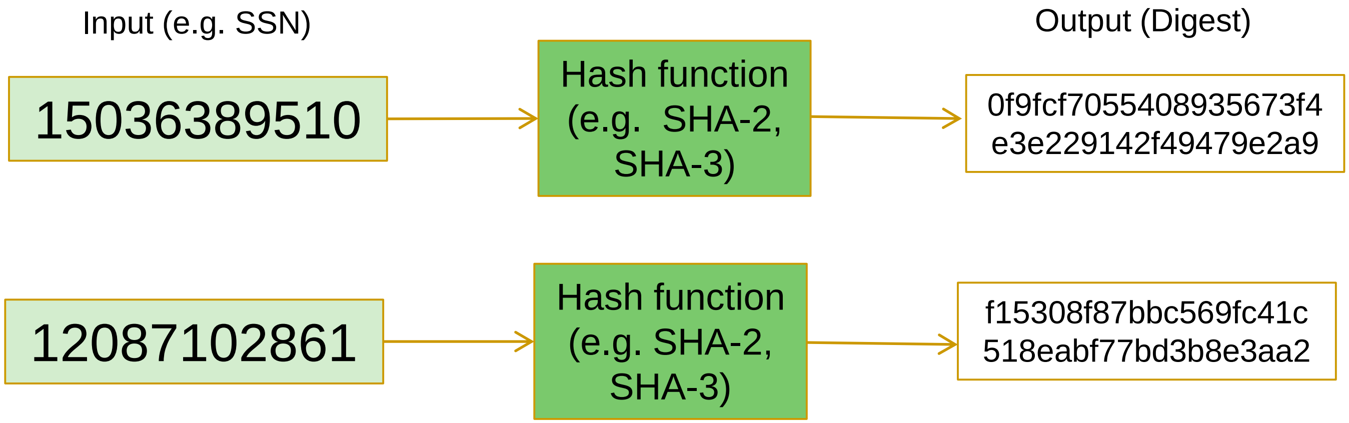
Why pseudonymisation?

- The GDPR makes about 15 references to pseudonymisation
  - Possible appropriate safeguard for:
    - *"purpose limitation balancing test" (art. 6, par. 4)*
    - *Data protection by design and by default (art. 25)*
    - *Security of processing (art. 32)*
    - *Processing of personal data for public interest, scientific or historical research purposes or statistical purposes (art. 89)*

- Pseudonymisation is also implied in several other places within GDPR
  - When the controller is able to demonstrate that is not in a position to identify the individual (data subject), Art. 15-20 shall not apply – i.e. right of access, right to rectification/erasure/restriction/portability *(art. 11)*
    - *Unless the data subject provides additional information enabling his/her identification*
  - Appropriately-implemented pseudonymisation can reduce the likelihood of individuals being identified in the event of a personal data breach

Some pitfalls in pseudonymisation

- A cryptographic hash function is a mathematically irreversible function, practically enabling “1-1” mapping between inputs and outputs (digests)



Input (e.g. SSN)
Output (Digest)
15036389510
Hash function (e.g. SHA-2, SHA-3)
0f9fcf7055408935673f4 e3e229142f49479e2a9
12087102861
Hash function (e.g. SHA-2, SHA-3)
f15308f87bbc569fc41c 518eabf77bd3b8e3aa2

- Many data controllers tend to use such an approach e.g. when tracking of users, but no identification, is needed in the context of the data processing
  - However, if the hash function is applied to an identifier that can be available via other sources, then re-identification is possible
    - Simply compute the hash value for the identifier and compare the result with the pseudonymised list

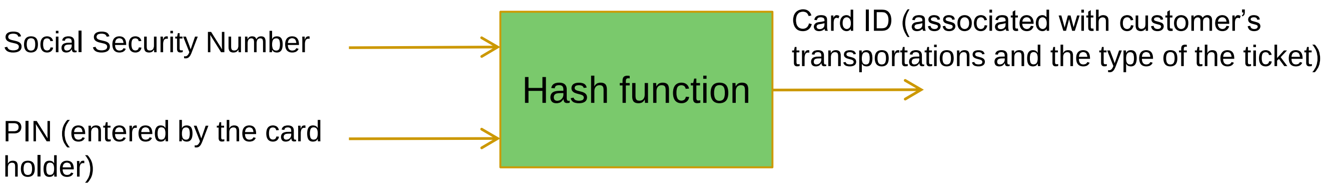
The case of e-ticketing system for public transports in Athens

- The initial inquiry of the Organisation of Domestic Transport in Athens (OASA) to the Hellenic DPA described a processing system which did not satisfy the data protection by design principle
  - To achieve all the desired purposes, OASA would store such information allowing to gain personalized information for a large proportion of the passengers
    - E.g. John Brown entered the metro station in Sintagma square at 8:00 at 8/10/2018 and arrived at Omonia square at 8:09 at 8/10/2018
  - Not proportionate with respect to the prescribed goals of the system

- The Hellenic DPA asked for an appropriate re-designing of the process (Opinion 1/2017)

The case of e-ticketing system for public transports in Athens (*cont.*)

- OASA adopted a system in which a pseudonymisation approach is being used



Social Security Number

PIN (entered by the card holder)

Hash function

Card ID (associated with customer's transportations and the type of the ticket)

- OASA, as well as any other party getting access to the card ID, will not be able to identify the user
  - Essential property for protecting privacy in transportations, since each transportation is associated with this ID
- Once the user looses his card, she/he will be able to prove that this specific card ID corresponds to her/him
- => Opinion 4/2017 of the Hellenic DPA

Conclusions - Discussion

- Adoption of data protection by design principle
  - Data should be relevant and limited to what is necessary in relation to the purposes for which they are processed
    - Decision to be made at an early stage of the design
  - Implement appropriate mechanisms
    - Pseudonymisation – Anonymisation
    - Encryption may have its own role, apart from ensuring data confidentiality

- See also:
  - Article 29 Working Party, *Opinion 05/2014 on Anonymisation Techniques*
  - ENISA, *Privacy and Data Protection by Design*, 2015.
  - ENISA, *Privacy by Design in Big Data*, 2015.
  - ENISA, Privacy and data protection in mobile applications, 2018.