Cryptography at the service of pseudonymisation

Konstantinos Limniotis

ICT Specialist, HDPA

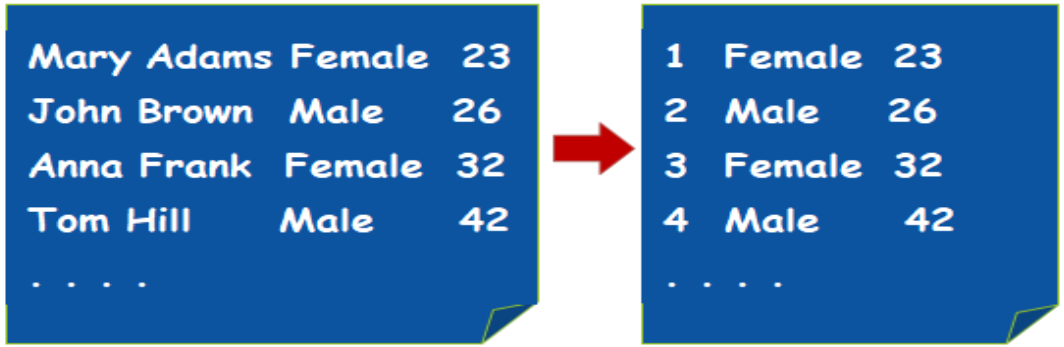
Adjunct Faculty Member – University of Athens & Open University of Cyprus

Overview

- The notion of pseudonymisation
  - Engineering and legal perspectives
- “Classical” cryptographic techniques as pseudonymisation tools
- Advanced cryptographic techniques as pseudonymisation tools
  - Indicative examples:
    - User-generated pseudonyms
    - Private Set Intersection
    - Secret sharing
- Conclusions

Pseudonymisation: Engineering perspective

- A pseudonym is defined as an identifier of a subject, which is different from the subject's *"real name"*
  - *More generally, the pseudonym replaces a data subject's identifier (i.e. an identifier allowing to explicitly identify the data subject within a specific context)*

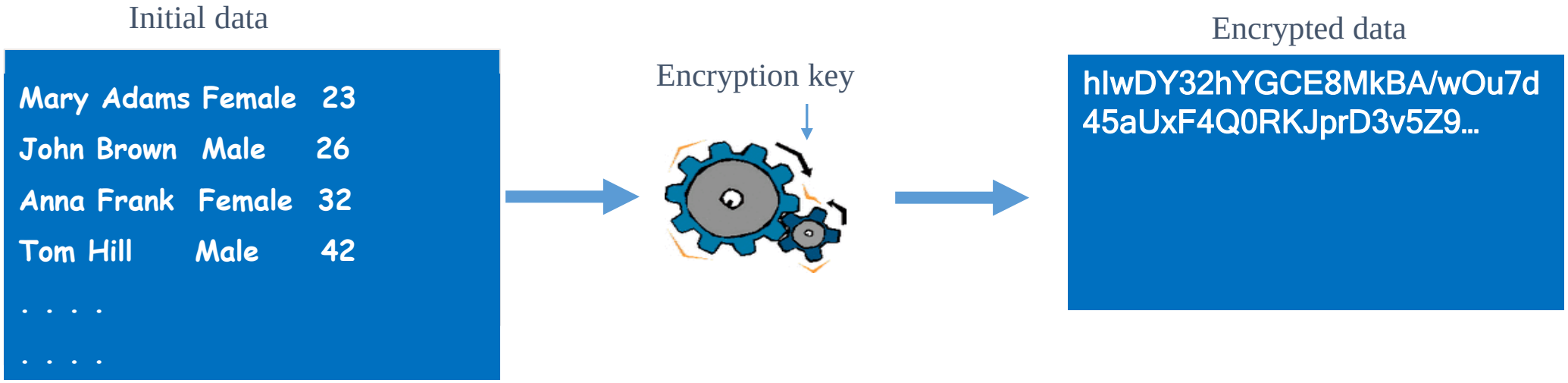


- See also, e.g., ISO 25237:2017 – *"Pseudonymisation is a particular type of de-identification that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms"*

Pseudonymisation: Legal perspective

- In the 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*".
- Pseudonymous data are personal (and not anonymous) data…
  - There exist additional data, being protected, that allow re-identification (in some context)

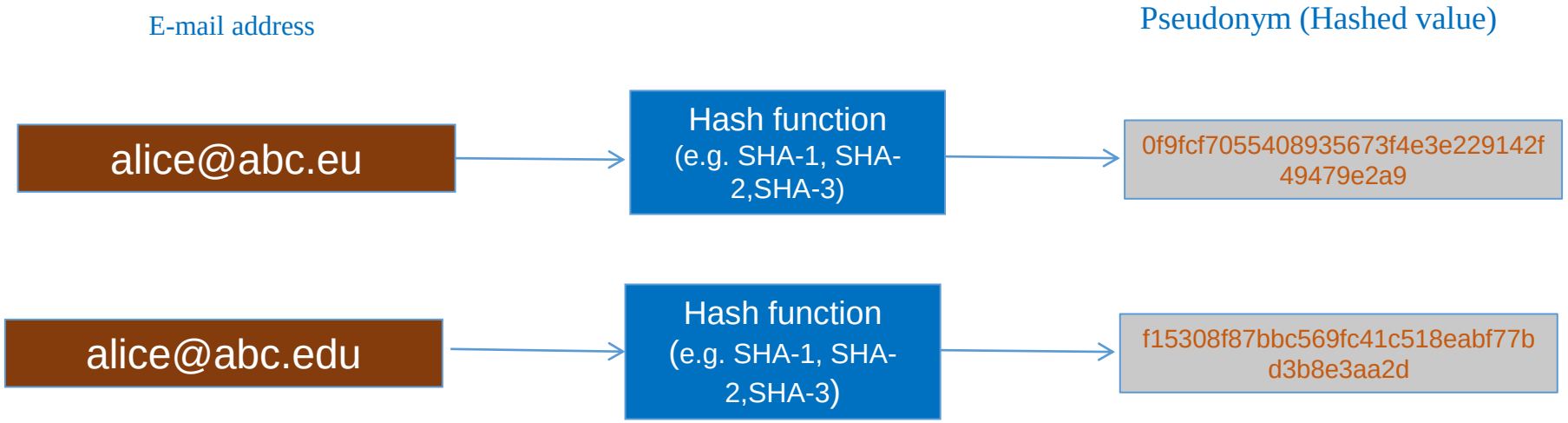
Pseudonymisation ≠ Encryption



- Encrypted data are unintelligible to anyone not having the decryption key (which inverses the encryption)
  - Not even (???) statistical analysis can be performed on encrypted data
    - This is general not the case in pseudonymisation
- Hence, the difference between pseudonymisation and encryption is obvious
  - However, appropriate use of cryptography may give rise to "good" pseudonymisation techniques…
  - The secret key could coincide with the "additional information needed for re-identification"

Pseudonymisation techniques based on (classical) cryptography

1. Cryptographic hash function



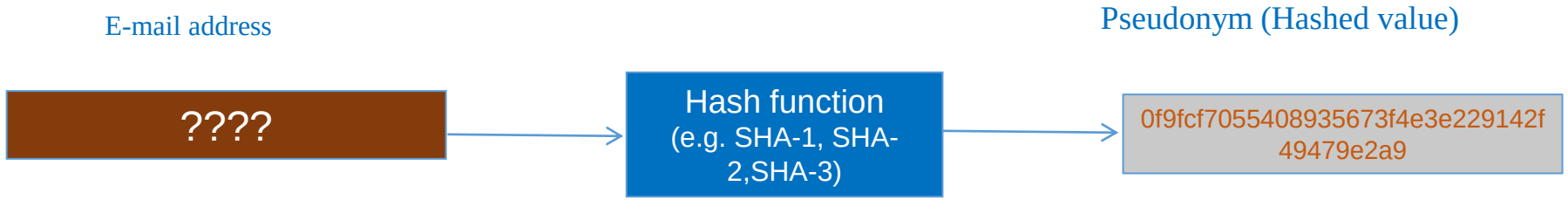
Properties of a hash function:

- Deterministic pseudonymisation (The same input always yields the same output)
- The output does not provide any information on the input
  - Mathematically irreversible (there is no reverse hashing)

People believe that hashing is a nice pseudonymisation technique. **But…**

Pseudonymisation techniques based on (classical) cryptography

1. Cryptographic hash function (Cont.)



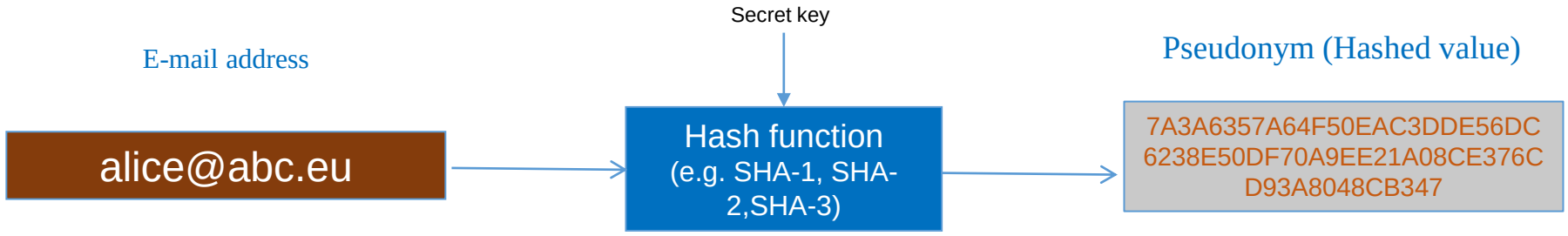
The “adversary” can easily verify whether any of the pseudonyms in the pseudonymised list corresponds to, e.g., alice@abc.eu

- Simply computes the hashed value of alice@abc.eu and checks…

- The size and the «predictability» of the input domain highly affects the level of protection (identity hiding) that a hash function provides as a pseudonymisation technique

Pseudonymisation techniques based on (classical) cryptography

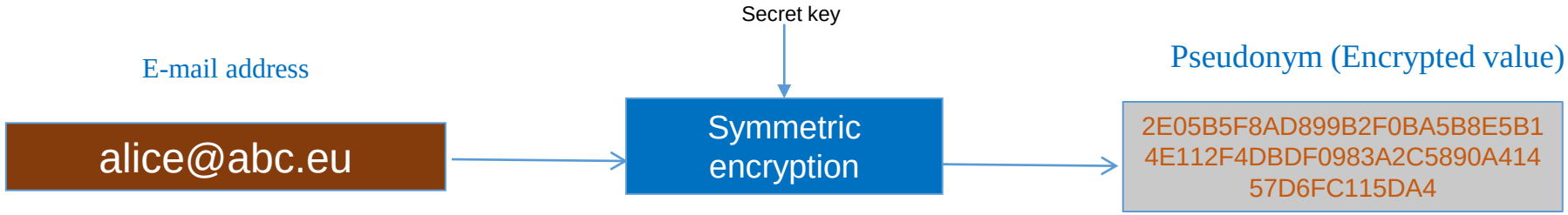
2. Cryptographic hash function with key



- Deterministic or randomised pseudonymisation, based on whether the secret key is fixed or not
- Knowledge of the pseudonym and the secret key does not allow direct estimation of the initial identifier
  - However, given an identifier and the secret key, it can be easily checked which is its corresponding pseudonym
- The key may be considered as *the additional information that allow re-identification* and should be secured.

Pseudonymisation techniques based on (classical) cryptography

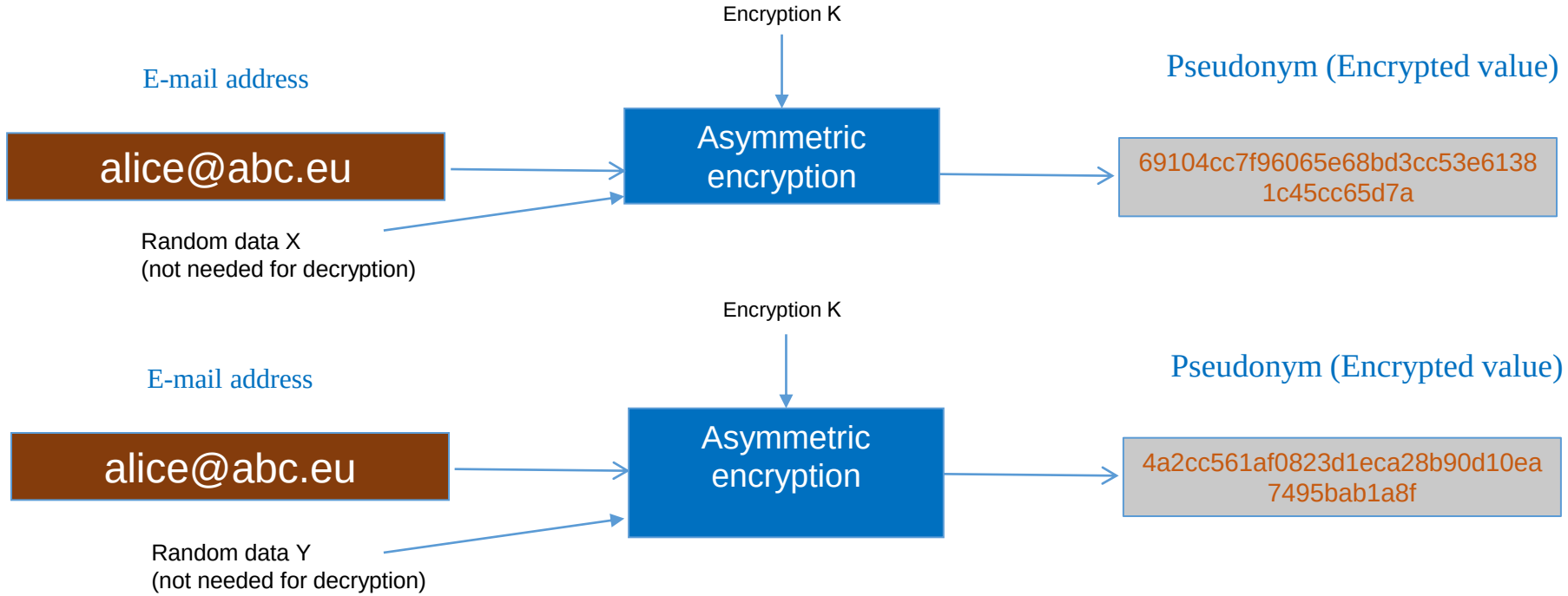
3. Encryption – the deterministic case



- Deterministic pseudonymisation, for fixed secret key (and no other randomisation in the input)
- Knowledge of the pseudonym and the pseudonymisation secret allows direct estimation of the initial identifier
- The secret key for encryption/decryption may be considered as *the additional information that allow re-identification* and should be secured.

Pseudonymisation techniques based on (classical) cryptography

4. Encryption – the probabilistic case

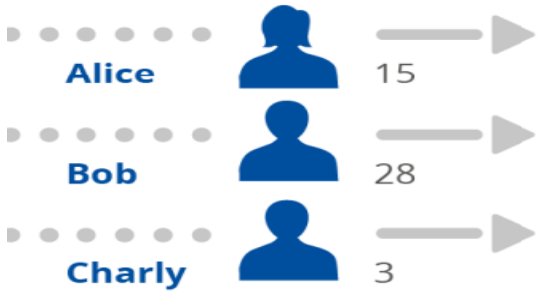


- Randomised pseudonymisation (different pseudonyms for the same identifier and the same encryption key)
- The decryption key (different from the encryption key, in the asymmetric encryption) may be considered as *the additional information that allow re-identification* and should be secured.

Results so far…

- Classical cryptographic techniques provide the means for transforming “identifiers” into “pseudonyms”
- The proper pseudoymisation approach depends on the specific scenario and needs
  - On a risk-based approach
- Finding out which is the proper approach, is not always an easy task
    - See also ENISA reports on pseudonymisation, 2018-2021
- Advanced cryptographic techniques may be prerequisite to address specific data protection challenges

Special case: User-generated pseudonyms



Organisation (e.g. data controller)

Free figure from www.freepik.com

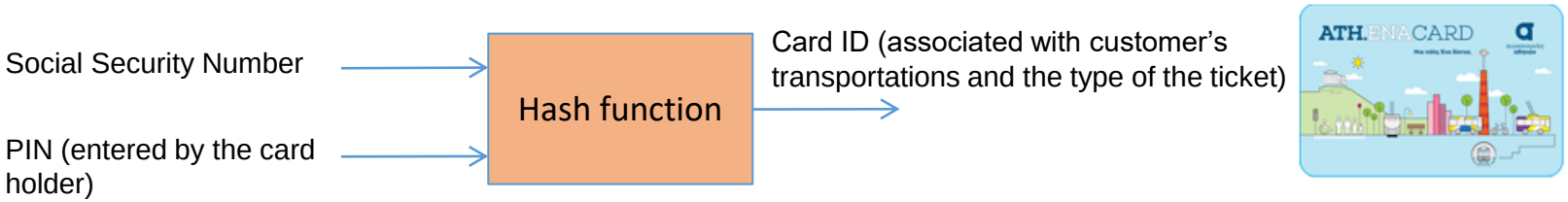
- Pseudonyms are being produced in a decentralized approach, in which the users actively participate in the generation of their pseudonyms
- The additional information needed to attribute a pseudonym (e.g. “15”) to an original user’s identifier (e.g. “Alice”) is under the control of the user
  - Enhancing user’s trust in the processing
- In several scenarios, such a property may be prerequisite to ensure the principle of data minimisation

Use case: e-ticketing system for public transports in Athens

- 2017: Initial inquiry of the Organisation of Domestic Transport in Athens (OASA) to the Hellenic DPA describing a processing system to support e-ticket services
  - Before the GDPR enters into force….
- To achieve all the desired (legal) 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-design of the process (Opinion 1/2017)
- With the GDPR terminology, the data protection by design principle was not present

Use case: e-ticketing system for public transports in Athens (*Cont.*)

- OASA finally adopted a system in which a user-based pseudonymisation approach is being used (HDPA Opinion 4/2017)



- The data controller (OASA), as well as any other party getting access to the card ID, will not be able to reverse it into the Social Security Number or into any other user's identifier
  - 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 – i.e. to prove ownership of the card ID (pseudonym)

Several research approaches

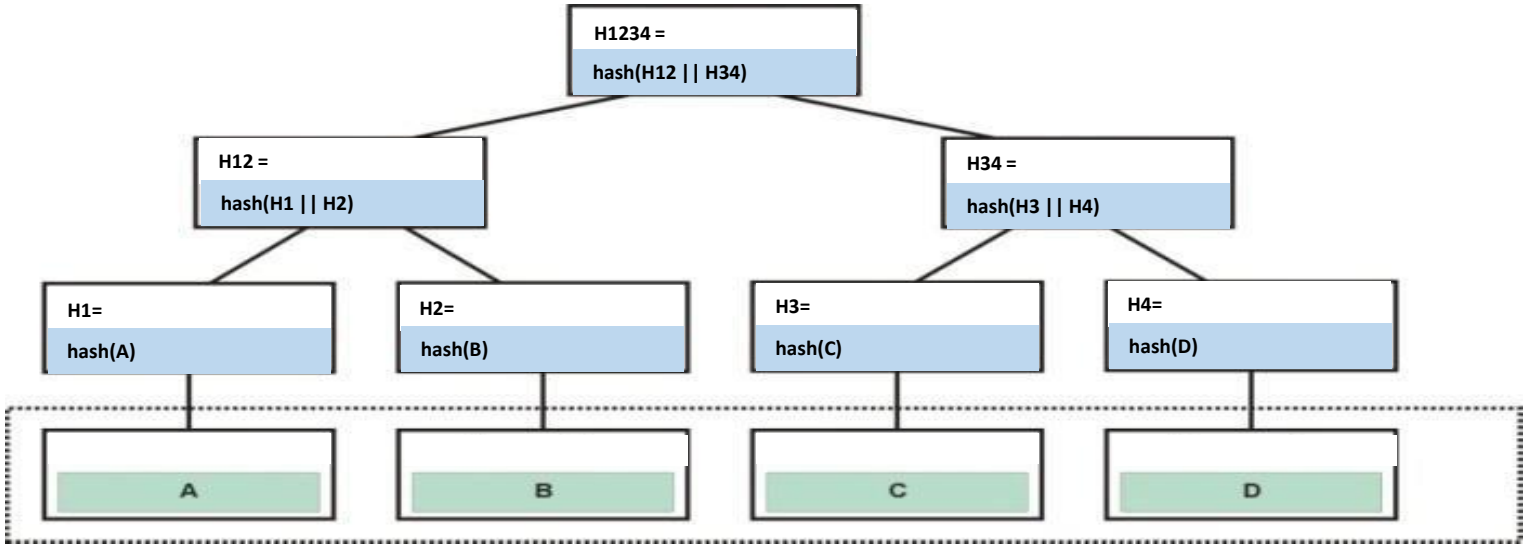
- Schartner et. al - 2005
- Lehnhardt et. al. – 2011
- Tunaru et. al. – 2015

Based on classical asymmetric cryptographic algorithms

- The main design are the following for user-generated pseudonyms (Lehnhardt et. al. – 2011):
  - "Hiding identities" - Linking a pseudonym to its owning user should not be possible for any other than the user herself, unless it is explicitly permitted
  - "Unlinkability" - In cases that users may have multiple pseudonyms, it should not be possible to identify different pseudonyms as belonging to the same user,
  - Injectivity - the pseudonym generation process should avoid duplicates
  - Flexibility - possible to add new pseudonyms to the user entities with minimal effort.
  - Ease of use

A recent approach (in Proc. of APF 2021)

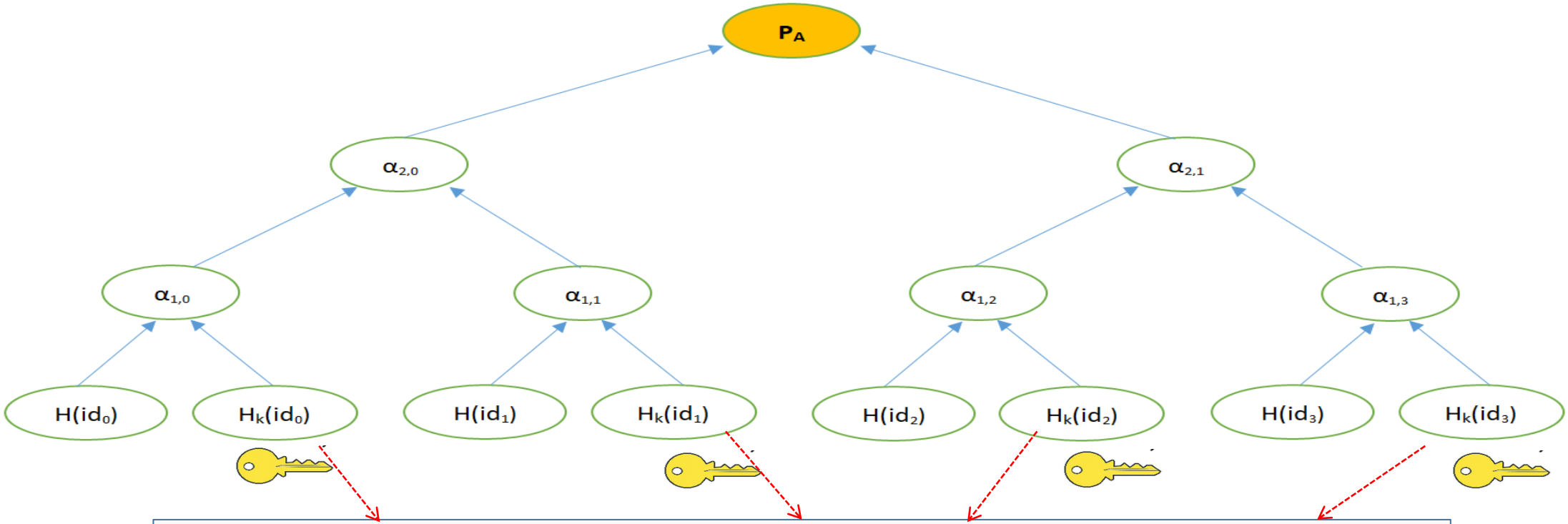
- Usage of cryptographic accumulators for pseudonym generation
  - Data structures allowing set membership operations
  - They allow to accumulate a finite set of values {x1, x2, . . . , xn} into a succinct value
- Focus on the Merkle trees (see also the recent ENISA report, 2021)



- Hash functions are much more efficient than public key encryptions and, moreover, post-quantum secure

Merkle trees as pseudonymisation technique – The main idea

Example: Derivation of a pseudonym **P_A** based on four identifiers of the user A:

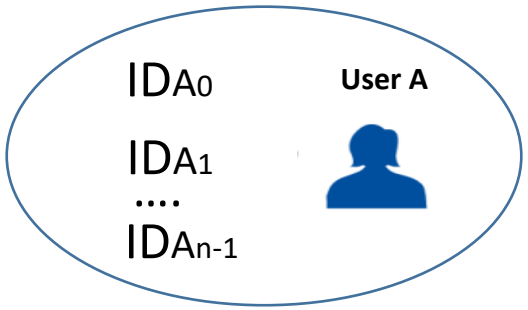


- **H:** Cryptographic hash function
- **Hk** : Keyed-hash function
- The same key k for all the leafs - known only to the user A (chosen by her)
- For another pseudonym, based on the same identifiers, simply change k

Properties of the Merkle tree-based pseudonymisation scheme

- Scenario: User A with some domain-specific identifiers **ID_{A0}**, **ID_{A1}**, … **ID_{An-1}** for n organisations *Org_0*, *Org_1*, … *Org_{n-1}*

The IDs could be of any form depending on the context, such as *Social Security Number, VAT Registration Number, e-mail address, device ID etc.*



Free figures from www.freepik.com

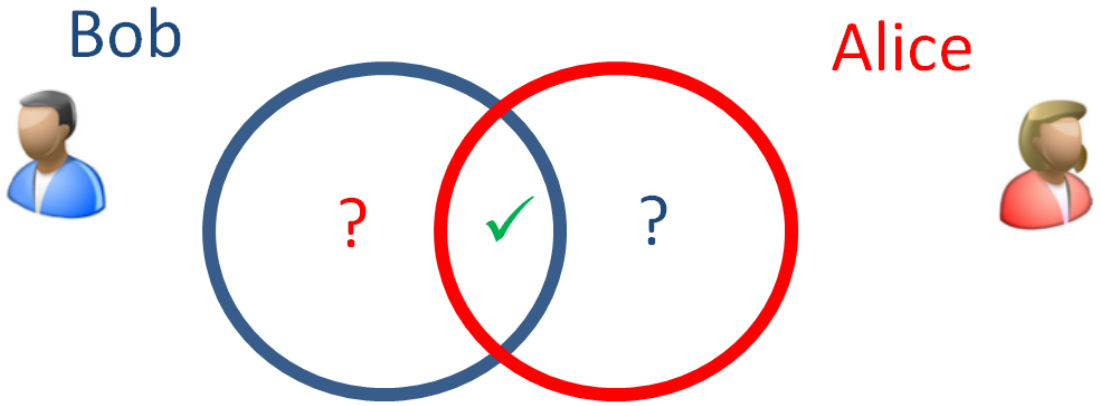
A ↔IDA0 **A** ↔IDA1 **A** ↔IDAn-1

User-generated pseudonyms: $PA^{(0)}$ $PA^{(1)}$ …. $PA^{(n-1)}$

- Each organisation Org_i can verify that the pseudonym $P_A^{(i)}$ stems indeed from the user with identifier ID_{Ai}
- **Important property**: The user A can prove to, e.g., Org_1 , that she has the pseudonym $PA^{(0)}$ in Org_0
  - Without revealing any other information on other identifiers of her (e.g. on ID_{A0})
- Prerequisite: The initial registration of each $PA^{(i)}$ to Org_i must be authenticated

- **Possible application area: Facilitate exchange of information between data controllers, with respect to data minimisation**
  - **Upon the user's request**

Private set intersection



- Problem: Find the common entries between two sets with personal data, without revealing anything more
- Naïve approach: Exchange hashed values and compare
  - Recall the weaknesses of hashes in terms of pseudonymisation….

Private set intersection (Cont.)

- Several protocols exist for efficient solution
- A simple approach, based on the classical Diffie-Hellman protocol:

$\{x_1, x_2, \ldots, x_n\}$

$\{y_1, y_2, \ldots, y_m\}$

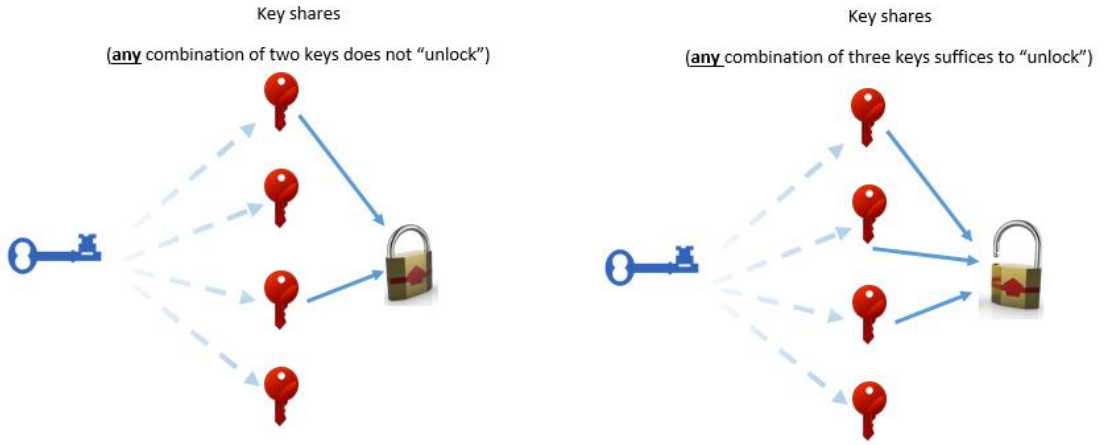
1. Choose random a
2. $H(x_1)^a, H(x_2)^a, \ldots, H(x_n)^a$
3. Exchange values
4. Compute $H(y_1)^{ba}, H(y_2)^{ba}, \ldots, H(y_m)^{ba}$
5. Compare each $H(y_i)^{ba}$ with each $H(x_j)^{ab}$

1. Choose random b
2. $H(y_1)^b, H(y_2)^b, \ldots, H(y_m)^b$
3. Exchange values
4. Compute $H(x_1)^{ab}, H(x_2)^{ab}, \ldots, H(x_n)^{ab}$
5. Send the values (with the same order)

- More efficient approaches exist, based on oblivious transfer protocols (see, e.g. Pinkas et. al. – 2018)
- Note that the values exchanged are fully compatible with the GDPR's notion on pseudonymisation!
  - The additional information needed for re-indentification is indeed protected….

Secret sharing

- The basic idea: “Split” a secret *s* into *n* shares
  - Re-construction of *s* is possible only if any *t + 1* parties exchange their information, but no less
- First proposed by a pioneering work from Shamir (1979)

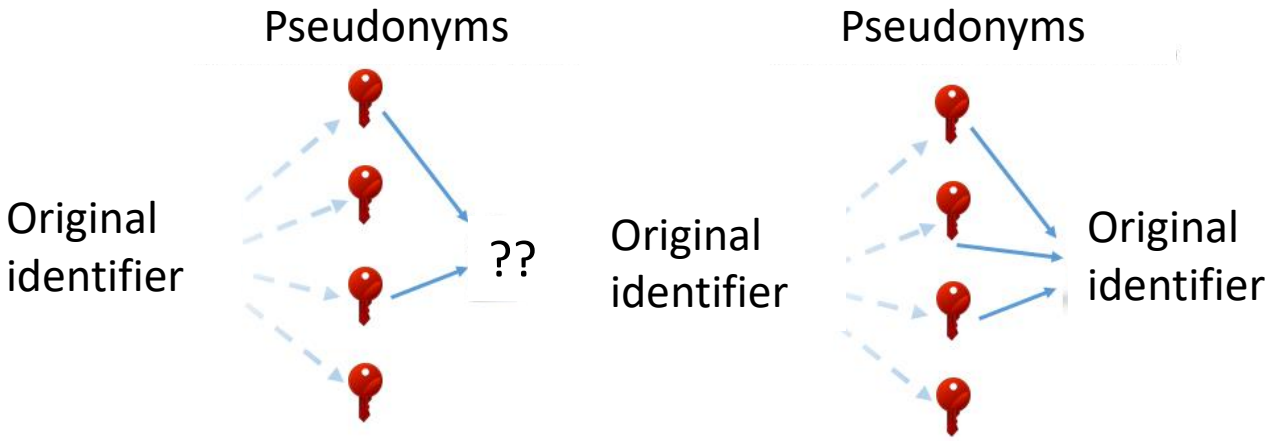


An example for n=4 and t=2

- The “obvious” application is to protect the secret key
- But it can also facilitate pseudonymisation

Secret sharing (*Cont.*)

- "Split" an identifier *ID* into *n* shares
  - Re-construction of *ID* is possible only if any *t + 1* parties exchange their information, but no less



An example for n=4 and t=2

- Several applications have been proposed in this context
- Again, this is compatible with the GDPR's notion on pseudonymisation!

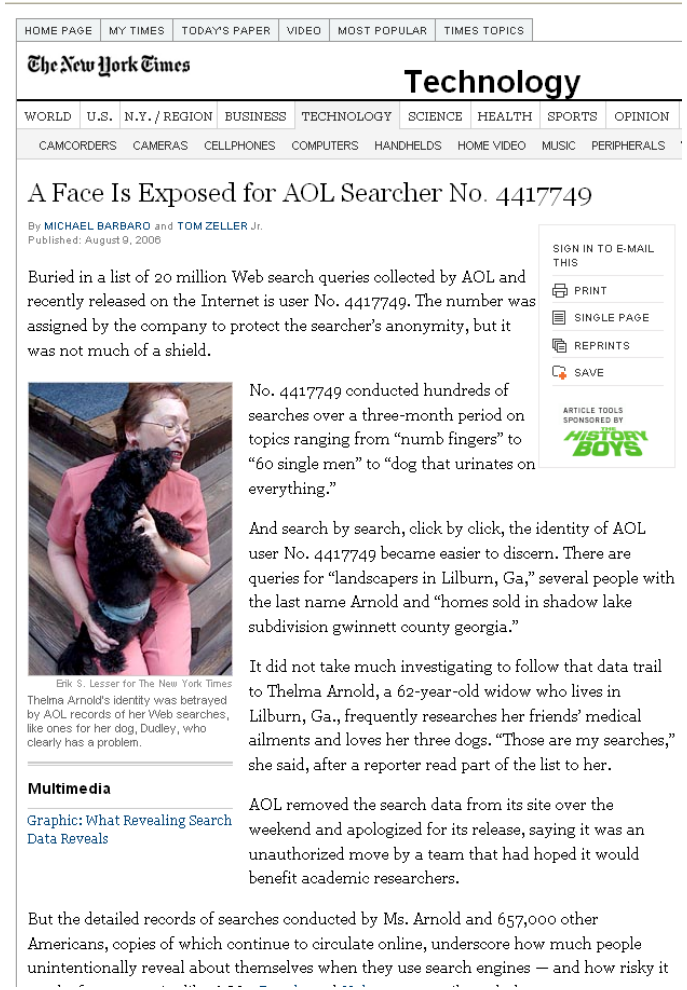
Other advanced cryptographic techniques

- Ring signatures
  - Allowing to verify that a signature is valid and stems from a member of a group, but without being able to explicitly identify which member of the group is the actual signer
- Zero-knowledge proofs
  - Allowing to prove that a statement is valid, without exposing anything else but the statement itself (i.e., such a statement is derived from secret information, but this piece of information is not revealed).
- Homomorphic encryption
  - Allowing to perform (some) operations on encrypted data
- […]
- The term Privacy Enhancing Cryptography has been recently introduced (NIST, 2021)

They may suffice to provide pseudonymisation (according to the GDPR) solutions for "challenging" scenarios

Note that it is not, at a first glance, obvious that these cryptographic techniques can also be pseudonymisation techniques

But, don’t forget…



HOME PAGE | MY TIMES | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS

The New York Times **Technology**

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

CAMCORDERS | CAMERAS | CELLPHONES | COMPUTERS | HANDHELDS | HOME VIDEO | MUSIC | PERIPHERALS

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

SIGN IN TO E-MAIL THIS
PRINT
SINGLE PAGE
REPRINTS
SAVE
ARTICLE TOOLS SPONSORED BY

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher’s anonymity, but it was not much of a shield.

Erik S. Lesser for The New York Times
Thelma Arnold’s identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from “numb fingers” to “60 single men” to “dog that urinates on everything.”

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for “landscapers in Lilburn, Ga,” several people with the last name Arnold and “homes sold in shadow lake subdivision gwinnett county georgia.”

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends’ medical ailments and loves her three dogs. “Those are my searches,” she said, after a reporter read part of the list to her.

Multimedia

Graphic: What Revealing Search Data Reveals

AOL removed the search data from its site over the weekend and apologized for its release, saying it was an unauthorized move by a team that had hoped it would benefit academic researchers.

But the detailed records of searches conducted by Ms. Arnold and 657,000 other Americans, copies of which continue to circulate online, underscore how much people unintentionally reveal about themselves when they use search engines — and how risky it

- Recall the “famous” AOL incident….
- The risk of re-identification is not only a matter of reversing pseudonyms…

Conclusion

- (Advanced) cryptographic techniques should be taken into account towards deciding which is the proper pseudonymisation technique
  - Depending on the context of the processing…
  - On a risk-based approach

Indicative references

- "Data pseudonymisation: Advanced techniques and use cases", A. Bourka and P. Drogkaris (Eds), ENISA Report (contributors: C. Lauradoux, K. Limniotis, M. Hansen, M. Jensen, P. Eftasthopoulos), January 2021
- "Pseudonymisation techniques and best practices", A. Bourka, P. Drogkaris and I. Agrafiotis (Eds), ENISA Report (contributors: M. Jensen, C. Lauradoux, K. Limniotis), December 2019
- "Recommendations on shaping technology according to GDPR provisions - An overview on data pseudonymisation", A. Bourka and P. Drogkaris (Eds), ENISA Report (contributors: K. Limniotis, M. Hansen), January 2019
- "Decentralized generation of multiple, uncorrelatable pseudonyms without trusted third parties", J. Lehnhardt, A. Spalka, Proc. Of TrustBus 2011, Springer.
- "Location-Based Pseudonyms for Identity Reinforcement in Wireless Ad Hoc Networks", B. Tunaru, B. Denis, B. Urguen, VTC 2015.
- "User-generated pseudonyms through Merkle trees", G. Kermezis, K. Limniotis and N. Kolokotronis, Annual Privacy Forum 2021 (APF), Springer.
- "Scalable Private Set Intersection Based on OT Extension", B. Pinkas, T. Schneider, M. Zohner, ACM Trans. Priv. Secur. 2018.
- "How to Share a Secret". A. Shamir, Commun. ACM, 1979.
- "Toward a PEC Use-Case Suite", Technical Report (Preliminary Draft), NIST, 2021.