X-Road Pseudonymization Service

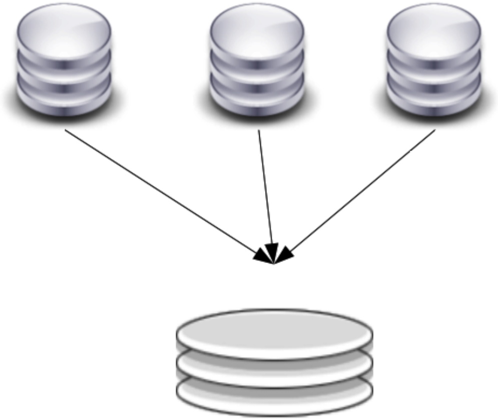
How (not) to design a security architecture

Jan Willemson

Cybernetica

February 6, 2011
Theory Days

Why Pseudonymization?



Why Pseudonymization?

- There are datasets containing sensitive, personally identifiable information
  - Medical, financial, social
- There is a need to perform statistical surveys and produce aggregated results based on several of those datasets
- The statistician is not granted to see the personal details, but standard IDs are needed for linking
- Sometimes, fully cryptographic methods (secure MPC, homomorphic encryption) are not applicable
  - Performance issues
  - High implementation costs
  - No need for strong security guarantees
  - Political fear of everything unknown
- So we will replace the IDs with *pseudonyms*

What are the Security Requirements?

- Who should be able the access the IDs?
  - Data donor. TTP?
- Who should be able to access the data fields?
  - Data donor. Researcher. A person him/herself? A relative? TTP?
- Is reidentification using the data fields a threat?
  - The Netflix/IMDB case
  - Usually this threat is ignored even though it renders most of the heavy-weight pseudonymization techniques void
- What are the "bad" guys/coalitions and what can they do?
  - Data donors? Researchers? Sysadmins? Users? TTP?
- Who and how should be able to grant linking?
  - Researcher? TTP?

What are the Security Requirements?

- Who should be able the access the IDs?
  - Data donor. TTP?
- Who should be able to access the data fields?
  - Data donor. Researcher. A person him/herself? A relative? TTP?
- Is reidentification using the data fields a threat?
  - The Netflix/IMDB case
  - Usually this threat is ignored even though it renders most of the heavy-weight pseudonymization techniques void
- What are the "bad" guys/coalitions and what can they do?
  - Data donors? Researchers? Sysadmins? Users? TTP?
- Who and how should be able to grant linking?
  - Researcher? TTP?

Conclusion:

There is no universal definition of security for pseudonymization

What Kind of Linking should be Allowed?

Researcher's view

Give us all the data so that we could link anything as we please to do a lot of research.

What Kind of Linking should be Allowed?

Researcher's view

Give us all the data so that we could link anything as we please to do a lot of research.

Regulator's view

Hey, guys, you are not here to please yourself, but to serve the society. We tell you when and what to link.

What Kind of Linking should be Allowed?

Researcher's view

Give us all the data so that we could link anything as we please to do a lot of research.

Regulator's view

Hey, guys, you are not here to please yourself, but to serve the society. We tell you when and what to link.

Public Information Act, §43^1(2):

A structured body of data processed within a database may consist exclusively of unique data contained in other databases.

What Kind of Linking should be Allowed?

Researcher's view

Give us all the data so that we could link anything as we please to do a lot of research.

Regulator's view

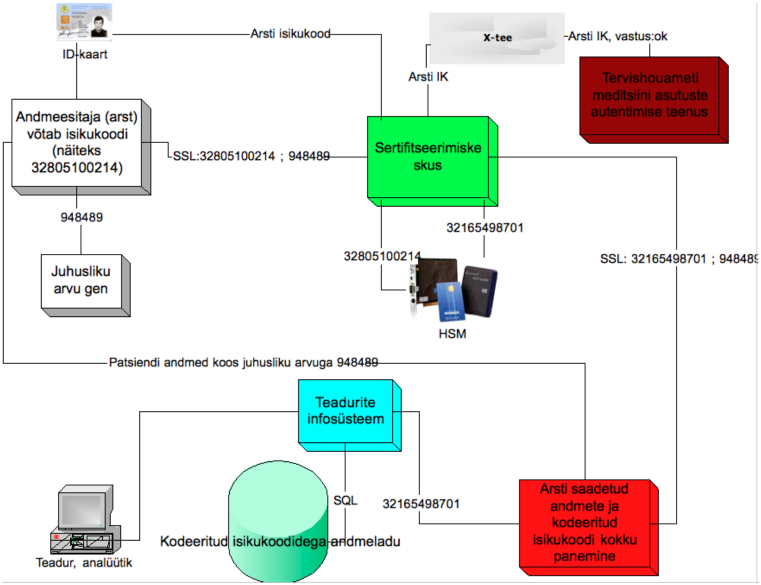
Hey, guys, you are not here to please yourself, but to serve the society. We tell you when and what to link.

Public Information Act, §43^1(2):

A structured body of data processed within a database may consist exclusively of unique data contained in other databases.

Read it as: Aggregated databases may not be used to create new aggregated databases. You will have to start from the original sources.

Pseudonymization in Estonia: the First Attempt



ID-kaart

Arsti isikukood

X-tee

Arsti IK, vastus:ok

Arsti IK

Tervishouameti meditsiini asutuste autentimise teenus

Andmeesitaja (arst) võtab isikukoodi (näiteks 32805100214)

SSL:32805100214 ; 948489

Sertifitseerimiske skus

948489

Juhusliku arvu gen

32165498701

32805100214

SSL: 32165498701 ; 94848

HSM

Patsiendi andmed koos juhusliku arvuga 948489

Teadurite infosüsteem

SQL

32165498701

Kodeeritud isikukoodidega andmeladu

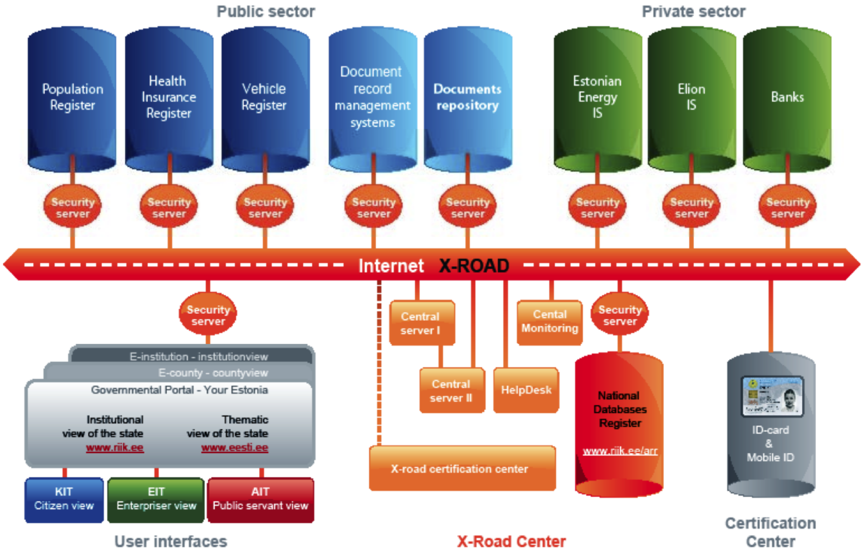
Arsti saadetud andmete ja kodeeritud isikukoodi kokku panemine

Teadur, analüütik

Pseudonymization in Estonia: the First Attempt Highlights

- In order to reconnect the pseudonymized IDs with data fields, random transport identifiers were used
- Pseudonymization was implemented via encryption by the HSM of Estonian national CA
    - Essentially, the CA acted as a TTP, seeing all the sensitive IDs
- Since the people at the CA only knew, how to perform public key operations on the HSM, they generated a key pair and threw half of it away
- During the first live tests it occurred that the HSM was unable to handle simultaneous encryption requests coming from different sources
- When a queueing mechanism was added, under certain circumstances the whole operation of the CA needed restarting

X-Road Infrastructure



X-Road Infrastructure: Characteristics

- Unified XML-based data exchange format
- Each database is supplied with a security server acting as a simple, but flexible HSM
- Minimal number of central services
  - Certification
  - Logging
  - Monitoring
- All the data exchange happens point-to-point and typically presumes an explicit agreement

X-Road Pseudonymization Service: General Principles

- No new TTP/centralized services, if possible
  - Instead, make full use of the existing infrastructure (security servers)
  - Since the security servers will hold the pseudonymization keys anyway, they may as well generate and distribute them
- Pseudonymization does not have massive performance requirements, but it should be as robust as possible
- No need for further actions with the aggregated database
  - Hence, no need for commutative cryptography or public key cryptography in general
  - We will use symmetric encryption
  - One-wayness based on public key encryption does not add much, since the ID space is small ($\approx 70 \cdot 10^6$ in case of Estonian IDs) and can be brute forced by the owner of the key anyway

X-Road Pseudonymization Service: Protocols

- Key generation and distriubution
  - (Security server of) data donor D_1 will generate an AES-256 key K_R
  - He will send a sigcrypted blob $Sig_1(Enc_i(K_R))$ to another data donor D_i
  - D_i will verify the signature and decrypt the key
- Database aggregation
  - When sending data from D_i to the aggregated researcher database R, the IDs are encrypted with the key K_R so that the records become $(Enc_{K_R}(ID), Data(ID))$
  - After all the pseudonymized datasets are transmitted, R links them based on the values $Enc_{K_R}(ID)$ as identifiers

X-Road Pseudonymization Service: Implementation and Benchmarks

- Key transmission is performed by a physical carrier
- Identifying the ID to pseudonymize is performed by standard XPath technology using pugiXML library
- Testing was done on security servers running Ubuntu Linux 10.04 LTS on Intel Core2 8200 processors
- Pseudonymization can happen in several parallel threads (8 in default settings)
- Data throughput achieved was 120MBps
- Memory requirement 45 . . . 55 MB per thread
- Our pseudonymization service was included into X-Road version 5, deployment of which in Estonia started on January 1st 2011

Thank you!

- Who asks a question may go to have lunch

Thank you!

- Who asks a question may go to have lunch
- Logically, I did not say anything about the people who do not ask questions. They can go to have lunch, too