CYBERNETICA

Fifty Shades of Personal Data

Partial Re-Identification and GDPR

Jan Willemson janwil@cyber.ee

A Million Euro Question

- What is personal data?

A Million Euro Question

- What is personal data?
  - GDPR Art. 4(1) gives a definition of personal data depending only on whether the person can be identified completely (even if this complete identification is indirect).

A Million Euro Question

- What is personal data?
  - GDPR Art. 4(1) gives a definition of personal data depending only on whether the person can be identified completely (even if this complete identification is indirect).
  - On the other hand, Recital 26 talks about identification as a process that depends on some likelihoods, costs, etc.

A Million Euro Question

- What is personal data?
  - GDPR Art. 4(1) gives a definition of personal data depending only on whether the person can be identified completely (even if this complete identification is indirect).
  - On the other hand, Recital 26 talks about identification as a process that depends on some likelihoods, costs, etc.
  - Additionally, Art. 11 together with Recital 57 describe a situation where the controller's ability to identify a person may change over time. However, it is left unclear whether this change is gradual or instantaneous.

Why do we care about identifiability at all?

- Because there are *attacks* that are enabled/made easier when someone can be identified.

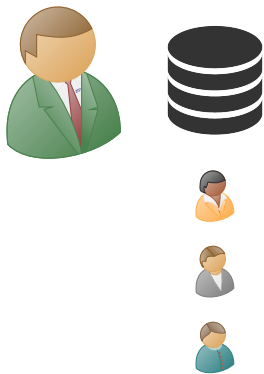
Why do we care about identifiability at all?

- Because there are *attacks* that are enabled/made easier when someone can be identified.
- However, identification does not have to be full in order for attacks to succeed.
- GDPR does not say directly whether identification has to be 100%, but e.g. five typologic categories of identification identified by Putrova all imply this.

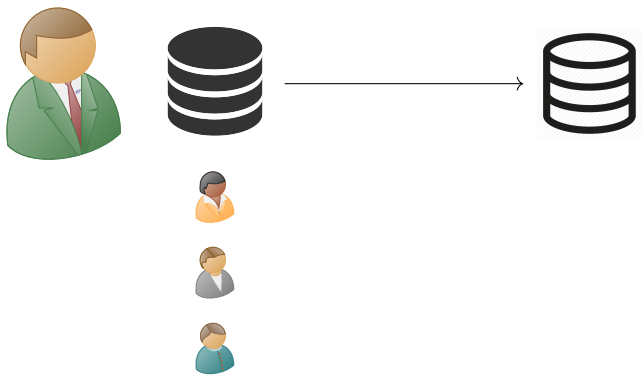
GDPR and risk assessment

- GDPR arguably takes a risk-based approach:
  - Art. 32 is concerned with security of processing,
  - Art. 35 states the data protection impact assessment,
  - several recitals contain further guidelines for risk assessment.
- However, none of them refers to any attacker model (attacker motivation, capabilities, etc.).

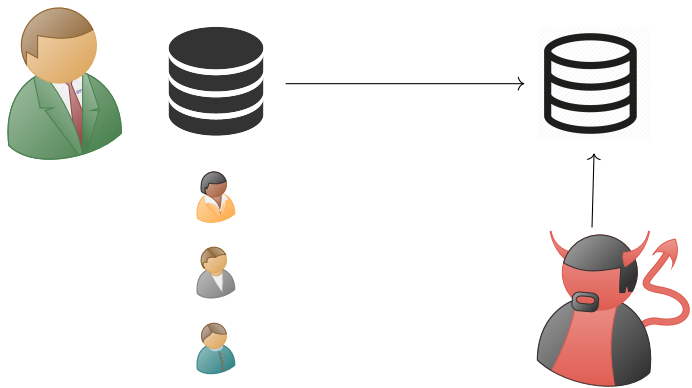
An example attack scenario (I)



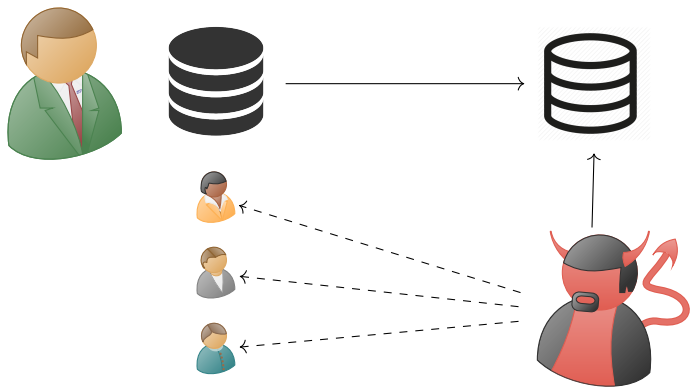
An example attack scenario (I)



An example attack scenario (I)



An example attack scenario (I)



An example attack scenario (II)

- Say, the  contains mobile positioning data without IDs.
- The  can visit the locations and see who passes there.
- Assume that t% of the people in the  who visit locations A, B, C, also visit the red light district.
- The  observes people in locations A, B, C, one of them being a famous politician.
- The  can blackmail the politician for p amount of money and has $\frac{t}{100} \cdot p$ as an expected monetary outcome of this "game".

Two observations

Observation 1

Re-identification does not have to be complete in order to facilitate successful attacks.

Two observations

Observation 1

Re-identification does not have to be complete in order to facilitate successful attacks.

Observation 2

Attacker's success in re-identification of the data subject(s) depends on the effort he is willing to invest.

Cost-benefit considerations (I)

- Let the attacker pay C as the cost of re-identification of a cohort of individuals.
- Assume every data subject S_i has a "fair price" p_i, and that the attacker can identify him/her as a member of group of size g_i. Then the expected outcome of the game is

$$T = \sum_i \frac{p_i}{g_i}.$$

- This game is profitable for the attacker if

$$T > C.$$

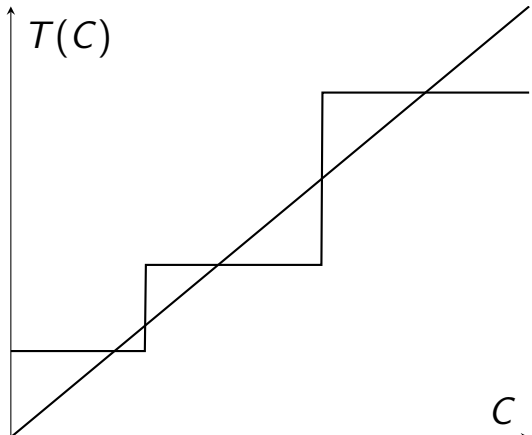
Cost-benefit considerations (II)

Lemma 1

The expected outcome T depends on the investment C in a monotonously increasing manner.

Proof. If the attacker has a strategy that gives him the outcome T as a result of investment C, he still has the same strategy available with resources $C_1 > C$. This means that with the optimal strategy, he can only get a better result $T_1 \geq T$.

An example $T(C)$ graph



Two more observations

Observation 3

The attacker does not necessarily have just one profitable strategy, but he may have several.

Observation 4

If the attacker is able to achieve even a marginal partial re-identification with zero cost, and there exists a subject S_i with positive attacker profit p_i, the attacker already has a profitable strategy.

Discussion

- The model presented can be developed further by e.g. also considering attacker penalties.
- Publishing sanitised datasets carries social benefits, whereas expenses of breaches are carried by individuals. Consequently, the society must offer compensation mechanisms to the suffering individuals.

Back to the GDPR

- Several aspects are not explicitly addressed in the GDPR, even though they should:
  - the issue of partial re-identification,
  - many possible levels of identifiability,
  - more clarity in the risk assessment, including
    - incentives of the attackers and cost-benefit considerations,
    - acknowledging that even a marginal success probability may be sufficient to mount a profitable attack.

Thank you!

- Questions?