ANONYMIZATION SYSTEMS AND UTILITY

MAURIZIO NALDI

IPEN WORKSHOP

ROME - 12 JUNE 2019

UNIVERSITA' degli STUDI di ROMA
TOR VERGATA

ipen

LUMSA
Università

WHAT IS ANONYMIZATION?

- A process by which personal data is (irreversibly) altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party

ANONYMIZATION TECHNIQUES

- Data removal (total or partial)
    - Attribute suppression
    - Record suppression
    - Character masking
- Data perturbation
    - Noise addition
    - Rounding
    - Permutation (a.k.a. swapping or shuffling)
- Data dilution
    - Record singularity removal (K-anonymity)
    - Attribute singularity removal (L-diversity)
    - Distribution closeness (T-closeness)
    - Aggregation (statistical databases)

ANONYMIZATION AND THE GDPR

- The GDPR does not apply to anonymous information ("Whereas" item no. 26)
- Anonymization allows not to apply further data protection measures

WHAT ANONYMIZATION IS NOT: DE-IDENTIFICATION

- Simply removing direct identifiers does not imply that data have been anonymized
- Acting on indirect identifiers individuals can be recognized by
  - Exploiting some identifiers (direct or indirect) left in the released data
  - Reversing pseudonymization by knowing the pseudonymization algorithm
  - Combining or linking datasets

DE-IDENTIFICATION FAILURE: THE NETFLIX CASE

- In 2006 Netflix released their movie-ranking data

- De-identification consisted in replacing individual names with random numbers and moving around personal details

- Two researchers at U. Texas (Arvind Narayanan and Vitaly Shmatikov) de-anonymized some of the data by comparing it with non-anonymous IMDb (Internet Movie Database) users’ movie ratings

- Simply knowing data about only two movies a user has reviewed, including the precise rating and the date of rating give or take three days allows for 68% re-identification success

A CASE OF DATASET COMBINATION

DE-IDENTIFICATION FAILURE: THE NEW YORK CITY TAXI CASE

NYC TAXI

- In 2014 the New York City Taxi and Limousine Commission released a dataset of all taxi trips taken in New York City that year
- The taxi cab medallion numbers and driver's license numbers were pseudonymized
- But the pseudonym was created with a one-way cryptographic hash
- The researchers iterated through all possible medallion numbers and license numbers, determining the cryptographic hash of each, and replacing the hash with the original number
- In addition, a data scientist intern at Neustar discovered he could find pictures taken of celebrities entering or leaving taxicabs with the medallion number in the picture
- By using indirectly identifying information - medallion number, time, and date - specific rides could be located in the dataset released by the New York City Taxi and Limousine commission, identifying pick up location, drop of location, amount paid, and even amount tipped

A CASE OF PSEUDONYM REVERSAL

DE-IDENTIFICATION FAILURE: GOV. WELD'S CASE

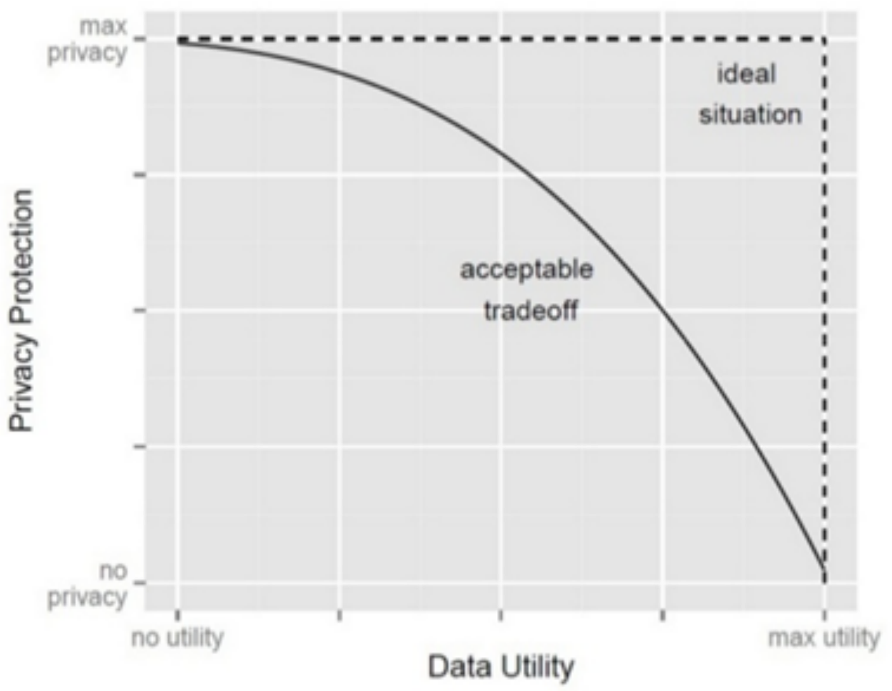
- In the mid-1990's, Massachusetts purchased health insurance for state employees and released records for every state employee's hospital visits
- Explicit identifiers such as name, address, and Social Security numbers were removed
- The record still contained almost a hundred unscrubbed attributes per patient
- Latanya Sweeney, then a graduate student, obtained the data and used the Massachusetts Governor's zip code, birthday, and gender to identify his medical history, diagnosis, and prescriptions

A CASE OF INSUFFICIENT DATA REMOVAL

WHAT'S LACKING IN CURRENT ANONYMIZATION PRACTICES

- Anonymization effectiveness
  - Anonymization techniques are applied without reference to their impact
  - An anonymization-by-design is needed
- Residual Utility
  - Anonymization techniques are applied without reference to the utility of the sanitized database

THE ANONYMIZATION-UTILITY TRADE-OFF



- We can achieve high level of anonymization but the resulting data may be useless
- Optimal trade-offs can be reached only if we are able to measure both features

A STEP FORWARD IN ANONYMIZATION QUANTIFICATION:

DIFFERENTIAL PRIVACY

- For statistical databases Differential Privacy has introduced a measure of anonymization through the change in the probability distribution of the response when the data to be anonymized are present or not
- This is embodied by the epsilon parameter that measures the degree of change
- Proposals have been put forward to relate the epsilon parameter to the design rules of the perturbation techniques involved in differential privacy

MEASURING THE RESIDUAL UTILITY

- Some metrics have been proposed :
  - Matrix norms
  - Correlation
  - Divergence measure
  - Database image quality
- The proper choice has to be related to the specific databases and the purpose of the analysis
- The economic value of the data should be incorporated in the process

CONCLUSIONS

- Moving towards a privacy-by-design approach requires
  - Measurement of anonymization effectiveness
  - Measurement of residual utility
- Engineering rules are needed to achieve optimality in the design of privacy measures
- Challenges ahead: guaranteeing anonymization not just for a single attribute but for the whole record (i.e., taking into account the intra-record correlation)