Cybernetica

Anonymisation and Pseudonymisation

Liina Kamm
Data processing

<table>
<thead>
<tr>
<th>ID</th>
<th>name</th>
<th>age</th>
<th>sex</th>
<th>cm</th>
<th>kg</th>
<th>Hgb</th>
<th>syst</th>
<th>diast</th>
</tr>
</thead>
<tbody>
<tr>
<td>49401223576</td>
<td>Valli Vaarikas</td>
<td>27</td>
<td>N</td>
<td>194</td>
<td>90</td>
<td>135</td>
<td>146</td>
<td>95</td>
</tr>
</tbody>
</table>
Methods for privacy-preserving data analysis

- Data export using pseudonymisation
- Data export using anonymisation
- Federated statistics and machine learning
- Using synthetic data
- Statistical analysis using secure multi-party computation
# Data de-identification techniques

<table>
<thead>
<tr>
<th>Applied privacy technology</th>
<th>No Privacy Technologies</th>
<th>ID code is pseudonymised</th>
<th>Rules-based pseudonymisation</th>
<th>Rules-based anonymisation</th>
<th>Noise-based anonymisation</th>
<th>Identifiable attributes suppressed</th>
<th>All data are encrypted</th>
<th>Data not processed at this organisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>49403136525</td>
<td>Mari Maasikas</td>
<td>Mari Maasikas</td>
<td>Kaire Kartul</td>
<td>Kaire Kartul</td>
<td>6574d303dbf63c66e</td>
<td>6574d303dbf63c66e</td>
<td>y.JspihvFT17m3k2Yv</td>
<td>-</td>
</tr>
<tr>
<td>Kraavi 27</td>
<td></td>
<td>6574d303dbf63c66e</td>
<td>6574d303dbf63c66e</td>
<td></td>
<td></td>
<td>6574d303dbf63c66e</td>
<td>XHDMuJa1GyGOE34</td>
<td>-</td>
</tr>
<tr>
<td>Põlva 63304</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NFDeEnijka2Oy24w</td>
<td>-</td>
</tr>
<tr>
<td>13. märts 1994</td>
<td>55 123 9876</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>rpxMCGqVFPAINLQ</td>
<td>-</td>
</tr>
<tr>
<td><a href="mailto:mari@maasik.as">mari@maasik.as</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>VEGlgTSeCVUtV0</td>
<td>-</td>
</tr>
<tr>
<td>9:00 - Põlva</td>
<td>9:00 - Põlva</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>UL86itWppitwU</td>
<td>-</td>
</tr>
<tr>
<td>9:15 - Valgemetsa</td>
<td>9:15 - Valgemetsa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>fb7glGltJy/jTuamR</td>
<td>-</td>
</tr>
<tr>
<td>9:30 - Ülenurme</td>
<td>9:30 - Ülenurme</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QYT1sMvyGyKKGP</td>
<td>-</td>
</tr>
<tr>
<td>9:45 - Tartu</td>
<td>9:45 - Tartu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SJWmMDWb7tatfavh</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ebxmXoJ0UFX0</td>
<td>-</td>
</tr>
</tbody>
</table>

- **Possibility of reidentification**
  - Directly, from identifiable records
  - Directly, from identifiable records
  - Indirectly, supported by auxiliary data sources
  - Indirectly, from mobility patterns, supported by auxiliary data sources
  - Indirectly, from mobility patterns, supported by auxiliary data sources
  - Indirectly, from mobility patterns, supported by auxiliary data sources
  - Impossible without access to the key
  - Impossible

- **Analytics tools**
  - Standard tools
  - Standard tools
  - Standard tools
  - Standard tools
  - Standard tools
  - Processing must be done by someone else

- **Risk of privacy breach**
  - High
  - Common
  - Low
  - Rare

- **Frequency of use**
  - High
  - Common
  - Low
  - Rare

21.04.2021
## Pseudonymisation

<table>
<thead>
<tr>
<th>ID</th>
<th>name</th>
<th>age</th>
<th>sex</th>
<th>cm</th>
<th>kg</th>
<th>Hgb</th>
<th>syst</th>
<th>diast</th>
</tr>
</thead>
<tbody>
<tr>
<td>49403136525</td>
<td>Mari Maasikas</td>
<td>27</td>
<td>N</td>
<td>165</td>
<td>75</td>
<td>110</td>
<td>122</td>
<td>75</td>
</tr>
<tr>
<td>39507082997</td>
<td>Ants Ōun</td>
<td>25</td>
<td>M</td>
<td>165</td>
<td>85</td>
<td>162</td>
<td>132</td>
<td>88</td>
</tr>
<tr>
<td>49401223576</td>
<td>Valli Vaarikas</td>
<td>27</td>
<td>N</td>
<td>194</td>
<td>90</td>
<td>135</td>
<td>146</td>
<td>95</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

## Pseudonymisation

<table>
<thead>
<tr>
<th>code</th>
<th>name</th>
<th>age</th>
<th>sex</th>
<th>cm</th>
<th>kg</th>
<th>Hgb</th>
<th>syst</th>
<th>diast</th>
</tr>
</thead>
<tbody>
<tr>
<td>HYP_01</td>
<td></td>
<td>27</td>
<td>N</td>
<td>165</td>
<td>75</td>
<td>110</td>
<td>122</td>
<td>75</td>
</tr>
<tr>
<td>HYP_02</td>
<td></td>
<td>25</td>
<td>M</td>
<td>165</td>
<td>85</td>
<td>162</td>
<td>132</td>
<td>88</td>
</tr>
<tr>
<td>HYP_03</td>
<td></td>
<td>27</td>
<td>N</td>
<td>194</td>
<td>90</td>
<td>135</td>
<td>146</td>
<td>95</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
# Anonymisation

<table>
<thead>
<tr>
<th>ID</th>
<th>name</th>
<th>age</th>
<th>sex</th>
<th>cm</th>
<th>kg</th>
<th>Hgb</th>
<th>syst</th>
<th>diast</th>
</tr>
</thead>
<tbody>
<tr>
<td>49403136525</td>
<td>Mari Maasikas</td>
<td>27</td>
<td>N</td>
<td>165</td>
<td>65</td>
<td>110</td>
<td>122</td>
<td>75</td>
</tr>
<tr>
<td>39507082997</td>
<td>Ants Ōun</td>
<td>25</td>
<td>M</td>
<td>165</td>
<td>85</td>
<td>162</td>
<td>132</td>
<td>88</td>
</tr>
<tr>
<td>49401223576</td>
<td>Valli Vaarikas</td>
<td>27</td>
<td>N</td>
<td>194</td>
<td>90</td>
<td>135</td>
<td>146</td>
<td>95</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

- **Anonymisation**

<table>
<thead>
<tr>
<th>age</th>
<th>sex</th>
<th>BMI</th>
<th>Hgb</th>
<th>syst</th>
<th>diast</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>N</td>
<td>23,9</td>
<td>110</td>
<td>122</td>
<td>75</td>
</tr>
<tr>
<td>25</td>
<td>M</td>
<td>31,2</td>
<td>162</td>
<td>132</td>
<td>88</td>
</tr>
<tr>
<td>27</td>
<td>N</td>
<td>23,9</td>
<td>135</td>
<td>146</td>
<td>95</td>
</tr>
<tr>
<td>27</td>
<td>N</td>
<td>23,9</td>
<td>135</td>
<td>146</td>
<td>95</td>
</tr>
</tbody>
</table>

- **Anonymisation**

<table>
<thead>
<tr>
<th>sex</th>
<th>kg</th>
<th>Hgb</th>
<th>syst</th>
<th>diast</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>75</td>
<td>110</td>
<td>122</td>
<td>75</td>
</tr>
<tr>
<td>N</td>
<td>90</td>
<td>135</td>
<td>146</td>
<td>95</td>
</tr>
<tr>
<td>N</td>
<td>75</td>
<td>110</td>
<td>122</td>
<td>75</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

- **Anonymisation**
Federated statistics

<table>
<thead>
<tr>
<th>ID</th>
<th>name</th>
<th>age</th>
<th>sex</th>
<th>cm</th>
<th>kg</th>
<th>Hgb</th>
<th>syst</th>
<th>diast</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>name</th>
<th>age</th>
<th>sex</th>
<th>cm</th>
<th>kg</th>
<th>Hgb</th>
<th>syst</th>
<th>diast</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Secure multi-party computation

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>49401223576</td>
<td>Valli Vaarikas</td>
<td>27</td>
<td>N</td>
<td>192</td>
<td>90</td>
<td>135</td>
<td>120</td>
</tr>
<tr>
<td>HYP_1</td>
<td>3</td>
<td>X</td>
<td>76</td>
<td>15</td>
<td>122</td>
<td>10</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>HYP_1</td>
<td>17</td>
<td>X</td>
<td>76</td>
<td>10</td>
<td>10</td>
<td>100</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>HYP_1</td>
<td>7</td>
<td>X</td>
<td>40</td>
<td>65</td>
<td>3</td>
<td>10</td>
<td>60</td>
<td></td>
</tr>
</tbody>
</table>
# Data synthesis

<table>
<thead>
<tr>
<th>code</th>
<th>name</th>
<th>age</th>
<th>sex</th>
<th>cm</th>
<th>kg</th>
<th>Hgb</th>
<th>syst</th>
<th>diast</th>
</tr>
</thead>
<tbody>
<tr>
<td>HYP_01</td>
<td></td>
<td>27</td>
<td>N</td>
<td>165</td>
<td>75</td>
<td>110</td>
<td>122</td>
<td>75</td>
</tr>
<tr>
<td>HYP_02</td>
<td></td>
<td>25</td>
<td>M</td>
<td>165</td>
<td>85</td>
<td>162</td>
<td>132</td>
<td>88</td>
</tr>
<tr>
<td>HYP_03</td>
<td></td>
<td>27</td>
<td>N</td>
<td>192</td>
<td>90</td>
<td>135</td>
<td>146</td>
<td>95</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>age</th>
<th>sex</th>
<th>cm</th>
<th>kg</th>
<th>Hgb</th>
<th>syst</th>
<th>diast</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>M</td>
<td>165</td>
<td>80</td>
<td>156</td>
<td>122</td>
<td>75</td>
</tr>
<tr>
<td>27</td>
<td>N</td>
<td>165</td>
<td>80</td>
<td>110</td>
<td>133</td>
<td>88</td>
</tr>
<tr>
<td>28</td>
<td>N</td>
<td>195</td>
<td>95</td>
<td>132</td>
<td>142</td>
<td>92</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Today on PETs

- Pseudonymisation
- Anonymisation
  - k-anonymity
  - l-diversity
  - t-closeness
Pseudonymisation is the **processing of personal data** in such a manner that the personal data **can no longer be attributed to a specific data subject 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.

(GDPR, art. 4(5))
Pseudonymisation

- Pseudonymised data cannot be attributed to a specific person without additional information
- These data can still be used for data linking between different datasets
- GDPR: pseudonymised data = personal data, processing is subject to data protection regulations
- Usually the identifier or personal information is encoded in some way (using a pseudonymisation secret)
- Decoding is a hard problem without the secret
- Decoding is simple with the secret
Examples of methods for pseudonymisation (1/3)

- Counter (substitute id by a number starting from s)
  - + Simple
  - + No connection to original identifier
  - - Reveals information about order of data
  - - Scalability

- Random number generator (RNG)
  - Choice of RNG
  - + No connection to original identifier
  - + Does not reveal information about order of data
  - - Extra effort for collisions
  - - Scalability
Examples of methods for pseudonymisation (2/3)

- Cryptographic hash function
  - Maps input strings of arbitrary length to fixed length outputs
  - + One-way
  - + Collision free
  - - Prone to brute force and dictionary attacks

- Message authentication code (MAC) (e.g., HMAC)
  - Keyed-hash function – a secret key is introduced to generate the pseudonym
  - Without this key, it is not possible to map the identifiers and the pseudonyms
  - + Reverting is infeasible, as long as the key has not be compromised
Examples of methods for pseudonymisation (3/3)

- Encryption
  - Mainly symmetric encryption is used (e.g., AES in different modes of operation)
  - Secret key is the pseudonymisation secret and the recovery secret
  - + Reverting is infeasible, as long as the key has not be compromised
  - - Need to deal with the block size (padding if the identifier is smaller than the block, compression or CTR mode if the identifier is larger than the block)
Pseudonymisation policies

- Deterministic pseudonymisation
  - ID is always replaced by the same pseudonym consistently between databases

- Document-randomised pseudonymisation
  - Each time ID appears, it is mapped to a different pseudonym
  - IDs are mapped to the same set of pseudonyms in different databases
  - Same result for two instances of pseudonymisation

- Fully-randomised pseudonymisation
  - Each time ID appears, it is mapped to a different pseudonym
  - IDs are mapped to different pseudonyms in different databases
  - Different results for two instances of pseudonymisation
  - + Best protection level, - prevents any comparison between databases
Adversaries

• Goal: to single out a specific individual from a group
• Insider adversary
  – Is able to get information about the pseudonymisation secret or other relevant additional information
• External adversary
  – Does not have direct access to the pseudonymisation secret or other relevant information
  – May have access to a pseudonymised dataset
  – May have access to a pseudonymisation oracle (can have arbitrary input data values pseudonymised)
Goals of an attack

- Recover the pseudonymisation secret
  - Re-identify the whole dataset
  - Perform further pseudonymisations
- Complete re-identification
  - Link one or more pseudonyms back to the identity of individuals
- Discrimination (recognising a distinction)
  - Identify at least one property of a pseudonym holder
  - May not directly lead to uncovering the identity, but may be enough to discriminate them
Example of discrimination 1

<table>
<thead>
<tr>
<th>ID</th>
<th>Gender</th>
<th>Exam Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>98</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>67</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>52</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>83</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>72</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>100</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>95</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>100</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>89</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>100</td>
</tr>
</tbody>
</table>
## Example of discrimination 2

<table>
<thead>
<tr>
<th>ID</th>
<th>Gender</th>
<th>Exam Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>98</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>67</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>52</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>83</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>72</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>98</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>95</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>100</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>89</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>100</td>
</tr>
</tbody>
</table>
Difficulty of the attack depends on

- The amount of data subject information contained in the pseudonym
- The background knowledge of the adversary
- The size of the identifier domain
- The size of the pseudonym domain
- The size of the pseudonymisation secret
- The choice and configuration of the pseudonymisation function
Guessing

- Background knowledge (e.g., probability distribution, side information) on
  - Some (or all) of the individuals,
  - The pseudonymisation function,
  - The dataset.
- Some identifiers may be more frequent than others
- Guesswork – exploiting the statistical characteristics of the identifiers
- Can be applied even when the identifier domain is huge as discrimination is possible by performing a frequency analysis of the observed pseudonyms.
Brute force attack: Complete re-identification

- Prerequisites
  - Adversary can compute the pseudonymisation function (i.e. there is no pseudonymisation secret) or
  - Adversary has access to a “black box” implementation of the pseudonymisation function.

- For complete re-identification, the identifier domain must be finite and relatively small

- Apply the pseudonymisation function on each value until adversary finds a match

- If the domain size is infinite, the brute force attack typically is infeasible
Brute force attack: Discrimination

- If re-identification is difficult, the adversary can try a discrimination attack.
- Consider a subdomain of the identifiers domain for which they can compute all the pseudonyms (e.g., specific names like Ann, Anna, Anne, Annie).
- If a pseudonymisation secret is used, these attacks are not possible as the attacker is not able to compute the pseudonymisation function.
- Then the attacker must exhaustively check all the possible secrets and computes the recovery function for each.
- To counteract this, the number of possible pseudonymisation secrets should be large.
Dictionary search

- Optimisation of brute force attack, which can save computation costs
- Large amount of pseudonyms are needed for re-identification or discrimination
- Adversary precomputes and saves a set of pseudonyms
- **Dictionary** contains a pseudonym and the corresponding identifier
- Re-identification of a pseudonym has only the cost of a lookup
- The dictionary search is essentially the computation and storage of the mapping table
Advanced pseudonymisation techniques

ENISA. Data Pseudonymisation: Advanced Techniques and Use Cases (January 28, 2021)


- Asymmetric encryption
- Ring signatures and group pseudonyms
- Secure multiparty computation
- Secret sharing
Anonymisation is 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.

(ISO/TS 25237:2017)
Pseudonymisation and anonymisation (1/2)

- Pseudonymised data is not anonymised data
- In an anonymised dataset, neither the controller nor a third party can identify any individual
- Anonymised data do not qualify as personal data
- In general it is not enough to simply remove an individual’s identifier
- **Quasi-identifiers** – combinations of attributes relating to an individual
- Pseudonymisation: existence of an association between personal identifiers and pseudonyms. Re-identification is possible, data is personal data
- Anonymisation: such an association should not be available by any means, re-identification is not possible, data is not personal data
Pseudonymisation and anonymisation (2/2)

• “Anonymous” in common language also describes cases where the identities of individuals are only hidden

• Even in the absence of personal identifiers, data are not necessarily anonymous

• Pseudonymisation often relies on anonymisation techniques for efficiency
  – In some cases it might be wise to use anonymisation techniques (e.g. attributes generalisation) in the pseudonymisation process, so as to reduce the possibility of third parties to infer personal data
Anonymised data (1/2)

• Anonymised data => not personal data => not subject to data protection regulations
• Impossible to identify individuals from anonymised data
• Data can be aggregated or converted into statistics so that individuals can no longer be identified from them
• This is permanent and it must be impossible for the controller or a third party to convert the data back into identifiable form
• Possibility of identification – costs of identification, time required and available technologies must be considered
Anonymised data (2/2)

• Whether an individual data item can be considered anonymous or not requires case-by-case evaluation
  – A dataset can contain detailed information on individuals which makes them indirectly identifiable (e.g., a rare attribute or a distinguishing set of attributes)
• If a controller discloses parts of a data set from which an individual can be identified, the dataset still contains personal data
• This kind of dataset remains subject to data protection regulations
Types of anonymisation techniques

• Randomisation
  – “Randomly” alter the data (e.g., via noise addition) to remove the link between the dataset and the individual

• Generalisation
  – Modify the scale or order of magnitude of an attribute (e.g., an exact age can be replaced by an age range)
Risks essential to anonymisation

- **Singling out** – identify some or all records of one individual in the dataset
- **Linkability** – ability to link (at least) two records of the same data subject or a group of data subjects
  - In the same database
  - In different databases
  - If an adversary finds two records that are assigned to the same group of individuals but cannot single out individuals in this group, the technique provides resistance against singling out but not against linkability
- **Inference** – possibility to deduce the value of an attribute from the values of a set of other attributes with significant probability
Randomisation

- Noise addition
- Permutation
- Differential privacy
Noise addition: description

- Modify attributes in the dataset so that they are less accurate
- Retain the overall distribution
- Example: change weight by +/-5 kg
- Commonly needs to be combined with other anonymisation techniques
- The level of noise depends on
  - the importance of the information in the analysis
  - the impact on individuals’ privacy as a result of disclosure of this attribute
Noise addition: guarantees and issues

- **Singling out**: Still possible, results less reliable
- **Linkability**: Still possible, a real record can wrongly be linked to noise
- **Inference**: May be possible, success rate will be lower, may lead to some false positives and false negatives

- If noise is not semantically viable, it can be filtered out
- It is a complementary measure, it is not a standalone solution
Failures of noise addition: Netflix Prize (1/3)


- Anonymised sample of the customers’ database of Netflix publicly released after for the Netflix Prize data mining contest:
  - 100 480 507 ratings on a scale 1-5
  - Over 18 000 movies
  - Created by 480 189 users between Dec 1999 and Dec 2005
  - All customer information removed except ratings and dates
  - Noise: ratings were slightly increased or decreased
Failures of noise addition: Netflix Prize (2/3)

• An adversary who knows only a little about an individual subscriber can easily identify this subscriber’s record in the dataset
• Even if it is hard to collect such information for a large number of subscribers, targeted de-anonymisation is a serious threat to privacy
• Example: A boss using the Netflix Prize dataset to find an employee’s entire movie viewing history after a casual conversation
• Using the Internet Movie Database as the source of background knowledge, they successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information
Failures of noise addition: Netflix Prize (3/3)

- 99% of user records could be uniquely identified in the dataset
  - Using 8 ratings (2 can be wrong)
  - Dates with 14-day errors as selection criteria
- 68% of user records could be uniquely identified in the dataset
  - Using 2 ratings
  - Dates with 3-day errors
- Sample was not uniformly random
- Very little noise had been added. Why?
Permutation: description

- Shuffling the values of attributes in a table
- Some will be artificially linked to different individuals
- Special form of noise addition
  - + Retains the range and exact distribution of each attribute within the dataset
  - - Will destroy the logical relationship or statistical correlation between attributes
- An attacker could identify the permuted attributes based on the correlation
- Permute a set of related attributes so the logical relationship does not break
- Example: reasons for hospitalisation, symptoms, department in charge
Permutation: guarantees and issues

- **Singling out**: Still possible, results are less reliable
- **Linkability**: May prevent the correct linking of attributes but still allow incorrect linkability (a real entry can be linked to a different individual)
- **Inference**: Probabilistic inference still possible. If the attacker does not know which attributes have been permuted, she has to consider that her inference is based on a wrong hypothesis

- Permuting non-sensitive attributes does not result in a significant gain
- Permuting strongly correlated attributes randomly is counterproductive
- It is a complementary measure, it is not a standalone solution
## Failures of permutation: permuted dataset

<table>
<thead>
<tr>
<th>Year of birth</th>
<th>Gender</th>
<th>Job</th>
<th>Income (permuted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>M</td>
<td>CEO</td>
<td>2500</td>
</tr>
<tr>
<td>1967</td>
<td>N</td>
<td>CFO</td>
<td>3500</td>
</tr>
<tr>
<td>1993</td>
<td>M</td>
<td>Engineer</td>
<td>6000</td>
</tr>
<tr>
<td>1979</td>
<td>M</td>
<td>Engineer</td>
<td>1200</td>
</tr>
<tr>
<td>1999</td>
<td>N</td>
<td>Intern</td>
<td>2400</td>
</tr>
</tbody>
</table>
## Failures of permutation: comparison to the original

<table>
<thead>
<tr>
<th>Year of birth</th>
<th>Gender</th>
<th>Job</th>
<th>Income (original)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>M</td>
<td>CEO</td>
<td>6000</td>
</tr>
<tr>
<td>1967</td>
<td>N</td>
<td>CFO</td>
<td>2500</td>
</tr>
<tr>
<td>1993</td>
<td>M</td>
<td>Engineer</td>
<td>2400</td>
</tr>
<tr>
<td>1979</td>
<td>M</td>
<td>Engineer</td>
<td>3500</td>
</tr>
<tr>
<td>1999</td>
<td>N</td>
<td>Intern</td>
<td>1200</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year of birth</th>
<th>Gender</th>
<th>Job</th>
<th>Income (permuted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>M</td>
<td>CEO</td>
<td>2500</td>
</tr>
<tr>
<td>1967</td>
<td>N</td>
<td>CFO</td>
<td>3500</td>
</tr>
<tr>
<td>1993</td>
<td>M</td>
<td>Engineer</td>
<td>6000</td>
</tr>
<tr>
<td>1979</td>
<td>M</td>
<td>Engineer</td>
<td>1200</td>
</tr>
<tr>
<td>1999</td>
<td>N</td>
<td>Intern</td>
<td>2400</td>
</tr>
</tbody>
</table>
Differential privacy

- Differential privacy generates anonymised views of a dataset for a third party while retaining a copy of the original data.
- Determines how much noise the data controller needs to add to get the necessary privacy guarantees for a query.
- Differential privacy techniques will not change the original data.
- As long as the original data remains, the data controller is able to identify individuals in results of differential privacy queries.
- Such results are considered personal data.
Differential privacy: benefits

- Datasets are provided to third parties in response to a specific query not as a single dataset
- Anonymisation techniques (e.g., addition of noise) can be used on a query
- Queries issued by an entity must be monitored to limit inference and linkability attacks
Differential privacy: guarantees and issues

- **Singling out**: Should not be possible (if only statistics are output and rules applied to the set are well chosen)
- **Linkability**: Might be possible (using multiple requests)
- **Inference**: Still possible (using multiple requests)

- It is difficult to generate the proper amount of noise, which will protect individuals’ privacy and retain the usefulness of released responses
Failures of differential privacy

- A combination of query results may allow disclosing secret information.
- An attacker may engineer multiple queries that progressively reduce the amplitude of the output until an individual is revealed, with high probability.
- As per regulation, it is not anonymisation.
Generalisation

- Aggregation and $k$-anonymity
- $l$-diversity/$t$-closeness
Aggregation and $k$-anonymity: description

- [link](http://spdp.di.unimi.it/papers/k-Anonymity.pdf)
- Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least $k$ individuals
- Group an individual with at least $k$ other individuals
- Useful when the correlation of detailed values may create quasi-identifiers
- The higher the parameter $k$, the stronger the privacy guarantees
- Attribute values are generalised (e.g., by lowering the granularity of an attribute, adding value ranges or intervals, grouping)
- $k$-anonymity is an NP-hard problem, algorithms have complexity exponential in the size of the quasi-identifier
Aggregation and $k$-anonymity: guarantees

- **Singling out**: Should not be possible
- **Linkability**: Still possible, but limited. In addition, within the group, the probability that two records correspond to the same individual is $1/k$
- **Inference**: Still possible (if all $k$ individuals are in the same group and if the adversary knows which group an individual belongs to)
Aggregation and $k$-anonymity: issues

- If $k$ is too big: Reducing the number of considered quasi-identifiers makes it easier to build clusters of $k$ individuals. However, these attributes can be used to single out an individual in a cluster of $k$ records.

- If $k$ is too small:
  - Any individual in a cluster is too significant and inference attacks have a higher success rate.
  - Individuals represent too important a fraction of the entries in a cluster.

- $k$-anonymisation does not work on high-dimensional datasets (Netflix Prize dataset, real-world datasets of individual recommendations and purchases).
Failures of $k$-anonymity

- Does not prevent inference attacks

<table>
<thead>
<tr>
<th>Year</th>
<th>Gender</th>
<th>Address</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1953</td>
<td>F</td>
<td>123</td>
<td>Heart attack</td>
</tr>
<tr>
<td>1953</td>
<td>F</td>
<td>123</td>
<td>Hypertension</td>
</tr>
<tr>
<td>1953</td>
<td>F</td>
<td>123</td>
<td>Hypertension</td>
</tr>
<tr>
<td>1959</td>
<td>F</td>
<td>123</td>
<td>Heart attack</td>
</tr>
<tr>
<td>1959</td>
<td>F</td>
<td>123</td>
<td>Heart attack</td>
</tr>
</tbody>
</table>

- Might leave interesting (and significant) records out of the analysis
$l$-diversity: description

- $l$-diversity extends $k$-anonymity
- Ensures that deterministic inference attacks are no longer possible
- Makes sure that in each equivalence class, every attribute has at least $l$ different values
- Limit the occurrence of equivalence classes with poor attribute variability
- An attacker with background knowledge on a specific individual is always left with a significant uncertainty
- Does not work if the attributes within a cluster are unevenly distributed or belong to a small range of values or semantic meanings
$t$-closeness: description

- Refinement of $l$-diversity
- Aims to create equivalent classes that resemble the initial distribution of attributes in the table
- At least $l$ different values should exist within each equivalence class, and
- Each value is represented as many times as necessary to mirror the initial distribution of each attribute
- Useful when it is important to keep the data as close as possible to the original dataset
/-diversity/t-closeness: guarantees and issues

• **Singling out**: Should not be possible

• **Linkability**: Still possible. The probability that the same entries belong to the same individual is higher than $1/N$ ($N$ – number of individuals in the dataset)

• **Inference**: Subject to probabilistic inference attacks, but not deterministic ones

• The distribution of sensitive values in each cluster should resemble the distribution of the values in the total population
Failures of $I$-diversity

- Vulnerable to probabilistic inference attacks

- 1959 Ann F 123
- 1959 Anna F 123
- 1959 Mary F 123

<table>
<thead>
<tr>
<th>Year</th>
<th>Gender</th>
<th>Address</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1953</td>
<td>F</td>
<td>123</td>
<td>Heart attack</td>
</tr>
<tr>
<td>1953</td>
<td>F</td>
<td>123</td>
<td>Hypertension</td>
</tr>
<tr>
<td>1953</td>
<td>F</td>
<td>123</td>
<td>Hypertension</td>
</tr>
<tr>
<td>1953</td>
<td>F</td>
<td>123</td>
<td>Heart attack</td>
</tr>
<tr>
<td>1959</td>
<td>F</td>
<td>123</td>
<td>Heart attack</td>
</tr>
<tr>
<td>1959</td>
<td>F</td>
<td>123</td>
<td>Heart attack</td>
</tr>
<tr>
<td>1959</td>
<td>F</td>
<td>123</td>
<td>Heart attack</td>
</tr>
<tr>
<td>1959</td>
<td>F</td>
<td>123</td>
<td>Hypertension</td>
</tr>
<tr>
<td>1959</td>
<td>F</td>
<td>123</td>
<td>Heart attack</td>
</tr>
</tbody>
</table>
Anonymisation vs machine learning and data mining


- Researchers fond that 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes

- Even heavily sampled anonymised datasets are unlikely to satisfy the modern standards for anonymisation set forth by GDPR

- Seriously challenge the technical and legal adequacy of the de-identification release-and-forget model

- Thy reject the claims that (1) re-identification is not a practical risk and (2) sampling or releasing partial datasets provide plausible deniability
De-identified hospital discharge data could be re-identified using basic demographic attributes [1]

Diagnostic codes, year of birth, gender, and ethnicity could uniquely identify patients in genomic studies data [2]

In 2016, the Australian Department of Health publicly released de-identified medical records for 10% of the population, researchers re-identified them 6 weeks later [3]
Re-identification (2/2)

- In 2016, journalists re-identified politicians in an anonymised browsing history dataset of 3 million German citizens, uncovering their medical information and their sexual preferences [4]
- Researchers were able to uniquely identify individuals in anonymized taxi trajectories in NYC [5]
- Bike sharing trips in London [6]
- Subway data in Riga [7]
- Mobile phone and credit card datasets [8,9]
Re-identification (references)


[6] Siddle, J. I know where you were last summer: London’s public bike data is telling everyone where you’ve been. https://vartree.blogspot.com/2014/04/i-know-where-you-were-last-summer.html (2014)


Further reading

- ENISA. Recommendations on shaping technology according to GDPR provisions – An overview on data pseudonymisation: https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-gdpr-provisions
- I. Dinur and K. Nissim. Revealing Information while Preserving Privacy http://www.cs.cornell.edu/~shmat/courses/cs6431/dn03.pdf
Thank you!

Cybernetica AS
Mäealuse 2/1, 12618 Tallinn, Estonia

info@cyber.ee  |  cyber.ee

@cybernetica  @Cybernetica  @CyberneticaAS