Assessing Disclosure Risk in Anonymized Datasets

Michele Bezzi (ATL) & Alexei Kounine (EPFL)
Outline

• Background & Motivation
• Anonymisation
• Disclosure Risk Estimation
  – Entropy measure
  – Properties
• Case Study: Flows
• Final remarks
Goal

• Problem:
  – The goal is to transform original data records so that no sensitive personal data are disclosed, whereas preserving the maximum amount of relevant information (anonymity vs. utility trade off), data integrity and consistency.

• Application
  – Creating datasets for application testing, whenever production DB contains sensitive data. (Our original goal)
  – Allowing researchers to share data and run analytical models on micro-data (e.g., log files), preserving privacy.
Goal

Risk-Utility Confidentiality Map

Data Utility

Disclosure Risk

Original Data

Research

Apply Anonymization

Data App Testing

Adapted from Duncan, et al. 2001
Anonymisation engine & risk estimation

Input data → Policy → Anonymisation → Risk Assessment → Utility Assessment → Output

Test
Implementation

• Using FLAIM (Framework for Log Anonymization and Information Management), developed by NCSA
• FLAIM anonymization engine (adapted) + risk module
Anonymisation primitives

IPs
• Black Marker (16 bits):
• Random Permutation (one-to-one mapping)
• Prefix-preserving (random permutation, but preserving structure)

<table>
<thead>
<tr>
<th>IP Address</th>
<th>Black Marker (16-bit)</th>
<th>Random Permutation</th>
<th>Prefix-preserving</th>
</tr>
</thead>
<tbody>
<tr>
<td>168.125.96.167</td>
<td>168.125.0.0</td>
<td>124.12.132.37</td>
<td>12.131.102.67</td>
</tr>
<tr>
<td>168.125.96.18</td>
<td>168.125.0.0</td>
<td>231.45.36.167</td>
<td>12.131.102.17</td>
</tr>
<tr>
<td>168.125.132.37</td>
<td>168.125.0.0</td>
<td>12.72.8.5</td>
<td>12.131.201.29</td>
</tr>
</tbody>
</table>

Port number
• Bilateral Classification: Replace with 0 or 65535 (the port smaller or larger than 1024): E.g., 27 -> 0, 2048->65535

Number of packets/bytes
• Add random noise (zero-average)
• Classification
**Attack scenario**

- The attacker aims at re-identifying released data by linking them with some background knowledge, which has some overlapping attributes with the released dataset.

- Estimating $P(r|s)$: knowing data masking transformations, distance based similarity

- More uncertain mapping is - lower risk

- Because the data holder does not know in advance which records and attributes might be available to the attacker, it must run the risk analysis on the whole released dataset and assume a set of key attributes the attacker might know and use for re-identification.

### Original data $S$

<table>
<thead>
<tr>
<th>SrcIP</th>
<th>SrcPort</th>
<th>DestIP</th>
<th>DestPort</th>
<th>Packets</th>
</tr>
</thead>
<tbody>
<tr>
<td>168.125.253.2</td>
<td>80</td>
<td>147.81.124.1</td>
<td>3157</td>
<td>40</td>
</tr>
<tr>
<td>39.109.219.43</td>
<td>7310</td>
<td>142.68.22.108</td>
<td>59959</td>
<td>120</td>
</tr>
<tr>
<td>35.187.130.82</td>
<td>161</td>
<td>213.48.19.68</td>
<td>22</td>
<td>83</td>
</tr>
</tbody>
</table>

### Anonymised data $R$

<table>
<thead>
<tr>
<th>SrcIP</th>
<th>SrcPort</th>
<th>DestIP</th>
<th>DestPort</th>
<th>Packets</th>
</tr>
</thead>
<tbody>
<tr>
<td>168.125.253.0</td>
<td>1023</td>
<td>10.1.1.1</td>
<td>65635</td>
<td>42</td>
</tr>
<tr>
<td>39.109.219.0</td>
<td>65535</td>
<td>10.1.1.1</td>
<td>65535</td>
<td>132</td>
</tr>
<tr>
<td>35.187.130.0</td>
<td>1023</td>
<td>10.1.1.1</td>
<td>0</td>
<td>81</td>
</tr>
</tbody>
</table>
Estimating risk

High Risk of re-identification

\[ H = 1.2 \]

Shannon entropy: Average # of binary questions to identify \( s \)

Small: risky

Low Risk of re-identification

\[ H = 3.7 \]

\[ H = 4.9 \]

Large: safe
Entropy as a risk measure

Shannon entropy: Average # of binary questions to identify a single $s$

$$H(R|s) = - \sum_{r \in R} P(r|s) \log_2 P(r|s)$$

Global risk:
Expected number of correct matches

$$E_{CM} = \sum_{s \in S} \frac{1}{2^{H(R|s)}}$$

k-anonymity condition
Some properties

• Directly linked to information loss (utility):

\[ I(S, R) = H(R) - \sum_{s \in S} P(s) H(R|s) \]

  – Minimal info loss:

\[ \sum_{s \in S} P(s) \log_2 H(R|s) \quad \text{with constraint } H(R|s) \geq h_{min} \]

• Additivity

\[ H(R_1, R_2|s) = H(R_1|s) + H(R_2|R_1, s) \]
Case study: flow

- ndfomp testing dataset provided by FLAIM group
- 10000 records
- Src/Dst IPs, Src/Dst ports, Bytes used
Risk as the percentage of expected correct matches

<table>
<thead>
<tr>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
<th>Scenario 4</th>
<th>Scenario 5</th>
<th>Scenario 6</th>
<th>Scenario 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM (32) IPs</td>
<td>BM (24) IPs</td>
<td>BM (16) IPs</td>
<td>BM (16) IPs</td>
<td>BM (16) IPs</td>
<td>BM (16) IPs</td>
<td>BM (16) IPs</td>
</tr>
<tr>
<td>C ports</td>
<td>C ports</td>
<td>Classify ports</td>
<td>Noise Bytes</td>
<td>Noise Bytes</td>
<td>Noise Bytes</td>
<td>Noise Bytes</td>
</tr>
<tr>
<td>Risk (%)</td>
<td>Risk (%)</td>
<td>Risk (%)</td>
<td>Risk (%)</td>
<td>Risk (%)</td>
<td>Risk (%)</td>
<td>Risk (%)</td>
</tr>
<tr>
<td>97.94%</td>
<td>97.94%</td>
<td>61.78%</td>
<td>53.20%</td>
<td>49.9%</td>
<td>3.98%</td>
<td>0.73%</td>
</tr>
</tbody>
</table>

- Original data
- Black Marker IPs (16 bits)
- BM (16) IPs
- Classify ports
- BM (16) IPs
- C ports
- Noise Bytes
- BM (24) IPs
- C ports
- Noise Bytes
- Random Perm IPs
- BM (32) IPs
- C ports
- Noise Bytes
Final remarks

Quantifying disclosure risk is essential for finding the optimal trade-off between privacy and utility.

Measure disclosure risk using entropy:

- General: applicable to any anonymization algorithm (unlike k-anonymity)
- Stable: depends on shape of the distribution
- Linked to Information Theory

Future works (a lot...):

- More realistic testing (larger dataset, correlation across fields/records)
- Utility, Optimisation, ...
Thanks for the attention

Michele.bezzi@accenture.com