An Open Source Tool for Game Theoretic Health Data De-Identification

Session #35: Oral Presentations - Privacy & De-identification

Fabian Prasser¹, James Gaupp², Zhiyu Wan², Weiyi Xia², Yevgeniy Vorobeychik², Murat Kantarcioglu³, Klaus Kuhn¹, Brad Malin²

¹ Technical University of Munich, Munich, Germany
² Vanderbilt University, Nashville, Tennessee, USA
³ University of Texas at Dallas, Richardson, Texas, USA
Disclosure

I disclose that neither I nor my partner have relevant financial relationships with commercial interests.
Motivation

Modern approaches to medical research
• Big data: high case numbers, detailed characterizations
• Re-use and sharing of data

Initiatives to improve reproducibility of research and availability of data
• NIH Statement on Sharing Research Data, Notice NOT-OD-03-032; 2003
• NIH Genomic Data Sharing Policy, Notice NOT-OD-14-124; 2014
• EMA Policy 0070 on Publication of Clinical Data for Medicinal Products for Human Use; 2014

De-identification is complex; involves many aspects
• Comprehensive and versatile tools are needed
### De-Identification of Structured Data

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>ZIP</th>
<th>Weight</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>Male</td>
<td>81539</td>
<td>71</td>
<td>C25.0 Malignant neoplasm of head of pancreas</td>
</tr>
<tr>
<td>76</td>
<td>Male</td>
<td>81675</td>
<td>80</td>
<td>C25.0 Malignant neoplasm of head of pancreas</td>
</tr>
<tr>
<td>66</td>
<td>Male</td>
<td>81929</td>
<td>85</td>
<td>C25.0 Malignant neoplasm of head of pancreas</td>
</tr>
<tr>
<td>81</td>
<td>Male</td>
<td>80802</td>
<td>79</td>
<td>C25.1 Malignant neoplasm of body of pancreas</td>
</tr>
<tr>
<td>74</td>
<td>Male</td>
<td>81249</td>
<td>88</td>
<td>C25.2 Malignant neoplasm of tail of pancreas</td>
</tr>
<tr>
<td>71</td>
<td>Female</td>
<td>80335</td>
<td>69</td>
<td>C18.2 Malignant neoplasm of ascending colon</td>
</tr>
<tr>
<td>64</td>
<td>Female</td>
<td>80339</td>
<td>71</td>
<td>C18.4 Malignant neoplasm of transverse colon</td>
</tr>
<tr>
<td>69</td>
<td>Male</td>
<td>80637</td>
<td>75</td>
<td>C18.7 Malignant neoplasm of sigmoid colon</td>
</tr>
<tr>
<td>55</td>
<td>Female</td>
<td>80638</td>
<td>77</td>
<td>C18.7 Malignant neoplasm of sigmoid colon</td>
</tr>
<tr>
<td>61</td>
<td>Male</td>
<td>81667</td>
<td>67</td>
<td>C18.7 Malignant neoplasm of sigmoid colon</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>ZIP</th>
<th>Weight</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>72.0</td>
<td>Male</td>
<td>81***</td>
<td>[80, 90]</td>
<td>C25.- Malignant neoplasm of pancreas</td>
</tr>
<tr>
<td>72.0</td>
<td>Male</td>
<td>81***</td>
<td>[80, 90]</td>
<td>C25.- Malignant neoplasm of pancreas</td>
</tr>
<tr>
<td>72.0</td>
<td>Male</td>
<td>81***</td>
<td>[80, 90]</td>
<td>C25.- Malignant neoplasm of pancreas</td>
</tr>
<tr>
<td>62.7</td>
<td>---</td>
<td>80***</td>
<td>[70, 80]</td>
<td>C18.- Malignant neoplasm of colon</td>
</tr>
<tr>
<td>62.7</td>
<td>---</td>
<td>80***</td>
<td>[70, 80]</td>
<td>C18.- Malignant neoplasm of colon</td>
</tr>
<tr>
<td>62.7</td>
<td>---</td>
<td>80***</td>
<td>[70, 80]</td>
<td>C18.- Malignant neoplasm of colon</td>
</tr>
</tbody>
</table>

**De-Identification Techniques**
- **Sampling**
- **Aggregation**
- **Deletion**
- **Masking**
- **Categorization**
- **Generalization**

---

**AMIA 2017**

Prasser et al.: An Open Source Tool for Game Theoretic Health Data De-Identification
The Risk-Utility Trade-Off

Two conflicting objectives need to be balanced
- Minimization of re-identification risks
- Minimization of reductions in data utility

Objectives need to be quantified
- **Risk**: several models, e.g. based on uniqueness
- **Utility**: several models, e.g. based on entropy

Traditional solution: simplification
- **Privacy model**: implements risk threshold
- Maximization of output data utility

![Diagram showing the trade-off between re-identification risk and data utility.](Image)
ARX Data Anonymization Tool

Comprehensive
• Many risk and transformation models
• (Almost) arbitrary combinations

Scalable
• Millions of records

Cross-platform
• GUI and library

Open source
• Apache 2

Semi-automated
• Four perspectives
Game Theoretic Data De-Identification

Stackelberg game models re-identification from an economic perspective

• Publishers gain *benefit* from sharing data at a certain degree of utility
• Publishers *lose* money for successful re-identification attacks
• Attackers *gain* benefit from re-identifying records
• Attackers need to bear *costs* for launching a re-identification attack

Publisher payout can be maximized by assuming a rational adversary

\[
Payout(r) = Benefit(r) - Attack(r) \cdot SP(r) \cdot Loss
\]

\[
Benefit(r) = (1 - IL(r)) \cdot Benefit
\]

\[
Attack(r) = \begin{cases} 
1, & \text{if } SP(r) \cdot Gain > Cost \\
0, & \text{otherwise}
\end{cases}
\]

[Optimization function]

Re-identification risk = success probability

Information loss = reduction in utility

Re-identification risk = success probability

[Wan, et al. PLOS ONE. 2015]
Objectives, Challenges and Results

Objective: to integrate the game theoretic model into ARX
Challenge 1: to transform the approach into ARX’s de-identification model

Privacy model
Captures profitability: introduces a threshold at zero payout

Utility model
Summarizes overall payout excluding non-profitable records

Also: Specific variants of the game can be solved by using traditional privacy models

Challenge 2: to achieve scalability

Solutions can be pruned from the search process by exploiting properties of publisher payout under the assumption that the adversary never attacks

Result: full integration into ARX
Enables new methods of solving the game
Evaluation

Populations-based risk  Dataset-based risk  Safe Harbor

Setup
- ~32k U.S. Census records
- Demographic attributes
- Adversary cost: $4
- Publisher benefit: $1,200

Results
- The game theoretic model typically outperforms HIPAA Safe Harbor
- Implementation details are less important when (a) a population table is used or (b) the benefit of publishing is significantly higher than the potential loss.
Thank you!

fabian.prasser@tum.de
arx.deidentifier.org
Question 1

Traditionally, de-identification is considered an optimization problem that directly balances privacy with data quality. The game theoretic model of de-identification indirectly achieves this balance using what proxy properties?

A. Research utility and record linkage
B. Adversary's cost, adversary's gain, publisher's benefit and publisher's loss
C. Encryption strength, processing time, hardware cost and network latency
D. Patient risk, provider risk and adversary’s success rate
Answer 1

A. Research utility and record linkage
B. Adversary's cost, adversary's gain, publisher's benefit and publisher's loss
C. Encryption strength, processing time, hardware cost and network latency
D. Patient risk, provider risk and adversary's success rate

Explanation: Game theory relies on the properties of adversary's cost, adversary's gain, publisher's benefit and publisher's loss to maximize the publisher's net payout over all published records. Increasing benefit and gain are proxies for increasing data quality, while increasing cost and loss are proxies for increasing privacy. Though these user defined properties are fixed prior to analysis, they are balanced across information loss and likelihood of attack that vary with levels of generalization and suppression.
Question 2

The solution space for a de-identification problem, with even a modest number of attributes, can be huge. As such, the processing time required to achieve de-identification can be considerable. What strategies have been implemented in ARX to improve scalability?

A. Full-domain generalization and search space pruning
B. Record level generalization and adversary cost analysis
C. l-Diversity
D. Differential privacy
Answer 2

A. Full-domain generalization and search space pruning
B. Record level generalization and adversary cost analysis
C. l-Diversity
D. Differential privacy

Explanation: Full-domain generalization applies the same level of generalization to all values of a given attribute. Efficiency comes from transforming groups of record as equivalence classes, rather than performing independent transformations on individual records. Search space pruning eliminates a given attribute generalization and all derived, increasing generalizations under the assumption that data quality monotonically decreases with increasing generalization.