SafePub: A Truthful Data Anonymization Algorithm With Strong Privacy Guarantees

Raffael Bild, Klaus A. Kuhn, and Fabian Prasser
Institute of Medical Informatics, Statistics and Epidemiology
Technical University of Munich
Barcelona, 25.07.2018
Background

- **Statistical Disclosure Control**
  - A posteriori approaches to data privacy
  - Extensively used in statistics
  - Methods include random sampling, modification, summarization, perturbation

- **Syntactical Data Anonymization**
  - Data is modified so that syntactic requirements are satisfied
  - „Traditional“ approach in computer science
  - Examples of syntactic privacy models: k-anonymity, l-diversity, t-closeness
  - Data anonymization algorithms balance privacy protection against utility (quantified by models)

- **Differential Privacy**
  - Not a property of a dataset, but of a data processing method
  - Strong degree of privacy protection
  - Gold standard in academia
  - Methods include the Laplace mechanism and the exponential mechanism
  - Increasingly used in practice, e.g. by Google and Apple
Motivation

- The ARX Data Anonymization Tool provides various privacy models, quality models and transformation techniques
- Release of microdata allows to perform flexible analyses
- Truthfulness of data desirable in many fields, including the medical domain
- **Goal:** Integrate differentially private data anonymization which
  - produces truthful microdata
  - integrates well with existing methods
Safe-Pub: High-level overview

- Based the mechanism \((k, \beta)\)-SDGS by Li et al.
- Satisfies \((\epsilon, \delta)\)-differential privacy

**Overview**

1. Random sampling (parameter \(\beta\))
2. K-Anonymization (parameter \(k\))
   - Attribute generalization
   - Record suppression

- Has only been studied from a purely theoretical perspective. Focus: Calculation of \(\epsilon\) and \(\delta\) resulting from \(\beta\) and \(k\)

Safe-Pub: Challenges

Challenge 1: How to calculate $\beta$ and $k$?

Challenge 2: How to determine a suitable scheme?

Challenge 3: Output data utility?
Challenge 1: Calculation of Parameters

Inversion of the following formulas:

\[
\varepsilon = - \ln(1 - \beta),
\]

\[
\delta = \max_{n: n \geq nm} a_n = \max_{n: n \geq nm} \sum_{j > \gamma n} \binom{n}{j} \beta^j (1 - \beta)^{n-j}
\]

where \( n_m = \left\lceil \frac{k}{\gamma} - 1 \right\rceil \) and \( \gamma = \frac{e^{\varepsilon - 1} + \beta}{e^{\varepsilon}} \)

Challenge: The sequence \( a_n \) is non-monotonic

Solution: Exploit sequence \( c_n = e^{-n(\gamma \ln \left( \frac{\gamma}{\beta} \right) - (\gamma - \beta))} \geq a_n \)

which is monotonic to determine \( \max_{n: n \geq nm} a_n \)
Challenge 2: Selection of a Generalization Scheme

$\varepsilon$-differentially private search strategy can be used

Challenges:
- No search strategy described

Solution:
- Differentially private implementation of a typical search-based anonymization algorithm
- Greedy search through all possible combinations of generalization levels (lattice)
- Repeated applications of the exponential mechanism guided by score functions capturing utility

Challenge 3: Utility of Data – Score Functions

Score functions tailored to general purpose quality models

- Data Granularity (cell-level)
- Generalization Intensity (cell-level)
- Discernibility (record-level)
- Group Size (record-level)
- Non-Uniform Entropy (attribute-level)

Workload-aware score function tailored to statistical classification

- Based on the special-purpose model proposed by Iyengar
Challenge 3: Utility of Data – Evaluation

Parameterization:

- A value of $\varepsilon$ in the order of one is recommendable
- A value of about 300 search steps is recommendable
- Small privacy budget in the order of 0.1 sufficient for the search
- It has been suggested to choose $\delta$ depending on the size $n$ of the dataset so that $\delta < \frac{1}{n}$ holds
Comparison of classification accuracies with prior work:
1-differential privacy: DiffGen, DiffP-C4.5, LDA, SDQ and DPNB
\((1, 10^{-\{9\ldots14\}})\)-differential privacy: SafePub, Fouad et al.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>DiffP-C4.5</th>
<th>LDA</th>
<th>DPNB</th>
<th>DPNB</th>
<th>SDQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>US Census</td>
<td>Nursery</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Competitor</td>
<td>82.1%</td>
<td>80.8%</td>
<td>82%</td>
<td>90%</td>
<td>79.9%</td>
</tr>
<tr>
<td>SafePub</td>
<td>80.9%</td>
<td>81.5%</td>
<td>81.2%</td>
<td>83.7%</td>
<td>83.8%</td>
</tr>
</tbody>
</table>

**Raffael Bild | The 18th PETS | SafePub: A Truthful Data Anonymization Algorithm With Strong Privacy Guarantees**

**Challenge 3: Utility of Data – Evaluations**

**US Census**

**Health interviews**

**Class attr. 1**

**Class attr. 2**
Conclusions

- SafePub can compete with state-of-the-art
- The method is simple and easy to parameterize
- To achieve truthfulness, $(\varepsilon, \delta)$-differential privacy must be implemented
- Various directions for future research:
  - Investigate more flexible data transformation techniques
  - Consider the effects of random sampling performed during data acquisition to reduce the amount of explicit random sampling