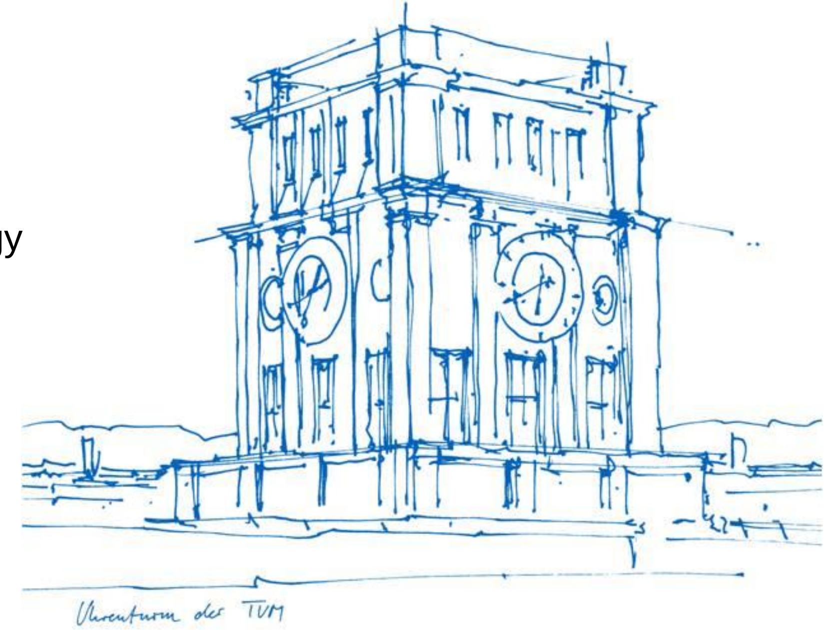
SafePub: A Truthful Data Anonymization Algorithm With Strong Privacy Guarantees

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

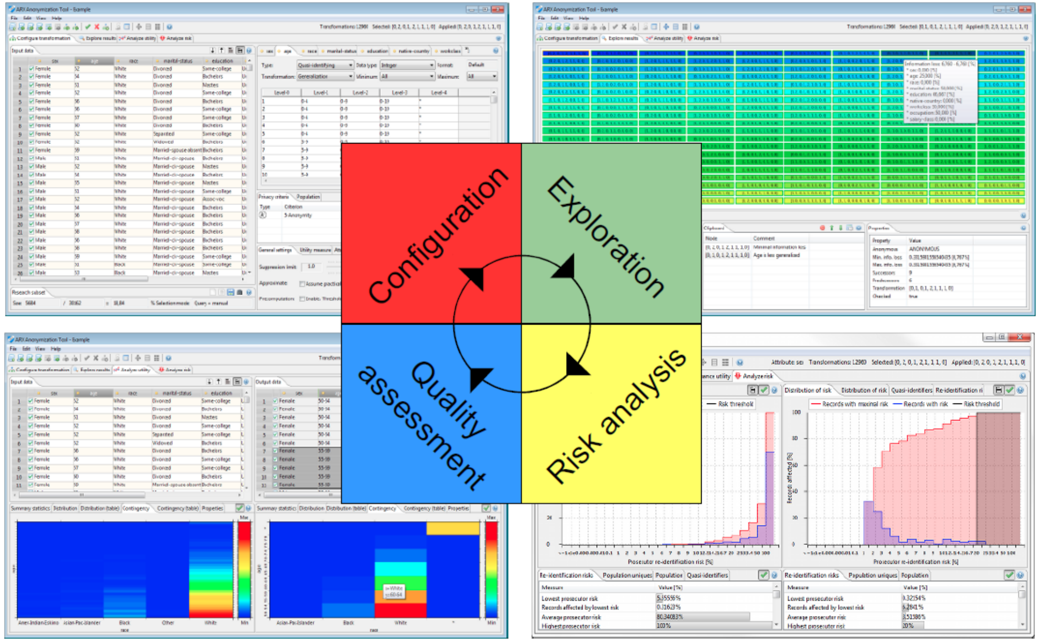
- **Syntactical Data Anonymization**
  - Data is modified so that syntactic requirements are satisfied
  - „Traditional“ approach in computer science
  - Examples of syntactic privacy models: k-anonymity, l-diversity, t-closeness
  - Data anonymization algorithms balance privacy protection against utility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

ARX Data Anonymization Tool



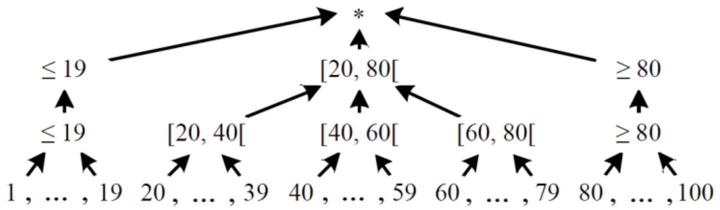
Safe-Pub: High-level overview

- Based the mechanism (k,β)-SDGS by Li et al.

- Satisfies (ε,δ)-differential privacy

- Overview
  1. Random sampling (parameter β)
  2. K-Anonymization (parameter k)
     – Attribute generalization
     – Record suppression

- Has only been studied from a purely theoretical perspective. Focus: Calculation of ε and δ resulting from β and k



Input data

Age	Gender	Zipcode	Income
20	Male	81667	2000
20	Female	81675	2000
50	Male	81925	2500
55	Male	81975	2800
37	Female	82567	3000
40	Female	81931	3000
40	Female	81931	3000

Random sampling

Random sample

Age	Gender	Zipcode	Income
20	Male	81667	2000
20	Female	81675	2000
50	Male	81925	2500
*	*	*	*
37	Female	82567	3000
*	*	*	*
40	Female	81931	3000

Attribute generalization

Generalized random sample

Age	Gender	Zipcode	Income
[20-40[	*	816--	≤ 2000
[20-40[	*	816--	≤ 2000
[40-60[	*	819--	> 2000
*	*	*	*
[20-40[	*	825--	> 2000
*	*	*	*
[40-60[	*	819--	> 2000

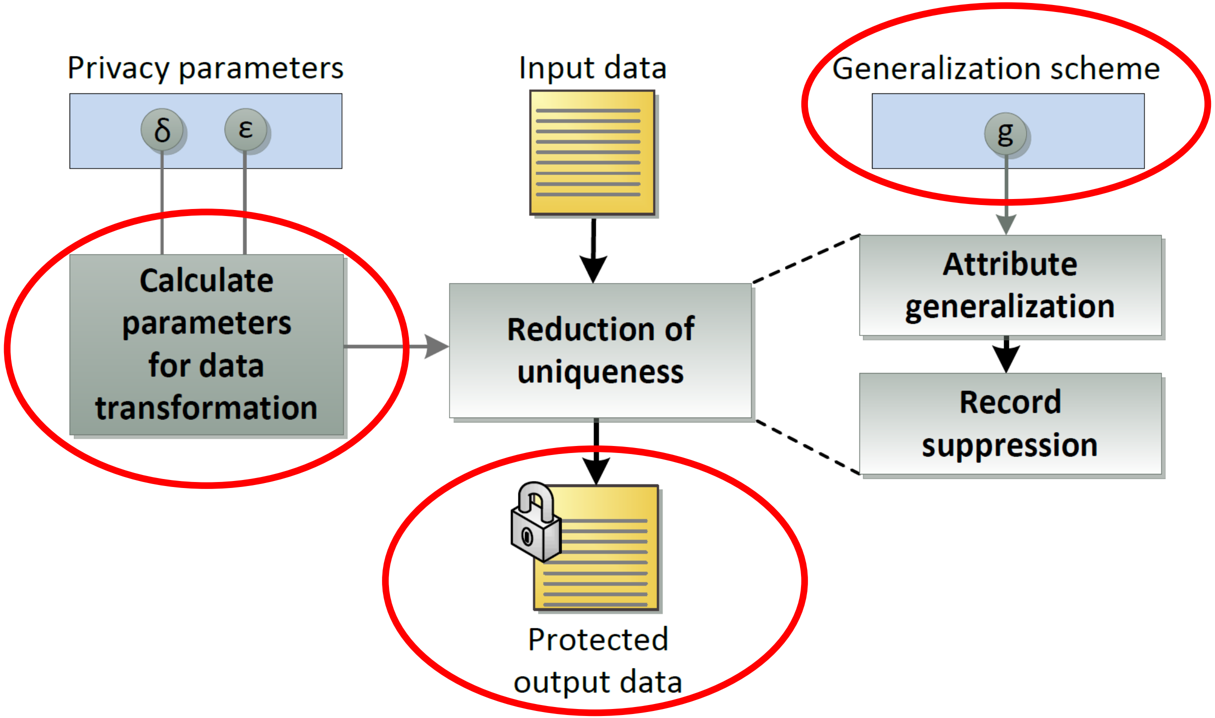
Record suppression

Output dataset

Age	Gender	Zipcode	Income
[20-40[	*	816--	≤ 2000
[20-40[	*	816--	≤ 2000
[40-60[	*	819--	> 2000
*	*	*	*
*	*	*	*
*	*	*	*
[40-60[	*	819--	> 2000

N. Li et al. On sampling, anonymization, and differential privacy: Or, k-anonymization meets differential privacy. In *ACM Symp. Information, Computer and Communications Security*, pages 32–33, 2012.

Safe-Pub: Challenges



Challenge 1:
How to calculate
β 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:

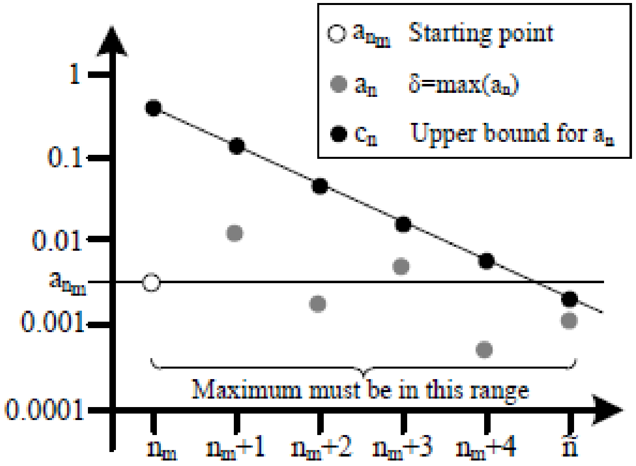
$$\varepsilon = -\ln(1-\beta),$$

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

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 $c_n = e^{-n\left(\gamma \ln\left(\frac{\gamma}{\beta}\right) - (\gamma - \beta)\right)} \geq an$ which is monotonic to determine $\max_{n:\, n \geq nm} a_n$



Challenge 2: Selection of a Generalization Scheme

ε-differentially private search strategy can be used

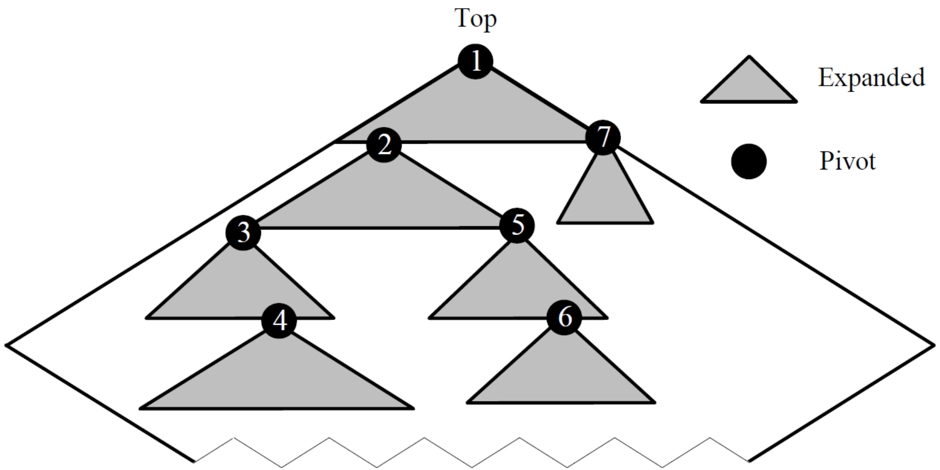
N. Li et al. On sampling, anonymization, and differential privacy: Or, k-anonymization meets differential privacy. In *ACM Symp. Information, Computer and Communications Security*, pages 32–33, 2012.

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

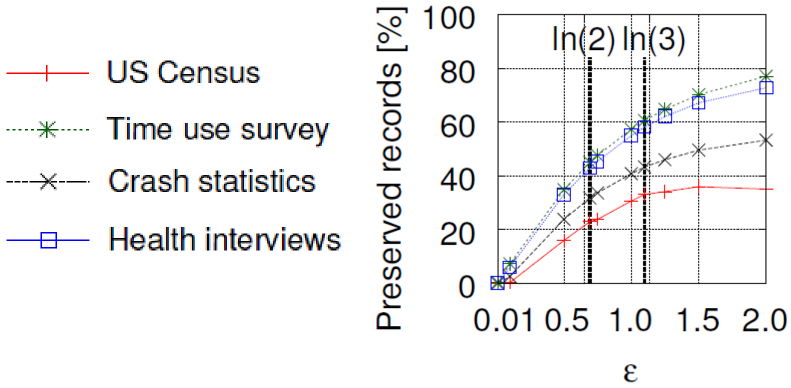
- Data Granularity (cell-level)
- Generalization Intensity (cell-level)
- Discernibility (record-level)
- Group Size (record-level)
- Non-Uniform Entropy (attribute-level)

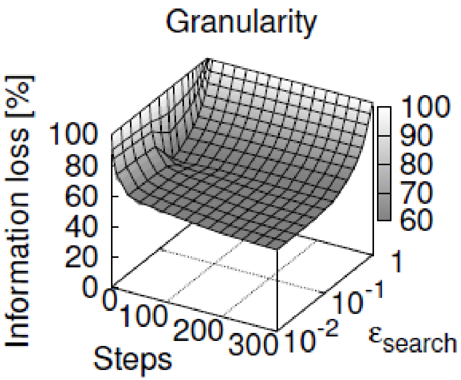
Workload-aware score function tailored to statistical classification

- Based on the special-purpose model proposed by Iyengar

Challenge 3: Utility of Data – Evaluation

Parameterization:





- A value of ε 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 δ depending on the size n of the dataset so that $\delta < \frac{1}{n}$ holds

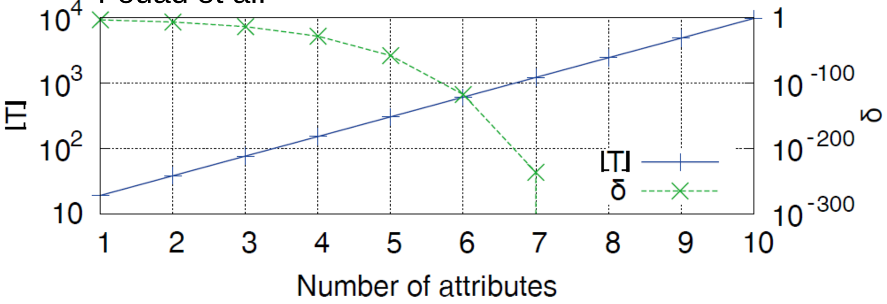
Challenge 3: Utility of Data – Evaluations

Comparison of classification accuracies with prior work:

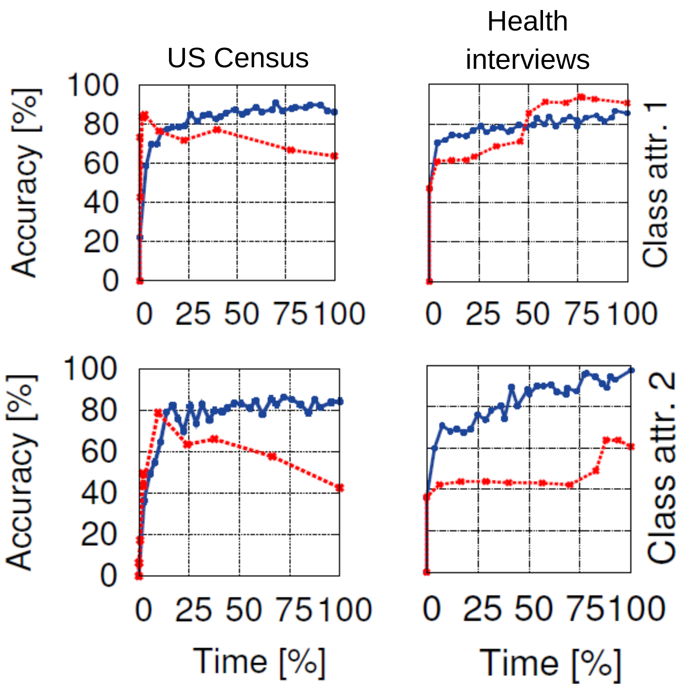
1-differential privacy: DiffGen, DiffP-C4.5, LDA, SDQ and DPNB

$(1,10^{-\{9\ldots14\}})$-differential privacy: SafePub, Fouad et al.

Fouad et al:



Algorithm	DiffP-C4.5	LDA	DPNB	DPNB	SDQ
Dataset	US Census			Nursery	
Competitor	82.1%	80.8%	82%	90%	79.9%
SafePub	80.9%	81.5%	81.2%	83.7%	83.8%



Conclusions

- SafePub can compete with state-of-the-art
- The method is simple and easy to parameterize
- To achieve truthfulness, (ε,δ)-differential privacy must be implemented
- Various direcetions 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