
SBA
Research

Anonymisation and Fingerprinting of Microdata:
A Genetic Algorithm for Finding Optimal Set for Data Distribution

Sarcevic T., Mayer R., *

SBA Research, Vienna, Austria

*Carmen Halbeisen, Mariusz Nitecki and Alexander Kurt

Bundesministerium Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie
Bundesministerium Digitalisierung und Wirtschaftsstandort
FFG Forschung wirkt.
wirtschaftsagentur wien Ein Fonds der Stadt Wien
European Commission
FWF Der Wissenschaftsfonds.
netidee OPEN INNOVATIONS

Motivation

- **Tracing the unauthorised usage of anonymised data**
- Our objective is to enable a system that allows:
  1. sharing *k*-anonymous data with multiple recipients
  2. tracing the shared copies in cases of unauthorised re-distribution or unauthorised usage
     - identify the recipient of the anonymised data copy
     - identify own signature to prove the ownership of the original data

Introduction

Ownership protection via a traceable marker

- **Fingerprinting**: embedding (hiding) a traceable marker into data that:
  - Identifies the owner
  - Identifies the recipient

- Fingerprinting of microdata:
  - Fingerprint generation $F(SK, id) \rightarrow [1,0]^l$
  - Fingerprint translates to a pattern of modifications of data values

Original data

Name	Sex	Birthdate	Disease
Bob	M	19.03.1970	Chest pain
Dave	M	20.03.1970	Short breath
Alice	F	18.04.1970	Obesity
Eve	F	21.04.1970	Short breath

Fingerprinted data 1

Name	Sex	Birthdate	Disease
Bob	M	21.03.1970	Chest pain
Dave	M	20.03.1970	Short breath
Alice	F	18.04.1970	Short breath
Eve	F	21.04.1970	Short breath

Fingerprinted data 2

Name	Sex	Birthdate	Disease
Bob	M	21.03.1970	Chest pain
Dave	M	20.03.1970	Short breath
Alice	F	18.04.1970	Chest pain
Eve	F	21.04.1970	Short breath

Introduction

Fingerprinting & anonymisation

- Requirements:
  - Allow sharing *k*-anonymised data
  - Trace unauthorised re-distribution of data

- Idea: „reuse“ the modification from anonymisation for tracing
- Fingerprint: *unique generalisation pattern*

Original data

Name	Sex	Birthdate	Disease
Bob	M	19.03.1970	Chest pain
Dave	M	20.03.1970	Short breath
Alice	F	18.04.1970	Obesity
Eve	F	21.04.1970	Short breath

2-anonymous set 1 [0,2]

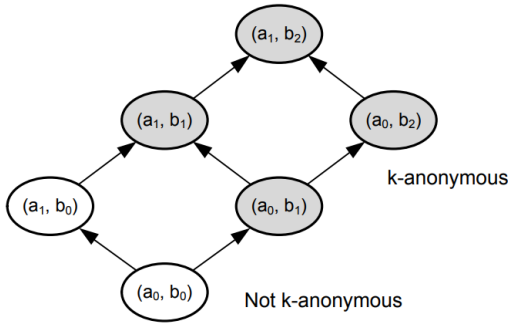
Name	Sex	Birthdate	Disease
*	M	1970	Chest pain
*	M	1970	Short breath
*	F	1970	Obesity
*	F	1970	Short breath

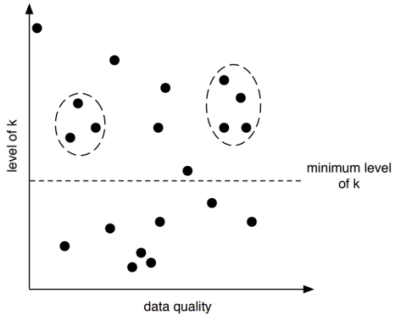
2-anonymous set 2 [1,1]

Name	Sex	Birthdate	Disease
*	*	03.1970	Chest pain
*	*	03.1970	Short breath
*	*	04.1970	Obesity
*	*	04.1970	Short breath

Fingerprinting & anonymisation: the algorithm

1. Obtain all generalisation patterns given the dataset and hierarchies
2. Discart those not satisfying *k*-anonymity
3. Cluster the fitting generalisation patterns on their max *k* and *some* utility metric
4. Treat the cluster as a group of equally good solutions for distribution
5. Collusion filter

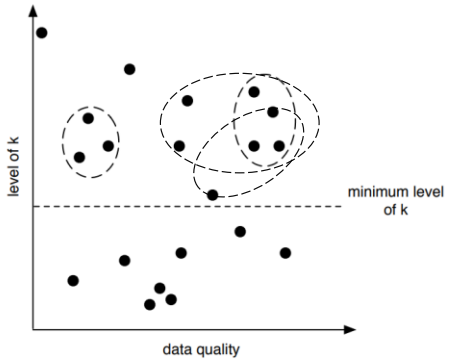




Kieseberg, Peter, et al. "An algorithm for collusion-resistant anonymization and fingerprinting of sensitive microdata." *Electronic Markets* 24.2 (2014): 113-124.

Fingerprinting & anonymisation: the algorithm

1. Obtain all generalisation patterns given the dataset and hierarchies
2. Discart those not satisfying *k*-anonymity
3. **Cluster the fitting generalisation patterns on their max *k* and *some* utility metric**
4. Treat the cluster as a group of equally good solutions for distribution
5. Collusion filter



Select a subset of generalisation patterns (anonymous data sets) where:

- All elements satisfy minimum privacy criterium
- The elements have similarly good utility
- The subset is large *enough* for sharing with recipients

Problem statement

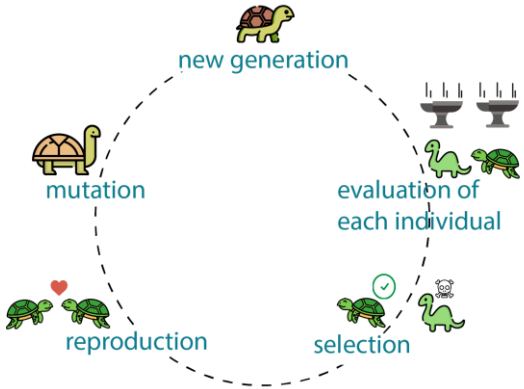
- **Problem statement**: find a subset of *n(S) k*-anonymous datasets $S = \{s_1, s_2, \dots, s_{n(S)}\}$ such that:
  - Total utility loss, $loss(S) = \frac{\sum_i^n s_i}{n}$ is minimised
  - Variation coefficient of the loss, $\mathrm{cv(S)} = \mathrm{cv}(loss(s_1), loss(s_2), \dots, loss(s_n))$ is minimised
  - Cardinality *n(S)* is maximised

- **Objective function**: $\min f(S) = 1 - \frac{n - n_{min}}{n_{max} - n_{min}} + loss(S) + cv(S)$

Genetic Algorithm

- **Genetic algorithm (GA)**: metaheuristic for solving optimisation problems inpired by natural selection

1. Initialisation
2. Evaluation using *fitness function*
3. Selection
4. Crossover / Reproduction
5. Mutation

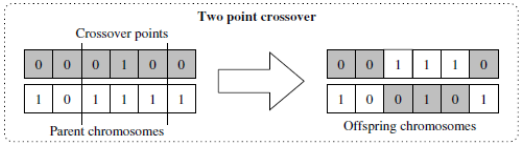


https://www.generativedesign.org/02-deeper-dive/02-04_genetic-algorithms/02-04-01_what-is-a-genetic-algorithm

Solution with GA: formulation

transformation	LM	sup. records	Precision	AES
[0,4,0,0,1,0,0,1,0]	0.2210	9.77%	0.2037	28.3195
...	...			

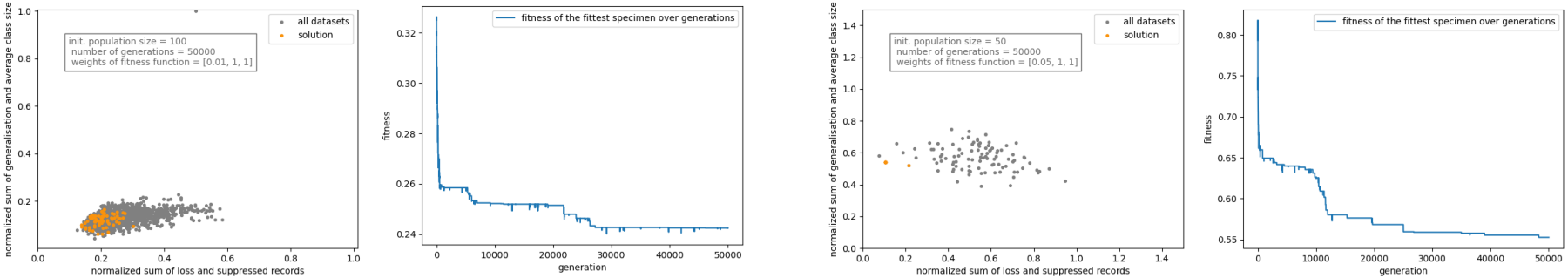
- **Fitness function**: $\min f(S) = 1 - \frac{n - n_{min}}{n_{max} - n_{min}} + loss(S) + cv(S)$
- **Genotype** (chromosomes) of a candidate solution $S = \{s_1, s_2, \dots, s_n\}$: $[1,0]^{n_{max}}$
- **Crossover**: two-point crossover



- **Mutations**:
  - Replacing gene within the candidate solution: with a probability of 40% we replace genes from the candidate solution with new data points from the dataset. The amount of replaced data points is between zero and a third of the size of the candidate solution (the amount is chosen randomly)
  - Adding/removing a gene: with a probability of 60% we add a new gene or delete one. The probability of choosing to add/remove a random data point from the dataset to the candidate solution is 50%.

Results

- Adult Census: search space of 1,197 3-anonymous data sets
- *Flash k*-anonymisation algorithm within ARX toolbox



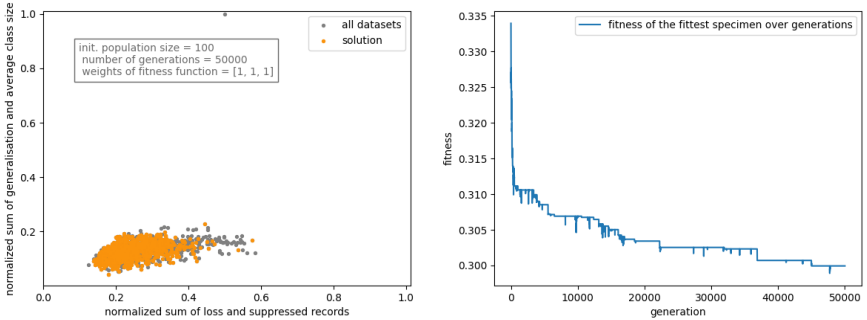
a) Full search space

b) Reduced search space: top 100 optimal datasets

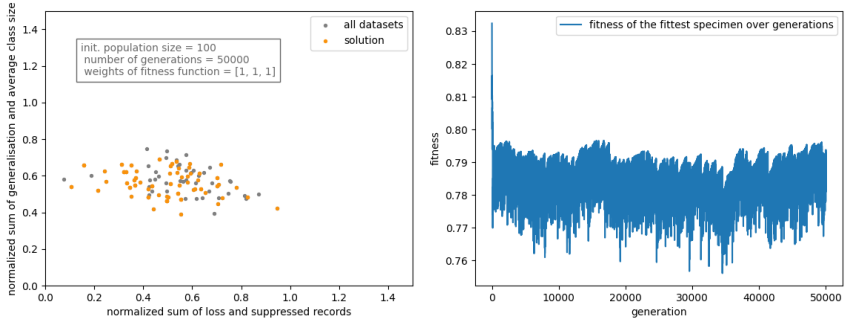
Fig1: Reduced weight of subset size $n(S)$ in the fitness funcion

Kohlmayer, Florian, et al. "Flash: efficient, stable and optimal k-anonymity." *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, 2012.

Results



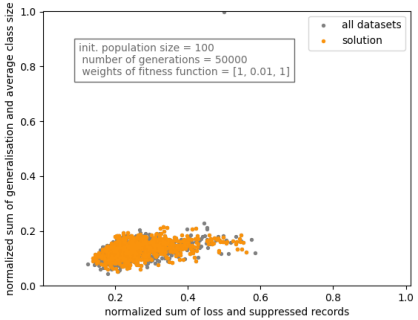
a) Full search space



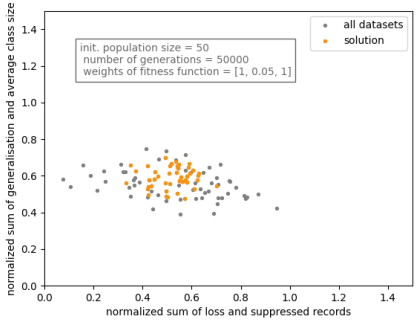
b) Reduced search space: top 100 optimal datasets

Fig2: Equal weights for subset size, total loss and variance of the subset

Results

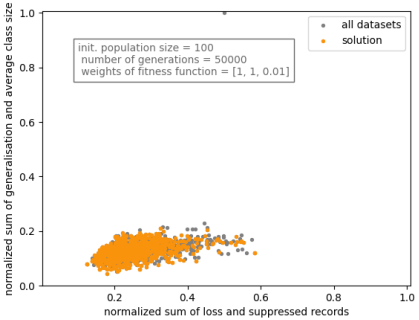


a) Full search space

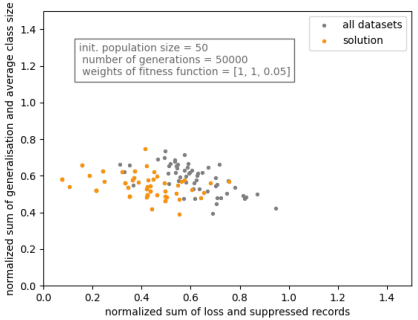


b) Reduced search space

Fig3: Less weight on total loss



a) Full search space



b) Reduced search space

Fig4: Less weight on variance coeffiecient

Concluding remarks and next steps

- Collusion filter after GA
  - Making sure that the recipients cannot avoid being traced by collaborating

- More privacy concerns for data subjects by sharing multiple anonymised copies?

- Comparison to sequential approach: *k*-anonymisation *then* fingerprinting

Merci!

Tanja Sarcevic

SBA Research, Floragasse 7, 1040 Vienna

TSarcevic@sba-research.org

Bundesministerium Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie

Bundesministerium Digitalisierung und Wirtschaftsstandort

FFG Forschung wirkt.

wirtschaftsagentur wien Ein Fonds der Stadt Wien

European Commission

FWF Der Wissenschaftsfonds.

netidee OPEN INNOVATIONS