# The Trade-off between Information Utility and Disclosure Risk in a GA Synthetic Data Generator

Yingrui Chen, Jennifer Taub, Mark Elliot
The University of Manchester

# Synthetic Data

- Statistical Disclosure Control Technique
- Publish an artificial dataset instead of the real dataset
- Having problem in finding the trade-off between information utility and disclosure risks

# Information Utility

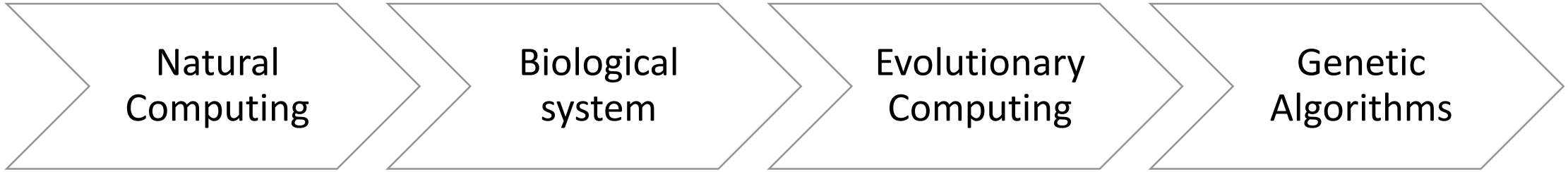
- Purdam and Elliot (2007):

  *"The loss of analytical validity as occurring when a disclosure control method has changed a dataset to the point at which a user reaches a different conclusion from the same analysis"* (p. 1102)
- Every statistical property can be considered as utility
- Utility is considered as the objective of the optimising program
- Objectives should be measureable and comparable

# Disclosure Risk

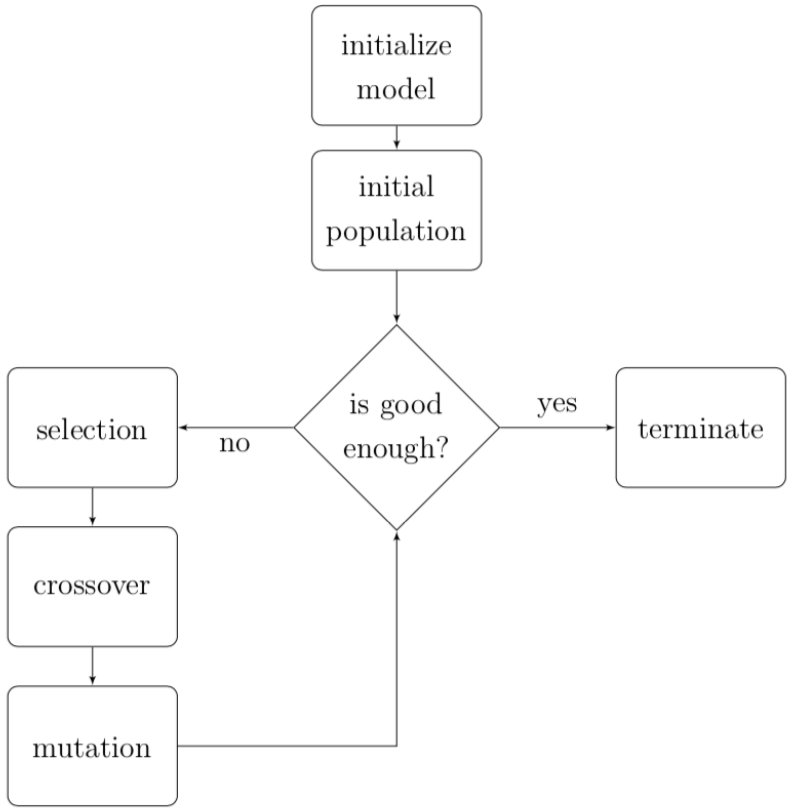
- Identification disclosure risk
  - A dataset contains identification risk if a data subject can be re-identified from the dataset.
- Attribute disclosure risk
  - A dataset contains attribution risk if sensitive information of any population units can be inferred from the dataset.

# Genetic Algorithms (GAs)




- Natural computing is computational systems inspired by natural systems
- simulating the process of organisms surviving from limited resources and predators
- It uses the principals of natural evolution and genotypic variation to solve complicated optimisation problems
- It simulates the process of natural evaluation including natural selection, crossover and mutation.

# Genetic Algorithms (GAs)




- GAs can cope with multiple objectives that may conflict with each other.
- GAs can explore the solution space of complex, high-dimensional problems like data synthesis.

# Model Design

- Initial Population
  - 100 candidates that are mutated from the original data, thus they are high in utility (and risk).
- Selection Operator:
  - Deterministic tournament selection operator with tournament size t = 2. i.e. 2 candidates are randomly selected into tournaments (with replacement) and only the winner can enter the crossover operator.

# Model Design

- Crossover Operator
  - Whole-Case Parallelised Crossover: it occurs on every case in the candidate, the case was chosen by determined crossover rate (0.1 in this paper) and it is then switched with the corresponding case in paired candidate.
- Mutation Operator
  - Uniform mutation: it gives every single element/cell in the candidate a chance (0.001 in this paper) to mutate.

# Objective Design

- Utility Objective
  - We use full contingency table to capture the variate structure for all variables in the data.
  - The distance between candidates and the original data is calculated by Jensen-Shannon Divergence:

    Suppose $P$ and $Q$ are two discrete probability distribution, $\mathrm{D}_{\mathrm{JS}}(P||Q)$ is defined by:

$$\mathrm{D}_{\mathrm{JS}}(P||Q) = \left(\frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M)\right)^{\frac{1}{2}}$$

    , where $M = \frac{1}{2}(P + Q)$ and $D_{KL}$ is Kullback-Leibler divergence.
  - The utility objective of GA is to minimise $\mathrm{D}_{\mathrm{JS}}$ between full contingency tables $CT_{FULL}$ of the synthetic and original data.

$$U(X, Y) = D_{JS}\left(\frac{CT_{FULL}(X)}{N}||\frac{CT_{FULL}(Y)}{N}\right)$$

# Objective Design

- Risk Objective
  - Elliot (2014) and Taub et al (2018) introduced a measure for disclosure risk of synthetic data called the Differential Correct Attribution Probability (DCAP), , which consists of a Correct Attribution Probability (CAP) score.

$$CAP_{s,j} = \Pr\left(T_{o,j} \middle| K_{o,j}\right)_s = \frac{\sum_{i=1}^{n}[T_{s,j} = T_{o,j}, K_{s,j} = K_{o,j}]}{\sum_{i=1}^{n}[K_{s,j} = K_{o,j}]}$$

    Where $o, s$ state original or synthetic data, the [] are Iverson brackets, $n$ is the number of records, $j$ is the index of case, and do is the original data and $K_o$ and $T_o$ as vectors for the key and target information.
  - In this paper, only the statistical uniques of the original dataset are used in calculating the CAP score for disclosure risk ($R$)

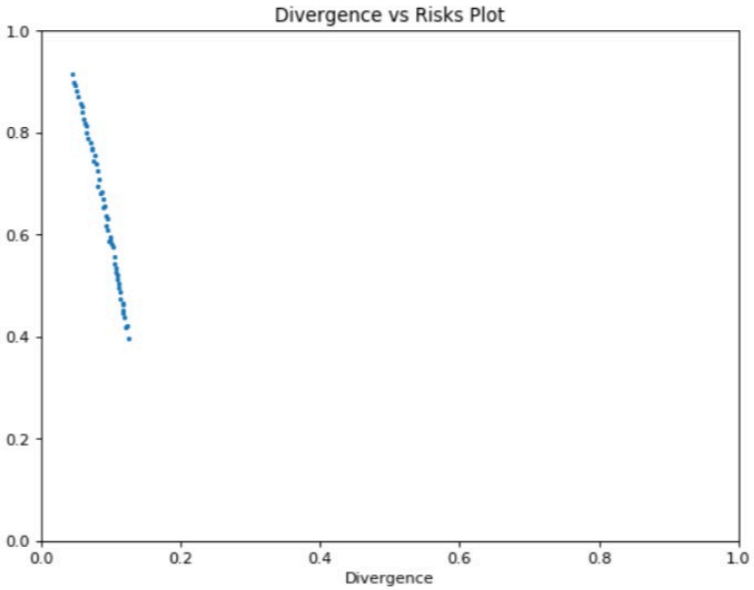
# Objective Design

- We composed the two objectives into a single function with same weights. The objectives must be normalised to avoid an objective with a larger numerical variance dominating the less varied one, otherwise the selection of weights is easy to get wrong and even a small difference in weighting can lead to different solutions.

$$F(X,Y) = \sqrt{U(X,Y) + R(X,Y)}/\sqrt{2}$$

# Experiment

- Data
  - The dataset is from the 1901 Scottish Census[1] and consists of 82,851 records.
- Process
  - As an output we synthesised a single synthetic dataset which has $R = 0.3964$ and $U = 0.1257$.
  - The process took 57 generations.
  - The right figure shows changing of risk and utility from the best candidates in every generation during optimising process.

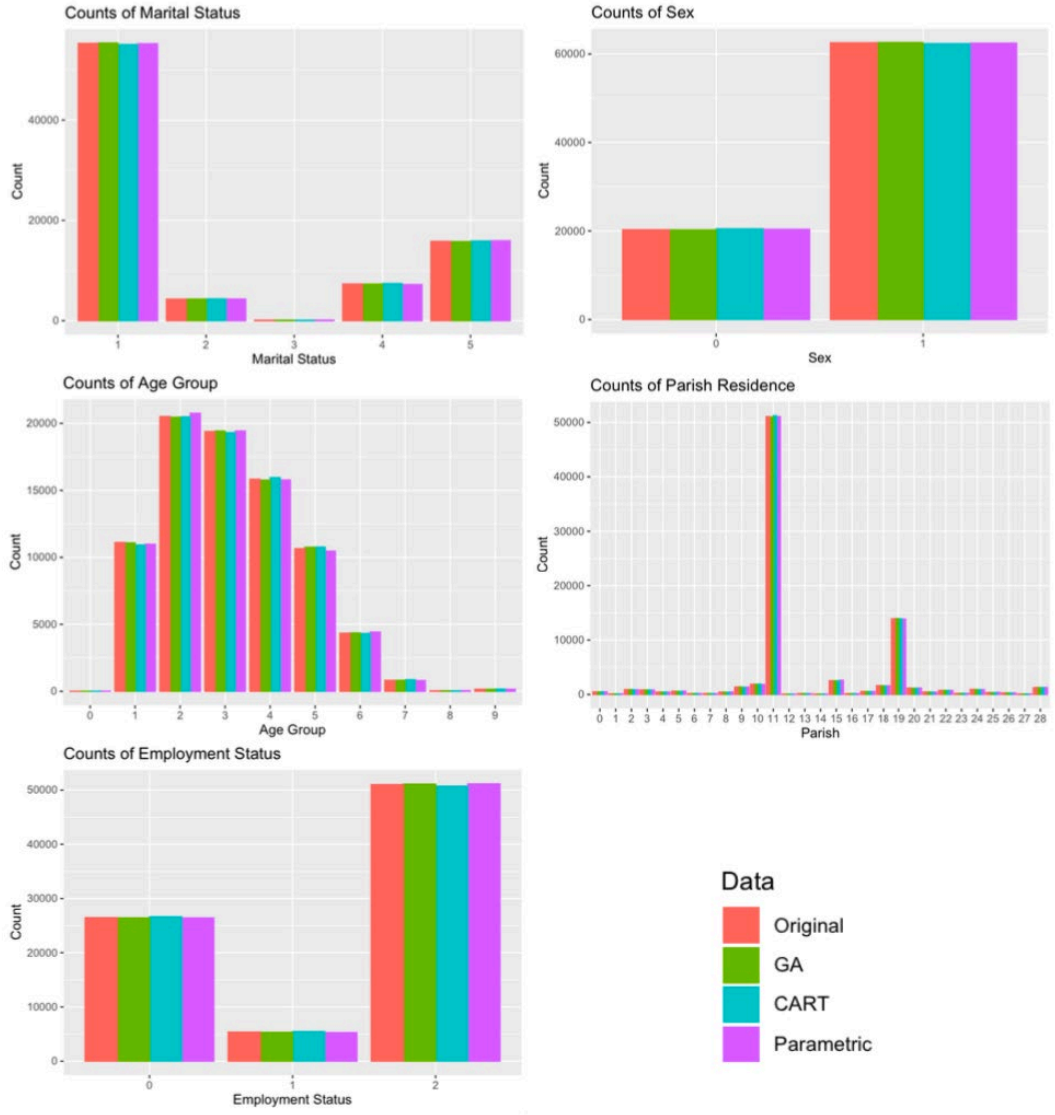



1. National Records of Scotland, (1901), 1901 Scottish Census.

# Comparisons to Other Synthesis Methods

We compared the output (GA synthetic data) with CART and parametric generated synthetic data

- Utility Comparison
  - Histogram comparison(right figure)



# Comparisons to Other Synthesis Methods

- propensity mean square error (pMSE) score comparison: the closer the pMSE is to 0 the better the data performs. In this instance all synthetic datasets have quite low pMSE scores, however the GA performs better than the CART and parametric datasets.

| Dataset | pMSE | Standardized pMSE | pMSE ratio |
|---|---|---|---|
| GA-synthetic Data | 5.44e-06 | -3.9221 | 0.1638 |
| CART-synthetic Data | 3.397e-05 | 0.1106 | 1.0236 |
| Parametric Synthetic Data | 3.17e-05 | 7.007e-06 | 0.9553 |

# Comparisons to Other Synthesis Methods

- Risk Comparison
  - Given that the baseline DCAP score for the univariate is 0.4154, the GA and parametric dataset would be considered no risk since they are below the baseline and the CART synthetic dataset would have minimal risk since it is very close to the baseline.

| Dataset | Risk |
|---|---|
| GA-synthetic Data | 0.3964 |
| CART-synthetic Data | 0.4168 |
| Parametric Synthetic Data | 0.3278 |

# Conclusion

- GAs are viable alternative to standard synthesisers.
- GAs are able to produce synthetic data that contains low disclosure risk and high information utility.
- Unlike other synthesisers, GAs could prove that it’s disclosure risk level can be pre-set, instead of being left as a post-hoc question.