Aggregation Equivalence Level for interpreting Synthetic Data Attribute Information

Lotte Pater & Sanne Smid, DUO

Paper, data & code: https://osf.io/rdpab/

Contact: lotte.pater@duo.nl; sanne.smid@duo.nl

Quantifying privacy risk: a main challenge in SynData research

REITER:

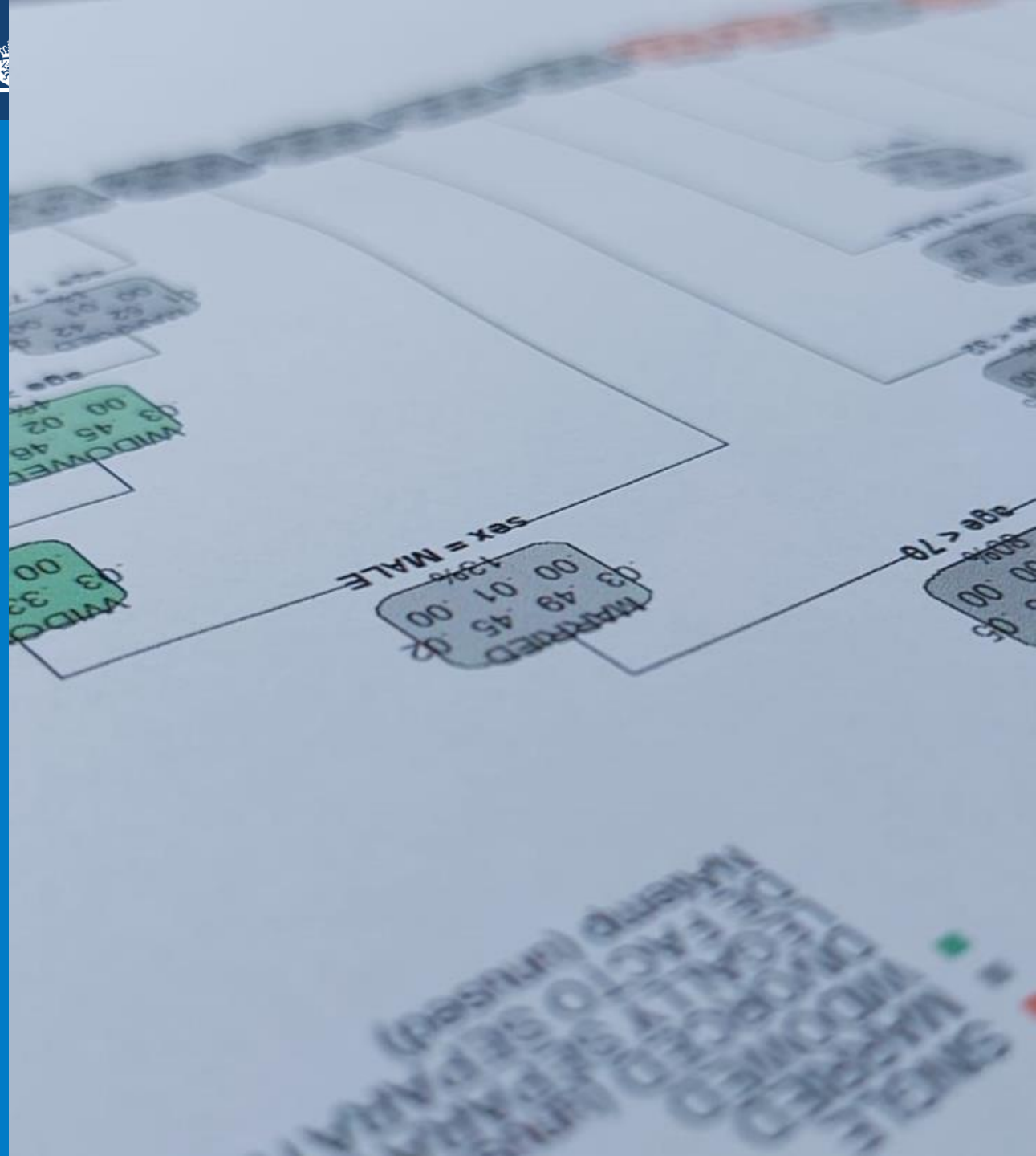
> While there is need to examine disclosure risks in synthetic data, there is no standard for doing so, especially in fully synthetic data. Instead, disclosure checks tend to be *ad hoc*. For

DRECHSLER:

> would be naïve to assume that fully synthetic data will never pose any threats of disclosing sensitive information. However, measuring these risks is challenging and research in this area is still limited.

Zooming in: Attribute Information

- Fully synthetic data: no 1-to-1 link
  - Identity disclosure often considered not relevant
- Attribute Information:
  - Here: any probabilistic information an attacker can infer about an individual based on a (synthetic) dataset in combination with some attribute(s)
  - E.g. Examination scores based on school
  - Note: definition confusion
  - Also note: strong part of statistical inference



Related problem: Privacy assessment needs to be interpretable

- Two statisticians; three opinions
  - But: many ethical judgements require broader input
- Fit into organization's statistical disclosure policy

Idea: make use of aggregated data

- Why?
  - Similar challenges as synthetic data
  - Already in organizations' way of working
- Here: suppressed at aggregation k
  - E.g. for k=5, replace {1,2,3,4} by "<5"
- Shoutout to Little et al. (2022)
  - Puts synthetic data in context of released subsamples

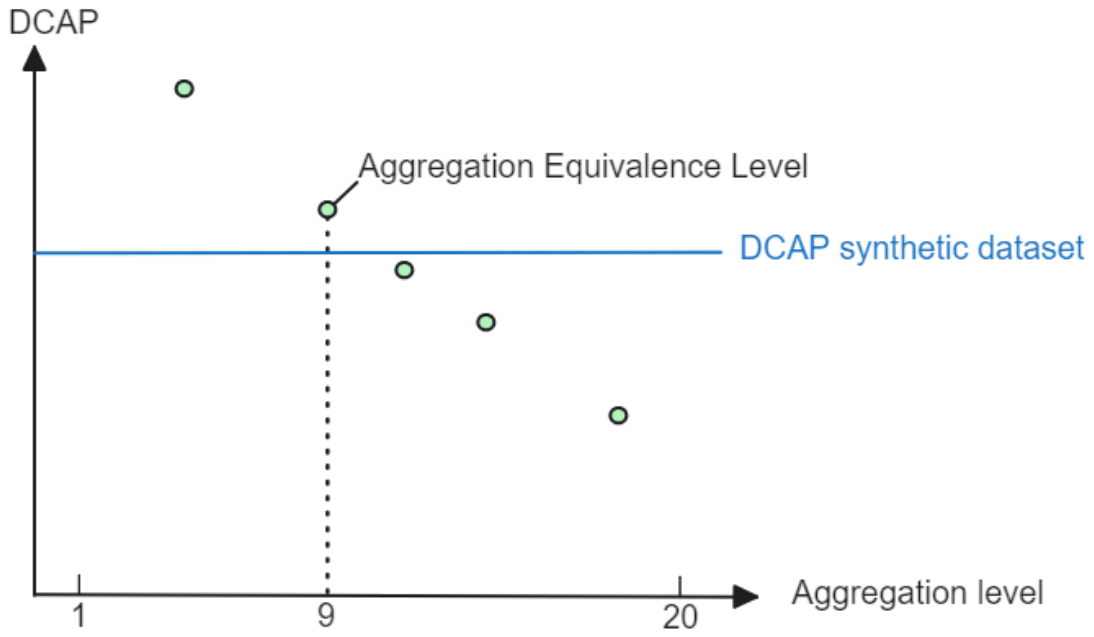
DCAP (Differential Correct Attribution Probability)

- Conceptualized as measure for aggregation information
- Measures how likely it is for someone to infer personal information about individuals in a synthetic dataset compared to a baseline
  - Lower DCAP = better protection
- Suboptimal measure
- Context-dependent
  - Hard to interpret

AEL (Aggregation Equivalence Level)

STEPS TO CALCULATE THE AEL

1. Create a synthetic dataset from the original data.
2. Create multiple aggregated datasets from the original data with different aggregation levels.
3. Calculate the average DCAP for each aggregated dataset and for the synthetic dataset.
4. Compare the DCAP scores and choose the aggregation level where the DCAP of the aggregated dataset is just above that of the synthetic dataset.



Communication

› Management summary (see OSF)

Aggregation Equivalence Level for interpreting Synthetic Data Attribute Information
https://osf.io/rdpab/

The corresponding article and R code can be found here: https://osf.io/rdpab
Authors: Sanne Smid, David van de Merwe, Lotte Pater
Contact: synthetische.data@duo.nl

What is synthetic data?

Synthetic data is the equivalent of data about real individuals that allows reliable research and analysis. A synthetic dataset contains practically the same statistical information as a real dataset but is created with computer-generated individuals, often referred to as "artificial persons." Analyzing this data yields nearly identical results as analyzing the original data, without the use of personal information.

Attribute information of synthetic data

When using synthetic data, the risk of re-identification disappears because the synthetic dataset consists of artificial individuals. However, synthetic data, like aggregated data, contains information about characteristics of individuals in the real data, such as the pass rate of a school – this is called attribute information. It is essential to quantify the degree of attribute information in a synthetic dataset to assess the privacy impact accurately. For this purpose, a statistical measure called the *Differential Correct Attribution Probability* (DCAP)[1] has been developed.

Differential Correct Attribution Probability (DCAP)

The DCAP indicates how likely it is for someone to infer personal information about individuals from a synthetic dataset, with a lower DCAP indicating better privacy protection. However, the DCAP is context-dependent and, therefore, challenging to interpret. We introduce the *Aggregation Equivalence Level* (AEL), which places the attribute information of synthetic data in the context of attribute information of aggregated data.

Aggregation Equivalence Level (AEL)

The idea behind the AEL is to find the aggregation level of an aggregated dataset that has the same or slightly more attribute information than the synthetic dataset. This makes it easier to interpret the privacy impact of a synthetic dataset. By using the AEL, existing policies regarding aggregated datasets can be applied to synthetic datasets, allowing informed decisions to be made about their privacy impact. The following four steps are required for this, and they can be performed by a data analyst.

Steps to calculate the AEL - to be performed by a data analyst

1. Create a synthetic dataset from the original data.
2. Create multiple aggregated datasets from the original data with different aggregation levels.
3. Calculate the average DCAP for each aggregated dataset and for the synthetic dataset.
4. Compare the DCAP scores and choose the aggregation level where the DCAP of the aggregated dataset is just above that of the synthetic dataset.

And now..?

- Look for alternatives to DCAP
- Perform a broader simulation study
- Also: interested to hear if your organizations would consider applying this