Synthetic Data
Privacy Evaluation and Disclosure Risk Estimates

Markus Hittmeir, Andreas Ekelhart, Rudolf Mayer

COMET
Competence Centers for Excellent Technologies
www.ffg.at/comet
SBA Research

Relation between Data Utility and Privacy

The poster on the *utility of synthetic data for machine learning* concerned three synthetic data generation tools: the Synthetic Data Vault (**SDV**), the DataSynthesizer (**DS**), and synthpop (**SP**). We obtained two main results:

- **SP** with standard settings tends to achieve better utility scores than SDV.
- For the **DS**, the results vary depending on the amount of noise injected by differential privacy.

Synthetic data with larger differences to the original (see Figures 1 and 2) tends to perform worse on certain tasks. On the other hand, it may provide better protection of the sensitive information in the original dataset. We hence complemented our utility evaluation with a privacy analysis.

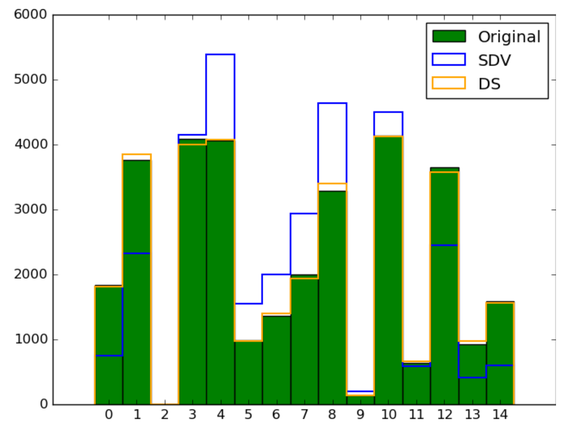
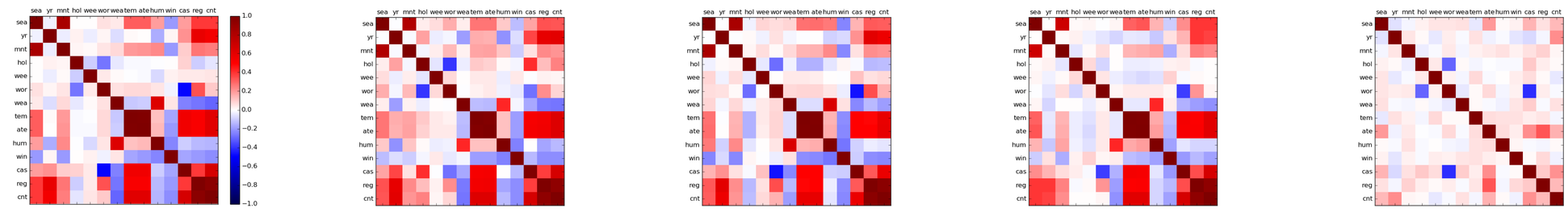


Figure 1: Distributions of the attribute 'occupation' on Adult Census data



(a) Real *(b)* DS *(c)* SP *(d)* SDV *(e)* DSP

Figure 2: Heatmaps showing correlations on Bike Sharing data; red is direct, blue indirect correlation. **SDV** and **DSP** show larger differences to the original than **DS** and **SP**.

Similarity Analysis and Distance Measures

For the assessement of privacy provided by the synthetic datasets, we were interested in similarities between original and synthetic data samples. For each row in the synthetic data, we hence computed the distance to the nearest neighboring sample in the real data (=minimum distance).

General Findings

In Figure 3, we have the minimum distance on the x-axis and the number of samples on the y-axis. We see that **DS** constructs many samples that are similar to original ones, whereas **SDV** and **DSP** appear to provide better protection of the privacy of individuals in the real data. Note that the histogram for **SP** looks very much like the one of **DS**.

(a) DS *(b)* SDV *(c)* DSP

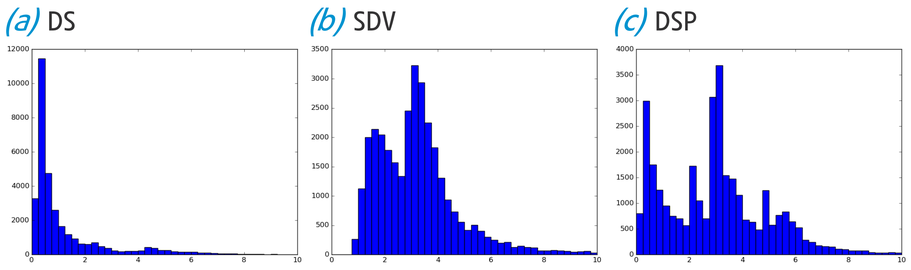


Figure 3: Minimum distances for the Adult Census dataset

We observed similar results on most of the datasets we considered. They appear to confirm that the better utility scores of **DS** and **SP** may be explained by the fact that these tools generate more synthetic samples that are close to original ones.

Disclosure Risk Estimates

In general, two possibilities of information disclosure are distinguished:

- **Identification disclosure** happens when a record in the dataset is linked to a certain individual.
- **Attribute disclosure** means that an attacker is able to infer someone's value of a sensitive attribute.

Our research focused on estimates for attribute disclosure risks. In Ref. 2, we generalized the Correct Attribution Probability (CAP) approach established in Ref. 1. This technique measures attribute disclosure risks for certain attack scenarios, consisting of:

1. A set of columns of the dataset, for which the attacker knows the values of their victim. Usually, these are quasi-identifying attributes like gender, age, or ZIP code.
2. A sensitive target column (e.g., health information or income).

We applied machine learning techniques to estimate the attacker's ability to retrieve the victim's value of target attribute.

General Findings

Table 1 shows the mean accuracy scores of one of said techniques in a certain scenario. Indeed, we observe lower risks on synthetic data than on the real data. However, the risks of **DS** and **SP** are higher then the risks of **SDV** and **DSP**.

	Real	DS	SP	SDV	DSP
Risk	51.7	46.8	48.8	39.0	45.2

Table 1: Attribute disclosure risk due to ENS from Scenario (1), Table 6 in Ref. 2.

Conclusion and Future Work

Synthetic data generation tools that performed better on utility tasks showed higher privacy and disclosure risks. All in all, we observed a trade-off between utility and privacy. As a consequence, one of our future goals is to optimize synthetic data for certain tasks and privacy requirements.

1. Taub et al., *Differential Correct Attribution Probability for Synthetic Data: An Exploration*, In: Proceedings of the 8th PSD, 2018.
2. M. Hittmeir, A. Ekelhart, R. Mayer, *A Baseline for Attribute Disclosure Risk in Synthetic Data*, In: Proceedings of the 20th CODASPY, 2020.

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

SBA Research (SBA-K1) is a COMET Centre within the framework of COMET – Competence Centers for Excellent Technologies Programme and funded by BMK, BMDW, and the federal state of Vienna. The COMET Programme is managed by FFG.