IAB INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

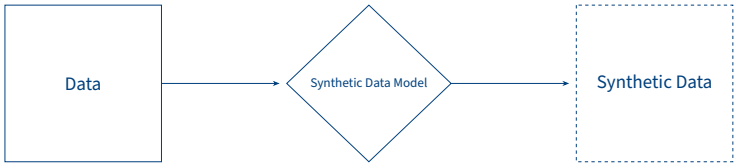
ON THE FORMAL PRIVACY GUARANTEES OF SYNTHETIC DATA

NBER Conference
Data Privacy Protection and the Conduct of Applied Research
Washington D.C., May 16th, 2024

Marcel Neunhoeffer
Jonathan Latner
Jörg Drechsler

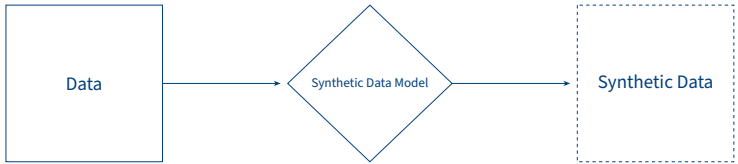
CONTENTS

SYNTHETIC DATA FOR STATISTICAL DISCLOSURE LIMITATION

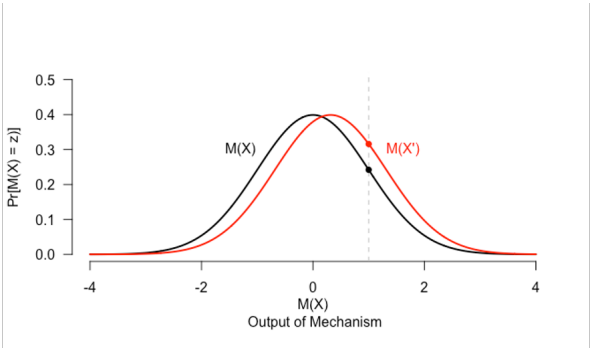


Can classical synthesizers based on Rubin's approach offer formal privacy guarantees?
And if so, under which circumstances?

SYNTHETIC DATA FOR STATISTICAL DISCLOSURE LIMITATION

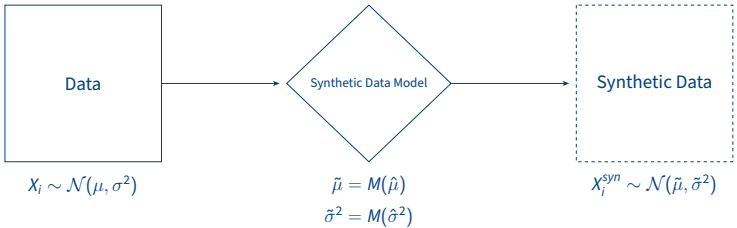


FORMAL PRIVACY GUARANTEES



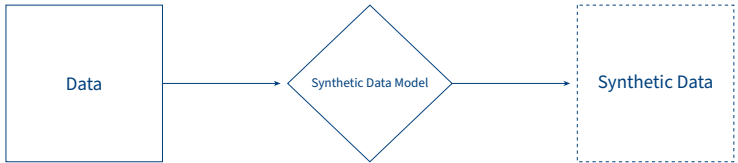
A randomized algorithm M is differentially private if for every pair of neighboring datasets $X, X \in \mathcal{X}$, the random variables $M(X)$ and $M(X')$ are similarly distributed.

SYNTHETIC DATA WITH FORMAL PRIVACY GUARANTEES

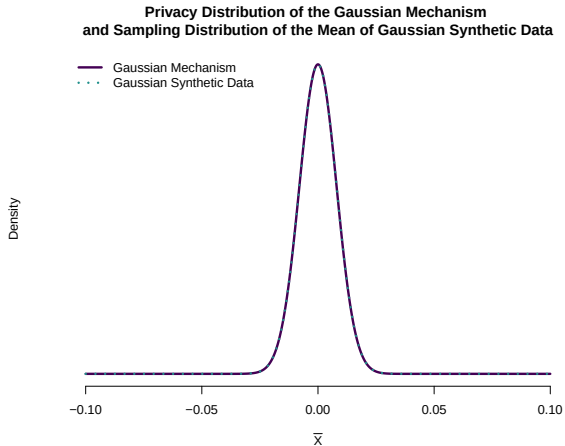


BUT THE SYNTHETIC DATA MODEL IS A RANDOMIZED ALGORITHM ALREADY

PLUG-IN SYNTHESIS WITH KNOWN VARIANCE



PLUG-IN SYNTHESIS WITH KNOWN VARIANCE



PLUG-IN SYNTHESIS WITH KNOWN VARIANCE

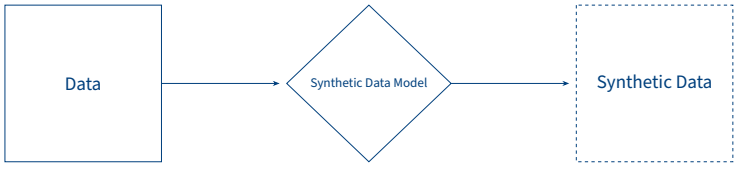
- Generating a single synthetic record offers ρ-zero concentrated DP with $\rho = \frac{4d^2}{2n^2\sigma^2}$.
- Using the composition property of ρ-zero concentrated DP, this implies that releasing m copies of synthetic data each containing n_{syn} records based on this model will imply a privacy loss of $\rho = \frac{mn_{syn}4d^2}{2n^2\sigma^2}$.
- The number of synthetic samples we can release for any desired level of privacy ρ is given by $n_{syn} = \lfloor \frac{2n^2\rho\sigma^2}{4d^2m} \rfloor$.

PLUG-IN SYNTHESIS WITH KNOWN VARIANCE

Some Observations:

- Synthetic data generated using this model can only have formal guarantees if $\sigma^2 > 0$.
- The number of synthetic samples n_{syn} determines the level of privacy.
- To get any meaningful formal privacy guarantees when releasing synthetic data, we need to limit the influence that any single observation can have on the parametric model's sufficient statistic(s).

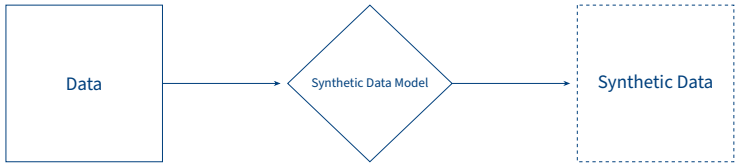
BAYESIAN SYNTHESIS WITH KNOWN VARIANCE



BAYESIAN SYNTHESIS WITH KNOWN VARIANCE

- The privacy loss of the Bayesian synthesis approach is $\rho = \frac{m n_{syn} 4d^2}{2n^2\sigma^2(1+1/n)}$.
- The number of synthetic samples that can be released for a fixed privacy level ρ is given by $n_{syn} = \lfloor \frac{2n^2\rho\sigma^2(1+1/n)}{4d^2 m} \rfloor$.
- The Bayesian approach allows for an alternative interpretation, in which the draw of μ^* is interpreted as the Gaussian mechanism, and the synthetic data generation is treated as post-processing. Under this setting, the privacy loss would be $\rho = \frac{4d^2 n}{2\sigma^2}$.

PLUG-IN SYNTHESIS WITH UNKNOWN VARIANCE



PLUG-IN SYNTHESIS WITH UNKNOWN VARIANCE

- Releasing m copies of synthetic data each containing n_{syn} records based on this model will imply a privacy loss of $\rho = \frac{m n_{syn} 4 d^2}{2 n^2 \hat{\sigma}^2}$ for the sample mean.
- Note that ρ is now data dependent and can no longer be released to the public without leaking information about the original data.
- For an overall formal privacy guarantee, we additionally need to analyze the sampling distribution of $\hat{\sigma}^2$ (see in the paper).

CONTACT

Marcel Neunhoeffer
marcel.neunhoeffer@iab.de