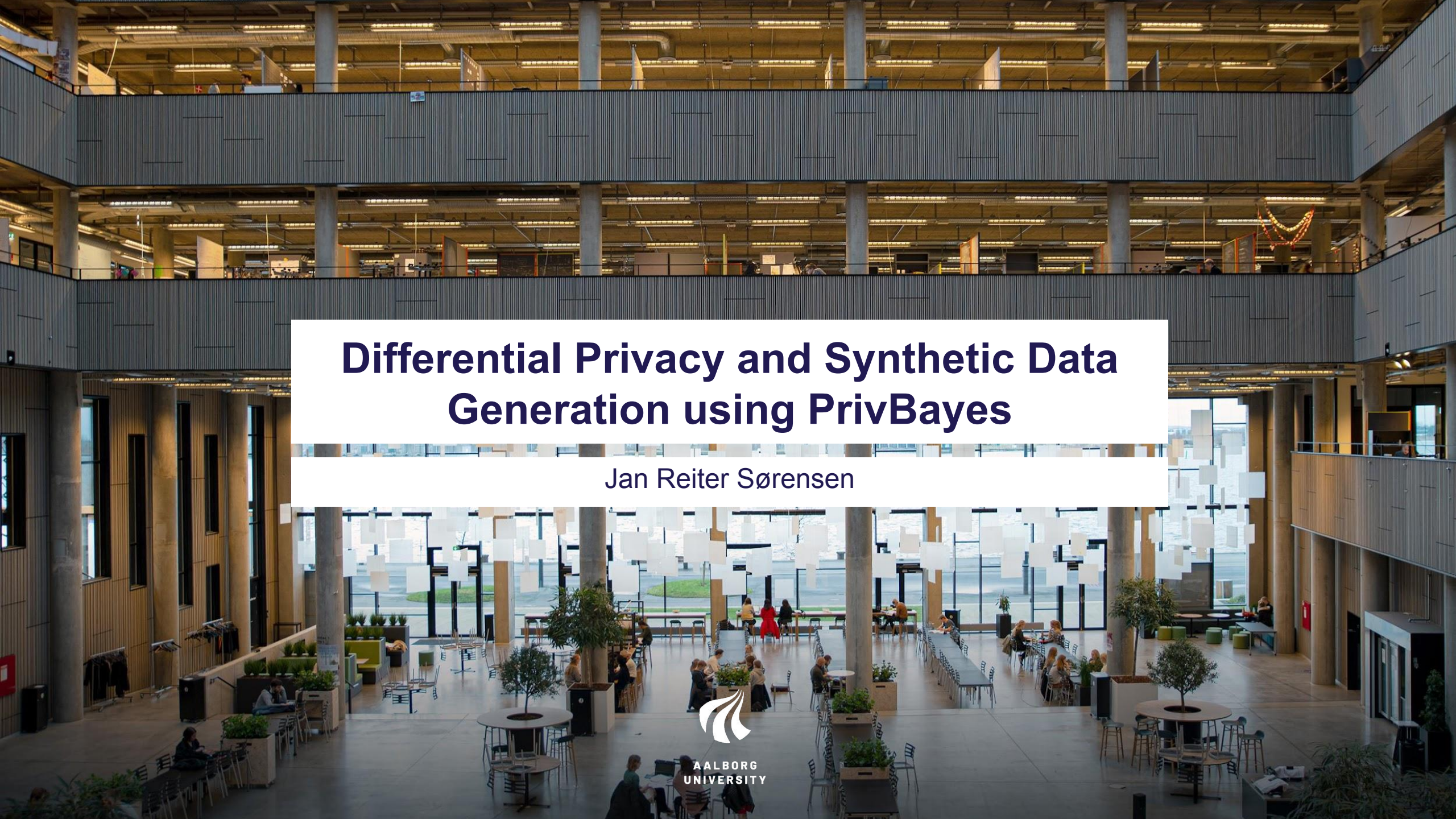


# Differential Privacy and Synthetic Data Generation using PrivBayes

Jan Reiter Sørensen

AALBORG
UNIVERSITY

# Agenda

- Problem statement
- Differential privacy
- Differentially private mechanisms introduced in this project
  - The smoothed histogram mechanism
  - The Laplace mechanism
  - The exponential mechanism
  - PrivBayes
- Investigating synthetic data by PrivBayes
  - Generating some data and sanity check for generating differentially private Bayesian networks
  - The mixture parameter
  - Overall privacy and power of Pearson’s test
- Conclusion

# Problem Statement boiled down

1. How are differentially private methods for synthesizing data defined and explained using a consistent and precise mathematical language?

2. How are differentially private synthesizers such as the smoothed histogram mechanism and PrivBayes implemented?
   a. In practice
   b. For instance using R?

3. What is the relationship between the privacy and the utility of synthetically generated data?

*"Differential privacy is a mathematical property of synthetic data generation methods, and rigorously showing that a synthesizer satisfies differential privacy needs thorough mathematical reasoning. Not all differentially private methods are introduced in a rigorous setting, and neither are they built on the same mathematical terminology. This leads to the following question. How are differentially private methods defined and explained using a consequent and precise mathematical language? There are several methods for synthesization of differentially private data such as the smoothed histogram mechanism and PrivBayes, but they are quite theoretical. How can these mechanisms, if possible, be implemented in practice using a programming language such as R? Furthermore, there is an inherent trade-off between the privacy and the utility of synthetically generated data. What is the relationship between the privacy and the utility of synthetic data?"*

# Differential Privacy

- The relative change of the conditional distribution with respect to a unit change in the original dataset is bounded by *exp(a)*.
- This guarantees a certain level of privacy even in the presence of outliers.
- For *a* close to zero the conditional distributions turn indistinguishable.

$$\sup_{\substack{B \in \mathcal{B}(\mathbb{R}^{k \times d}) \\ (x, x') \in N(X)}} \frac{Q(B|X = x)}{Q(B|X = x')} \leq \exp(\alpha)$$

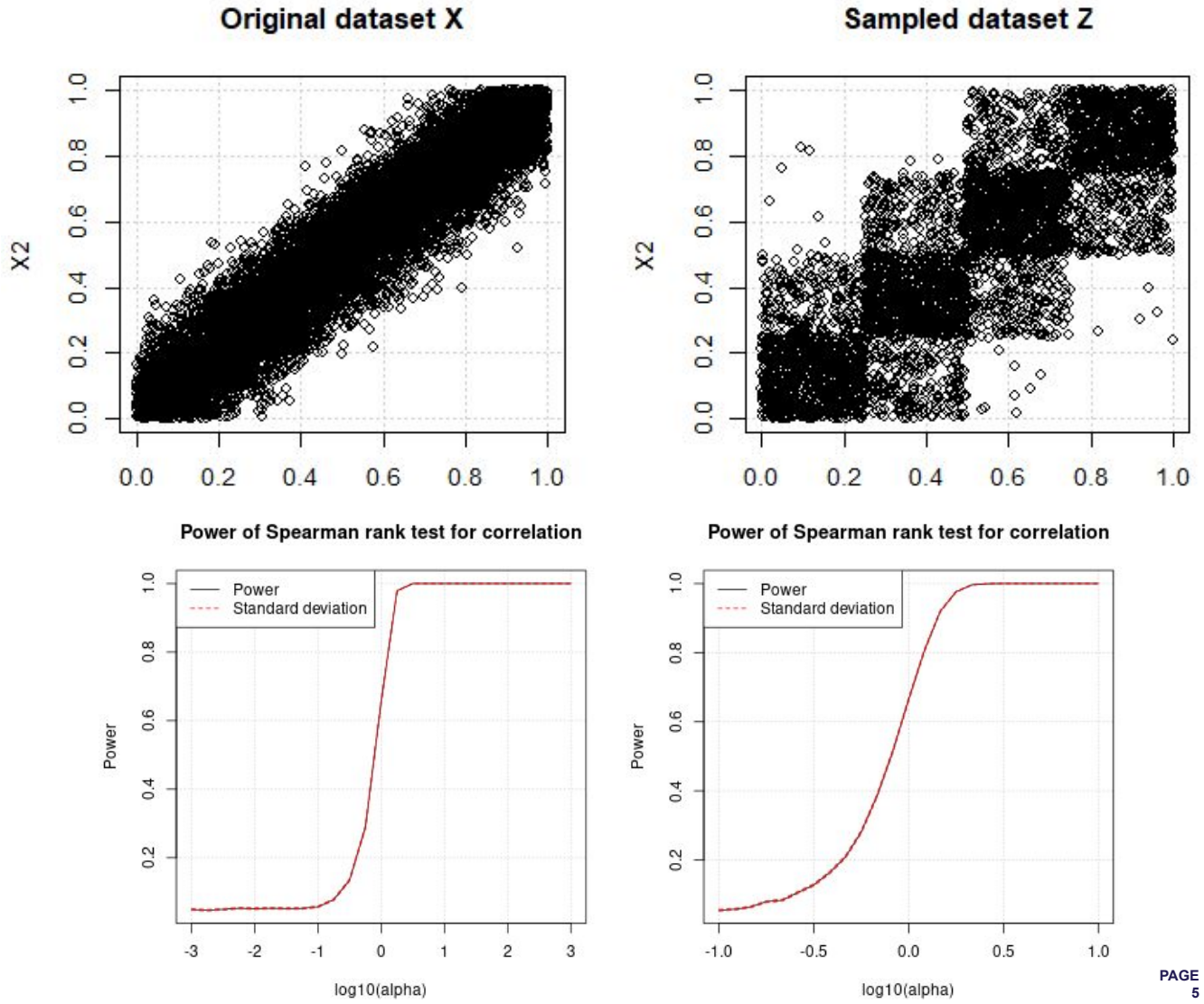
$$\frac{Q(B|X = x)}{Q(B|X = x')} \in \left[\exp(-\alpha), \exp(\alpha)\right]$$

# The smoothed histogram mechanism

- Generates a histogram on the basis of a dataset.
- Smoothing is applied in order to introduce differential privacy.

$$\alpha = k \ln \left( \frac{(1-\delta)m}{n\delta} + 1 \right)$$

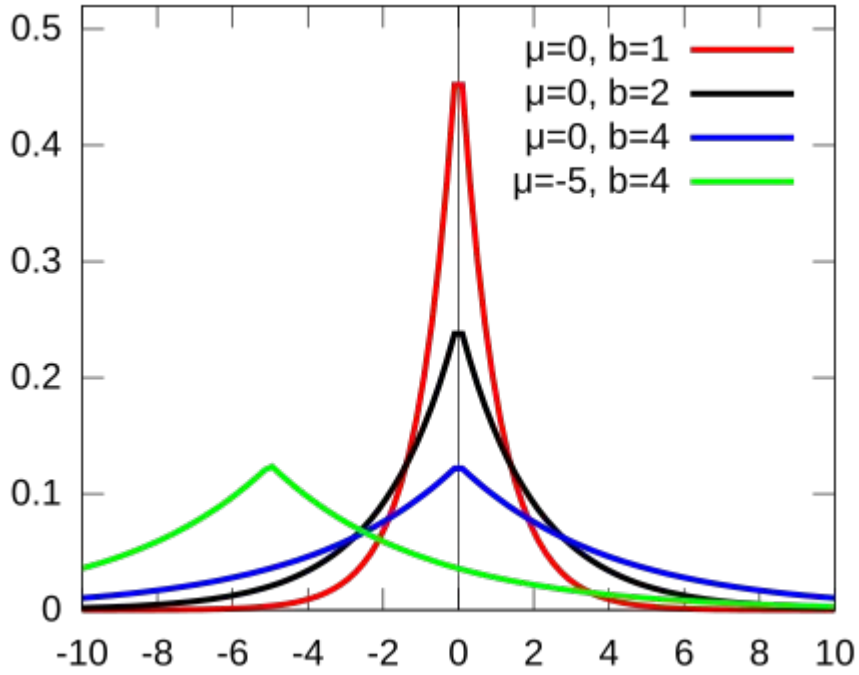
- Can easily be applied to any dataset consisting of numeric or categorical variables, **but scales poorly due to the curse of dimensionality.**



# The Laplace mechanism and the exponential mechanism

- The Laplace mechanism:
  - Add Laplace distributed noise to each observation.
  - The scale parameter depends on the sensitivity and the wanted level of differential privacy.
  - Method can only be applied to continuous variables

- The Exponential mechanism:
  - The function *q* scores a potential output to an input, and outputs are then sampled such that high scores are exponentially more likely to be picked than lower scores.
  - Can be applied to any type of data, but can be difficult to implement.
  - Postulated in the literature to be a foundational differentially private mechanism, meaning that **all differentially private mechanisms are variants of the exponential mechanism.**

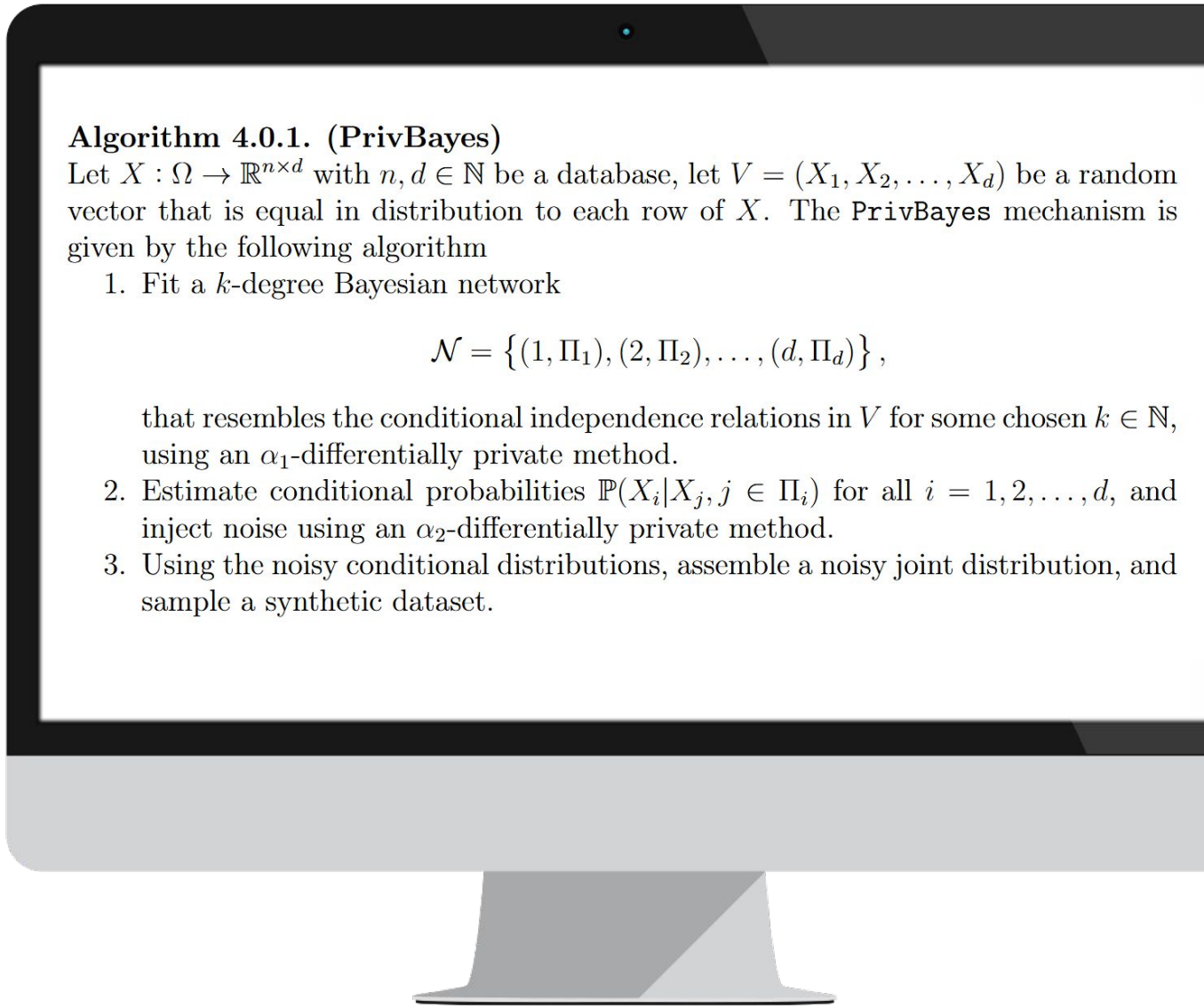



$$g_x(r) \propto \exp\left(\alpha q(x,r)\right),$$

$$Q(B|X=x) := \int_B g_x(r) d\mu(r)$$

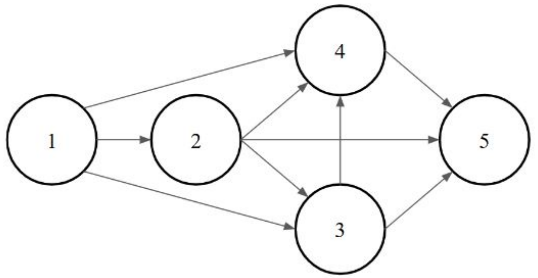
# PrivBayes

- Utilizes Bayesian networks.
- We can implement this using the exponential mechanism for generating Bayesian networks differentially private.
- The Laplace mechanism can be used for injecting noise into the conditional probabilities.

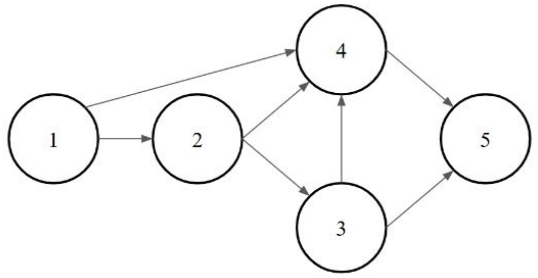


# Generating some data and sanity checking the Bayesian network generating mechanism

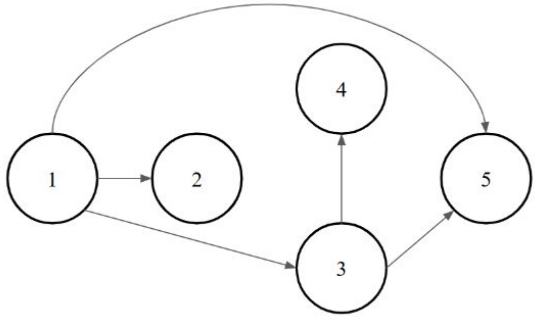
$$\begin{bmatrix} 1 & 0.6 & 0.5 & 0.4 & 0.3 \\ 0.6 & 1 & 0.6 & 0.5 & 0.4 \\ 0.5 & 0.6 & 1 & 0.6 & 0.5 \\ 0.4 & 0.5 & 0.6 & 1 & 0.6 \\ 0.3 & 0.4 & 0.5 & 0.6 & 1 \end{bmatrix}$$



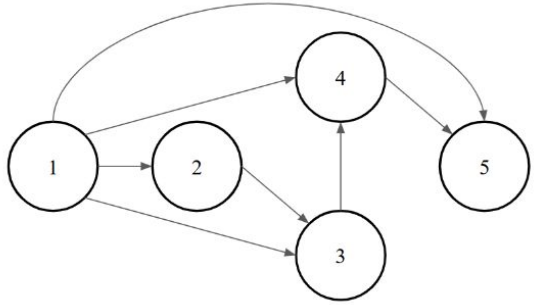

(a) Generated under 50-differential privacy.




(b) Generated under 5-differential privacy.




(c) Generated under 0.5-differential privacy.




(d) Generated under 0.05-differential privacy.

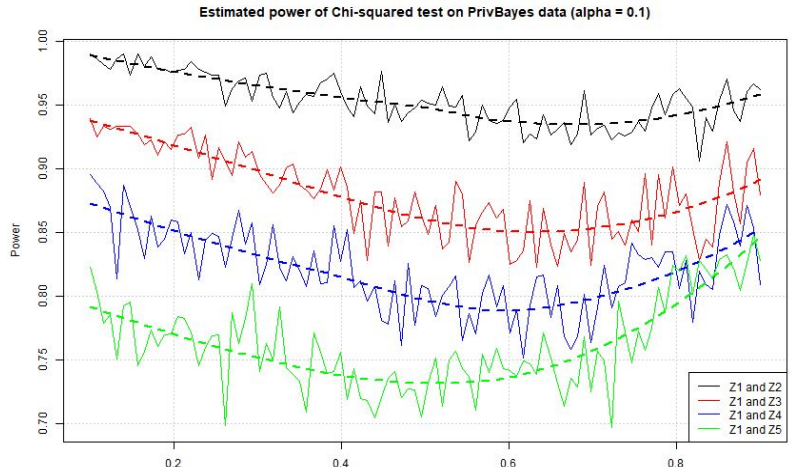
# The mixture parameter

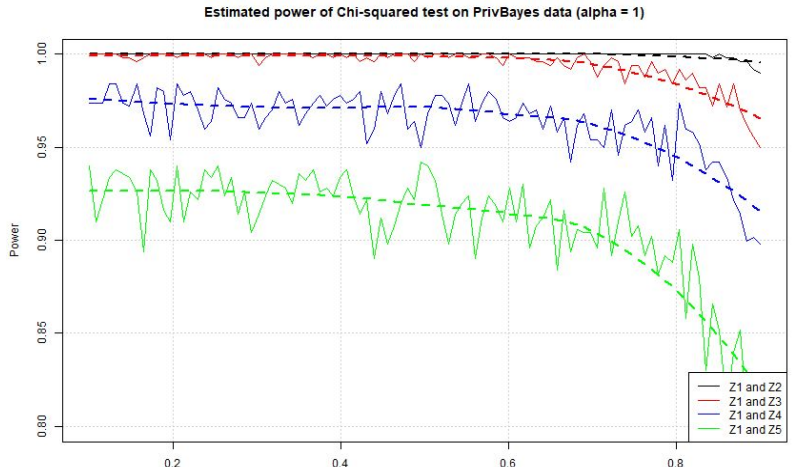
$$\alpha := (\alpha_1 + \alpha_2)$$

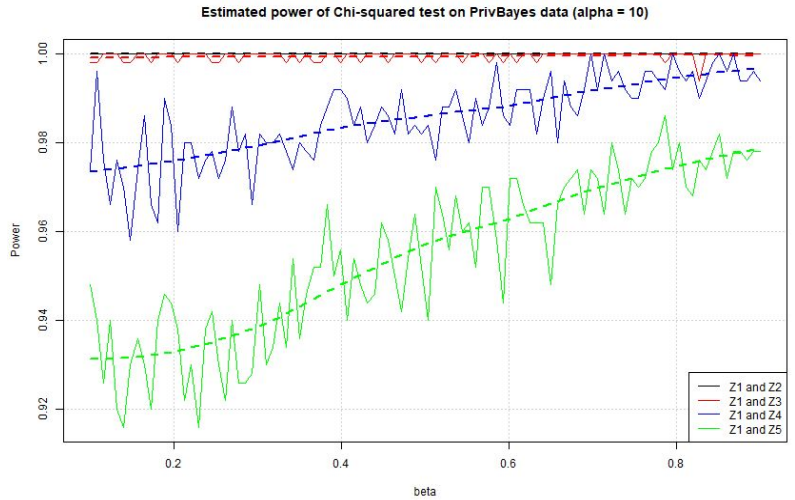
$$\beta := \frac{\alpha_1}{\alpha} \qquad \beta \in [0, 1]$$

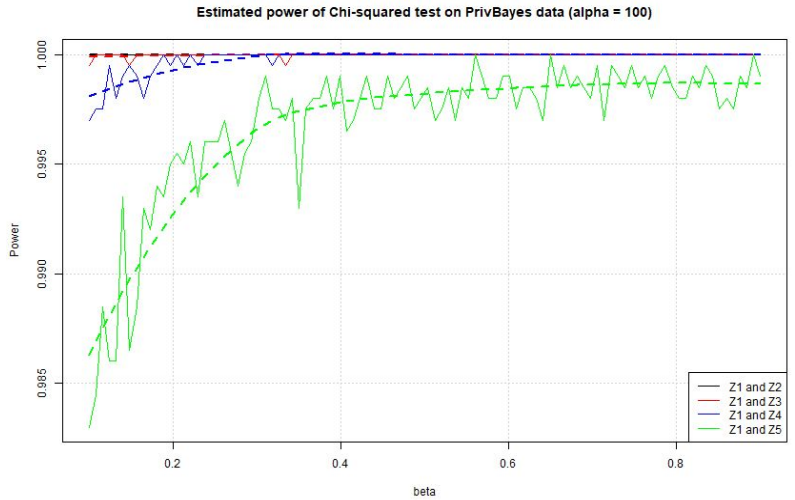
$$\alpha_1 = \beta\alpha$$

$$\alpha_2 = (1 - \beta)\alpha$$

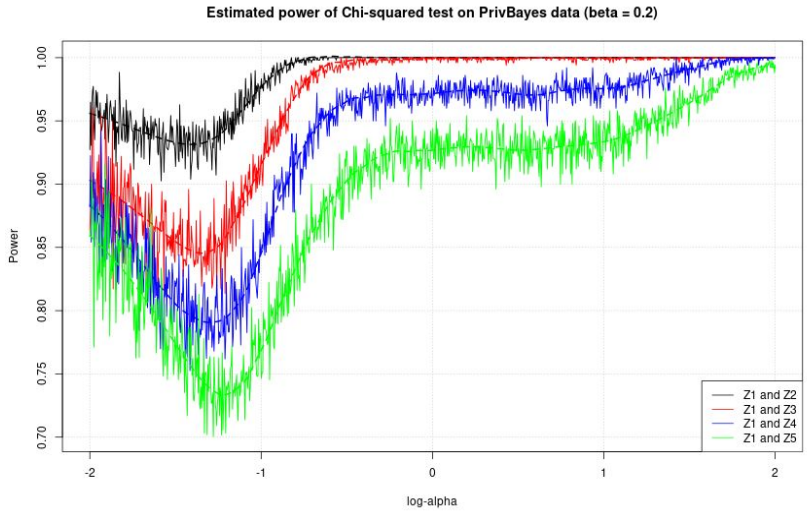


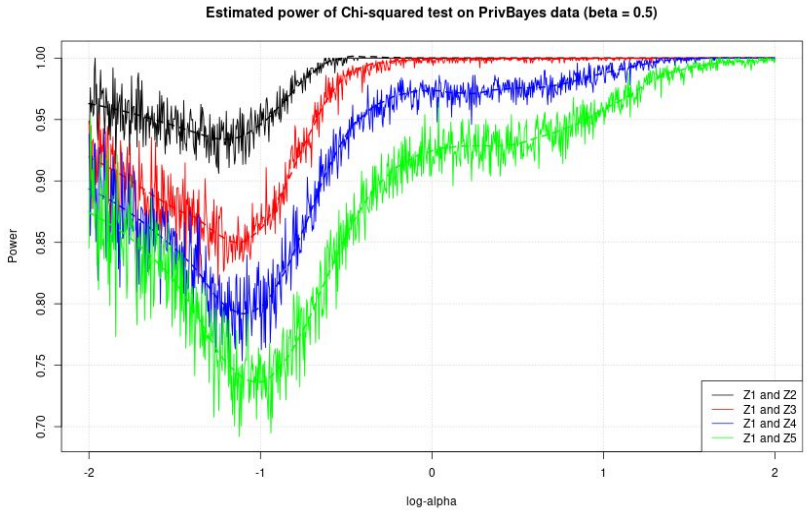



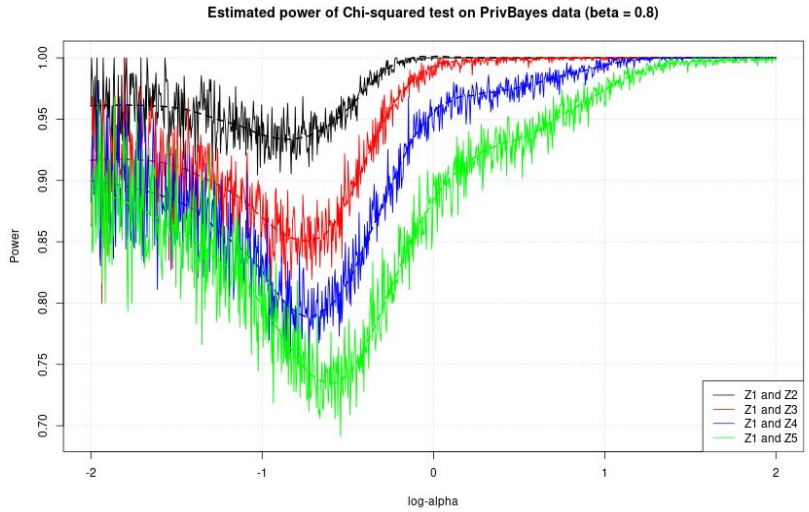



# Overall privacy and power of Pearson's test



# In conclusion

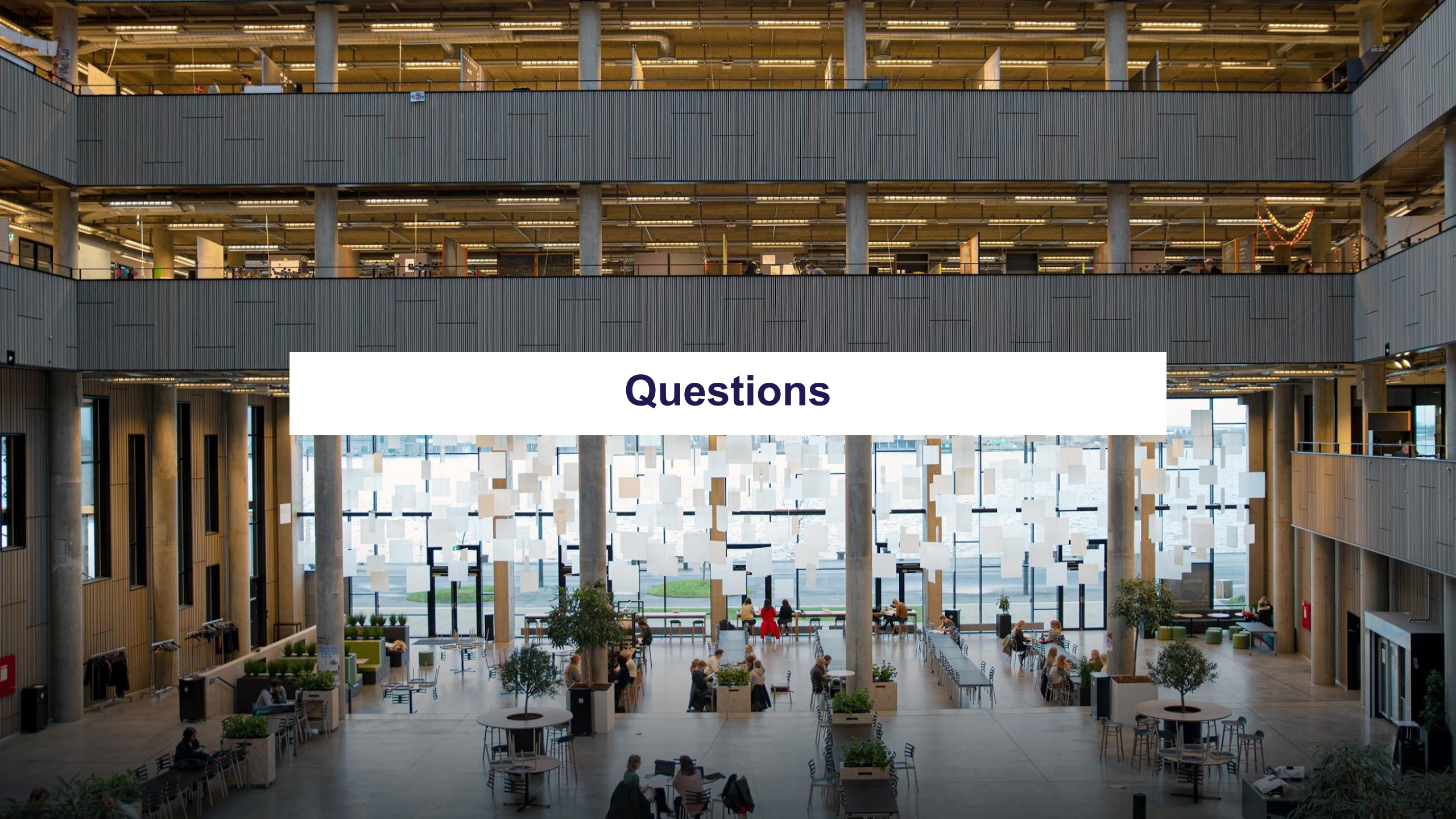
- We saw that the relationship between the mixture parameter and power depended on the overall privacy budget.
  - Extreme privacy: Either high or low beta
  - Slightly less privacy: Important to not have too much privacy on the conditionals.
  - Low privacy: Important to not have too much privacy on the Bayesian network generating mechanism.

- We saw that the power decreased as the overall privacy budget decreased, that is a higher level of privacy implied lower power. Comparing PrivBayes to the histogram method, we saw a much better privacy-utility trade off for PrivBayes data.

# Acknowledgements


- For supervision
  - Rasmus Plenge Waagepetersen
  - Martin Bøgsted and Heidi Søgaard

- For good advice and considerations
  - Jakob Skelmose and Rasmus Rask



# Questions