Karlsruhe Institute of Technology

PRIVACY AND SECURITY

KASTEL

Paper

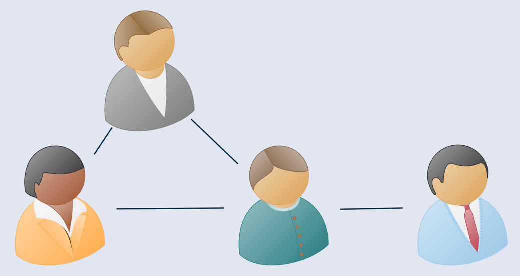
Balancing Privacy and Utility in Correlated Data

A Study of Bayesian Differential Privacy

Martin Lange[KIT] | **Patricia Guerra-Balboa**[KIT] | Javier Parra-Arnau[UPC] | Thorsten Strufe[KIT]

Motivation

- **Dependencies** among data records are present **in most of real world scenarios**.
- **Bayesian DP** extends DP to account for the **effect of correlations in privacy leakage**.
- Current BDP mechanisms suffer from **poor utility** and **lack applicability**.



Without further assumptions

- Privacy decreases linearly proportional to number of correlated records:

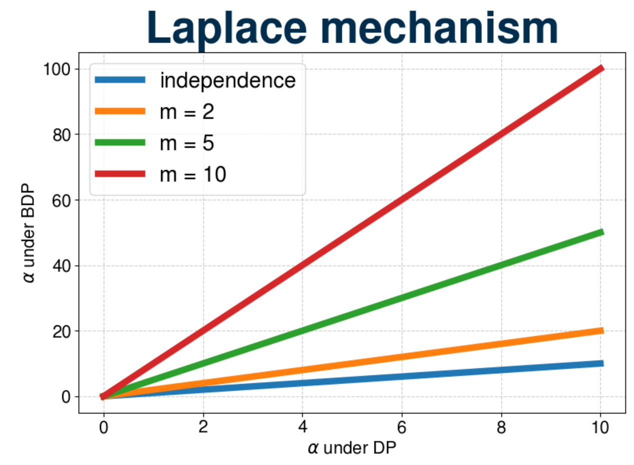
$$\varepsilon\text{-DP} \Rightarrow m\varepsilon\text{-BDP}$$

This result is tight! Even if $\rho \to 0$.

Impossibility Result: We cannot protect against arbitrary correlations and provide utility at the same time.

For the same confidence level, the upper bound on the query error α increases sharply:

Laplace mechanism



⇒ We need to focus on specific distributions. In this paper we analyze Gaussian and Markov models.

Multivariate Gaussian Correlation

Main result

- Let $\mathcal{M}$ be an $\varepsilon\ell_1$-private mechanism,
- input data drawn from a multivariate Gaussian distribution
- $\rho(m-2) < 1$ is the maximum correlation coefficient.

Then, $\mathcal{M}$ is Bayesian d-private with

$$d(x, x') \leq \left(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1\right) \varepsilon |x' - x|.$$

- We extend metric privacy to **Bayesian metric privacy**.
- Using clipping as preprocessing step, $c_I(D)_i = \max(a, \min(b, D_i))$, we recover BDP:

$$\left(\frac{m^2}{4(\frac{1}{\rho} - m + 2)} + 1\right) M\varepsilon\text{-BDP}.$$

where M is the diameter of the interval $I = [a, b]$.

Markov Model

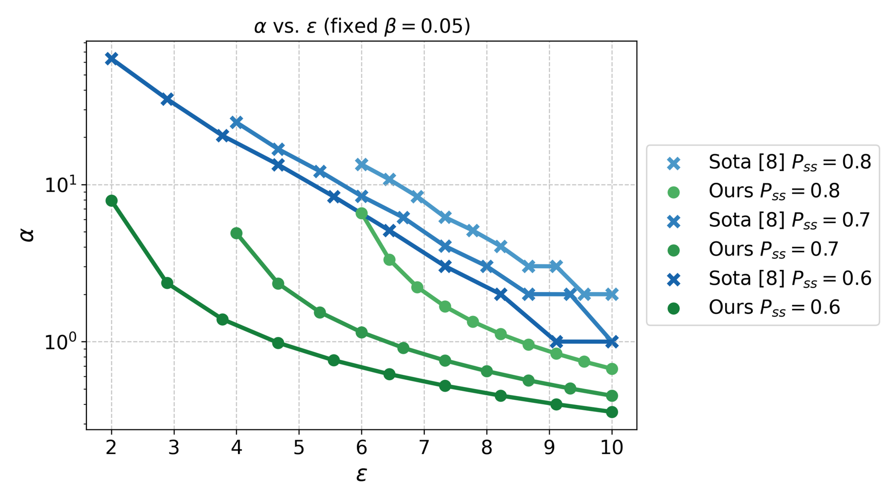
Main result

- Let $\mathcal{M}$ be an ε-DP mechanism,
- input data sampled form Markov chain with transition matrix $P \in \mathbb{R}^{S \times S}$ and initial distribution $w \in \mathbb{R}^S$ with the following properties:

(H1) For all $x, y \in \mathcal{S}$ we have $P_{x,y} > 0$ and, (H2) $wP = w$.

Then, $\mathcal{M}$ is an $(\varepsilon + 4 \ln \gamma)$-BDP mechanism where $\gamma = \frac{\max_{x,y \in \mathcal{S}} P_{xy}}{\min_{x,y \in \mathcal{S}} P_{xy}}$.

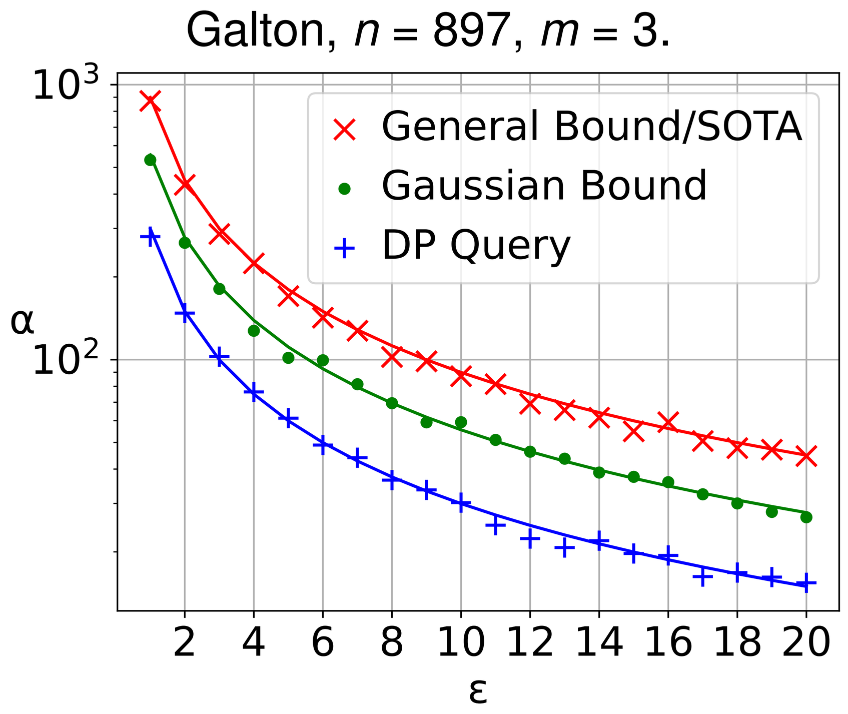
Previous mechanism	Ours
$P_{xy} > 0$	$P_{xy} > 0$
stationary	stationary
lazy	
binary	
symmetric	
$\varepsilon' > 0$	$\varepsilon' > 4\ln(\gamma)$



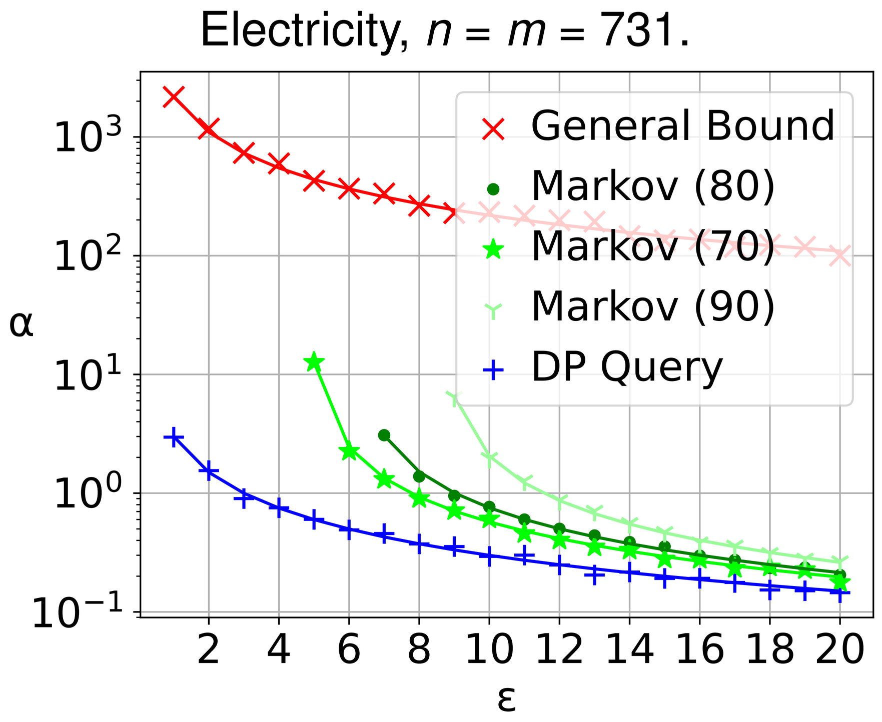
Impact on Utility for Real Databases

- From our theorems: **Noise recalibration** of the Laplace mechanism ⇒ **BDP**.
- **Substantial utility gains** compared to the standard bound.
- Markov bound independent of n ⇒ **huge improvement for large datasets**.

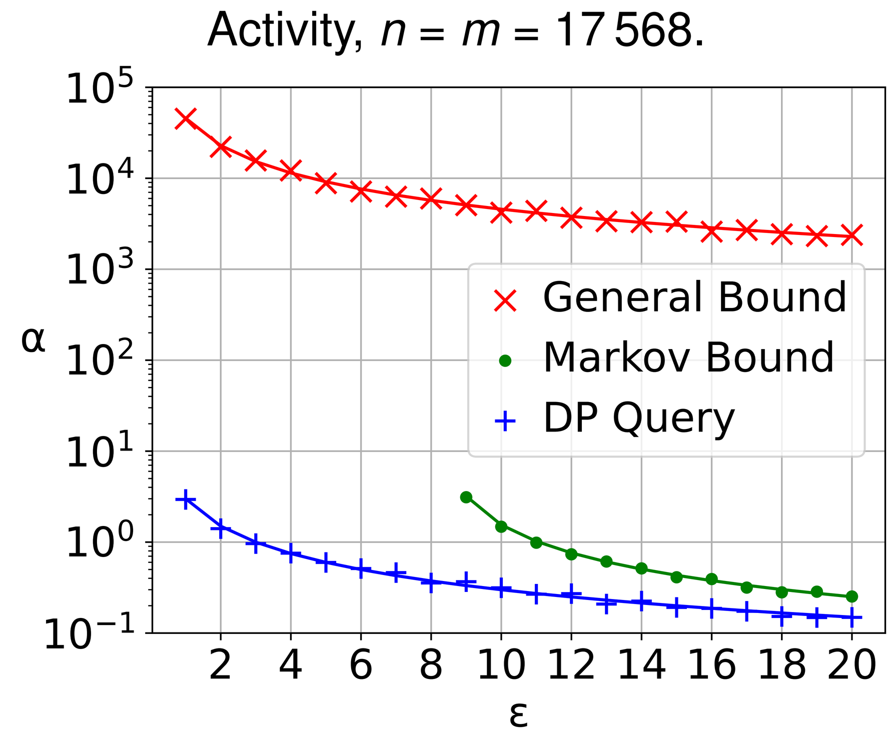
Galton, $n = 897$, $m = 3$.



Electricity, $n = m = 731$.



Activity, $n = m = 17\,568$.



– **Theoretical Utility Metric:** (α, β)-accuracy, i.e., $\Pr[|q(D) - \mathcal{M}(q(D))| \geq \alpha] \leq \beta$. Specifically, $\beta = 0.05$, i.e., 95% confidence interval.
× **Empirical Utility Metric:** The upper bound of a $(1 - \beta)$ confidence interval for the absolute query error.

Conclusion

We prove that BDP is a suitable solution for privacy-preserving data analysis when correlations are structured, e.g., small groups, weak Gaussian correlations, or time-series data.

Bin Yang, Issei Sato, and Hiroshi Nakagawa. "Bayesian Differential Privacy on Correlated Data". In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. 2015. DOI: 10.1145/2723372.2747643.
Darshan Chakrabarti et al. *Optimal Local Bayesian Differential Privacy over Markov Chains*. 2022. arXiv: 2206.11402.