UNIVERSITÄT ZU LÜBECK

WP 3.12: Anonymity Guarantees Against Attackers with Partial Background Knowledge

Dennis Breutigam

29/02/24

IM FOCUS DAS LEBEN

How to Model Privacy

Personal data is needed for all kind of research.
Tradeoff: Utility of the data versus privacy of the individuals.

How to Model Privacy

Personal data is needed for all kind of research.
Tradeoff: Utility of the data versus privacy of the individuals.

Publish Modified Database

- Data reduction
- Microaggregation
  - k-anonymity

How to Model Privacy

Personal data is needed for all kind of research.
Tradeoff: Utility of the data versus privacy of the individuals.

Publish Modified Database

- Data reduction
- Microaggregation
  - k-anonymity

Secret Database - Allow Queries

- Give modified answers.
- Use entropy of the data.

Common Models

- Noiseless Privacy (NP)
- Differential Privacy (DP)

Formal Structure

Example Database

ID	Name	Weight	Age	Height
1	Bob	72	37	177
2	Alice	57	44	154
3	Maja	78	91	162
...	...	...	...	...

Formal Structure

Example Database

ID	Name	Weight	Age	Height
1	Bob	72	37	177
2	Alice	57	44	154
3	Maja	78	91	162
...	...	...	...	...

Database

- An individual I is a vector of the space $W = (W_i)_{i=1}^d$.
- A database D^n of n individuals is a sequence of individuals.
- The universe of possible databases $\mathbf{D}^n \subseteq W^n$.
- Assume individuals as independent identical distributed.

Formal Structure

Queries

- A query is a deterministic function $F\colon W^n \to A$.
- Where A denotes the set of possible answers.

Formal Structure

Queries

- A query is a deterministic function $F\colon W^n \to A$.
- Where A denotes the set of possible answers.

Example Queries

- Average income of inhabitants.
- Number of patients with disease…
- Number of young smokers with high blood pressure.

Formal Structure

Queries

- A query is a deterministic function $F\colon W^n \to A$.
- Where A denotes the set of possible answers.

Example Queries

- Average income of inhabitants.
- Number of patients with disease…
- Number of young smokers with high blood pressure.

Nessessary Properties of Queries

- Not tailored to specific entries of the database.
- Symmetric functions

Formal Structure

Property Queries

For arbitrary $U \subseteq W$ query F_U asks for the percentage of individuals that have property U.

- π_U a priori probability to have property U.

Formal Structure

Property Queries

For arbitrary $U \subseteq W$ query F_U asks for the percentage of individuals that have property U.

- π_U a priori probability to have property U.

Extreme Probabilities

- Problems arise for π_U close to 0 or 1.
  - Rare diseases

Formal Structure

Noise Mechanisms

Adding noise to an answer to hide personal information.

- For this consider a random mechanism M.
  - Adding gaussian noise to the average income of inhabitants.
- (F, M) is the query F complemented by M.

Formal Structure

How do we distinguish between databases that contain the sensitive elements and those that do not?

[1] Calibrating noise to sensitivity in private data analysis. - Dwork et. al.

Formal Structure

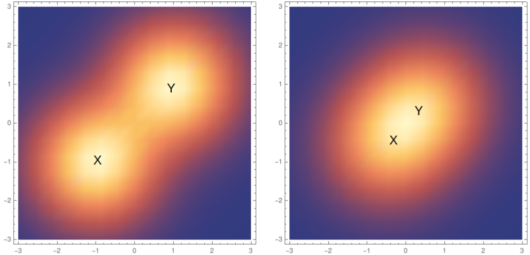
How do we distinguish between databases that contain the sensitive elements and those that do not?

(ϵ, δ) – indistinguishability [1]

Two random variables X, Y are indistinguishable $X \approx_{\epsilon,\delta} Y$ if

$$Pr[X \in S] \leq e^{\epsilon} Pr[Y \in S] + \delta$$
$$Pr[Y \in S] \leq e^{\epsilon} Pr[X \in S] + \delta$$

for all measurable sets S.



[1] Calibrating noise to sensitivity in private data analysis. - Dwork et. al.

Formal Structure

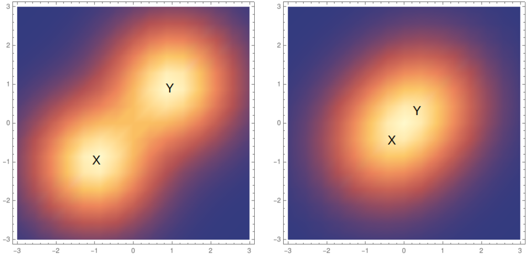
How do we distinguish between databases that contain the sensitive elements and those that do not?

(ϵ, δ) – indistinguishability [1]

Two random variables X, Y are indistinguishable $X \approx_{\epsilon,\delta} Y$ if

$$Pr[X \in S] \leq e^{\epsilon} Pr[Y \in S] + \delta$$
$$Pr[Y \in S] \leq e^{\epsilon} Pr[X \in S] + \delta$$

for all measurable sets S.



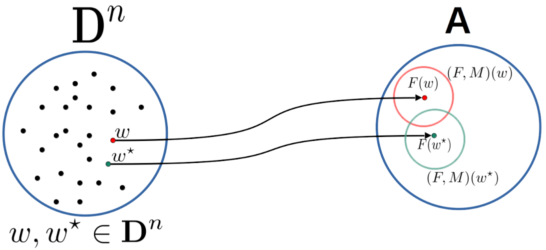
Privacy Ratio

$$R_Y^X(S) := \frac{Pr[X \in S]}{Pr[Y \in S]}.$$

[1] Calibrating noise to sensitivity in private data analysis. - Dwork et. al.

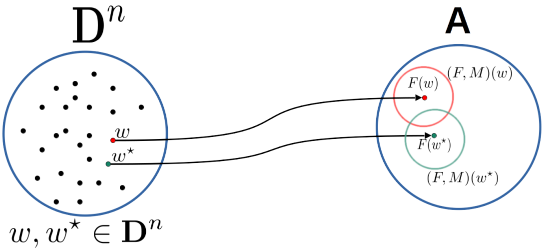
Underlying privacy notions

(ϵ, δ) – Differential Privacy[1]



Underlying privacy notions

(ϵ, δ) – Differential Privacy[1]

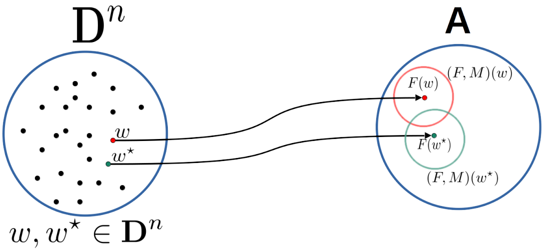


- Uses neighborhood relationship on $\mathbf{D}^n$.
- For all adjacent databases $w, w^\star \in \mathbf{D}^n$

$$(F, M)(w) \approx_{\epsilon,\delta} (F, M)(w^\star).$$

Underlying privacy notions

(ϵ, δ) – Differential Privacy[1]



- Uses neighborhood relationship on $\mathbf{D}^n$.
- For all adjacent databases $w, w^\star \in \mathbf{D}^n$

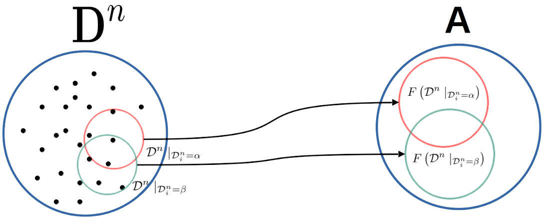
$$(F, M)(w) \approx_{\epsilon,\delta} (F, M)(w^\star).$$

Attacker knows the full database but the one sensitive entry.

- Strong privacy guarantees.
- Estimates needed noise.
- Important impact on the utility.

Underlying privacy notions

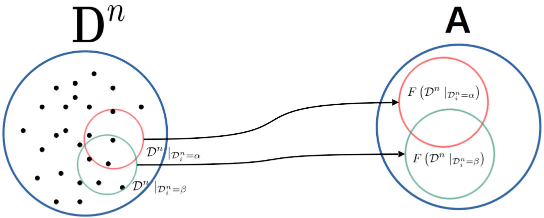
(ϵ, δ) – Noiseless Privacy[2]



[2]Noiseless Database Privacy - Bhaskar et. al.

Underlying privacy notions

(ϵ, δ) – Noiseless Privacy[2]



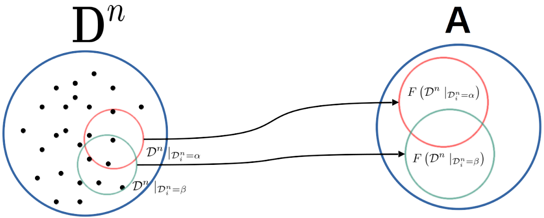
- Distribution $\mathcal{D}^n$ on $\mathbf{D}^n$.
- Condition the distribution, such that an individual i has different properties α, β.
  - $\mathcal{D}^n_{i\leftarrow\alpha} := \mathcal{D}^n \,|_{\mathcal{D}_i^n=\alpha}$

$$F(\mathcal{D}^n_{i\leftarrow\alpha}) \approx_{\epsilon,\delta} F(\mathcal{D}^n_{i\leftarrow\beta})$$

[2] Noiseless Database Privacy - Bhaskar et. al.

Underlying privacy notions

(ϵ, δ) – Noiseless Privacy[2]



- Distribution $\mathcal{D}^n$ on $\mathbf{D}^n$.
- Condition the distribution, such that an individual i has different properties α, β.
  - $\mathcal{D}^n_{i\leftarrow\alpha} := \mathcal{D}^n \mid_{\mathcal{D}_i^n=\alpha}$

$$F(\mathcal{D}^n_{i\leftarrow\alpha}) \approx_{\epsilon,\delta} F(\mathcal{D}^n_{i\leftarrow\beta})$$

$\mathcal{D}^n$ parameters are public knowledge.

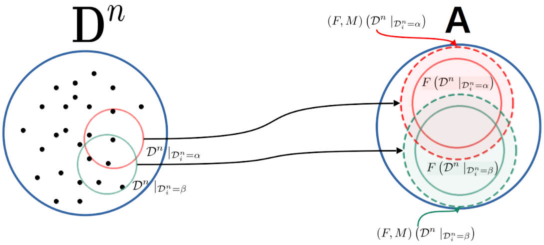
→ Attackers knowledge as condition.

- Utilizes entropy in the data.
- Analyzes deterministic queries.

[2]Noiseless Database Privacy - Bhaskar et. al.

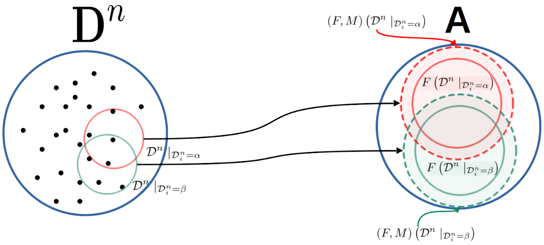
Underlying privacy notions

(ϵ, δ) – Distributional Privacy



Underlying privacy notions

(ϵ, δ) – Distributional Privacy

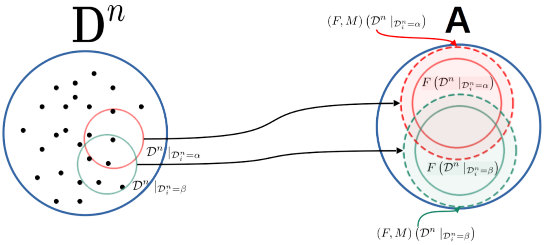


- Distribution $\mathcal{D}^n$ on $\mathbf{D}^n$.
- For all $i \in \{1, \dots, n\}$ individuals and possible values $\alpha, \beta \in W$

$$(F, M)(\mathcal{D}^n_{i \leftarrow \alpha}) \approx_{\epsilon,\delta} (F, M)(\mathcal{D}^n_{i \leftarrow \beta})$$

Underlying privacy notions

(ϵ, δ) – Distributional Privacy



- Distribution $\mathcal{D}^n$ on $\mathbf{D}^n$.
- For all $i \in \{1, \dots, n\}$ individuals and possible values $\alpha, \beta \in W$

$$(F, M)(\mathcal{D}^n_{i \leftarrow \alpha}) \approx_{\epsilon,\delta} (F, M)(\mathcal{D}^n_{i \leftarrow \beta})$$

- Utilizes entropy in the data.
- Estimates needed noise.
- Complex interactions of Distributions.

Analyzing Distributional Privacy

Analysis method

1. Comparison of different methods for adding noise.
2. Compare the utility loss.
3. Take π_U into account.

Analyzing Distributional Privacy

Analysis method

1. Comparison of different methods for adding noise.
2. Compare the utility loss.
3. Take π_U into account.

Quality measurment

- Utility loss: Amount of noise used.
  - Variance of noise ψ.
- Privacy parameters: (ϵ, δ)

Noise Sources

Direct Addition

Using the mechanism M, which works as follows: $(F, M) = F + N_M$

- A common mechanism in Differential Privacy.

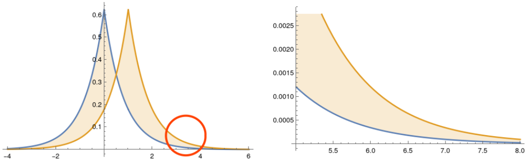
Noise Sources

Direct Addition

Using the mechanism M, which works as follows: $(F, M) = F + N_M$

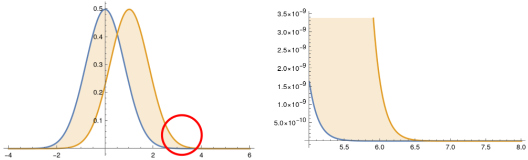
- A common mechanism in Differential Privacy.

$N_M^{\mathcal{L}ap} \sim \mathcal{L}ap(0, \psi)$ - Laplace Noise



- Gives pure DP guarantees.

$N_M^{\mathcal{N}} \sim \mathcal{N}(0, \psi)$ - Gaussian Noise



- Convenient properties.
- Commonly used.

Noise Sources

ID	Name	Attribute	...
1	*Andrew*	*187*	
2	*Anna*	*153*	
3	*Bob*	*162*	
4	*Berta*	*178*	
5	*Clemens*	165	
6	*Cathrine*	*192*	
...	...	...	
n	Xena	*192*	

Subsampling

ID	Name	Attribute	...
1	*Anna*	*153*	
2	*Berta*	*178*	
3	*Clemens*	165	
4	*Darius*	*184*	
...	...	...	
m	Wilca	*158*	

Subsampling

Mechanism S_m draws subset uniformly.

- m size of subset.
- $\lambda = m/n$ selection probability.

[3]Privacy Amplification by Subsampling: Tight Analyses via Coupling and Divergences, Balle et. al.

Noise Sources

ID	Name	Attribute	...
1	Andrew	187	
2	Anna	153	
3	Bob	162	
4	Berta	178	
5	Clemens	165	
6	Cathrine	192	
...	...	...	
n	Xena	192	

Subsampling

ID	Name	Attribute	...
1	Anna	153	
2	Berta	178	
3	Clemens	165	
4	Darius	184	
...	...	...	
m	Wilca	158	

Subsampling

Mechanism S_m draws subset uniformly.

- m size of subset.
- $\lambda = m/n$ selection probability.

Characteristics

- Enhances DP mechanisms[3].
- Low interaction with underlying distribution.

[3] Privacy Amplification by Subsampling: Tight Analyses via Coupling and Divergences, Balle et. al.

Subsampling and Differential Privacy

Parameters Obtained

- Property queries are $(0,1)$-DP.
- Indistinguishability must hold for any database pair.
- Consider databases w, w' where either non or one has the U property.
  - F_U only possible answers are 0 and $1/m$.

$$R_{(F_U,S)(w')}^{(F_U,S)(w)} = \frac{\lambda Pr\left[(F_U,S)(w) = 1/m \mid x_1 \in S\right] + (1-\lambda)Pr\left[(F_U,S)(w) = 1/m \mid x_1 \notin S\right]}{\lambda Pr\left[(F_U,S)(w') = 1/m \mid x_1' \in S\right] + (1-\lambda)Pr\left[(F_U,S)(w') = 1/m \mid x_1' \notin S\right]}$$

$\rightarrow$ This can take the form $1/0$, in which case its probability mass is λ.

It further holds that (F_U, S_m) is $(0,\delta)$-DP.

Subsampling and Distributional Privacy

Find the events for which the ratio is not bounded!

Query Distribution

- $\mathcal{B}in(n, \pi_U)$ the binomial distribution with n trials and probability π_U.
- $F_U(\mathcal{D}^n) \sim (1/n)\mathcal{B}in(n, \pi_U)$
- Subsampling of size m independent of $\mathcal{D}^n$.
- $(F_U, S_m)(\mathcal{D}^n) \sim (1/m)\mathcal{B}in(m, \pi_U)$

Conditional Propabilitys

Two cases for the sensitive entry i:

1. $\mathcal{D}_i^n \in U$
2. $\mathcal{D}_i^n \notin U$

Thus we consider

$(F_U, S_m)(\mathcal{D}_{i \in U}^n) := (F_U, S_m)(\mathcal{D}^n \mid_{\mathcal{D}_i^n \in U})$.

Subsampling and Distributional Privacy

Find the events for which the ratio is not bounded!

Query Distribution

- $\mathcal{B}in(n, \pi_U)$ the binomial distribution with n trials and probability π_U.
- $F_U(\mathcal{D}^n) \sim (1/n)\mathcal{B}in(n, \pi_U)$
- Subsampling of size m independent of $\mathcal{D}^n$.
- $(F_U, S_m)(\mathcal{D}^n) \sim (1/m)\mathcal{B}in(m, \pi_U)$

Conditional Propabilitys

Two cases for the sensitive entry i:

1. $\mathcal{D}_i^n \in U$
2. $\mathcal{D}_i^n \notin U$

Thus we consider

$(F_U, S_m)(\mathcal{D}_{i\in U}^n) := (F_U, S_m)(\mathcal{D}^n \mid_{\mathcal{D}_i^n \in U})$.

Ratio

$$R_{(F_U,S_m)(\mathcal{D}_{i\in U}^n)}^{(F_U,S_m)(\mathcal{D}_{i\notin U}^n)} = \frac{\lambda j + (1-\lambda)m\pi_U}{\lambda(m-j)\left(\frac{\pi_U}{1-\pi_U}\right) + (1-\lambda)m\pi_U}$$

Subsampling and Distributional Privacy

Ratio Bound

- Ratio is monotone.
- Consider the ratio as continuous function.
- Find γ such that $e^{\epsilon} = R^{(F_U, S_m)(\mathcal{D}^n_{i \notin U})}_{(F_U, S_m)(\mathcal{D}^n_{i \in U})}((1+\gamma)\pi_U)$.

Solved by

$$\gamma^{\star} = \lambda^{-1} \frac{e^{\epsilon} - 1}{1 + e^{\epsilon} \frac{\pi_U}{1 - \pi_U}}.$$

Subsampling and Distributional Privacy

Ratio Bound

- Ratio is monotone.
- Consider the ratio as continuous function.
- Find γ such that $e^{\epsilon} = R^{(F_U,S_m)(\mathcal{D}^n_{i\notin U})}_{(F_U,S_m)(\mathcal{D}^n_{i\in U})}((1+\gamma)\pi_U)$.

Solved by

$$\gamma^{\star} = \lambda^{-1}\frac{e^{\epsilon}-1}{1+e^{\epsilon}\frac{\pi_U}{1-\pi_U}}.$$

Bound δ

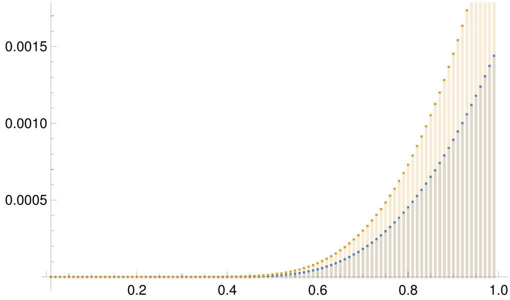
Since the ratio is symmetric we have

$$\delta \leq Pr\left[(F,M)(\mathcal{D}^n_{i\leftarrow\alpha}) \geq \lambda^{-1}\frac{e^{\epsilon}-1}{1+e^{\epsilon}\frac{\pi_U}{1-\pi_U}}\right].$$

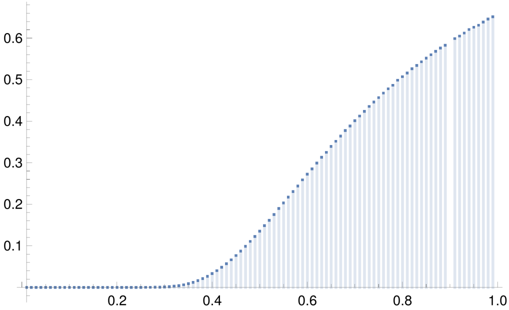
This can be used to calculate δ exactly.

Subsampling and Distributional Privacy

Setting: $n = 1000, \epsilon = 0.1$



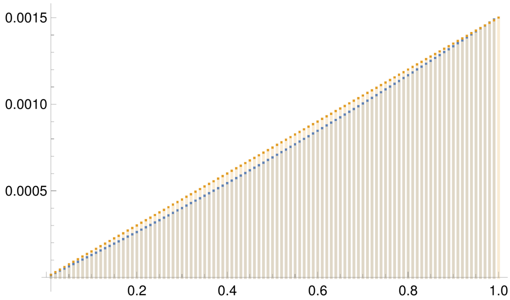
For $\pi_U = 0.5$ and $\pi_U = 0.75$
and small steps of $\lambda \in [0, 1]$.



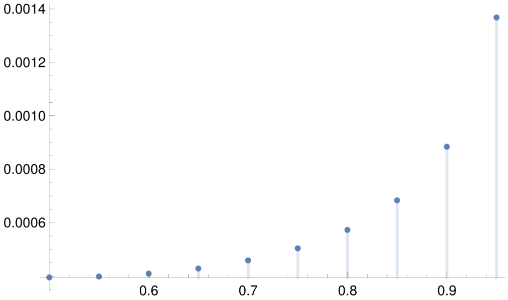
Small steps of $\lambda \in [0, 1]$,
the step ratio where $\pi_U = 0.5$.

Subsampling and Distributional Privacy

Subsampling and Distributional Privacy in Realtion to π_U



Comparison DP δ and distributional δ for the same ϵ and variable λ.



Changes in δ for $\epsilon = 0.001$ and small steps $\pi_U \in [0.5, 1]$.

Subsampling boosts privacy of property queries!

Added Noise and Distributional Privacy

Calculating δ

The goal is to bound the ratios in dependence of the variance ψ.

- $(F_U, M)(\mathcal{D}^n)$ mixed distributions.
  - Since the noise sample space is **R**.
  - Consider all outcomes of $F_U(\mathcal{D}^n)$.

Added Noise and Distributional Privacy

Calculating δ

The goal is to bound the ratios in dependence of the variance ψ.

- $(F_U, M)(\mathcal{D}^n)$ mixed distributions.
  - Since the noise sample space is **R**.
  - Consider all outcomes of $F_U(\mathcal{D}^n)$.

Laplace Noise

We archive pure ϵ-Distributional Privacy for

$$\epsilon \geq \frac{1}{\psi \cdot n}.$$

Gaussian Noise

The curve of (ϵ, δ) can be computed but as bound we get

$$\delta \leq Pr\left[(F_U, M) \geq \epsilon \cdot n \cdot \psi^2 + \frac{1}{2n}\right].$$

Added Noise and Distributional Privacy

Computing δ for Gaussian Noise

Take ratios of probability density functions $\mathrm{d}R^{(F_U,M)(\mathcal{D}^n_{i\notin U})}_{(F_U,M)(\mathcal{D}^n_{i\in U})}$ and compute the zero $x^\star$ of

$$\frac{\sum_{j=0}^{n-1} e^{-\frac{1}{2}\left(\frac{x-((j+1)/n)}{\psi}\right)^2}\binom{n-1}{j}\pi_F^j(1-\pi_F)^{n-j-1}}{\sum_{j=0}^{n-1} e^{-\frac{1}{2}\left(\frac{x-(j/n)}{\psi}\right)^2}\binom{n-1}{j}\pi_F^j(1-\pi_F)^{n-j-1}} - e^{\epsilon}.$$

Then compute the integral

$$\delta = \int_{x^\star}^{\infty}\left(1 - \frac{e^{\epsilon}}{\mathrm{d}R^{(F_U,M)(\mathcal{D}^n_{i\notin U})}_{(F_U,M)(\mathcal{D}^n_{i\in U})}(s)}\right)\sum_{j=0}^{n-1} e^{-\frac{1}{2}\left(\frac{x-((j+1)/n)}{\psi}\right)^2}\binom{n-1}{j}\pi_F^j(1-\pi_F)^{n-j-1}\mathrm{d}x.$$

Error Estimation

How does the mechanism affect the quality of the queries answer?

Method

- Consider (F_U, M) as estimator for π_U.
  - Calculate its quadratic error.
- The expected quadratic difference of (F_U, M) to F_U.

Error Estimation

How does the mechanism affect the quality of the queries answer?

Method

- Consider (F_U, M) as estimator for π_U.
  - Calculate its quadratic error.
- The expected quadratic difference of (F_U, M) to F_U.

Subsampling Estimator

- $(F_U, M) \sim (1/m)\mathcal{B}in(m, \pi_U)$
- Therefore (F_U, M) is unbiased.

Error Estimation

How does the mechanism affect the quality of the queries answer?

Method

- Consider (F_U, M) as estimator for π_U.
  - Calculate its quadratic error.
- The expected quadratic difference of (F_U, M) to F_U.

Subsampling Estimator

- $(F_U, M) \sim (1/m)\mathcal{B}in(m, \pi_U)$
- Therefore (F_U, M) is unbiased.

The variance is known as

$$MSE((F_U, M)) = \frac{\pi_U(1 - \pi_U)}{m}.$$

Error Estimation

Expected Difference Subsampling

$$\mathbf{E}_{\mathcal{D}^n}\left[\left((F_U, M) - F_U\right)^2\right] = \frac{\pi_U(1-\pi_U)}{m} - \frac{\pi_U(1-\pi_U)}{n}.$$

Error Estimation

Expected Difference Subsampling

$$\mathbf{E}_{\mathcal{D}^n}\left[((F_U, M) - F_U)^2\right] = \frac{\pi_U(1-\pi_U)}{m} - \frac{\pi_U(1-\pi_U)}{n}.$$

Added Noise

Assuming mean-free independent noise:

- Expected difference equals the variance.

Error Estimation

Expected Difference Subsampling

$$\mathbf{E}_{\mathcal{D}^n}\left[((F_U, M) - F_U)^2\right] = \frac{\pi_U(1-\pi_U)}{m} - \frac{\pi_U(1-\pi_U)}{n}.$$

Added Noise

Assuming mean-free independent noise:

- Expected difference equals the variance.

The mean square error:

$$MSE((F_U, M)) = \frac{\pi_U(1-\pi_U)}{n} + \mathrm{Var}\,(N_M)$$

Privacy comparison under fixed error

Comparison of the amplifying effect of different mechanisms.

Fixing the Error

- Added noise error equals the variance.
- Determine the variance with respect to the selection probability.
- Calculate the privacy parameters.

Privacy comparison under fixed error

Comparison of the amplifying effect of different mechanisms.

Fixing the Error

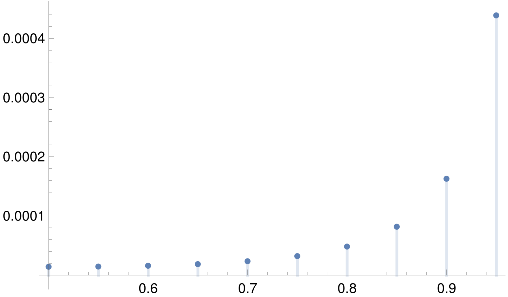
- Added noise error equals the variance.
- Determine the variance with respect to the selection probability.
- Calculate the privacy parameters.

Subsampling with selection probability λ has the same utility as added noise with variance:

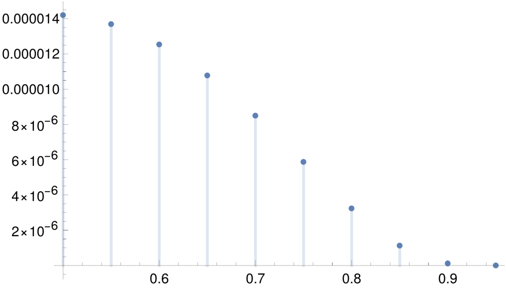
$$\psi = \frac{\pi_U(1-\pi_U)}{\lambda n}(1-\lambda)$$

Privacy Comparison under Fixed Error

Setting: $n = 1000, \lambda = 1/\sqrt{n}$



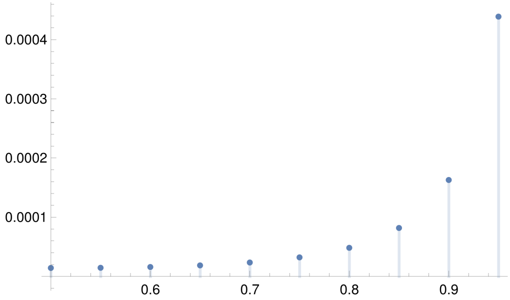
ϵ of Laplace-Noise for $\delta = 0$
in small π_U steps.



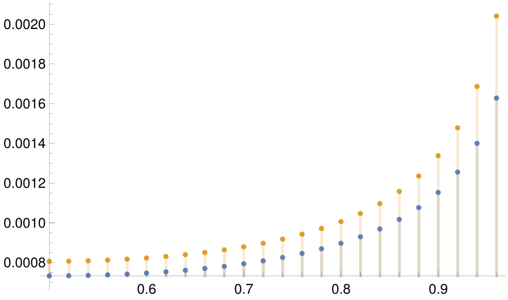
δ of (F_U, S_m) to the
right side $\epsilon/\sqrt{n}, \pi_U$ combination.

Privacy comparison under different π_U Values

Setting: $n = 1000, \lambda = 1/\sqrt{n}$



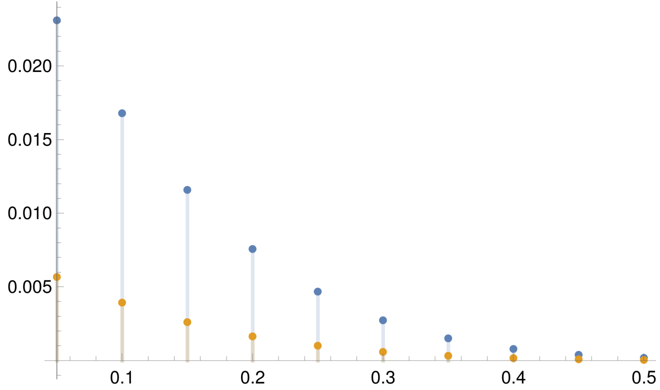
ϵ of Laplace-Noise for $\delta = 0$.



Here adjust ϵ/n and ϵ/n^2.

Privacy Comparison Gauss and Subsampling

Setting: $n = 100, \lambda = 1/\sqrt{n}$



δ comparison for gauss $\epsilon \in [0.05, 0.5]$
for subsampling the ϵ was amplified by 0.1.

Further Work

Subsampling

- General privacy amplification theorem.
- Composition queries.
  - Handling knowledge growth.
  - Reduce dependencies between queries.
  - Handling of privacy budget.

Model Extensions

- Handling knowledge growth/change.
  - Composition queries.