Differential Privacy with Dynamic Budget Allocation

Student: CHAN, Lo Yuet
Supervisor:LOUKIDES, Grigorios
Module: 7CCSMMCPJ
Date: 4-Sep-2017

OUTLINE

- Differential Privacy
- AHP and Static privacy budget allocation
- Simple Dynamic Privacy Budget Allocation on AHP (SDPA-AHP)
- Flexible Dynamic Privacy Budget Allocation on AHP (DPA-AHP)

DIFFERENTIAL PRIVACY

- Private, sensitive information on social media platform

- Life pattern and behaviour
- Branding strategy, business intelligence
- Privacy infringement

- Blurring while preserving important pattern in datasets

DIFFERENTIAL PRIVACY

- Assumes a differential privacy algorithm $\mathcal{A}$
- Assumes two neighbouring datasets $\mathbf{H}_1$, $\mathbf{H}_2$

- $\mathcal{A}$ is ϵ-differentially private only if

$$\Pr(\mathcal{A}(\mathbf{H}_1) = S) \leq \exp(\epsilon) \cdot \Pr(\mathcal{A}(\mathbf{H}_2) = S)$$

- ϵ is the privacy budget

DIFFERENTIAL PRIVACY

- Differential Privacy with clustering on histograms
  - E.g. AHP
- Node Differential Privacy
  - E.g. database with graphs / networks
- High Dimension Data
  - E.g. time-series data

ACCURATE HISTOGRAM PUBLICATION

- Suggested by Zhang in 2014
- Minimise Approximation error by clustering bins with similar counts
- Beats competing algorithms (PHP, NoiseFirst etc)
- Able to handle multi-dimension datasets

ACCURATE HISTOGRAM PUBLICATION

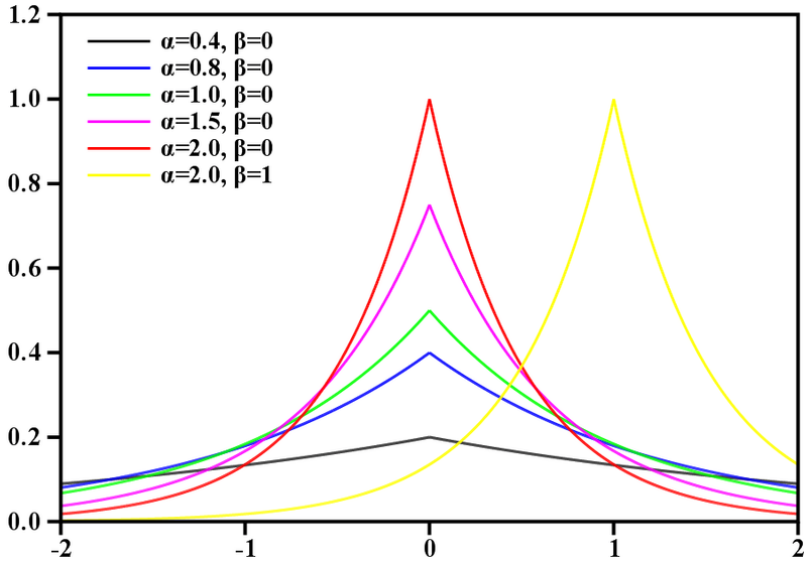
1. Allocate equal privacy budget to each bin
2. Apply i.i.d. Laplace noise to each bin based on allocated budget
3. Apply greedy clustering algorithm
4. Apply Laplace noise onto each bin based on cluster mean
5. Output sanitised data

ACCURATE HISTOGRAM PUBLICATION

1. **Allocate equal privacy budget to each bin**
2. **Apply i.i.d. Laplace noise to each bin based on allocated budget**
3. Apply greedy clustering algorithm
4. Apply Laplace noise onto each bin based on cluster mean
5. Output sanitised data

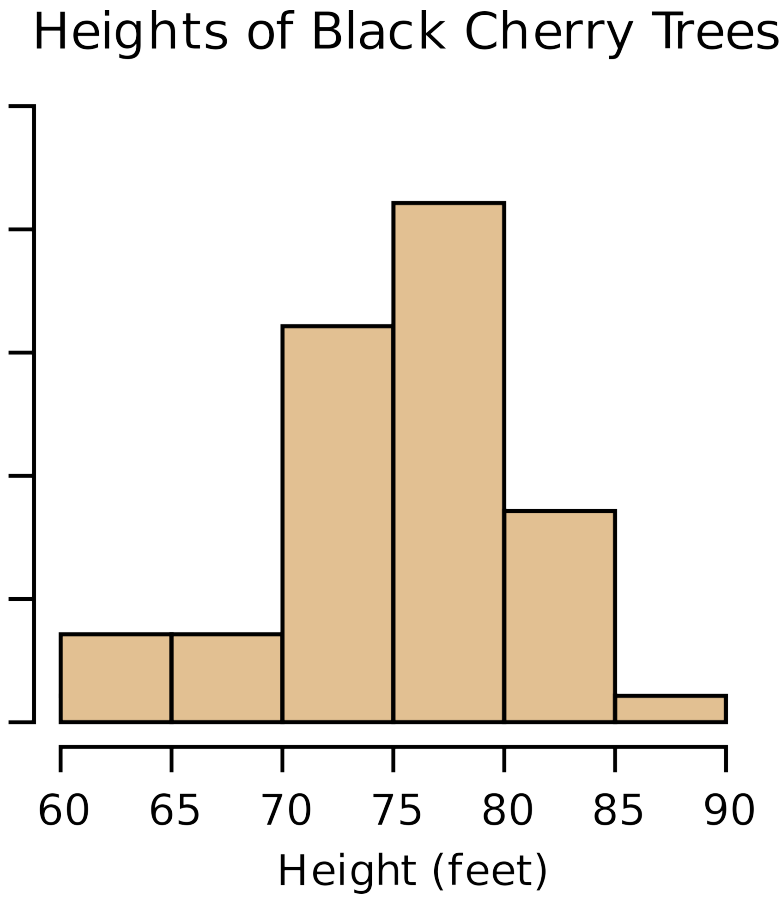
LIMITATION OF AHP

- Static privacy budget allocation
- Apply i.i.d. Laplace noise
- Lap (μ, b)



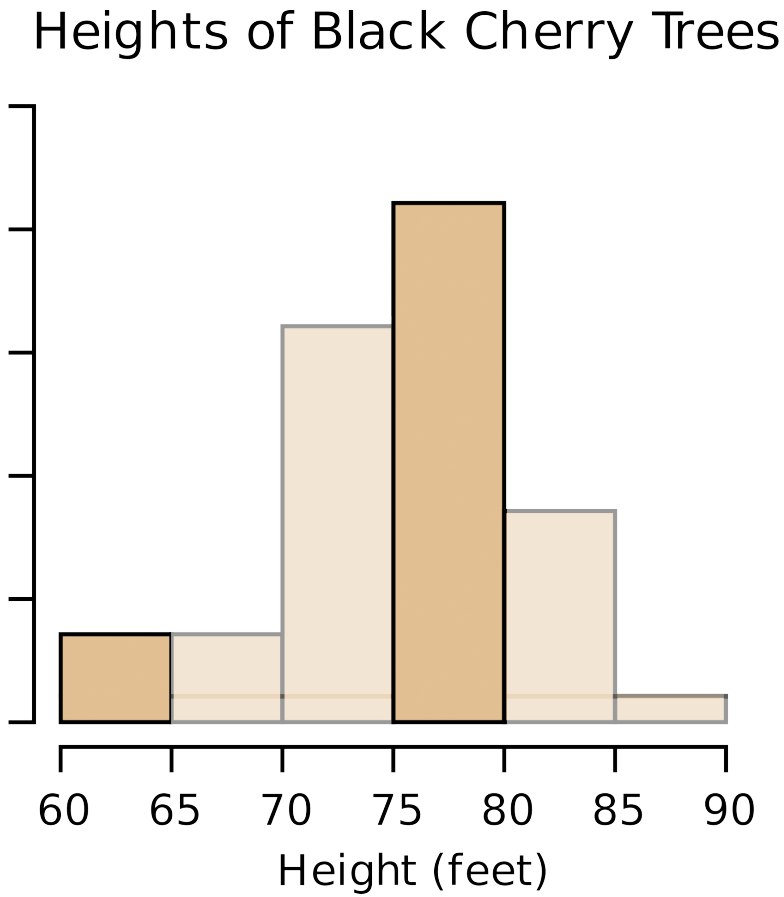
LIMITATION OF AHP

- Static privacy budget allocation
- Apply i.i.d. Laplace noise



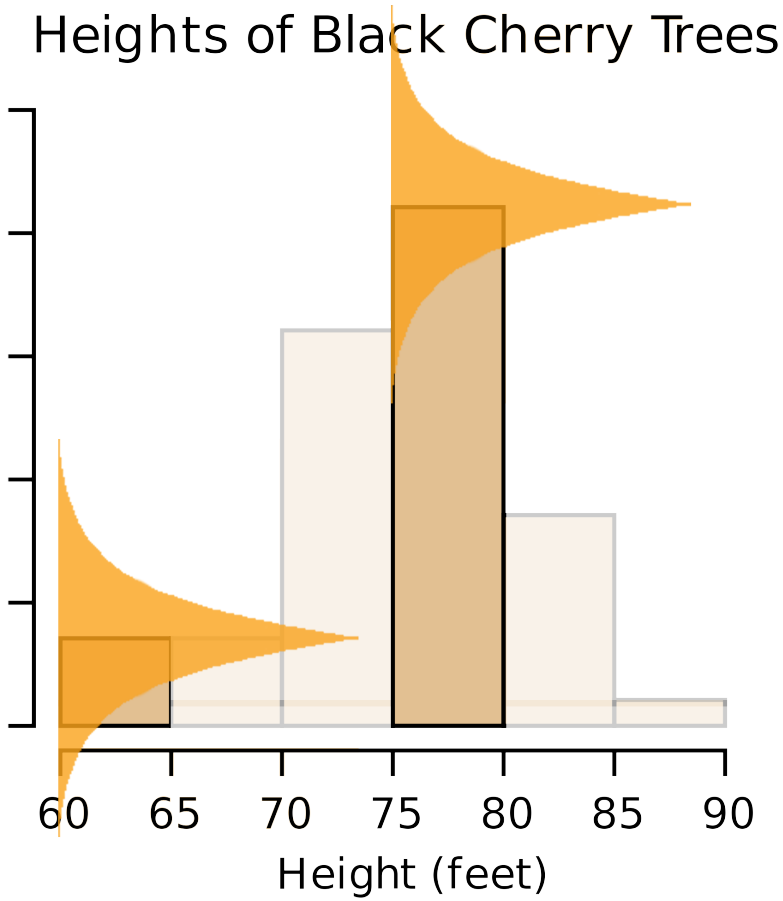
LIMITATION OF AHP

- Static privacy budget allocation
- Apply i.i.d. Laplace noise



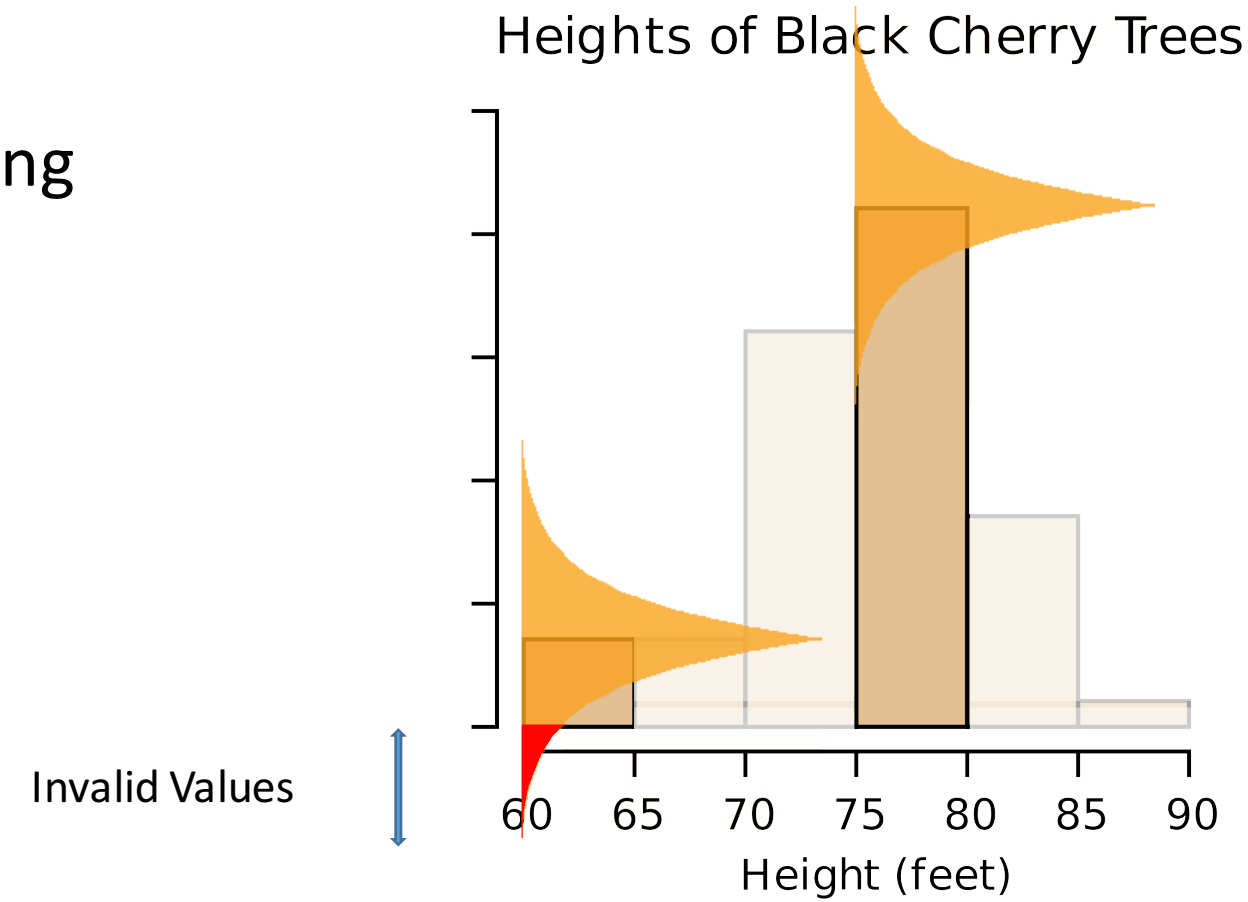
LIMITATION OF AHP

- Static privacy budget allocation
- Apply i.i.d. Laplace noise



LIMITATION OF AHP

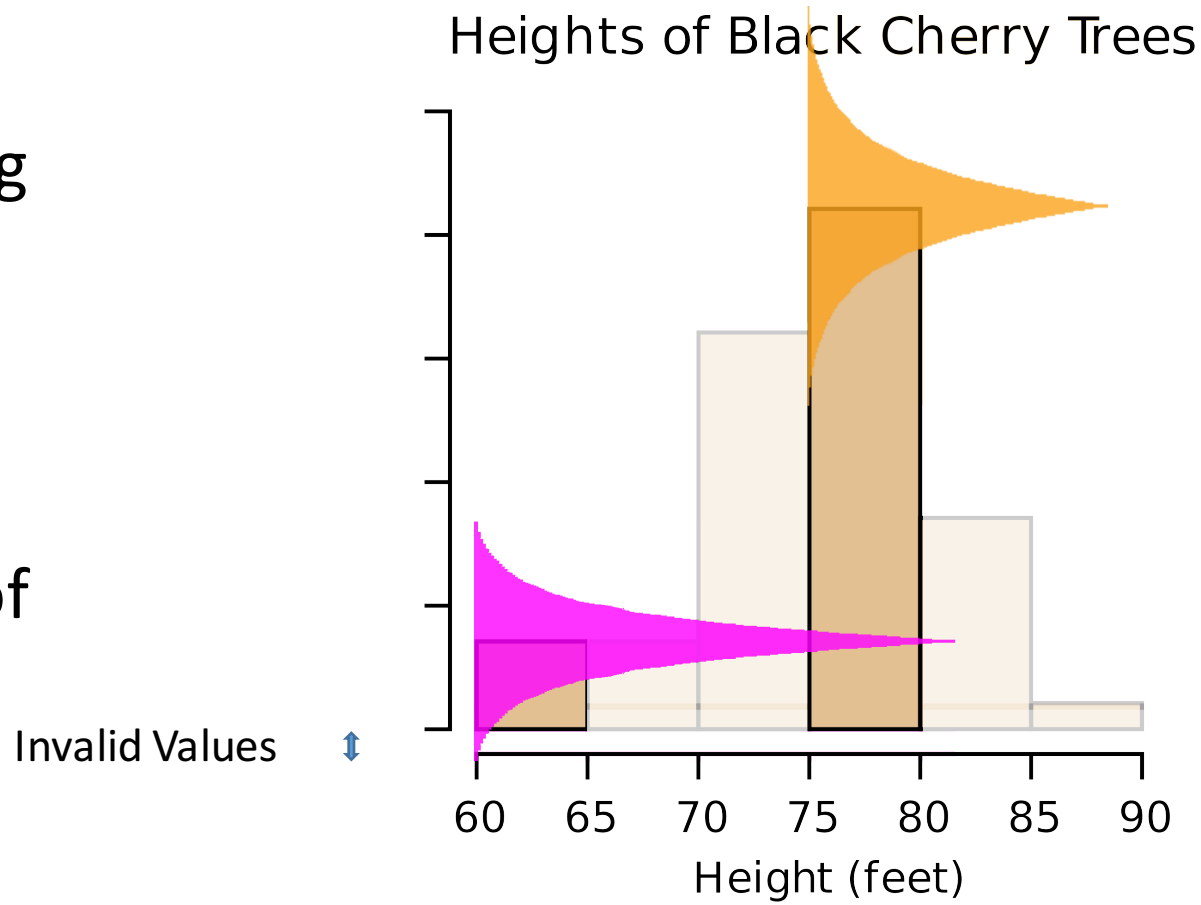
- Static privacy budget allocation
- Apply i.i.d. Laplace noise
- Result in high chance of getting invalid values



LIMITATION OF AHP

- Static privacy budget allocation
- Apply i.i.d. Laplace noise
- Result in high chance of getting invalid values

- Apply i.d. Laplace noise
- Minimise/reduce the chance of getting invalid values



NOVEL MECHANISMS

- Based on **ascendingly** sorted histograms

1. Simple Dynamic Privacy Budget Allocation on AHP (SDPA-AHP)

2. Flexible Dynamic Privacy Budget Allocation on AHP (DPA-AHP)
   - User can specific how dynamic the privacy budget is allocated
   - SDPA-AHP, and AHP are two special cases

SDPA-AHP

$$f(i,n) = \frac{n-i}{n \times \frac{n+1}{2}} \qquad for\ 0 \leq i \leq n-1$$

- where n is the number of bins, i as the index of the bin.

SDPA-AHP

$$f(i,n) = \frac{n-i}{n \times \frac{n+1}{2}} \qquad for\ 0 \leq i \leq n-1$$

- where n is the number of bins, i as the index of the bin.

Rank	Score
0	5
1	4
2	3
3	2
4	1

SDPA-AHP

$$f(i,n) = \frac{n-i}{n \times \frac{n+1}{2}} \qquad for\ 0 \leq i \leq n-1$$

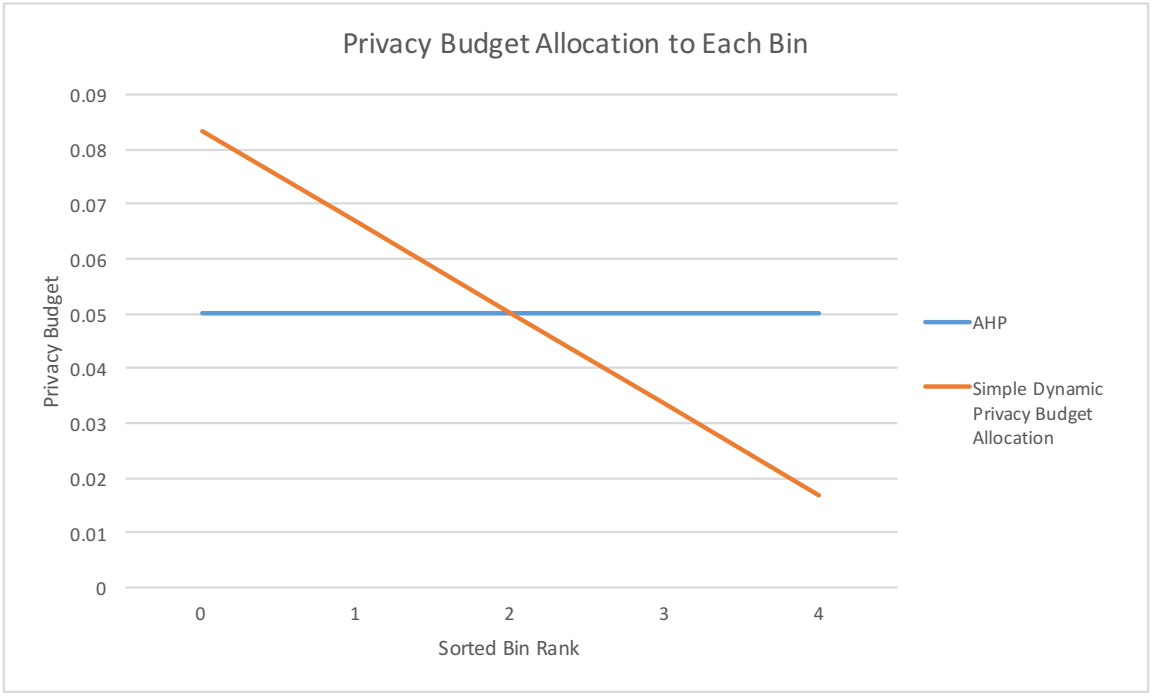
- where n is the number of bins, i as the index of the bin.

Rank	Score	Proportion of Score
0	5	0.333
1	4	0.267
2	3	0.200
3	2	0.133
4	1	0.067

SDPA-AHP

$$\epsilon_i = f(i,n) \times n \times \epsilon$$

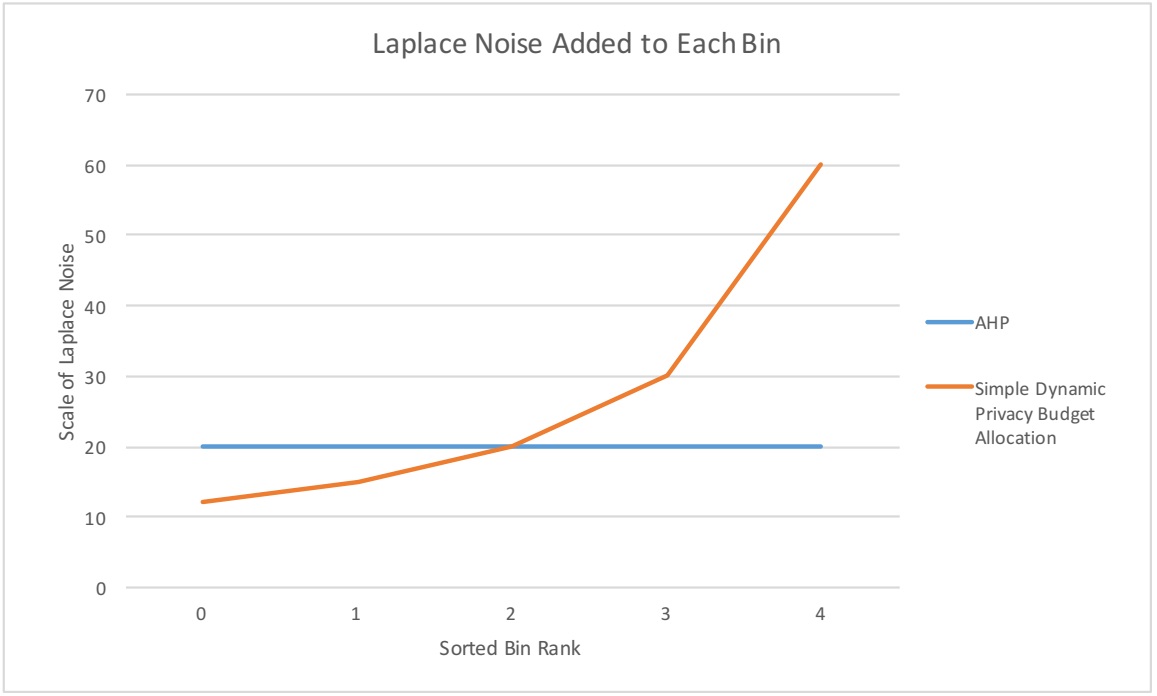
- where n is the number of bins, i as the index of the bin.



SDPA-AHP

$$\epsilon_i = f(i,n) \times n \times \epsilon$$

- where n is the number of bins, i as the index of the bin.



DPA-AHP

$$v(i,n,\delta) = \left(\left\lceil\frac{n}{2}\right\rceil + \frac{n-2i-1}{2}\times\delta\right) \qquad for\ 0 \le i \le n-1$$

$$f(i,n,\delta) = \frac{v(i,n,\delta)}{\sum_{i=0}^{n-1} v(i,n,\delta)} \qquad for\ 0 \le i \le n-1$$

- where n is the number of bins, i as the index of the bin.
- δ specifyies how the privacy budget is allocated

DPA-AHP

- Recall in SDPA-AHP

Rank	Score	$f(i,n)$
0	5	0.333
1	4	0.267
2	3	0.200
3	2	0.133
4	1	0.067

DPA-AHP

- Steps / Score difference are adjustable (δ)

Rank	$f(i,n,0)$	$f(i,n,0.5)$	$f(i,n,1)$
0	0.200	0.267	0.333
1	0.200	0.233	0.267
2	0.200	0.200	0.200
3	0.200	0.167	0.133
4	0.200	0.133	0.067

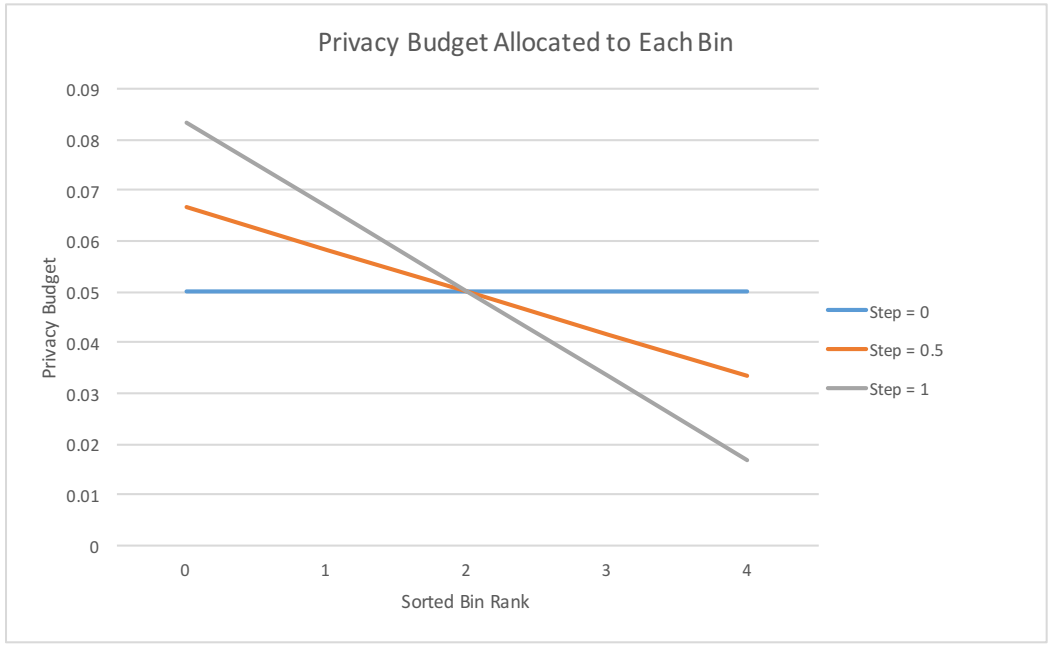
DPA-AHP

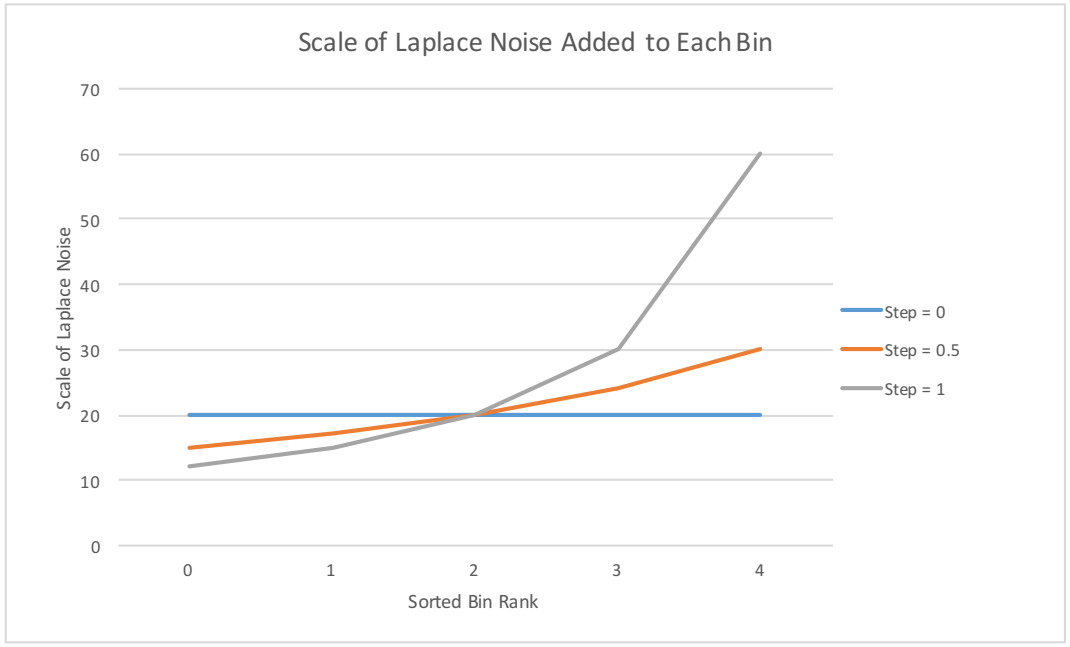
- Steps / Score difference are adjustable (δ)

Rank	$f(i,n,0)$	$f(i,n,0.5)$	$f(i,n,1)$
0	0.200	0.267	0.333
1	0.200	0.233	0.267
2	0.200	0.200	0.200
3	0.200	0.167	0.133
4	0.200	0.133	0.067
	AHP		**SDPA-AHP**

DPA-AHP

- Scale of noise increases less exponentially with smaller δ (Step)





TESTS & EXPERIMENTS

- Six testing datasets
  - Four real datasets
  - Two synthetic datasets

Dataset Name		Rows	Histograms
TSMC_NYC	Real	227428	1083
TSMC_TKY	Real	57303	2293
UBICOMP	Real	27149	2060
CONPOLBLOGS	Real	14409	1
LOGNORMAL	Synthetic	3544	66
EXPONENTIAL	Synthetic	25929	51

TESTS & EXPERIMENTS

- Two evaluation metrics
  - Kullback-Leibler divergence (KLD) – measures difference in distribution
  - Mean Squared Error (MSE) – measures errors of range queries

$$KLD(\boldsymbol{H}, \widetilde{\boldsymbol{H}}) = \sum_{i=1}^{n} H_i \ln \frac{H_i}{\widetilde{H_i}}$$

$$MSE(\boldsymbol{H}, \widetilde{\boldsymbol{H}}, \boldsymbol{Q}) = \frac{\sum_{i=1}^{m} \left(Q_i(\boldsymbol{H}) - Q_i\left(\widetilde{\boldsymbol{H}}\right) \right)^2}{m}$$

TESTS & EXPERIMENTS

- DPA-AHP beats AHP by about 10% on KLD
- Experiments repeated with $\epsilon = 0.01$, $\epsilon = 0.1$, and $\epsilon = 1$

- $\delta = 0.075$ is suggested

Dataset	$\delta = 0$ (AHP)	$\delta = 0.025$	$\delta = 0.05$	$\delta = 0.075$	$\delta = 0.1$	$\delta = 0.15$
TSMC_NYC	8.625	8.484	**7.934**	8.588	8.620	8.606
TSMC_TKY	9.252	9.082	8.861	**8.346**	8.768	9.009
UBICOMP	0.711	0.666	0.643	**0.604**	0.703	0.737
CONPOLBLOGS	7.787	7.630	7.565	**6.582**	7.184	7.528
LOGNORMAL	5.012	4.895	4.526	**4.454**	4.829	4.843
EXPONENTIAL	11.783	11.597	11.449	**11.364**	10.666	11.356

Table KLD results for $\epsilon = 0.01$

TESTS & EXPERIMENTS

Dataset	$\delta = 0$ (AHP)	$\delta = 0.025$	$\delta = 0.05$	$\delta = 0.075$	$\delta = 0.1$	$\delta = 0.15$
TSMC_NYC	8.602	8.567	7.937	**7.879**	8.490	8.654
TSMC_TKY	8.813	8.719	8.464	**8.034**	8.511	8.594
UBICOMP	0.698	0.681	**0.622**	0.708	0.734	0.747
CONPOLBLOGS	6.429	6.303	**6.201**	6.304	6.429	6.604
LOGNORMAL	5.038	4.973	**4.573**	4.671	4.841	4.883
EXPONENTIAL	7.504	7.325	**6.928**	7.208	7.293	7.292

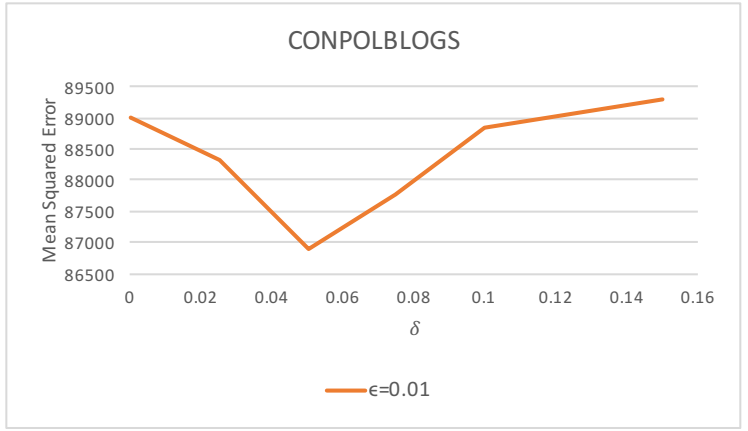
Table KLD results for $\epsilon = 0.1$

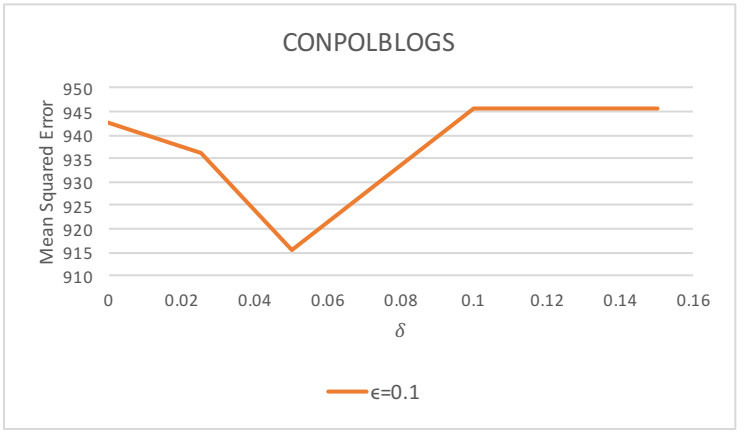
Dataset	$\delta = 0$ (AHP)	$\delta = 0.025$	$\delta = 0.05$	$\delta = 0.075$	$\delta = 0.1$	$\delta = 0.15$
TSMC_NYC	2.782	2.694	**2.560**	2.582	2.689	2.757
TSMC_TKY	2.831	2.808	**2.655**	2.658	2.705	2.794
UBICOMP	1.046	1.046	1.029	**0.958**	0.978	1.01
CONPOLBLOGS	1.220	**1.186**	1.196	1.221	1.221	1.270
LOGNORMAL	2.994	2.940	2.740	**2.617**	2.677	2.898
EXPONENTIAL	0.904	0.837	0.826	**0.822**	0.842	0.851

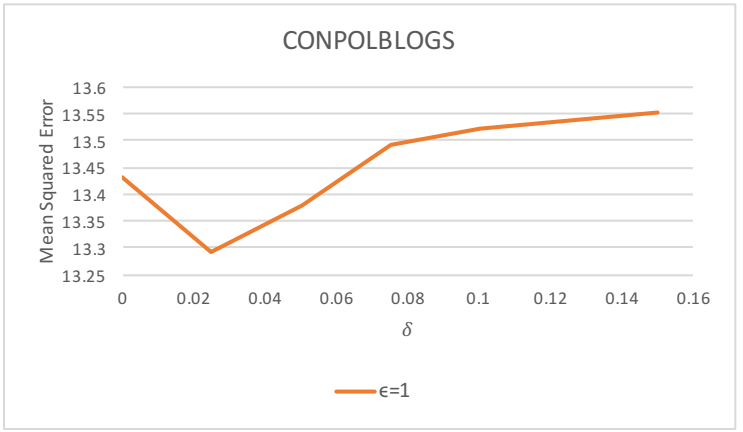
Table KLD results for $\epsilon = 1$

TESTS & EXPERIMENTS

- Dip in MSE is also observed
- $\delta = 0.05$ to $\delta = 0.075$ is suggested

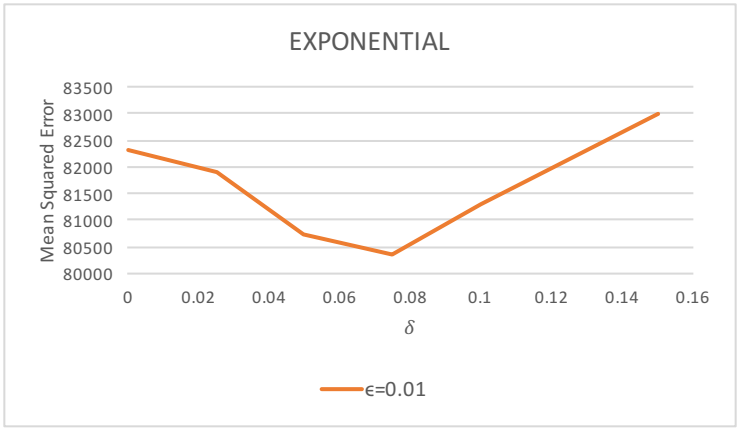


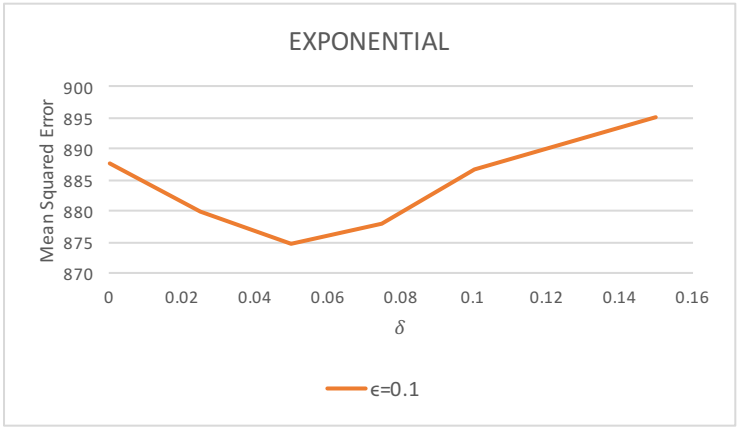


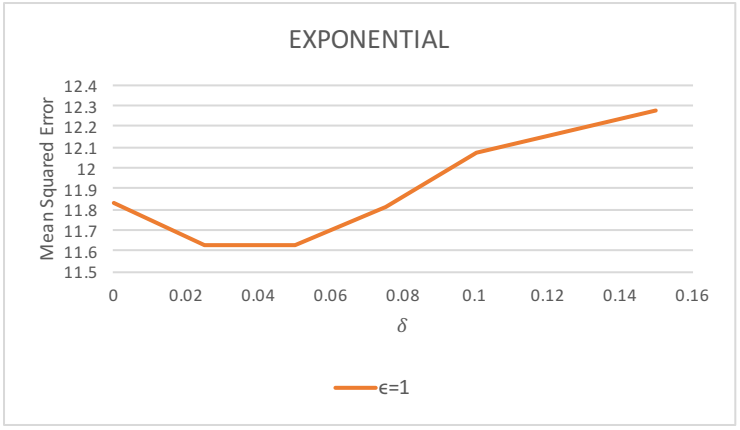


TESTS & EXPERIMENTS

- Dip in MSE is also observed
- $\delta = 0.05$ to $\delta = 0.075$ is suggested

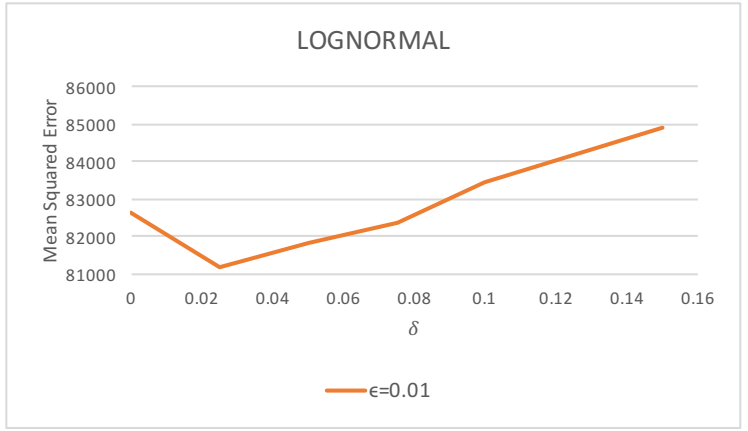


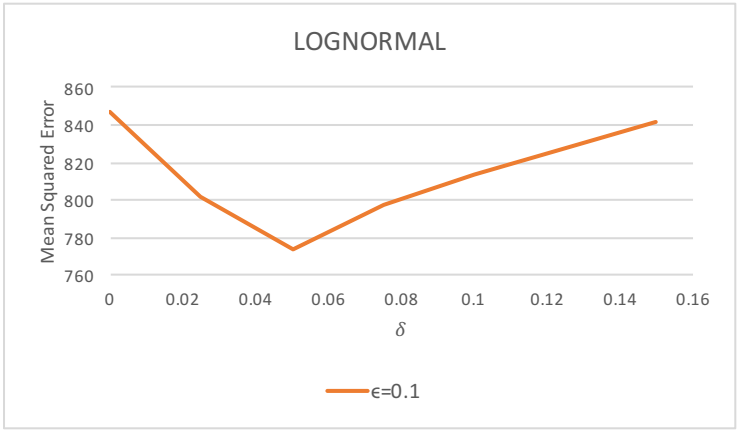


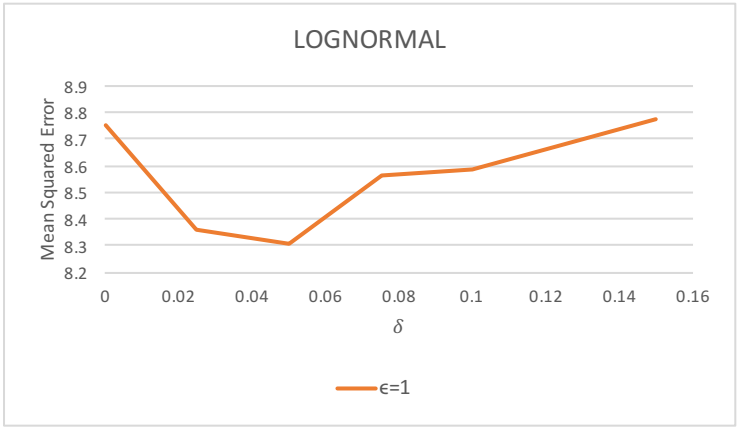


TESTS & EXPERIMENTS

- Dip in MSE is also observed
- $\delta = 0.05$ to $\delta = 0.075$ is suggested

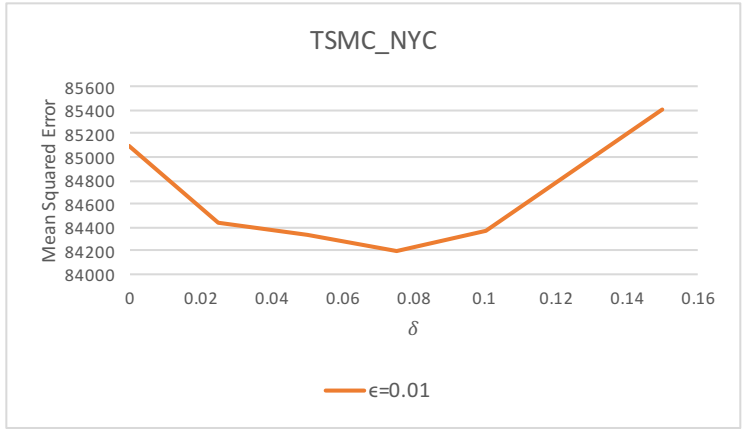


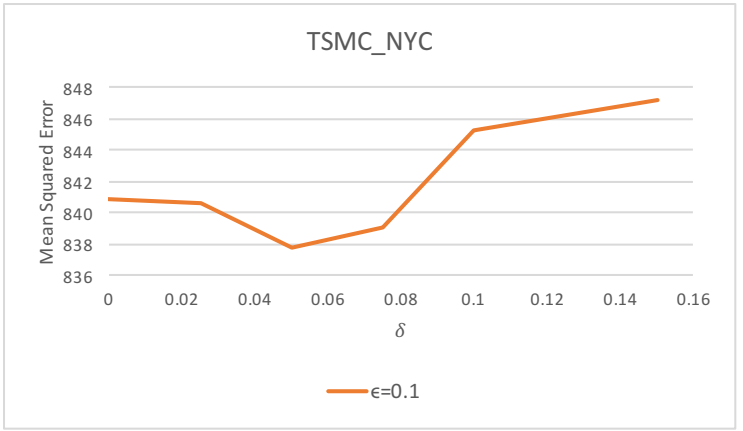


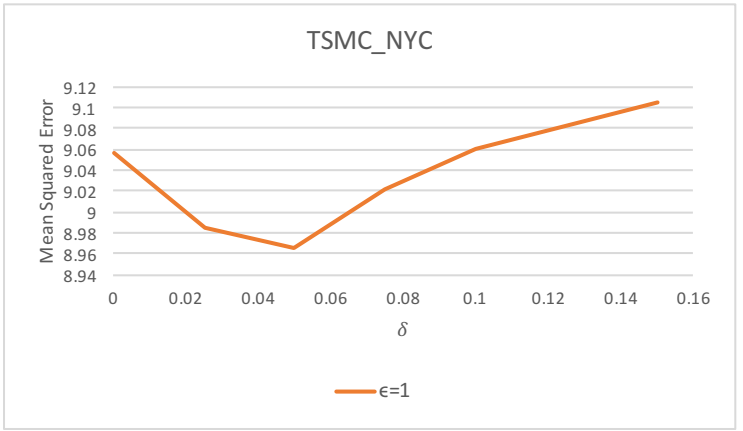


TESTS & EXPERIMENTS

- Dip in MSE is also observed
- $\delta = 0.05$ to $\delta = 0.075$ is suggested

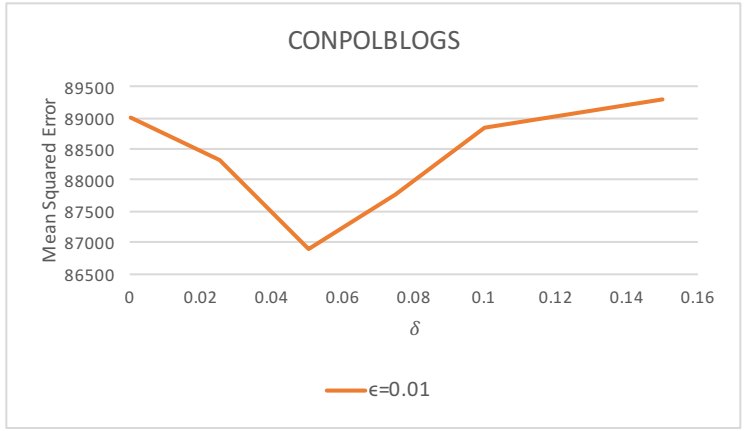


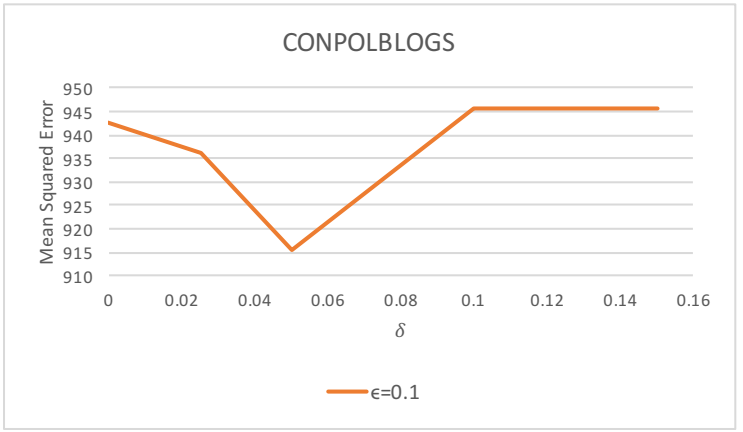


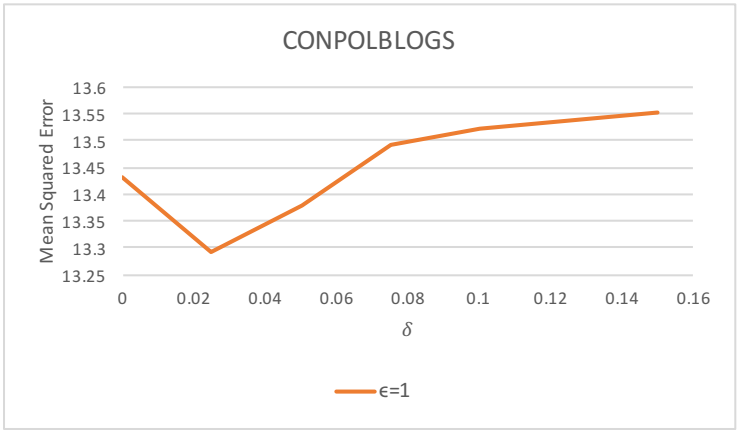


TESTS & EXPERIMENTS

- Dip in MSE is also observed
- $\delta = 0.05$ to $\delta = 0.075$ is suggested

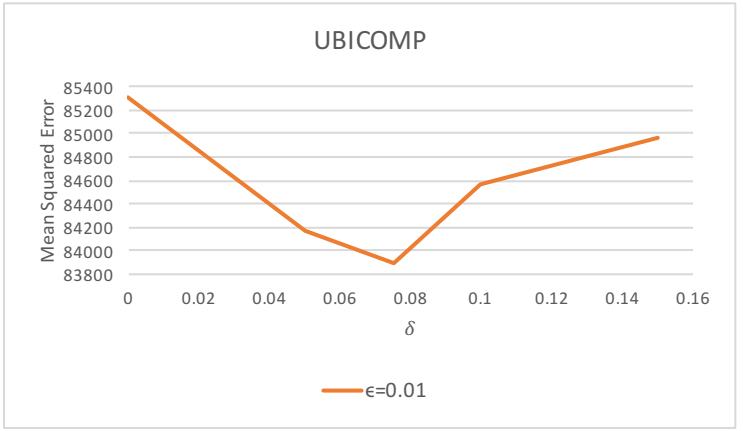


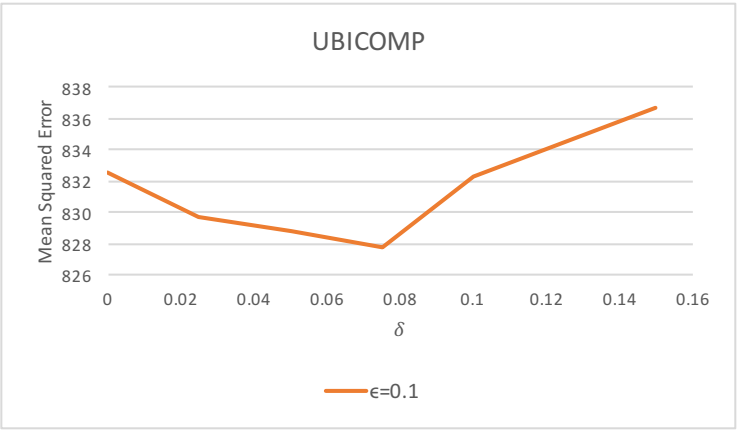


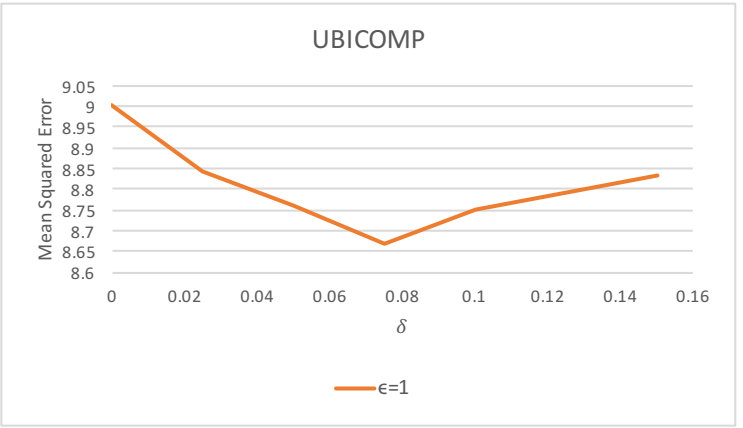


TESTS & EXPERIMENTS

- Dip in MSE is also observed
- $\delta = 0.05$ to $\delta = 0.075$ is suggested



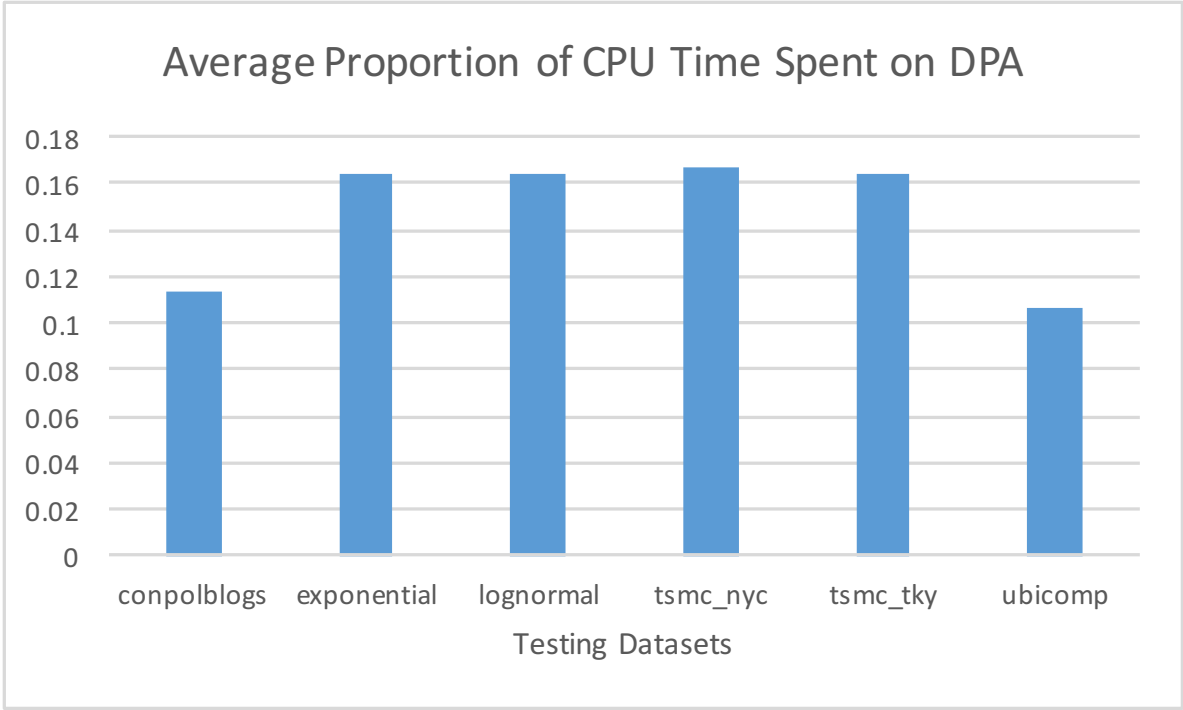




TESTS & EXPERIMENTS

- $O(n)$ time complexity
- Align with AHP

- DPA takes 20% of CPU time



CONCLUSION

- Two novel ϵ-differentially private mechanisms
  - SDPA-AHP, DPA-AHP
- Dynamic privacy budget allocation
- Bested AHP on both KLD and MSE
- Time complexity aligns with AHP

DEMO SECTION

Q & A SECTION