AINA2023

Expectation-Maximization Estimation for Key-Value Data Randomized with Local Differential Privacy

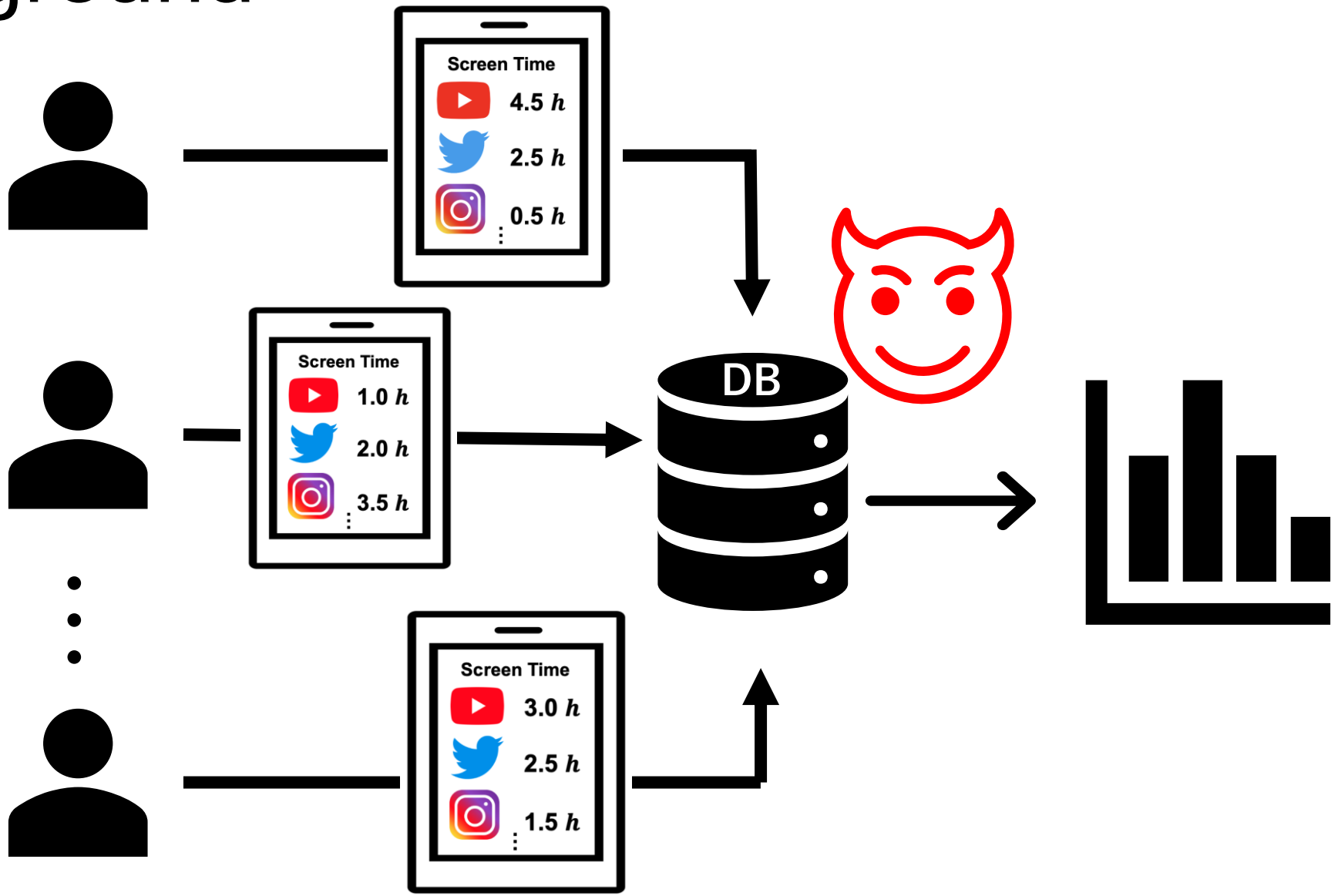
Hikaru Horigome, Hiroaki Kikuchi

(Meiji University)

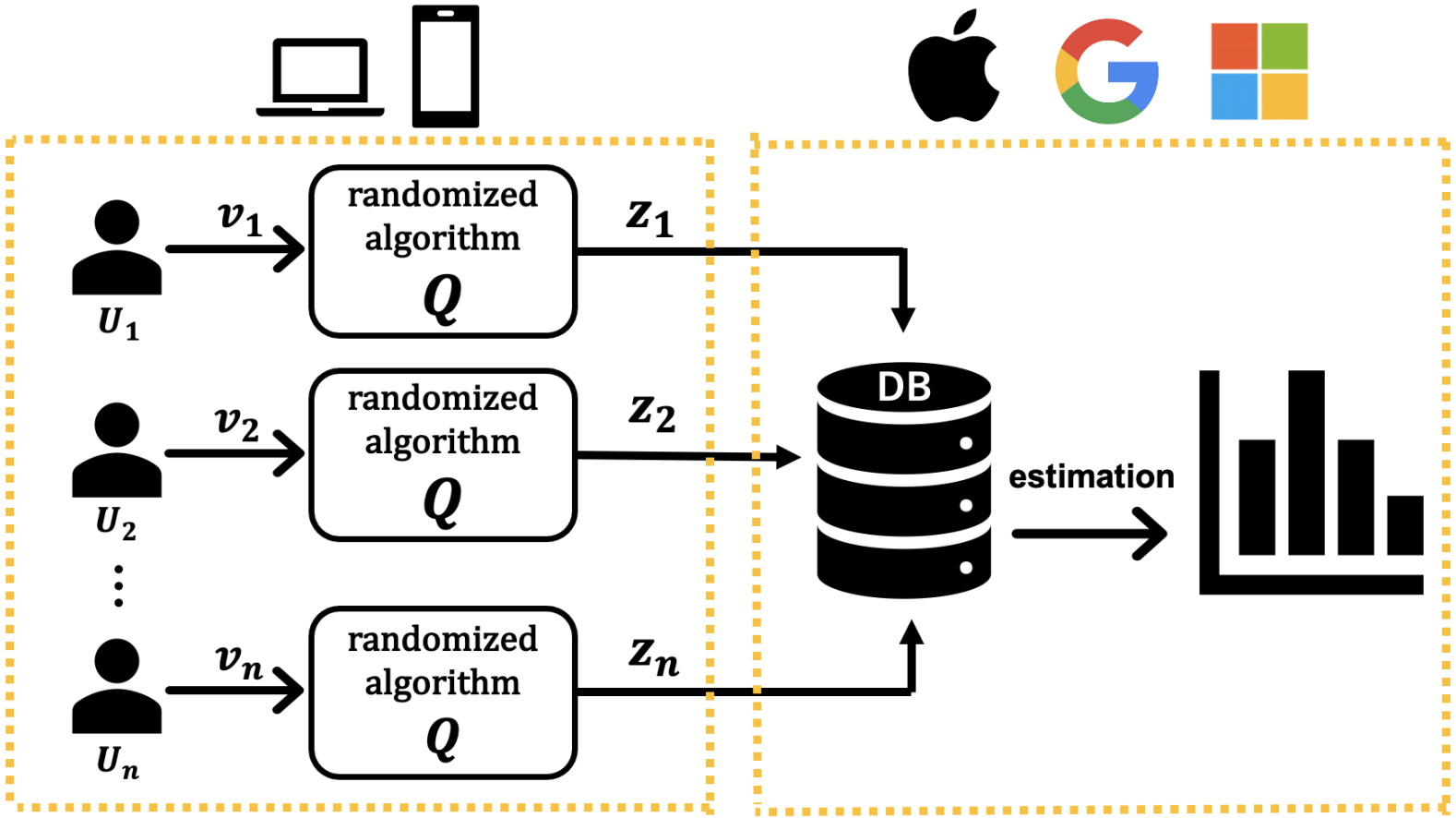
Chia-Mu Yu

(National Yang Ming Chiao Tung University)

Background


Screen Time
4.5 h
2.5 h
0.5 h
Screen Time
1.0 h
2.0 h
3.5 h
Screen Time
3.0 h
2.5 h
1.5 h
DB

Local Differential Privacy (LDP) [Duchi, et al., 2013]



A randomized algorithm Q satisfies ε-LDP if for all pair of values $v, v' \in V$ and all subset S of range Z ($S \subset Z$)

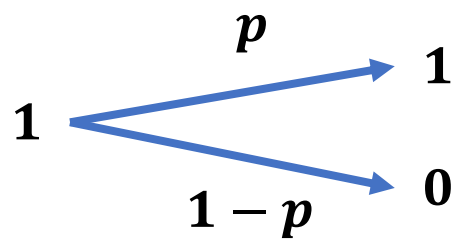
$$\frac{Pr[Q(v) \in S]}{Pr[Q(v') \in S]} \leq e^{\varepsilon}$$

PrivKV [Q.Ye , et al., 2019]

$< k_{YouTube}, v_{YouTube} >$

$< k'_a, v'_a > = < 1 , 0.5 >$

Randomized Response(RR)



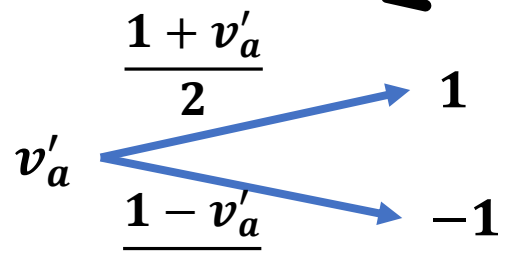
Value Perturbation Primitive(VPP)

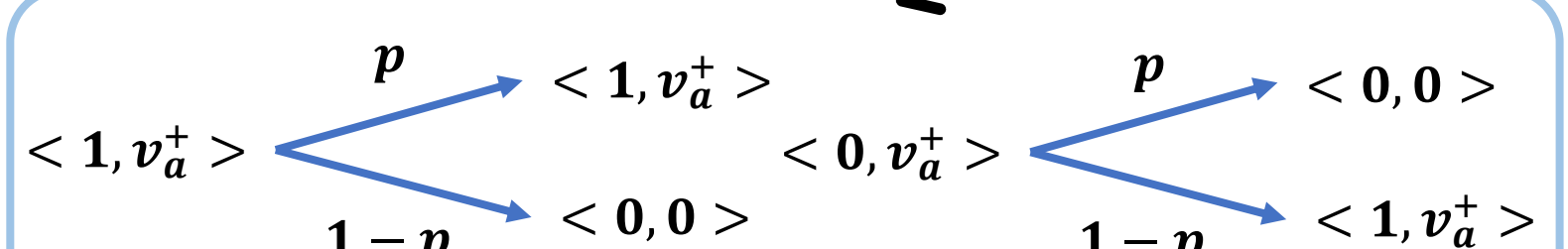
$< k'_a, v'_a > \rightarrow$ **discretization**$(v'_a) \rightarrow < k'_a, v^*_a > \rightarrow RR(v^*_a) \rightarrow < k'_a, v^+_a > \rightarrow RR(k'_a) \rightarrow < k^*_a, v^+_a >$

$v'_a = [-1, 1]$

$v^*_a = \{-1, 1\}$

$v^+_a = \{-1, 0, 1\}$

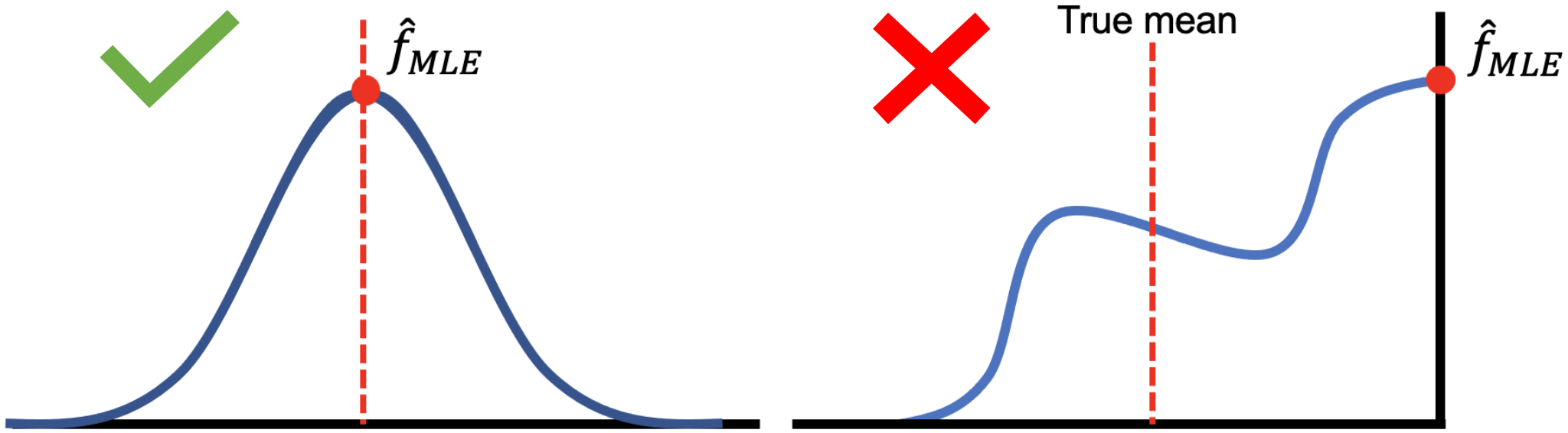




Q. Ye, H. Hu, X. Meng, H. Zheng, “PrivKV : Key-Value Data Collection with Local Differential Privacy”, IEEE S&P, pp. 294-308, 2019.

Problems of PrivKV

- Most Likelihood Estimation(MLE) : Low estimation accuracy



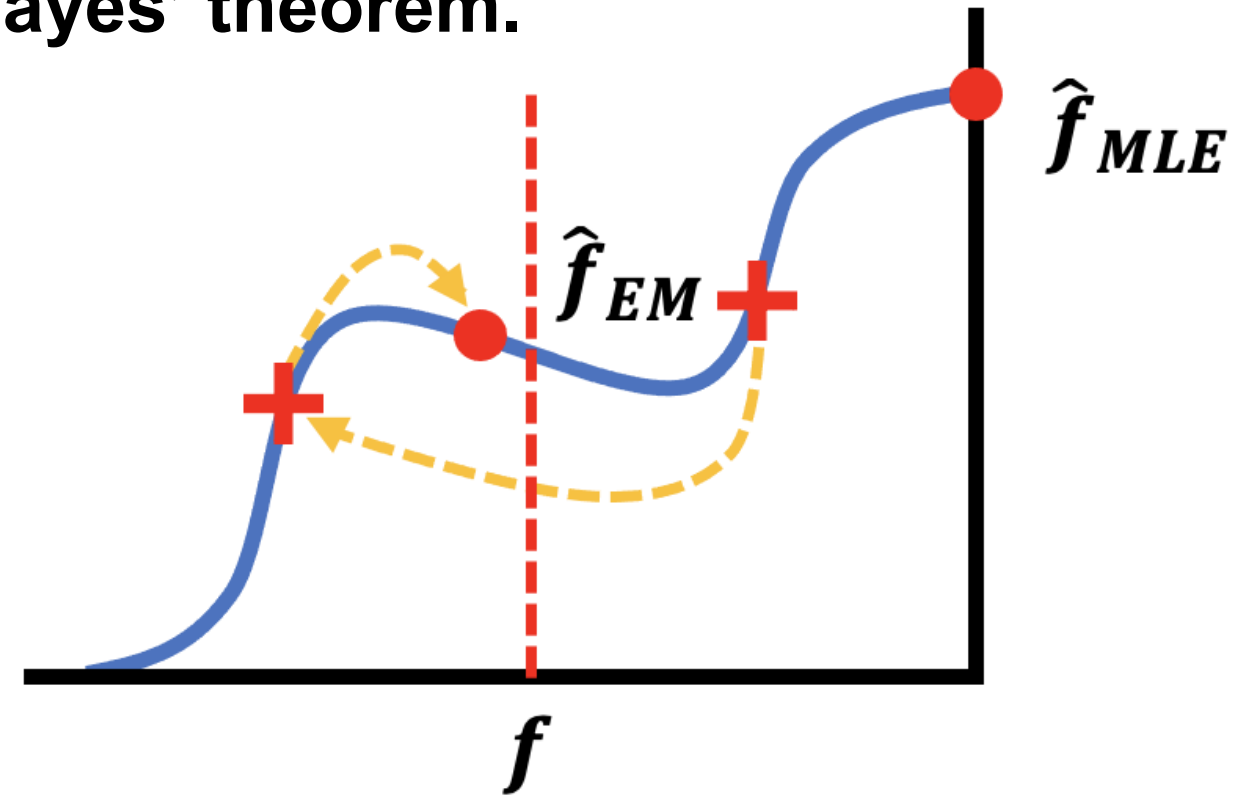
Our idea

- Our Approach
  - Expectaion Mazimization (EM) algorithm estimation
- Proposal
  - We apply the EM algorithm to the randomization used in PrivKV

	Randomization	Estimation
PrivKV(M)	Randomized Response (RR)	MLE
Ours	+ Value Perturbation Primitive(VPP)	EM

Why EM algorithm works better?

- **EM algorithm [Dempster, et al., 1997]**
- **Iterative process for which posterior probabilities are updated based on Bayes' theorem.**



Our idea

VPP(randomized value)

randomized key

$< k'_a, v'_a > \rightarrow$ **discretization**$(v'_a) \rightarrow < k'_a, v^*_a > \rightarrow RR(v^*_a) \rightarrow < k'_a, v^+_a > \rightarrow RR(k'_a) \rightarrow < k^*_a, v^+_a >$

$v'_a = [-1, 1]$ $\quad v^*_a = \{-1, 1\}$ $\quad v^+_a = \{-1, 0, 1\}$

Estimating the marginal probability of $< k'_a, v^*_a >$ using the output $< k^*_a, v^+_a >$

✗ $v^+_a \rightarrow v'_a$ $\quad$ ✓ $v^+_a \rightarrow v^*_a$

$< k^*_a, v^+_a > = \{< 1, 1 >, < 1, -1 >, < 0, 0 >\}$

$< k'_a, v^*_a > = \{< 1, 1 >, < 1, -1 >, < 0, 1 >, < 0, -1 >\}$

$< k_a , v_a > = < 0 , \_ * >$

Instead of independently calculating the frequency and mean values, we consider the key-value data as a whole

Experiment 1

- **Datasets**

3 synthetic datasets

Distribution	#users	#items	value range
Gaussian	10^5	100	$-1\ -1$
Power-law			
Linear			

2 open datasets

Datasets	#ratings	#users	#items	value range
MovieLens	10,000,054	69,877	10,677	$0.5\ -5$
Clothing	192,544	96,57	3,831	$1\ -10$

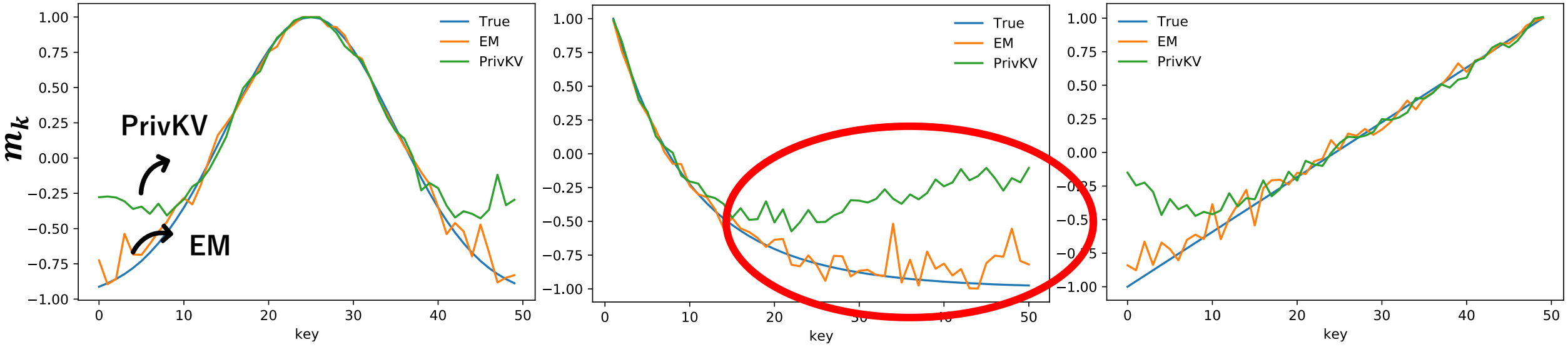
- **Evaluation metrics**
  - **Mean Square Error(MSE)**

$$MSE(f) = \frac{1}{|K|}\sum_{i=1}^{|K|}\left(\frac{\hat{f}_i}{n} - \frac{f_i}{n}\right)^2$$

$$MSE(m) = \frac{1}{|K|}\sum_{i=1}^{|K|}(\hat{m}_i - m_i)^2$$

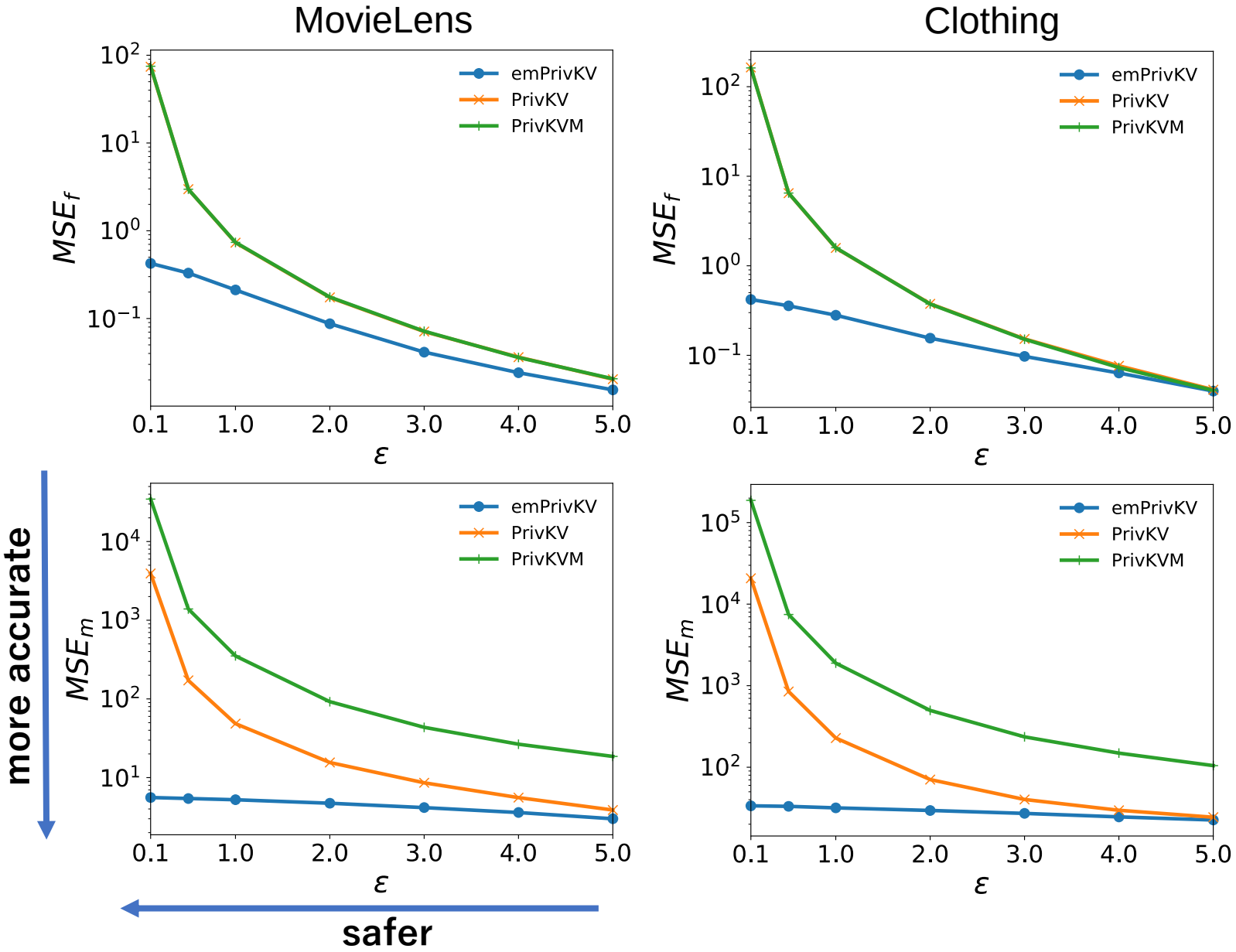
Result1 Estimated mean distribution

$$(\epsilon = 4,\ \ \mathbf{n} = 10^5)$$



EM is more accurately when frequencies are small

Result 2 Estimation error MSE

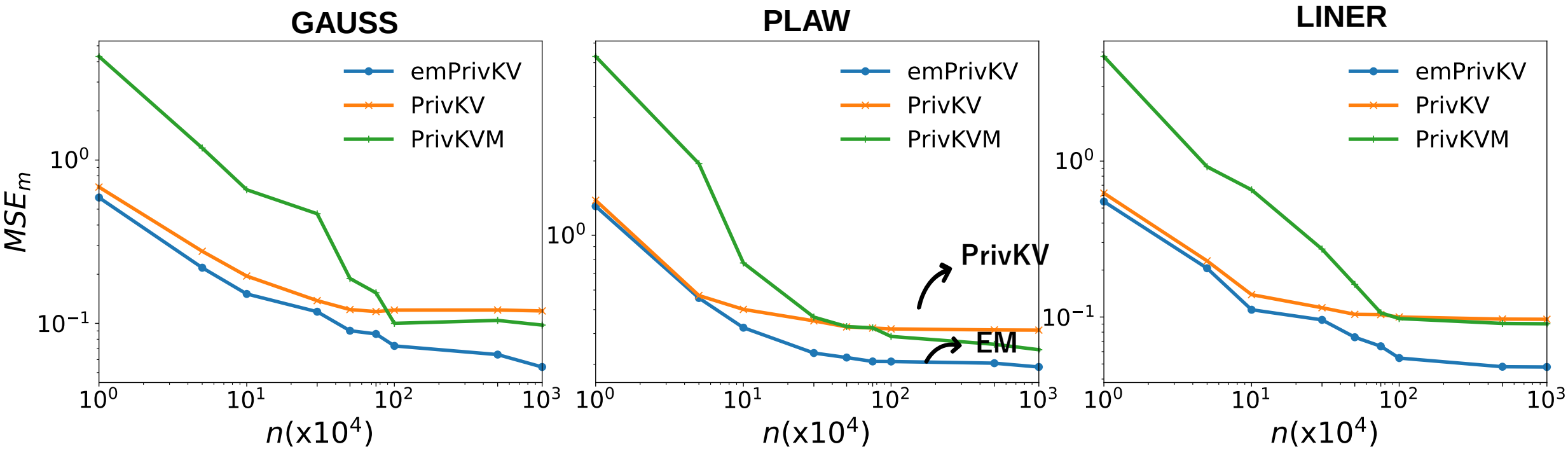


- A proposed method is more accurate than the other two methods regardless to ε.
- Especially effective for smaller values of ε.

Result 3

MSE_m with regard to the number of users n

#key=50, $\varepsilon = 2$



EM's MSE_m is the smallest for all n

Conclusion

- **We studied the LDP algorithms for key-value data that estimate the key frequencies and the mean values.**
- **We propose an algorithm based on the EM algorithms to improve the estimation accuracy perturbed in the LDP algorithms PrivKV.**
- **The EM algorithm estimates the frequency and mean well even when the frequency of the key-value record is limited.**
- **Hence, we conclude that the proposed EM algorithm is appropriate for private data analysis dealing with rare item.**