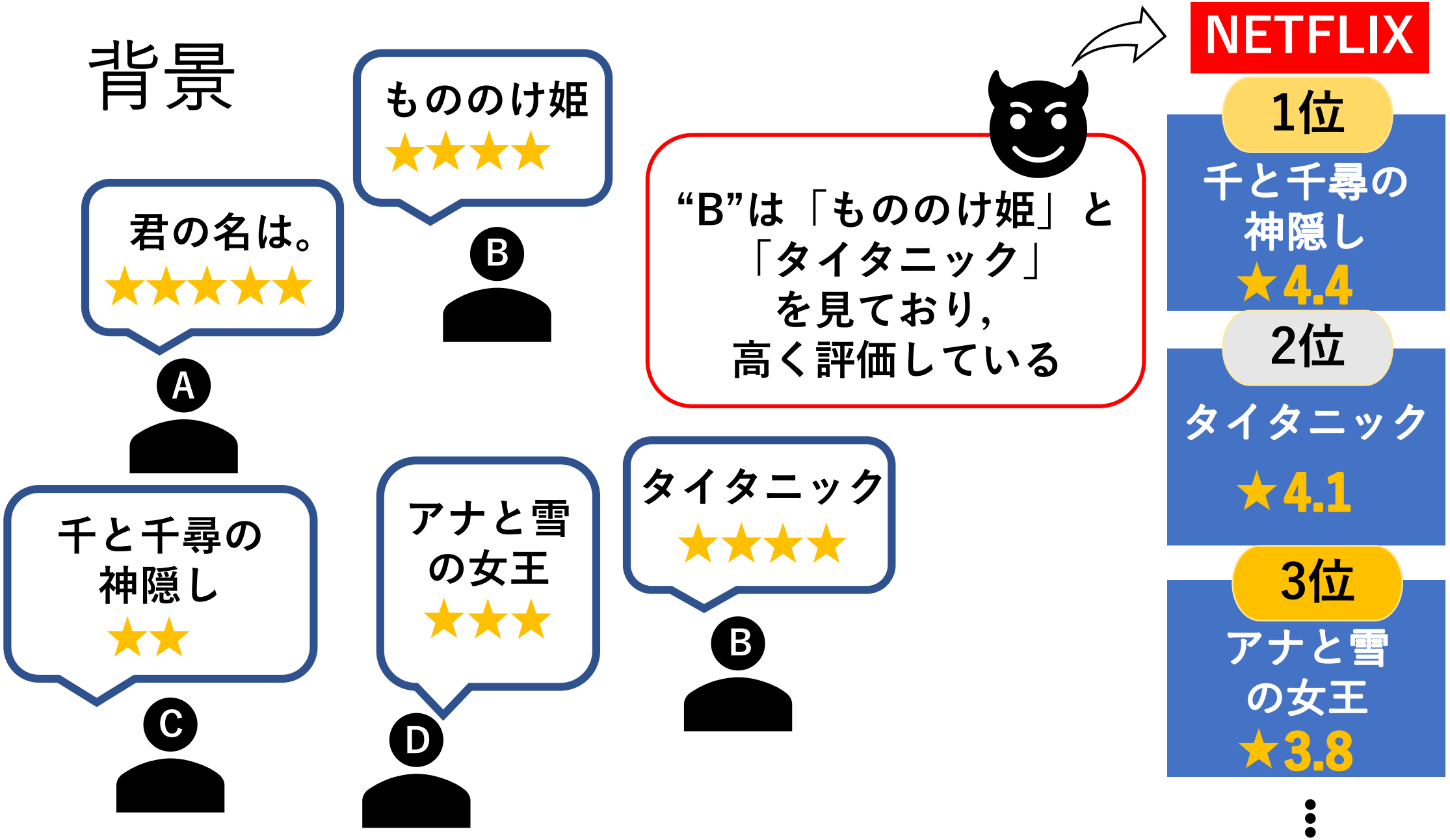
DICOMO 2022

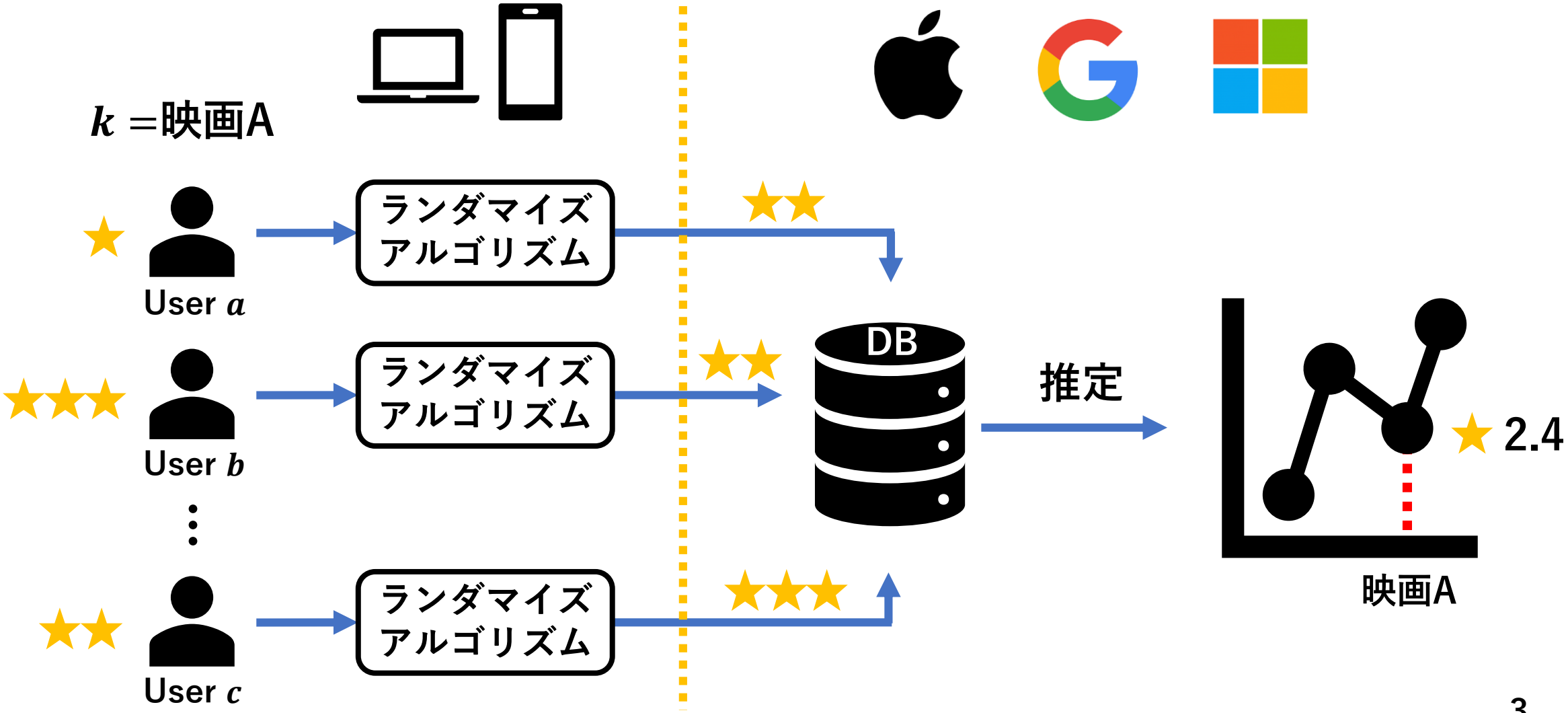
key-valueデータにおける局所差分プライバシアルゴリズムPrivKVの改良

堀込光，菊池浩明（明治大学）
Chia-Mu Yu（National Yang Ming Chiao Tung University）

背景


ものの け姫
★★★★
B
君の名は。
★★★★★
A
千と千尋の
神隠し
★★
C
アナと雪
の女王
★★★
D
タイタニック
★★★★
B
“B”は「もののけ姫」と
「タイタニック」
を見ており，
高く評価している
NETFLIX
1位
千と千尋の
神隠し
★4.4
2位
タイタニック
★4.1
3位
アナと雪
の女王
★3.8

局所差分プライバシー(LDP)



先行研究：PrivKV（1. 摂動）

key-valeデータ

k	v
千と千尋の神隠し	2
タイタニック	*
アナと雪の女王	*
君の名は。	4
もののけ姫	3

サンプリング $< k, v >$
$<$ 君の名は。$, 4 >$

PrivKV

出力

a	k^* (フラグ)	v^+
君の名は。	1	-1

$a =$ “君の名は。”

VPP (valueのランダマイズ)

RR(keyのランダマイズ)

$< k'_a, v'_a > \rightarrow v'_a$の2値化 $\rightarrow < k'_a, v^*_a > \rightarrow RR(v^*_a) \rightarrow < k'_a, v^+_a > \rightarrow RR(k'_a) \rightarrow < k^*_a, v^+_a >$

$< 1, 0.5 > \longrightarrow < 1, 1 > \longrightarrow < 1, -1 > \longrightarrow < 1, -1 >$

先行研究：PrivKV（2. 推定）

- **MLE（Maximum Likelihood Estimation）**

集計(a =君の名は。)

	k_a^*	v_a^+
u_1	1	-1
u_3	0	0
u_{20}	1	1
	⋮	
u_{51}	1	1
u_{77}	0	0

平均値推定

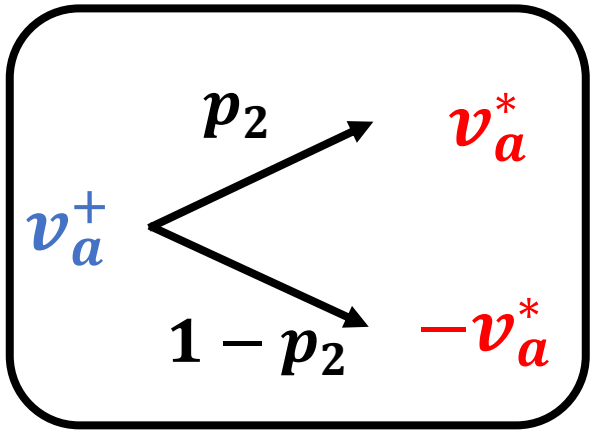
$$n'_1 = count(v_a^+ = 1)$$
$$n'_2 = count(v_a^+ = -1)$$
$$N = n'_1 + n'_2$$

$$\mathrm{L}(\hat{n}_1) = \frac{N(p_2-1)+n'_1}{2p_2-1}$$
$$\mathrm{L}(\hat{n}_2) = N - \hat{n}_1$$

key= aの推定平均値$\hat{m}_a$は,

$$\hat{m}_a = \frac{n'_1 - n'_2}{n'_1 + n'_2}$$

$RR(v_a^*)$



$$E(n'_1) = n_1 p_2 + n_2(1-p_2)$$

MLEの問題点

k	v
君の名は。	*

VPP（valueのランダマイズ）

$< k'_a, v'_a > \rightarrow v'_a$の2値化 $\rightarrow < k'_a, v^*_a > \rightarrow RR(v^*_a) \rightarrow < k'_a, v^+_a >$

$< 0, 0.4 > \rightarrow v'_a$の2値化 $\rightarrow < 0, -1 > \rightarrow RR(-1) \rightarrow < 0, -1 >$

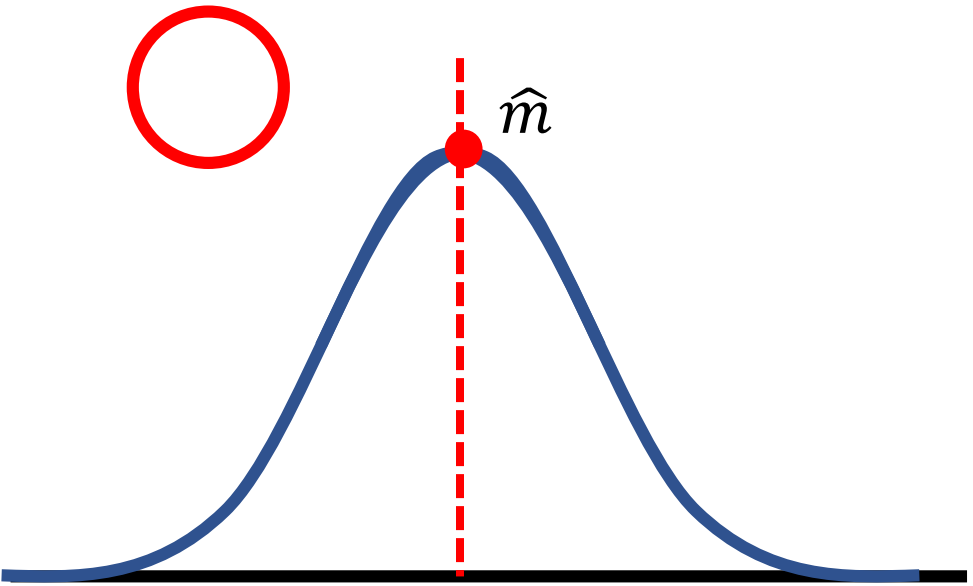
$< k'_a, v'_a > = < 0, random[-1, 1] >$

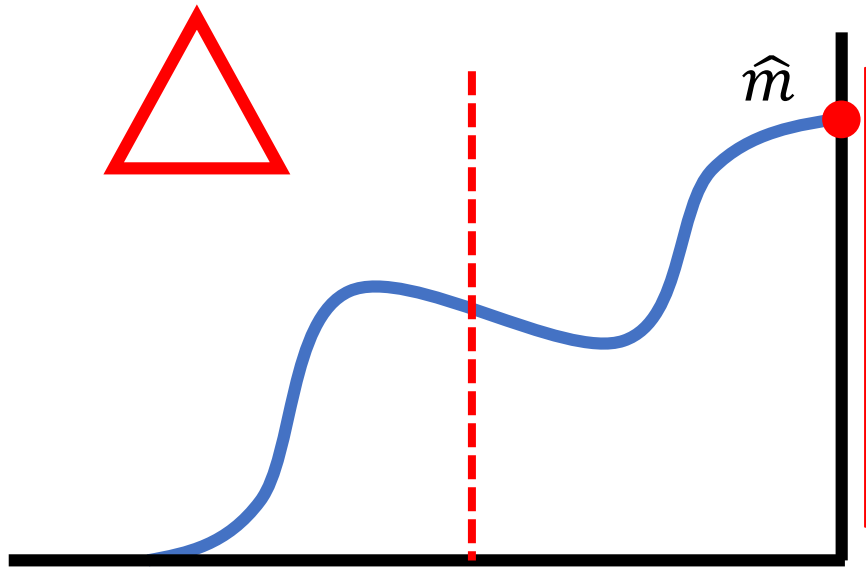
評価者の少ないkeyの遷移

多

$< k, v >$ → $< k^*_a, v^+_a >$

$< $君の名は。$, * >$ → $< 1, 1 >$ or $< 1, -1 >$





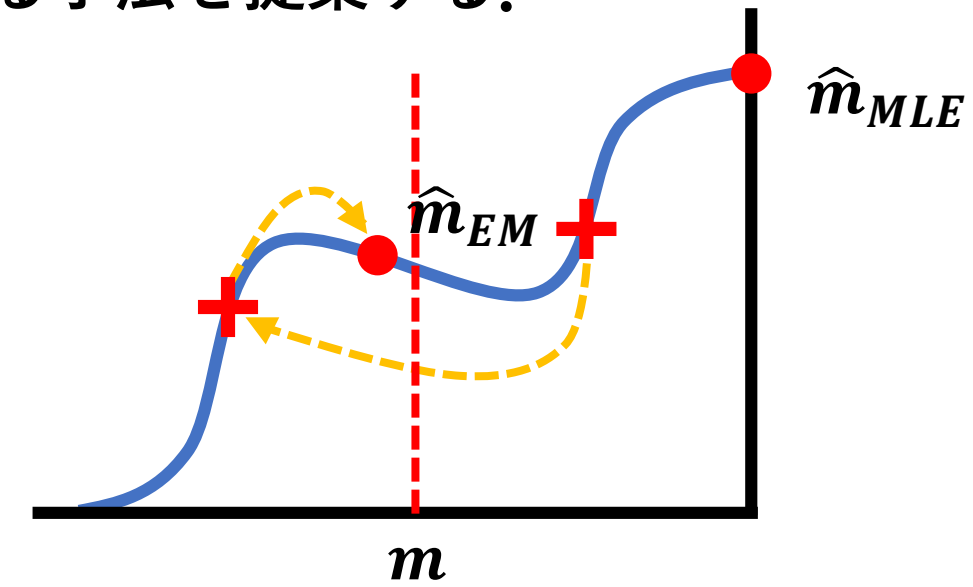
$v^+_a = v^*_a$であるか
$v^+_a = -v^*_a$であるか
の推定を考えており，
$v =$ * を考慮して
いない．

研究概要

	摂動化	推定
PrivKV	VPP+RR	MLE
本提案		EM

- 解決手法

PrivKVにEM(Expectation Maxmization)アルゴリズム[宮川雅巳,1987]を適用し，推定する手法を提案する．



提案手法
EM(Expectation Maximization)アルゴリズムの適用

摂動の流れ

a = “君の名は。”　　VPP (valueのランダマイズ)　　keyのランダマイズ

$< k'_a, v'_a > \rightarrow v'_a$の2値化 $\rightarrow < k'_a, v^*_a > \rightarrow RR(v^*_a) \rightarrow < k'_a, v^+_a > \rightarrow RR(k'_a) \rightarrow < k^*_a, v^+_a >$

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

事前確率を求める対象の設定

案1　$< k'_a, v'_a > \longrightarrow < k^*_a, v^+_a >$　×

案2　$< k'_a, v^*_a > \longrightarrow < k^*_a, v^+_a >$　○

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

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

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

・出力 $< k^*_a, v^+_a >$ を用いて $< k'_a, v^*_a >$ の事前確率を推定する．

RQ.

- **RQ1. 提案手法はPrivKVよりも高精度か？**
- **RQ2. 安全性εによって推定誤差に影響はあるのか？**
- **RQ3. データ規模nによって推定誤差に影響はあるのか？**

評価実験の概要

実験目的

- 3つのRQ.を調査する

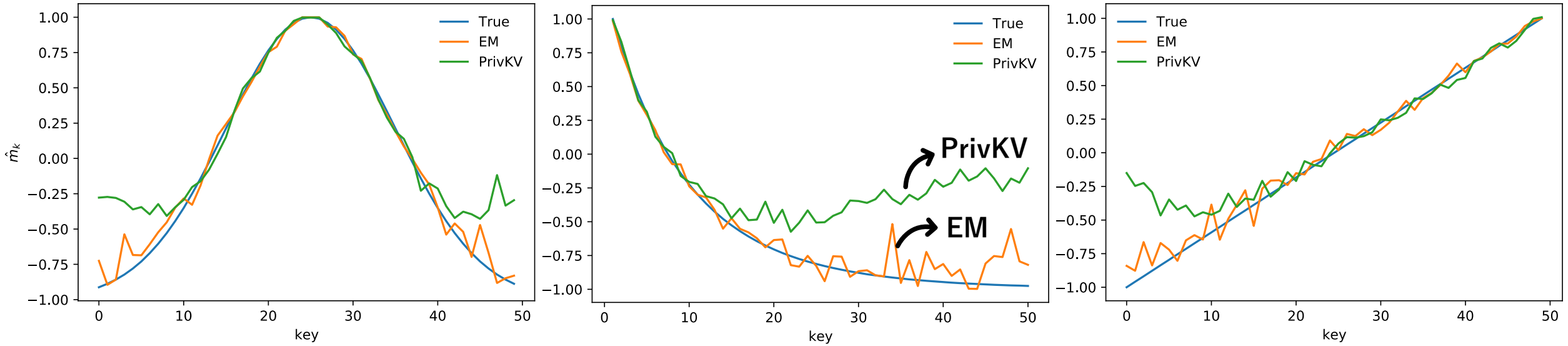
方法

- **データ**
  - keyとvalueがガウス分布，べき分布，線形分布に従う合成データ
- **評価実験**
  - PrivKVと提案手法でkey-valueデータの度数と平均値を推定し，推定誤差MSEを算出する.
  - この試行を10回行いMSEの平均値をアルゴリズムの評価値とする.

実験結果1　推定平均値の分布

RQ1. 提案手法はPrivKVよりも高精度か？

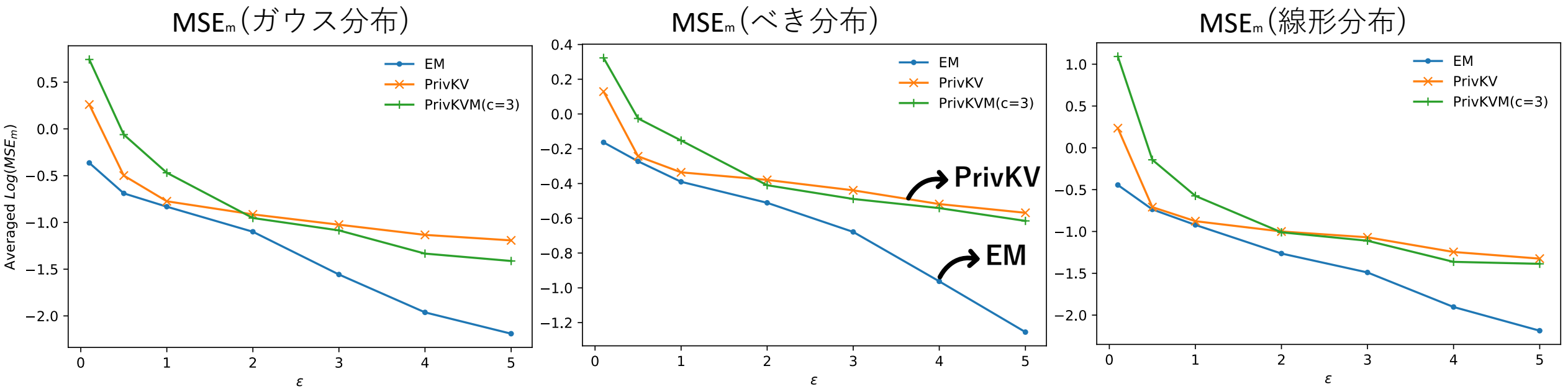
平均値の推定分布（$\epsilon = 4,\ \mathbf{n} = 10^5$）



提案手法は低頻度のkeyに対しても高い推定精度

実験結果2
RQ2. 安全性εによって推定誤差に影響はあるのか？

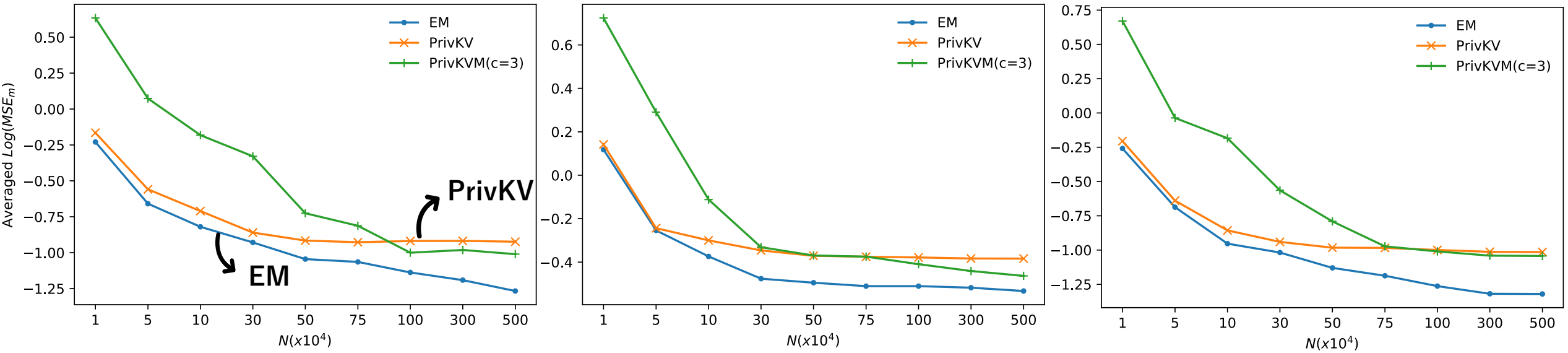
key数=50, ユーザ数 $n = 10^5$



いかなる安全性εでも提案手法の誤差が小さい

実験結果3
RQ3. データ規模nによって推定誤差に影響はあるのか？

key数=50, $\varepsilon = 2$



データ規模に関わらず提案手法の誤差が小さい

まとめ

- PrivKVでは推定に最尤推定法が用いられており，度数が0.2以下の小さなkeyに関して，平均値の推定誤差が大きい.

- PrivKVで摂動化したkey-valueデータにEMアルゴリズムを適用して推定する手法を提案した.

- その結果，平均値推定では合成データの実験の平均で 85.2% の改善が見られた.