Relation Mining Under Local Differential Privacy

Kai Dong, Zheng Zhang, Chuang Jia, Zhen Ling*, Ming Yang, Junzhou Luo, Xinwen Fu

Background

- More and more data are collected by centralized institutions
- Data mining can fully unleash the value of data

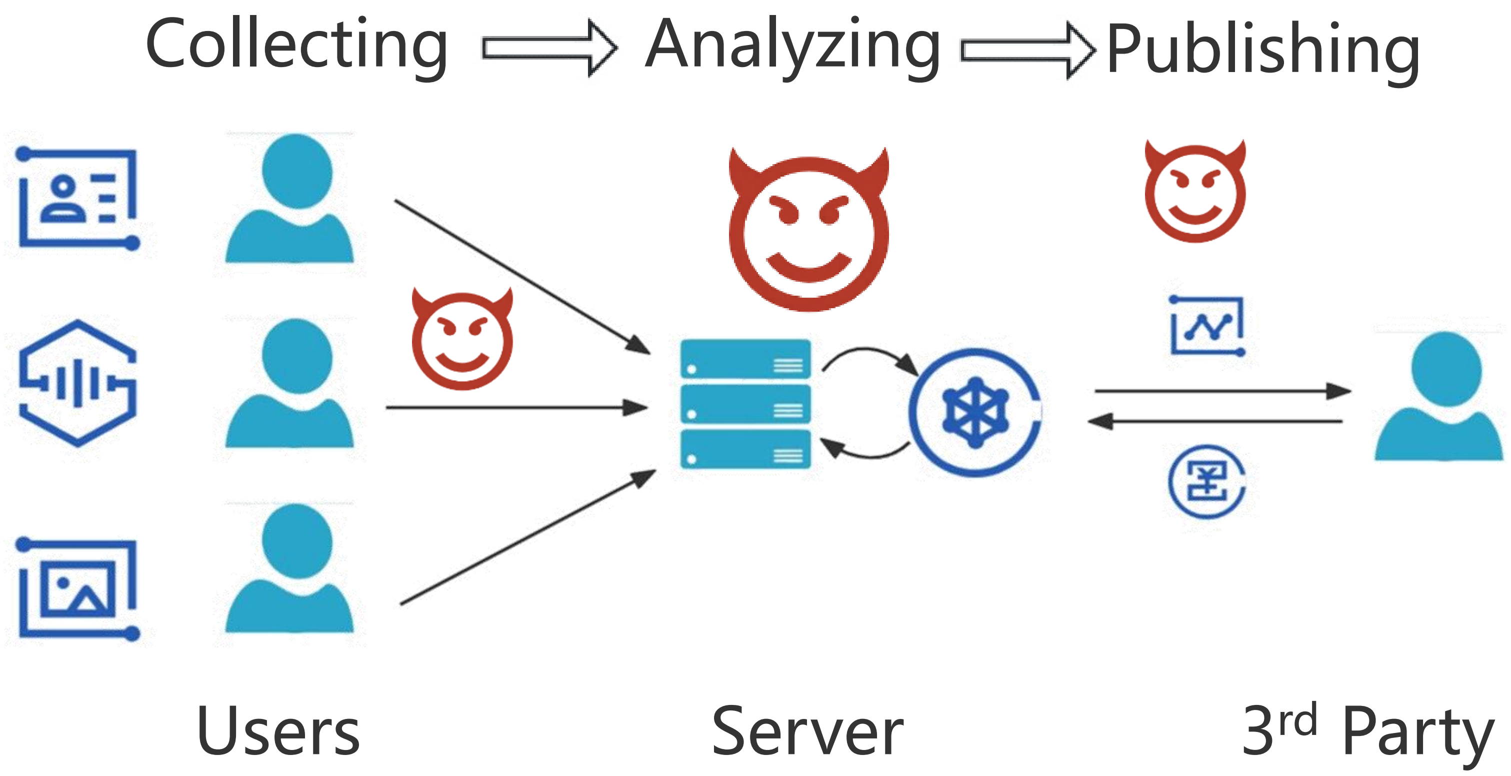
Market Analysis

Popular Emojis Discovery

Healthcare Insurance

Background

- There is a risk of data leakage throughout its entire lifecycle, especially during data analyzing, as central servers may be untrustworthy



Local Differential Privacy (LDP)

- ϵ-Local Differential Privacy (LDP) : An algorithm Ψ satisfies ϵ-LDP [FOCS'13], if and only if for any two values $x_1, x_2 \in \mathbf{X}$, we have:

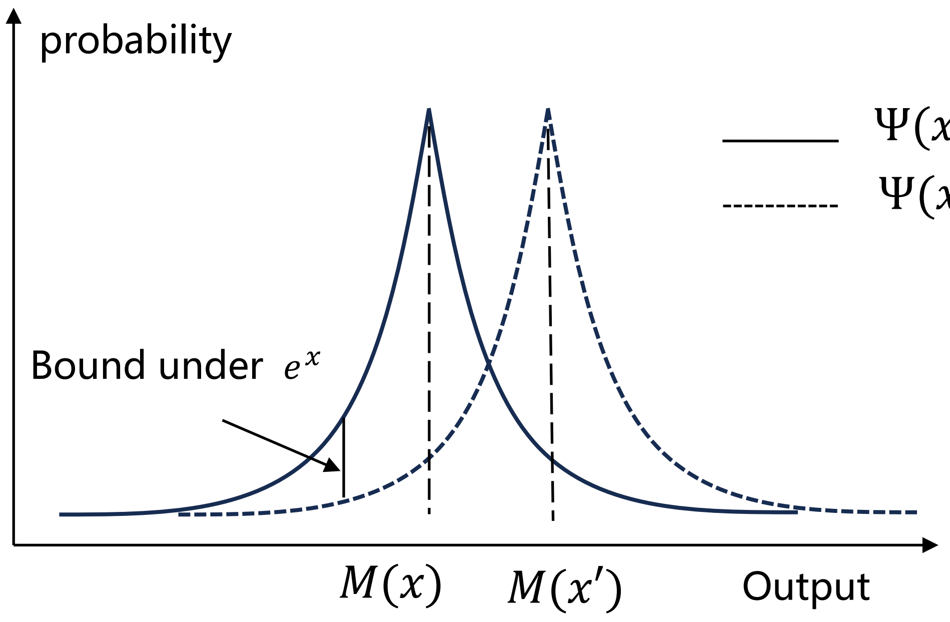
$$\forall T \in Range(\Psi) : \Pr(\Psi(x_1) = T) \leq e^{\varepsilon} \cdot \Pr(\Psi(x_2) = T),$$

where Range(Ψ) denotes the range of Ψ.

privacy budget ϵ reflects the trade-off between data privacy and utility in the LDP algorithm.

$\epsilon\downarrow$, privacy $\uparrow$, utility $\downarrow$

$\epsilon\uparrow$, privacy $\downarrow$, utility $\uparrow$



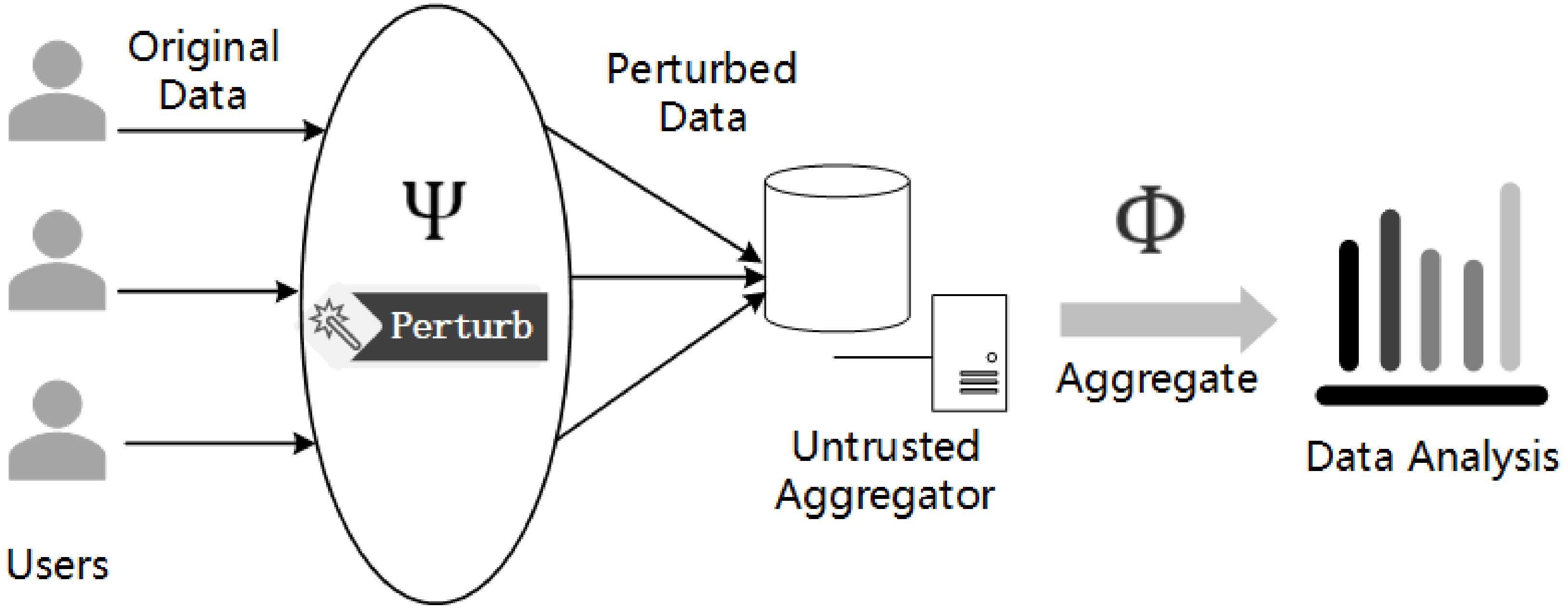
Data Mining under LDP

- Workflow

  A data mining task under LDP can be formalized as an LDP protocol $\mathcal{T}$ consisting of a pair of algorithms $\langle\Psi,\Phi\rangle$, defined as follows:

$$\mathcal{T}(\varepsilon) \triangleq \langle\Psi,\Phi\rangle.$$

  Ψ : perturbation algorithm to perturb local data
  Φ : aggregation algorithm to extract useful knowledge



Data Mining under LDP

- Existing LDP protocols

  1、Statistical estimation

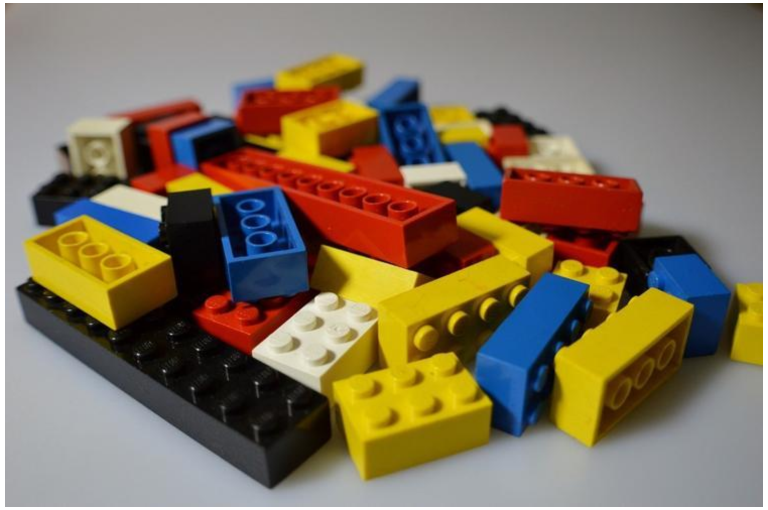
  - Mean Estimation：Duchi, PM, HM[ICDE'19]
  - Frequency Estimation (FO)：

    GRR, OLH [Security'17]

  2、Item-level data mining

  Set-Valued Item Mining (SVIM)[S&P ' 18]

  SVSM[S&P'18]、CALM[CCS'18]、 PCKV [Security'20] 、



Relation Mining (RM) under LDP

- ## Problem Definition

  - Relation level knowledge holds significant importance
    - ✓ Association rule mining Walmart
    - ✓ Temporal relation mining amazon
  - The usefulness of relations are measured by two criteria:
    - **Support** indicates popularity $s(w) \triangleq \frac{1}{n}\sum_{j=1}^{n}\mathbb{I}(w, \mathbf{w}^{\mathrm{j}})$
    - **Confidence** indicates reliability $c\big((x_a, x_b)\big) \triangleq \frac{s\big((x_a, x_b)\big)}{s(x_a)}$
  - Settings: Users $\mathbf{U} = \{u_1, u_2, \dots, u_n\}$, Items $\mathbf{X} = \{x_1, x_2, \dots, x_d\}$,
    A relation w is denoted as (x_a, x_b).
  - Objective: First identify the top-k_s relations in support, and then, from these k_s relations, find the top-k_c relations in confidence.

Relation Mining (RM) under LDP

- Challenges How to ensure accuracy

1、Curse of Dimensionality

- LDP noise is positively correlated with the domain dimensionality
- The domain dimensionality of relations is at least the square of items

2、Conflict between high-support and high-confidence

Existing technology prefers "high support but low confidence ".

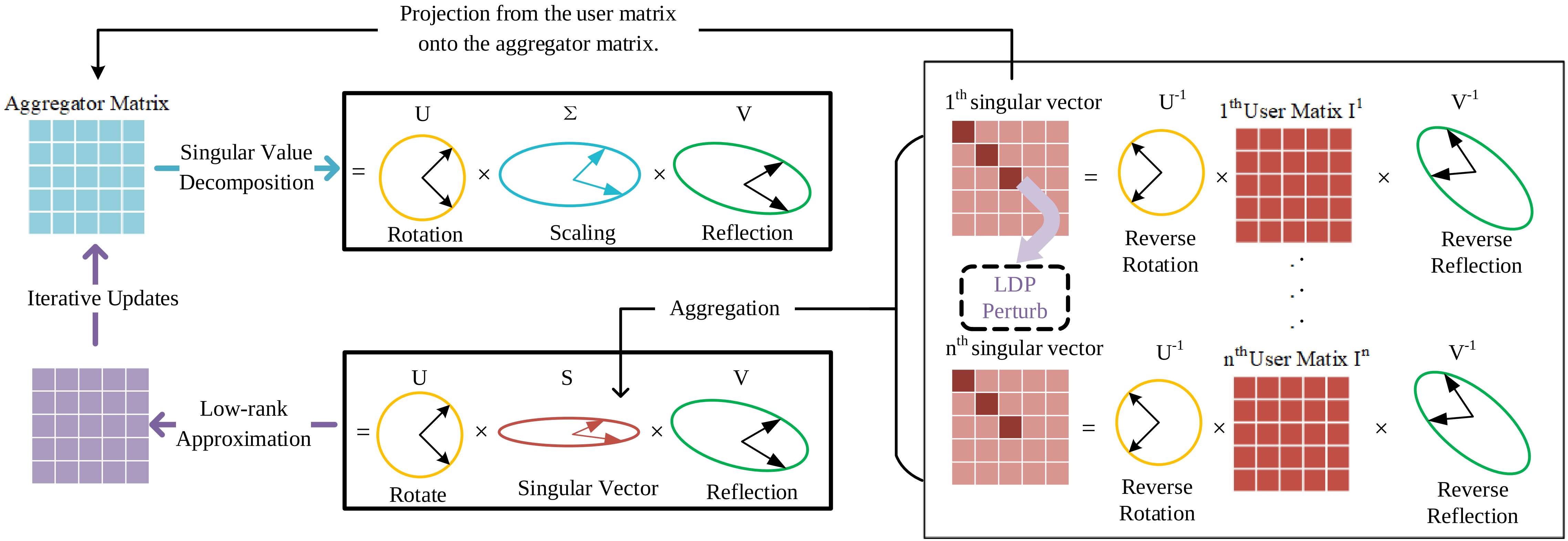
$$\begin{array}{c c} & \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\ \begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{array} & \begin{bmatrix} 40 & 20 & 20 & 0 & 0 & 0 \\ 20 & 30 & 10 & 0 & 0 & 0 \\ 20 & 10 & 30 & 0 & 0 & 0 \\ 0 & 0 & 0 & 25 & 25 & 25 \\ 0 & 0 & 0 & 25 & 25 & 25 \\ 0 & 0 & 0 & 25 & 25 & 25 \end{bmatrix} \end{array}$$

LDP-RM

- Basic Idea **Reduce data dimensionality to reduce LDP noise**

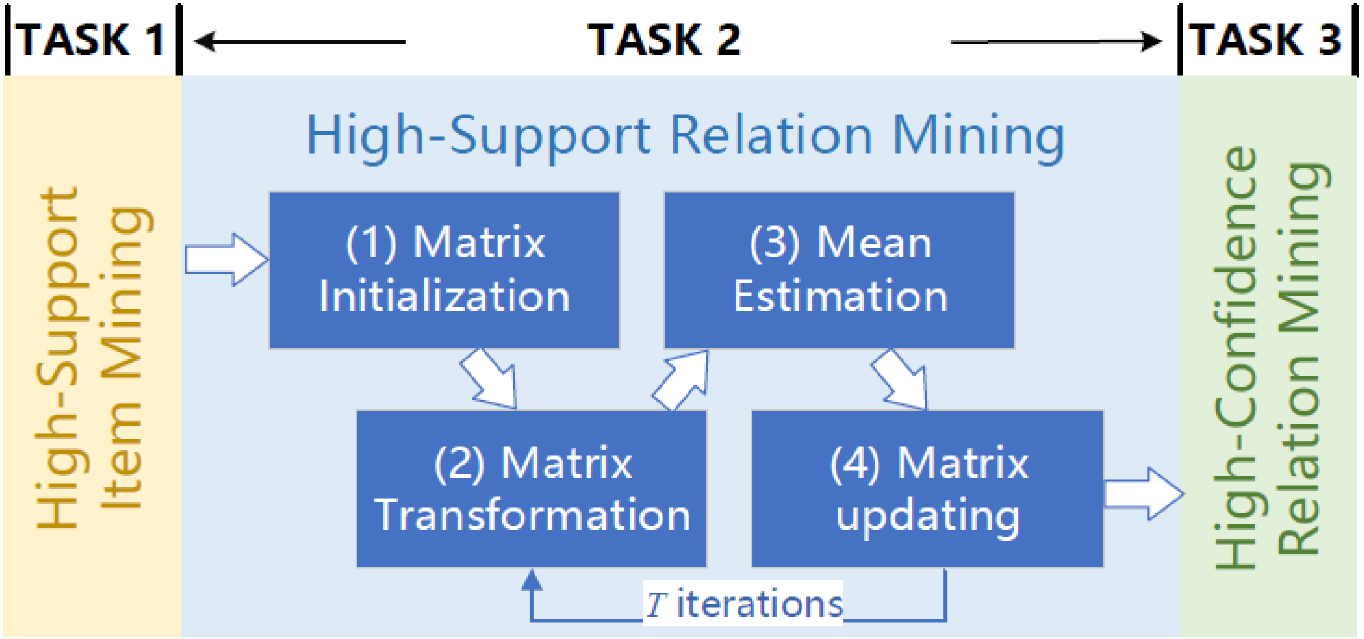
1. Pre-estimation： Identify the top-k items in support (k < d)
2. Projection: LDP noise level $O(k^2) \rightarrow O(r)$
3. Iterative: Updating the estimation of the aggregator matrix

LDP noise level
$O(d^2) \rightarrow O(k^2)$



LDP-RM

- Workflow



Grouping Strategy

- To save privacy budget, all users are randomly divided into 3 groups corresponding to three tasks, with each user queried only once in each task.

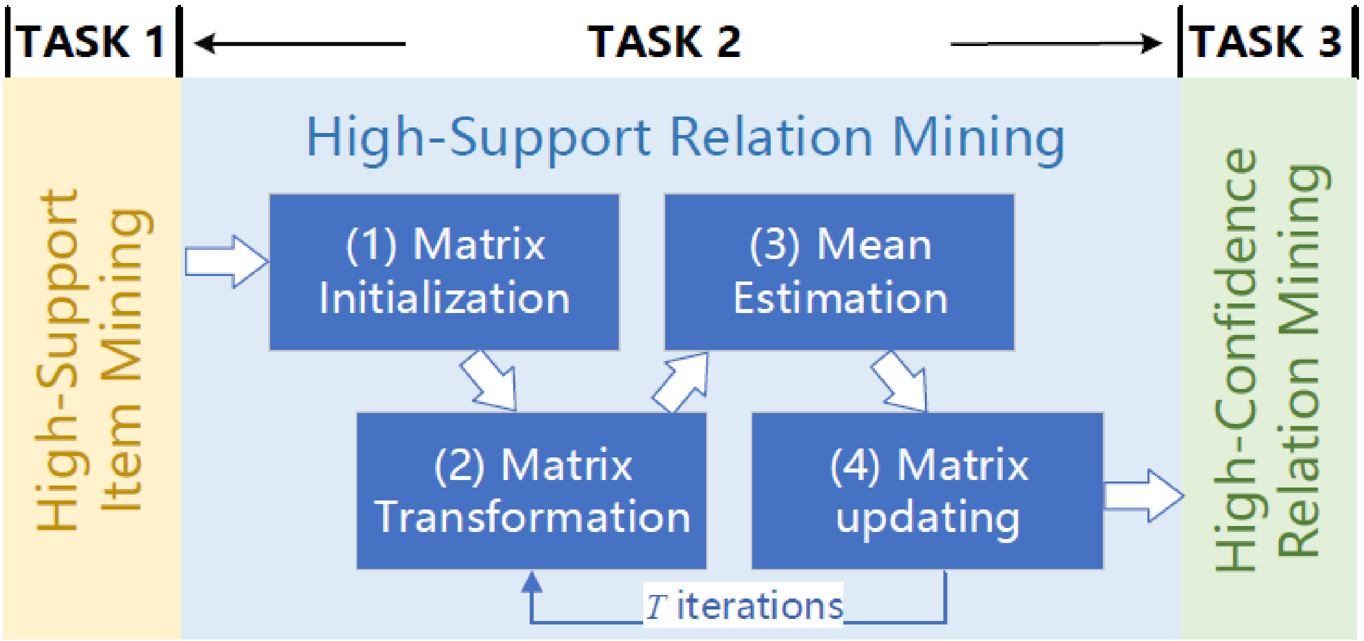
(1) High-Support Item Mining

The aggregator interacts with the first group of users and employs the SVIM protocol, finds the top-k items in support and estimate their support.

The size of the relation domain is reduced from $O(d^2)$ to $O(k^2)$.

LDP-RM

- Workflow



Grouping Strategy

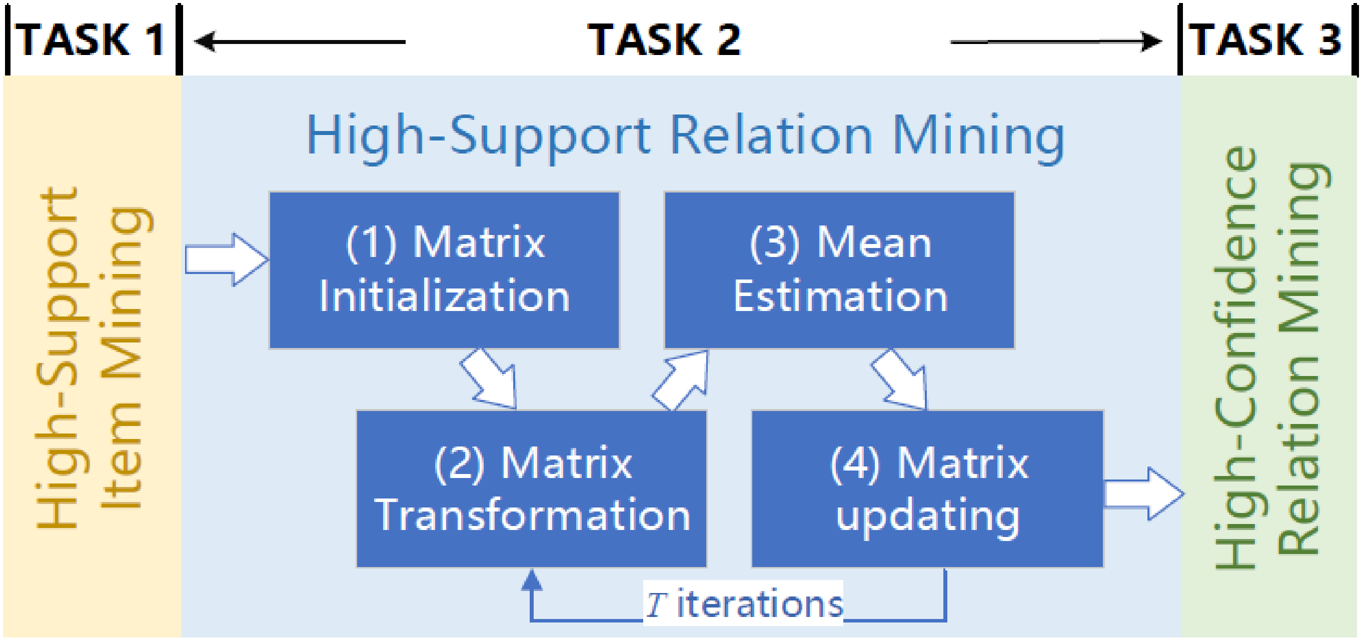
- To save privacy budget, all users are randomly divided into 3 groups corresponding to three tasks, with each user queried only once in each task.

(2) ★ High-Support Relation Mining

The aggregator interacts with the second group of users through 4 stages to finds the top-k_s relations in support which constitute a candidate set. The size of the relation domain is reduced from $O(k^2)$ to $O(r)$.

LDP-RM

- Workflow



Grouping Strategy

- To save privacy budget, all users are randomly divided into 3 groups corresponding to three tasks, with each user queried only once in each task.

(3) High-Confidence Relation Mining

The aggregator interacts with the third group of users finds the top-kc relations in confidence from the candidate set.

Analysis

- Privacy

  LDP-RM satisfies ε-LDP defined on users' items and relations.

- Utility

■ Estimation Error

Method	Estimation Error
SVIM	$O_1\left(\ell\sqrt{\log(\ell/\beta)}/\epsilon\sqrt{n}\right)$
HM	$O_2\left(\sqrt{r\log(r/\beta)}/\epsilon\sqrt{n}\right)$

■ Estimation Bias

The best rank-r approximation introduces a degree of bias in the process of recovering matrix.

According to the Eckart-Young-Mirsky Theorem, we strike a balance by selecting a relatively small r:

$$\min r \ \ \mathrm{s.t.} \sum_{i=1}^{r} \sigma_i \Big/ \sum_{i=1}^{k} \sigma_i \geq \theta$$

■ Error vs Bias

$r\downarrow$, bias $\uparrow$, error $\downarrow$

$r\uparrow$, bias $\downarrow$, error $\uparrow$

Unbiased variant: HM-RM
To demonstrate the bias has a minimal impact on the results.

Analysis

- Computational Overhead

Methods	User Side			Aggregator Side
	Group ♯1 (n_1)	**Group ♯2 (n_2)**	**Group ♯3 (n_3)**	
LDP-RM	$O(\log d)$	$O(k^2 + \log r)$	$O(\log k_s)$	$O(n_1 \log d + n_2(k^2 + \log r) + 2Tk^2 + n_3 \log k_s)$
SVSM [43]	$O(\log d)$	$O(\log k_s)$		$O(n_1 \log d + (n_2 + n_3) \log k_s)$
CALM [50]	$O(\log d)$	$O(2^l)$		$O(n_1 \log d + (n_2 + n_3) 2^l)$
PCKV [18]	$O(\log d)$	$O(\log k_s)$		$O(n_1 \log d + (n_2 + n_3) \log k_s)$

The computational overhead of LDP-RM is mainly due to the performance bottleneck of the second group of users, which brings the overhead of $O(k^2 + \log r)$.

Generalizing LDP-RM

- Relations Comprising More Items
  - Modification1: if support of guessed relation surpasses support of an item in candidate, then substitute into candidate
  - Modification2: if support of relation in matrix surpasses support of item in candidate, then reconstruct candidate by selecting the top-k frequent items/relations
- Frequency Oracle on Large Domains

  LDP-RM can serve as a fundamental Frequency Oracle, called SVD-FO

  - Encode: original value domain $\mathrm{X} = \{x_1, x_2, \cdots,, x_d\}$

    $\rightarrow$ virtual value domain V^2. ($\mathrm{V} = \{v_1, v_2, \cdots,, v_{\lceil d^{1/2} \rceil}\}$)
  - LDP-RM: Estimate matrix $(M_{a,b}) \in \mathbf{R}^{\lceil d^{1/2} \rceil \lceil d^{1/2} \rceil}$

Evaluation

- Datasets

Dataset	Domain Size	Users	Scenario
IFTTT	354	300k	relation mining between 2 items
Movie	5000	400k	relation mining between 2 items
Modified IFTTT	354	300k	relation mining among 3 items
Retail	2603	300k	association rule mining
Kosarak	41,270	990k	item mining on a large domain

- Metrics

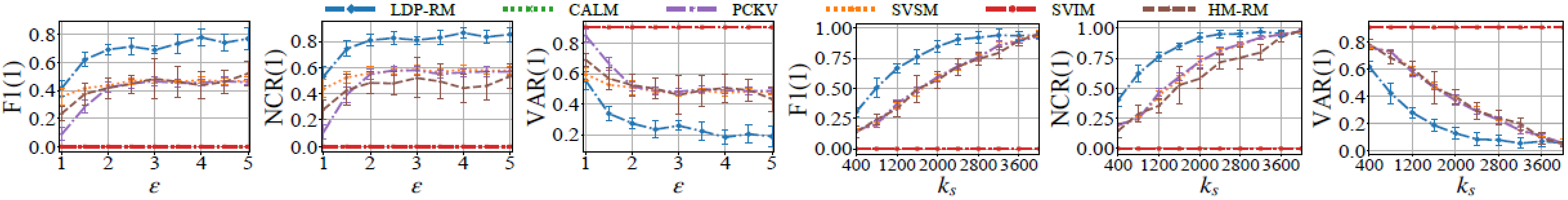
$$\mathrm{F1} = \frac{2}{1/P+1/R} = \frac{2PR}{P+R} \qquad \mathrm{NCR} = \sum_{w\in W_e\cap W_t} q(w) / \sum_{w\in W_t} q(w) \qquad \mathrm{VAR} = \frac{1}{|W_t|}\sum_{w\in W_t}\big(\rho(w)-\varphi(w)\big)^2$$

- Compared Methods

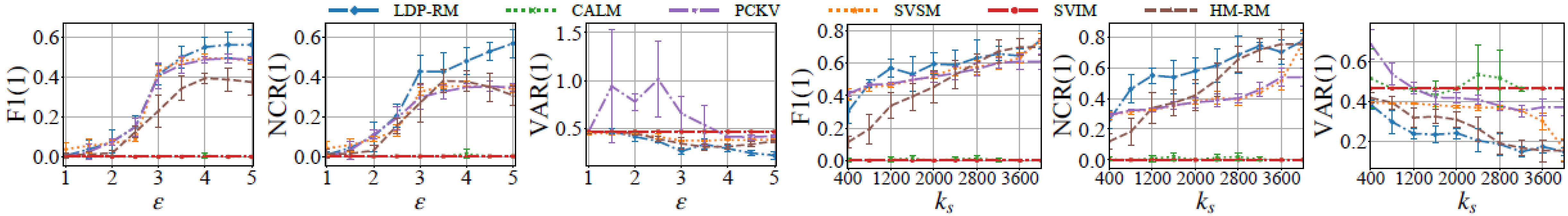
SVIM、SVSM、CALM、PCKV、HM-RM

Evaluation

- Experiments in mining relations between items



(a) F1 on ε (b) NCR on ε (c) VAR on ε (d) F1 on k_s (e) NCR on k_s (f) VAR on k_s

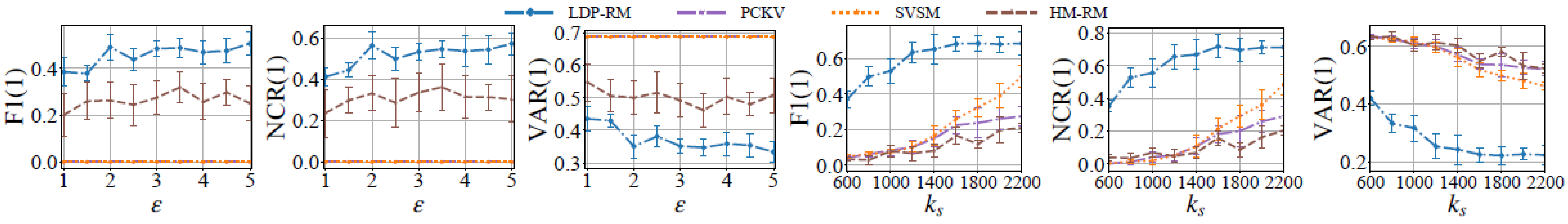


(a) F1 on ε (b) NCR on ε (c) VAR on ε (d) F1 on k_s (e) NCR on k_s (f) VAR on k_s

ε↑， F1↑， NCR↑， VAR↓； k_s↑, F1↑， NCR↑， VAR↓

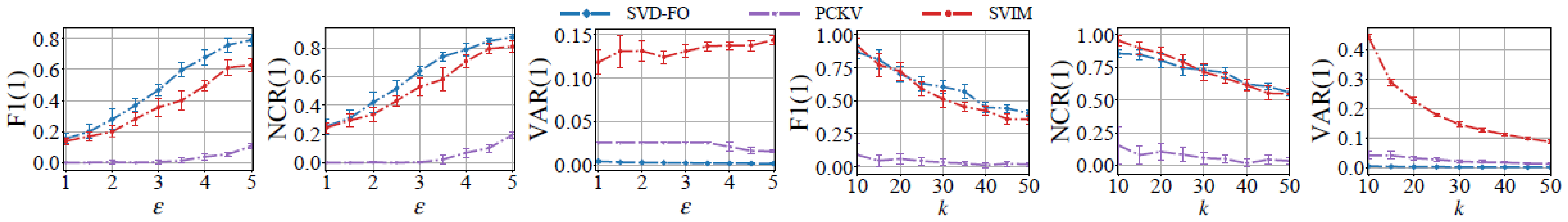
Evaluation

- Experiments in mining relations among items



(a) F1 on ε (b) NCR on ε (c) VAR on ε (d) F1 on k_s (e) NCR on k_s (f) VAR on k_s

- Experiments in item mining



(a) F1 on ε (b) NCR on ε (c) VAR on ε (d) F1 on k (e) NCR on k (f) VAR on k

Evaluation

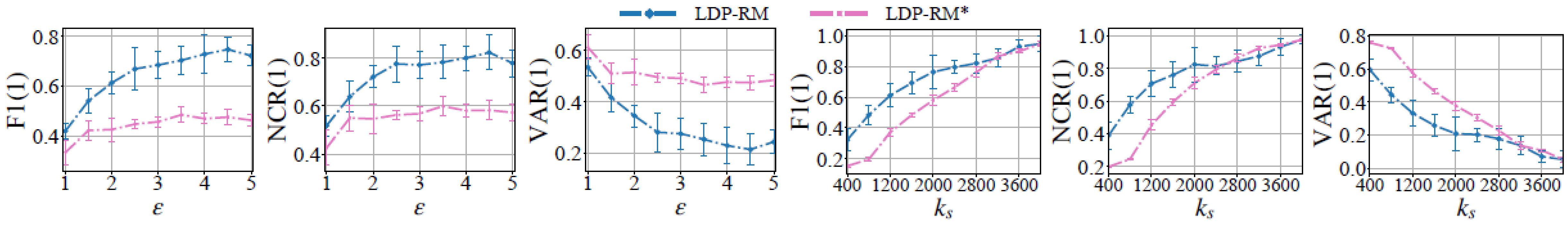
- Experiments in mining association rules

Method	Retail			Retail*		
	F1	NCR	VAR	F1	NCR	VAR
LDP-RM	0.558	0.640	0.246	0.546	0.654	0.180
SVSM	0.554	0.435	0.317	0.463	0.396	0.308
CALM	0.183	0.195	0.568	0.108	0.159	0.461
SVIM	0	0	0.603	0	0	0.491
PCKV	0	0	0.603	0	0	0.491
HM-RM	0.325	0.329	0.414	0.379	0.437	0.285

Retail*: slightly modify the Retail dataset by excluding data related to the top-8 items in support

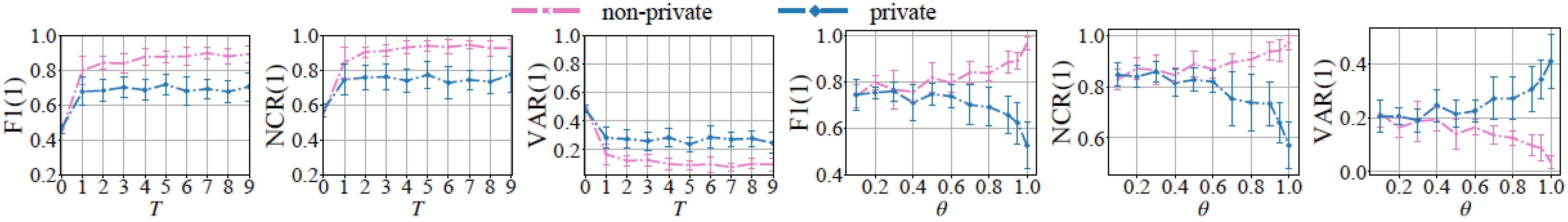
Evaluation

- Comparison of iterative and non-iterative algorithms.



(a) F1 on ε (b) NCR on ε (c) VAR on ε (d) F1 on k_s (e) NCR on k_s (f) VAR on k_s

- Comparison of private and non-private algorithms.



(a) F1 on T (b) NCR on T (c) VAR on T (d) F1 on θ (e) NCR on θ (f) VAR on θ

Conclusions

- First introduce and investigate the problem of relation mining under LDP.
  - A fundamental problem
  - Implementing LDP in RM is challenging

- Propose LDP-RM, the first relation mining method under LDP.
  - Discover high support and high confidence relations
  - Utilize SVD and low rank approximation
  - Generalize to Item mining

THANK YOU!

Questions: dk@seu.edu.cn