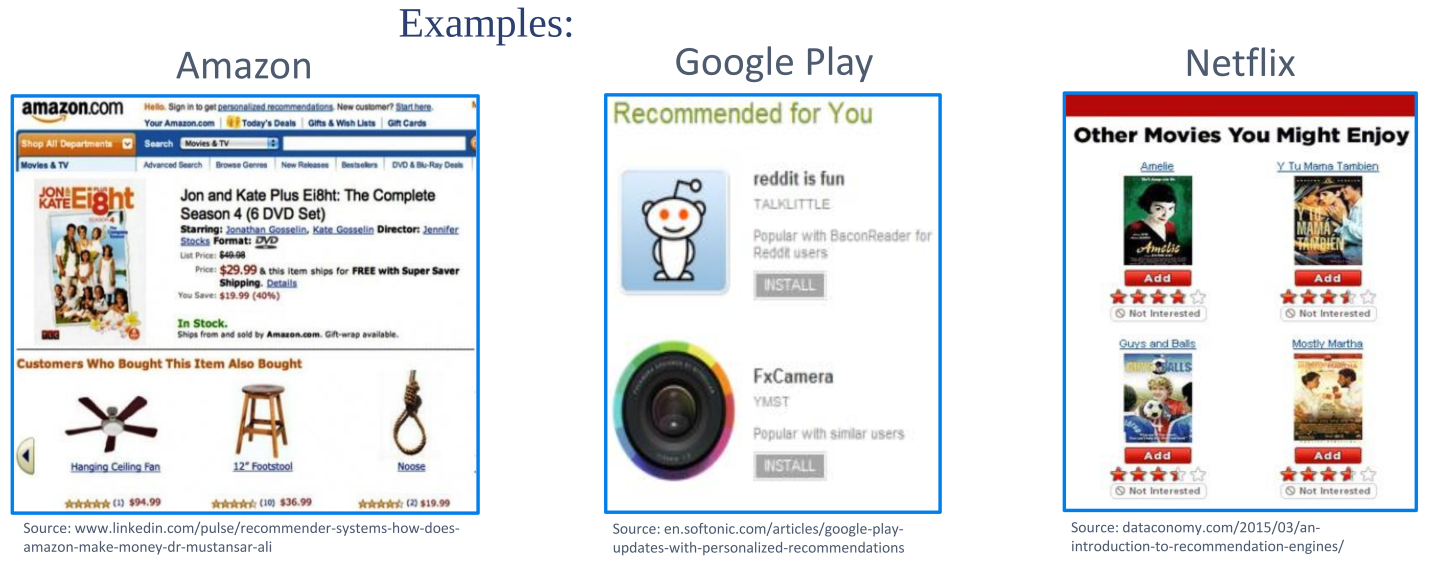
Differentially Private Matrix Completion, Revisited

Prateek Jain[1], **Om Thakkar[2]**, Abhradeep Thakurta[3]

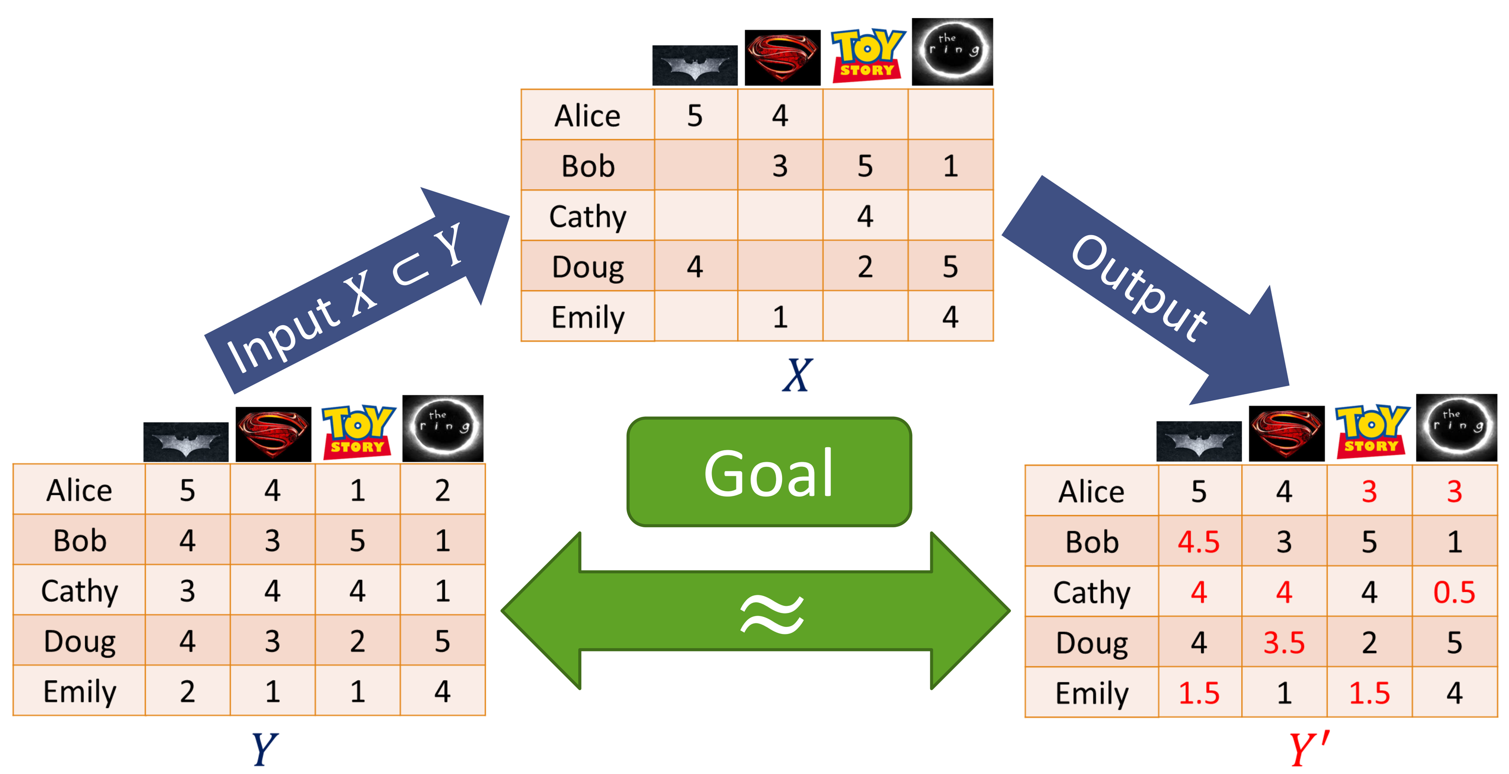
[1]Microsoft Research, [2]The Pennsylvania State University, [3]University of California-Santa Cruz

INTRODUCTION

Collaborative Filtering: To provide personalized recommendations via crowdsourcing. Examples:



Matrix Completion: Given an incomplete matrix $X \subset Y$, output Y', such that $Y' \approx Y$.



Alice	5	4		
Bob		3	5	1
Cathy			4	
Doug	4		2	5
Emily		1		4

X

Alice	5	4	1	2
Bob	4	3	5	1
Cathy	3	4	4	1
Doug	4	3	2	5
Emily	2	1	1	4

Y

Alice	5	4	3	3
Bob	4.5	3	5	1
Cathy	4	4	4	0.5
Doug	4	3.5	2	5
Emily	1.5	1	1.5	4

Y'

- Assumption: Y has **low-rank (or, bounded nuclear norm)**.

THE NEED FOR PRIVACY

Users may **not** prefer to reveal what movies they saw

- Or how much they liked them!

Bob	5

Our Goal: To **provide personalized recommendations** using crowdsourced data while **ensuring user-level differential privacy (DP)**.

PRIVACY MODEL

Joint DP [KPRU'14]: Mechanism $A: D^m \to T$ is (ϵ, δ)-Joint DP if for all *neighboring* datasets $x, x^* \in D^m$, and for all $i \in [m]$, for all sets of outcomes $S_{-i} \subseteq T_{-i}$,

$$P(A_{-i}(x) \in S_{-i}) \le e^{\epsilon} P(A_{-i}(x^*) \in S_{-i}) + \delta.$$

Here, $A_{-i}(X)$ = output of A on input X **without** user i's output.

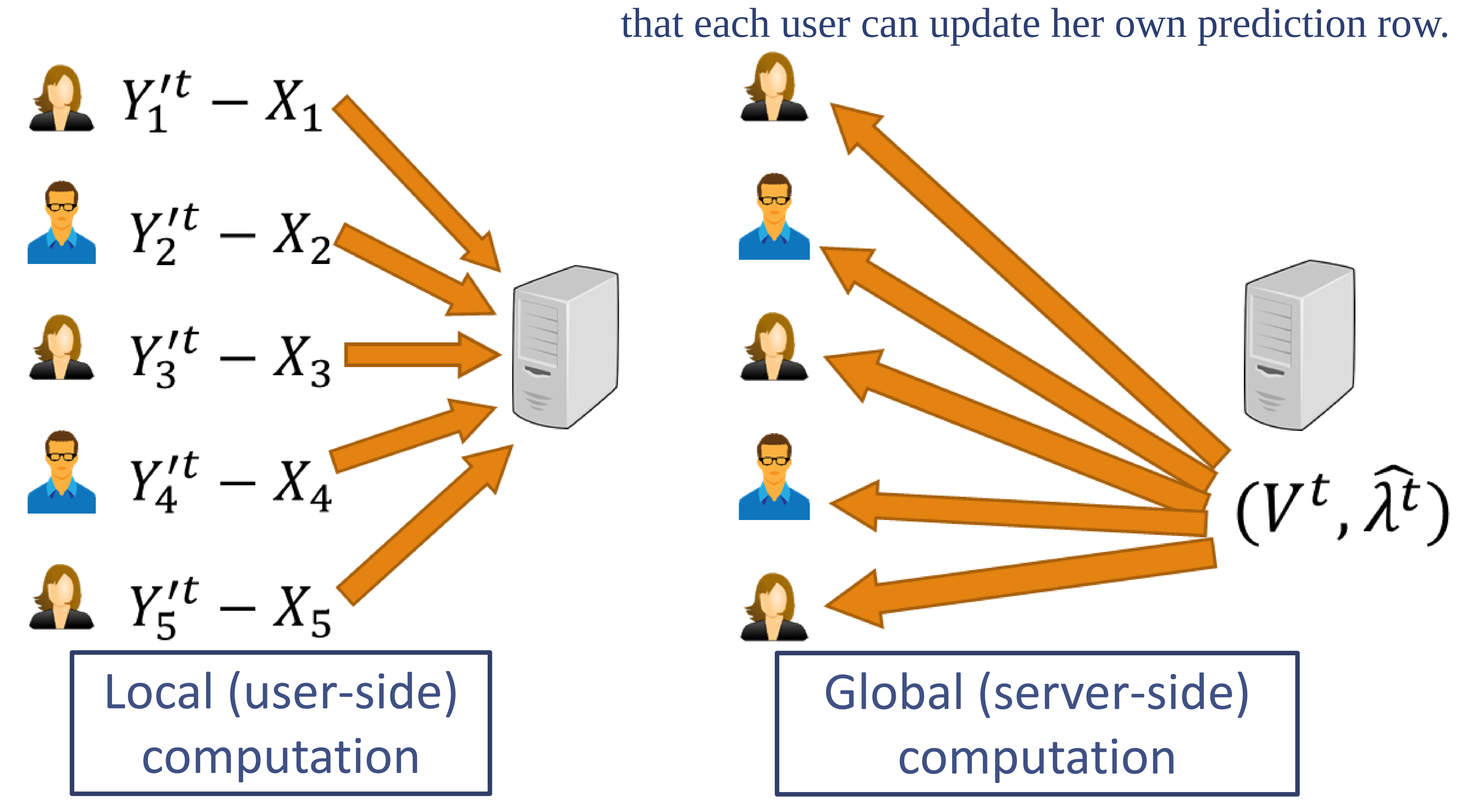
Distinction from standard DP [DMNS'06]:

- Under Joint DP, A's output for user i can depend **arbitrarily** on i's input. Consequence: Better personalized recommendations!

JOINT DIFFERENTIALLY PRIVATE FRANK-WOLFE

An iterative process, having two major steps in every iteration t:

1. *Local (user-side) computation:* Each user $i \in [m]$ computes the error of her incomplete row X_i from her predicted row Y_i^t, and sends the covariance of the error to a central server.
2. *Global (server-side) computation:* The server adds Gaussian noise $N(0, \sigma \mathbb{1}^{n \times n})$ to the sum of the error covariances, computes a global rank-1 update via SVD, and releases it publicly so that each user can update her own prediction row.



$$Y_i'^{t+1} = \left(1 - \frac{1}{T}\right) Y_i'^t - \frac{1}{T}\left(\frac{k(Y_i'^t - X_i)V^t(V^t)^{tr}}{\hat{\lambda}^t}\right)$$

User i's update step

T = Total number of iterations, k = Nuclear norm bound on Y

THEORETICAL RESULTS

1. *Privacy guarantee:*

- If $\sigma = \frac{L^2}{\epsilon}\sqrt{64 \cdot T\log\left(\frac{1}{\delta}\right)}$, then the Frank-Wolfe algorithm above is (ϵ, δ)-Joint DP.

2. *Utility guarantee:*

- If $\|Y\|_{nuc} \le k$, $\max_{i \in [m]} \|X_i\|_2 \le L$, and we run (ϵ, δ)-Joint DP Frank-Wolfe (FW) algorithm for T iterations, then with high probability:

$$\textit{Empirical Risk} = \frac{1}{|\Omega|}\sum_{i,j \in \Omega}(Y'_{ij} - X_{ij})^2 = \tilde{O}\left(\frac{k^2}{|\Omega| T} + \frac{kL(nT)^{1/4}}{|\Omega|\sqrt{\epsilon}}\right).$$

Here, Ω is the set of non-zero indices in X.

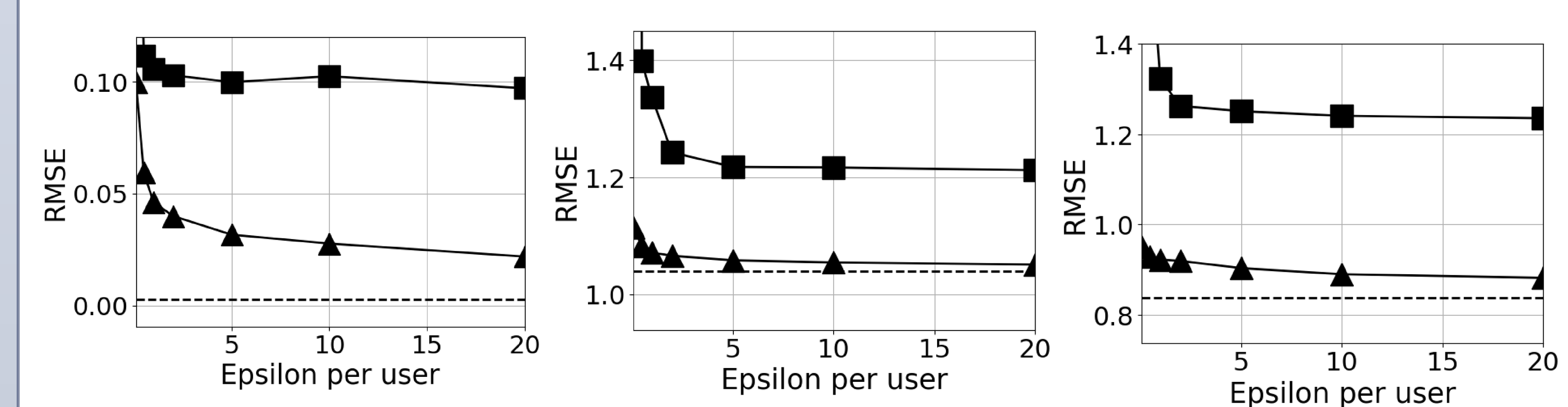
Standard Frank-Wolfe convergence error

Error due to Privacy

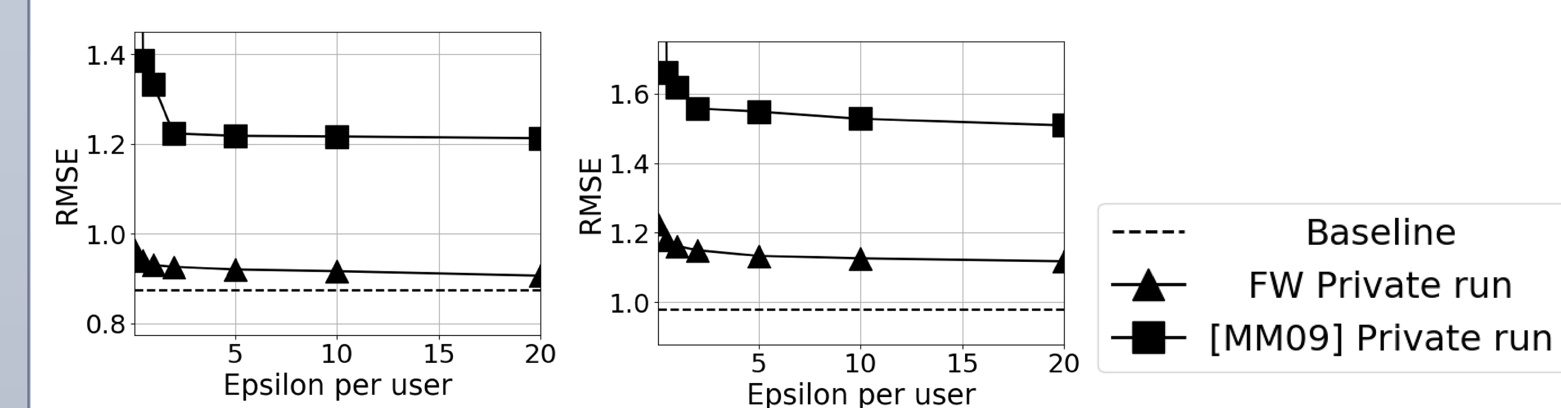
Additionally, $\textit{Empirical Risk} = \tilde{O}\left(\frac{1}{|\Omega|}\left(\frac{k^6 n L^4}{\epsilon^2}\right)^{1/5}\right)$ for $T = \tilde{O}\left(\left(\frac{k^4 \epsilon^2}{n L^4}\right)^{1/5}\right)$.

EXPERIMENTAL RESULTS

- m = number of users, n = number of items
- Unless specified, we sample ≈ 80 ratings per user, and each rating ∈ [0,5]



Synthetic dataset	Jester dataset	MovieLens10M (Top 400)
$Y = uv^{tr}$, where $u = [0,1]^{m \times 1}$, $v = [0,1]^{n \times 1}$, $m = 500k$, $n = 400$, each rating ∈ [0,1]	$m \approx 73k$, $n = 100$ jokes, No sampling of ratings	$m \approx 70k$, $n = 400$ most rated movies



Netflix (Top 400)	Yahoo (Top 400)	Legend for all the plots
$m \approx 474k$, $n = 400$ most rated movies	$m \approx 995k$, $n = 400$ most rated songs	

CONCLUSIONS

- We design a variant of the Frank-Wolfe algorithm for matrix completion.
  - We make it amenable for **user-level** Joint DP by splitting the iterative update step into 2 parts, local (user-side) computation and global (server-side) computation.
  - We provide the privacy and utility guarantees for it.
  - We demonstrate its performance on a variety of benchmark datasets, showing that
    - it provides nearly the same accuracy as the state-of-the-art non-private algorithm, and
    - it outperforms the existing state-of-the-art private matrix completion method [MM'09] by as much as 30%.

REFERENCES

[DMNS'06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. In *TCC*, 2006.

[KPRU'14] Michael Kearns, Mallesh Pai, Aaron Roth, and Jonathan Ullman. In *ITCS*, 2014.

[MM'09] Frank McSherry and Ilya Mironov. In *KDD*, 2009.