



Poisoning Attacks to Local Differential Privacy Protocols for Key-Value Data

Yongji Wu, Xiaoyu Cao, Jinyuan
Jia, Neil Zhenqiang Gong

Background

- Companies are collecting more and more data...
- Key-value data is pervasive data form, widely used in:

Recommender Systems

(item_id, rating)

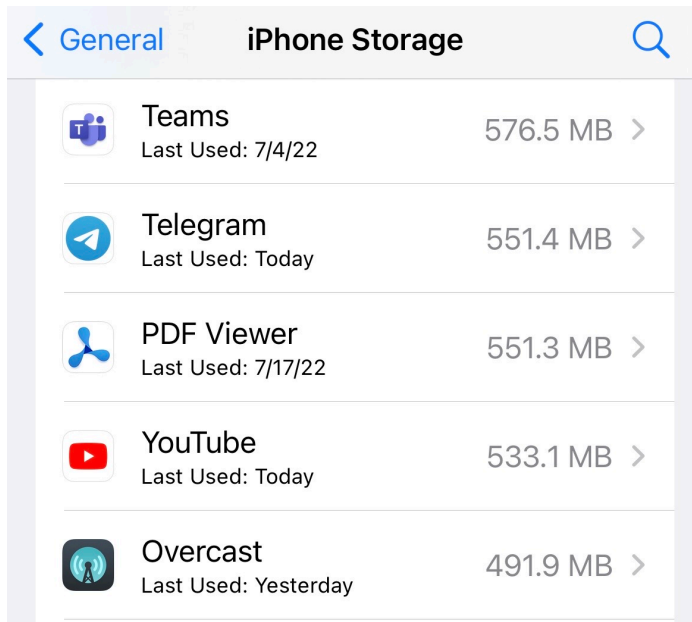
Internet of Things

(sensor_id, data)

Application Usage Analytics

(func_id, timestamp)

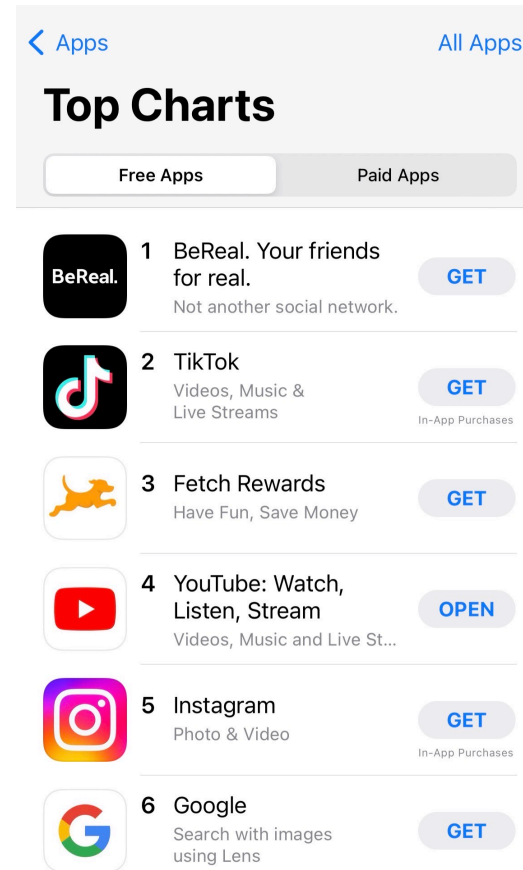
Key-Value Data Collection - RecSys



A screenshot of the 'iPhone Storage' settings page. It shows a list of installed apps with their names, last used dates, and storage sizes. The apps listed are Teams (576.5 MB), Telegram (551.4 MB), PDF Viewer (551.3 MB), YouTube (533.1 MB), and Overcast (491.9 MB).

App	Last Used	Size
Teams	7/4/22	576.5 MB
Telegram	Today	551.4 MB
PDF Viewer	7/17/22	551.3 MB
YouTube	Today	533.1 MB
Overcast	Yesterday	491.9 MB

What apps have you installed?
How frequently you use them?



A screenshot of the 'Top Charts' section in the App Store, showing a list of popular apps. The apps are ranked from 1 to 6, with their names, descriptions, and download buttons (GET or OPEN).

Rank	App	Description	Action
1	BeReal.	Your friends for real. Not another social network.	GET
2	TikTok	Videos, Music & Live Streams	GET
3	Fetch Rewards	Have Fun, Save Money	GET
4	YouTube: Watch, Listen, Stream	Videos, Music and Live St...	OPEN
5	Instagram	Photo & Video	GET
6	Google	Search with images using Lens	GET

What's the most popular apps?

Ratings & Reviews

[See All](#)

4.7

out of 5



26,770,382 Ratings

How about their average ratings?

What about User Privacy?

PRO CYBER NEWS
Marriott Reports Data Breach

BUSINESS

Capital One Reports Data Breach Affecting 100 Million Customers, Applicants

Alleged hacker, a former employee of Amazon Web Services, arrested by federal agents in Seattle

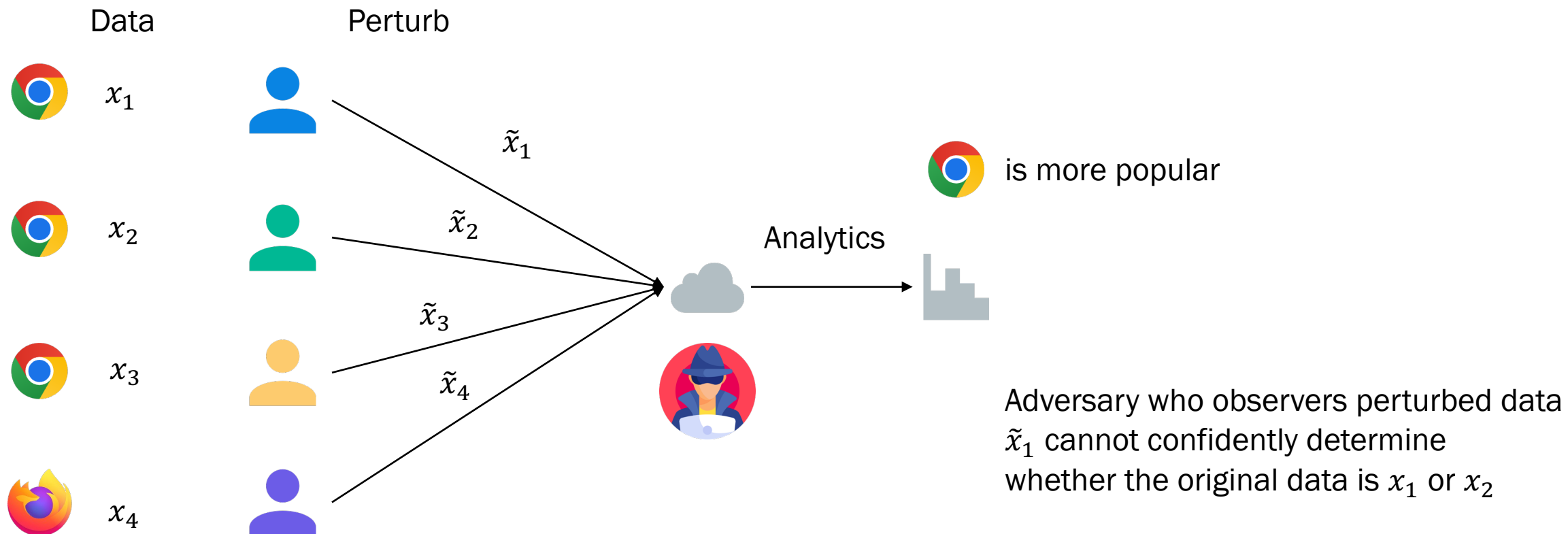
A 21-year-old
of customer

Data on 50
Is

Solution

- Locally-private data collection
- Raw data never leaves user's device

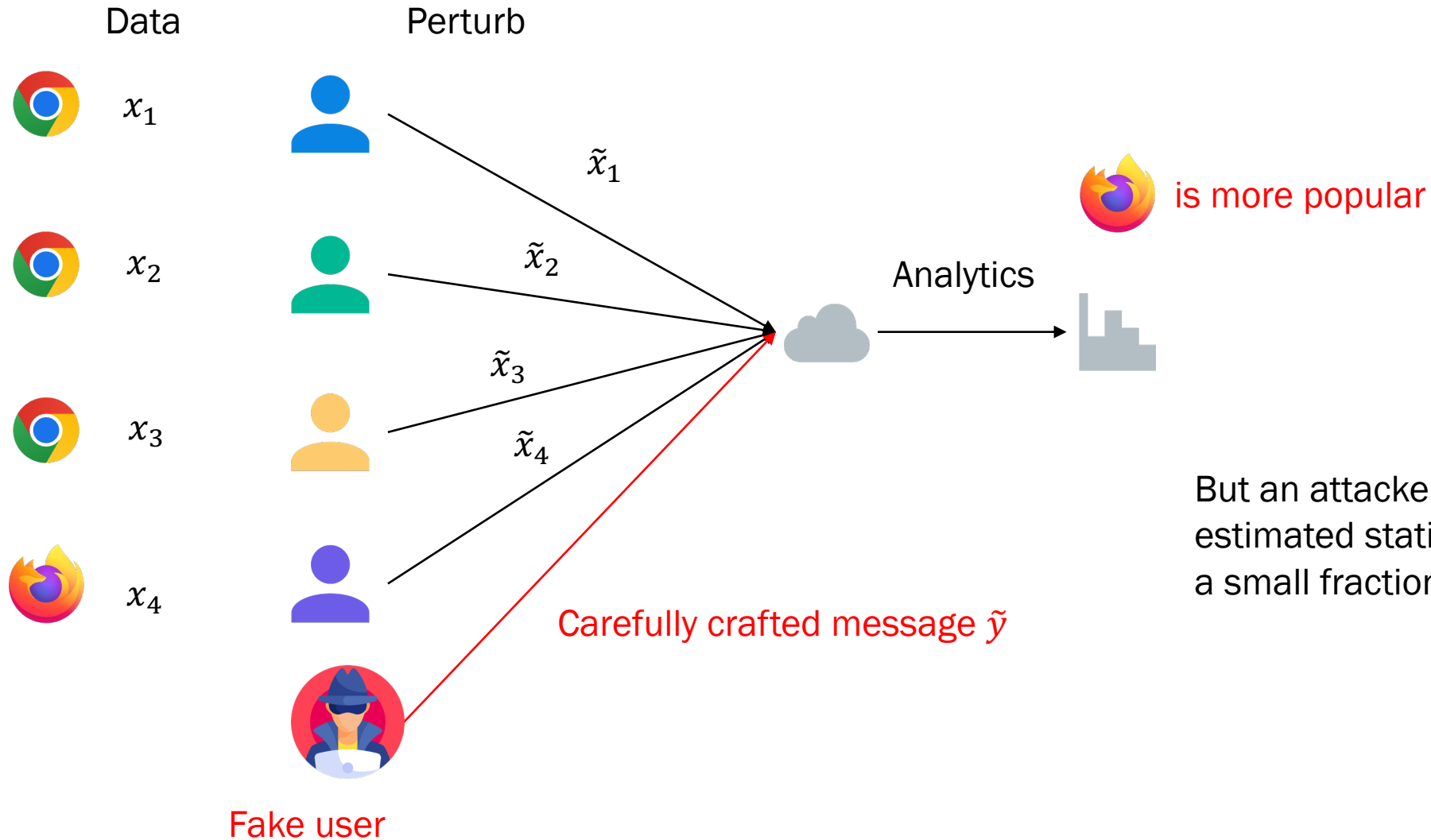
Local Differential Privacy (LDP)



Protocols for KV Data Collection

- PrivKVM [Ye, et al; S&P 19]
- PCKV-UE [Gu, et al; USENIX Security 20]
- PCKV-GRR [Gu, et al; USENIX Security 20]

LDP is Vulnerable to Attacks



But an attacker can greatly compromise the estimated statistics of an LDP protocol, with a small fraction of fake users

LDP Protocols for Key-Value Data

- We have a dictionary of d keys
- Each user has a set of KV pairs $\langle k, v \rangle$, where v is normalized into $[-1, 1]$
- We want to estimate the frequency and mean of each key

Threat Model

Attacker's goal

Promote frequency and mean estimation of some target keys

Attacker's knowledge

LDP protocol, including the parameter settings

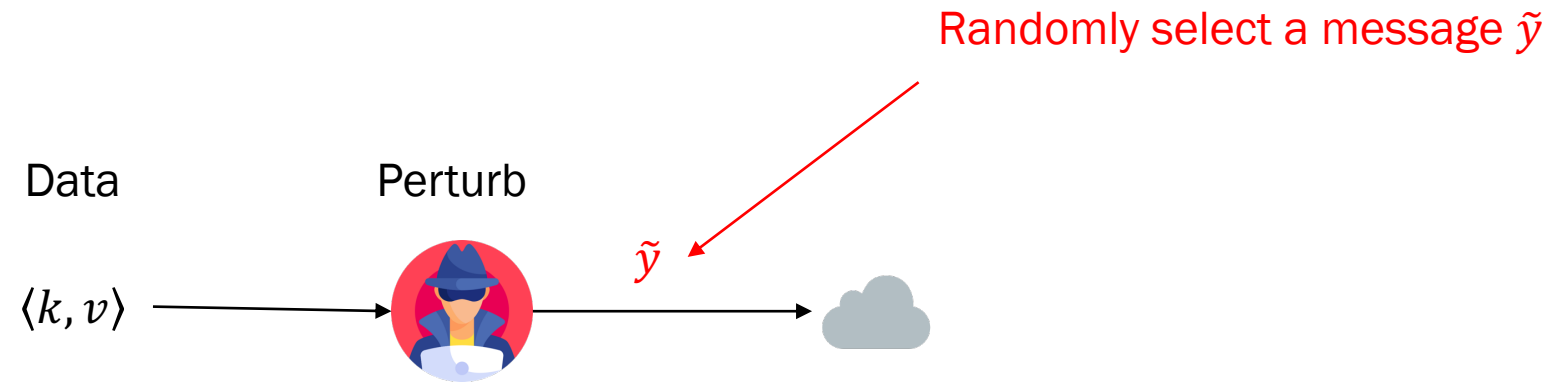
Attacker's capability

- Insert a small fraction of fake users
- Craft their messages

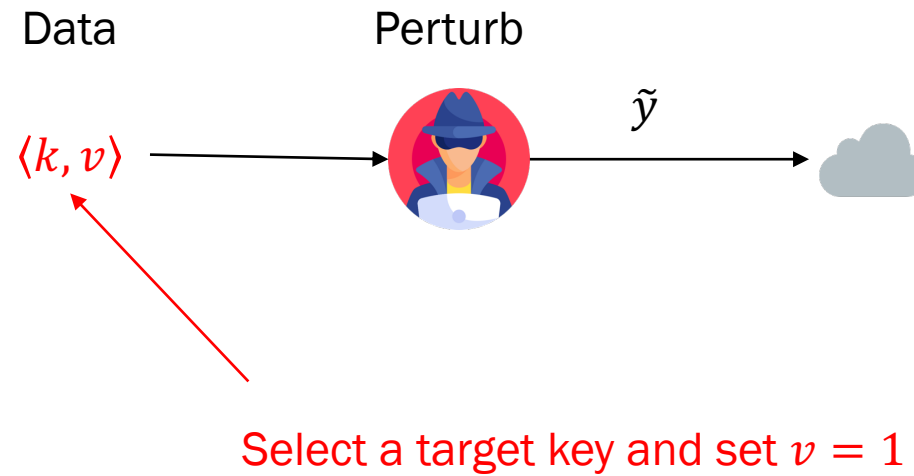
Our Three Attacks

- Baselines
 - Random Message Attack (RMA)
 - Random Key-Value Pair Attack (RKVA)
- Maximal Gain Attack (M2GA)

Random Message Attack (RMA)



Random Key-Value Pair Attack (RKVA)



Maximal Gain Attack (M2GA)

- Maximize the gains
- Solve the two-objective optimization problem:

$$\max_{\mathbb{Y}} \begin{bmatrix} G_f(\mathbb{Y}) \\ G_m(\mathbb{Y}) \end{bmatrix}$$

\mathbb{Y} : crafted messages for the fake users

G_f : frequency gain

G_m : mean gain

Theoretical evaluation

	PrivKVM	PCKV-UE	PCKV-GRR
M2GA	$\frac{\beta}{1+\beta} \left[1 - f_{\mathbb{T}} + \frac{2-r}{e^{\varepsilon/2}-1} \right]$	$\frac{\beta\ell}{1+\beta} \left[2r - f_{\mathbb{T}} + \frac{4r}{e^{\varepsilon}-1} \right]$	$\frac{\beta}{1+\beta} \left[(1 - f_{\mathbb{T}})\ell + \frac{2(d'-r)}{e^{\varepsilon}-1} \right]$
RMA	$\frac{\beta}{1+\beta} \left[\frac{(e^{\varepsilon/2}-2d+1)r}{2(e^{\varepsilon/2}-1)d} - f_{\mathbb{T}} \right]$	$\frac{\beta\ell}{1+\beta} \left[\frac{4e^{\varepsilon}r}{3(e^{\varepsilon}-1)} - f_{\mathbb{T}} \right]$	$\frac{\beta(r-f_{\mathbb{T}}d')\ell}{(1+\beta)d'}$
RKVA	$\frac{\beta}{1+\beta} \left[1 - f_{\mathbb{T}} + \frac{1-r}{e^{\varepsilon/2}-1} \right]$	$\frac{\beta\ell}{1+\beta} (1 - f_{\mathbb{T}})$	$\frac{\beta\ell}{1+\beta} (1 - f_{\mathbb{T}})$

We can theoretically analyze the frequency and mean gains

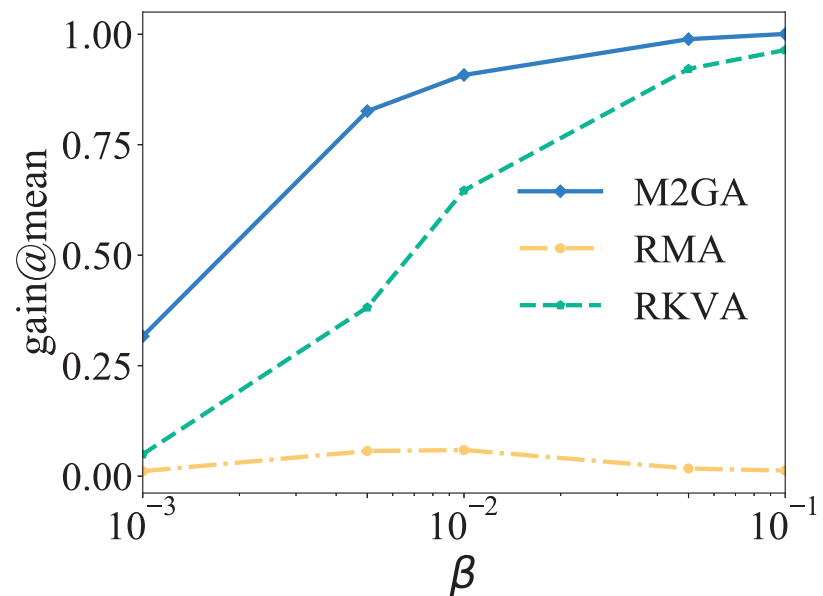
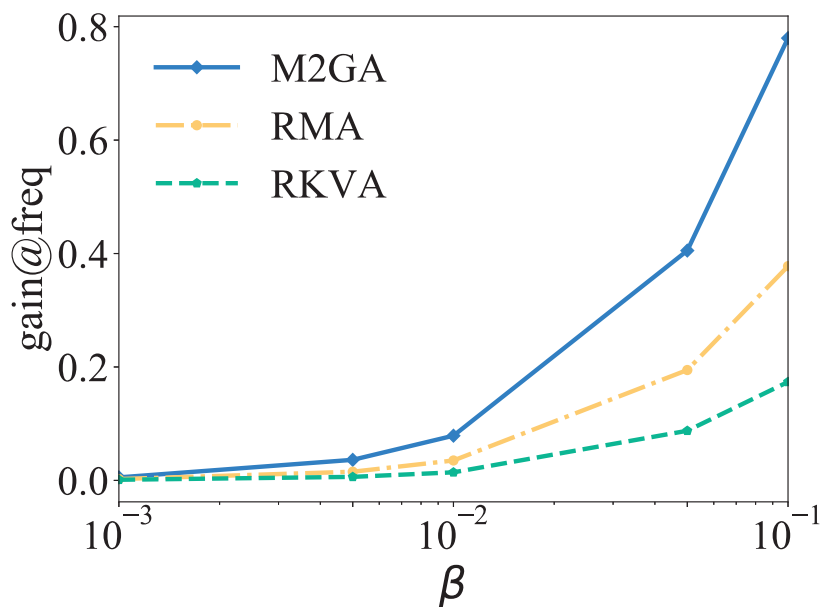
Read our paper for more details

Theoretical evaluation - takeaways

- M2GA is the best-performing attack;
- The frequency gain of an attack increases as # of fake users increases;
- The smaller the true mean value is, the larger the (approximate) mean gain is.

Empirical Evaluation

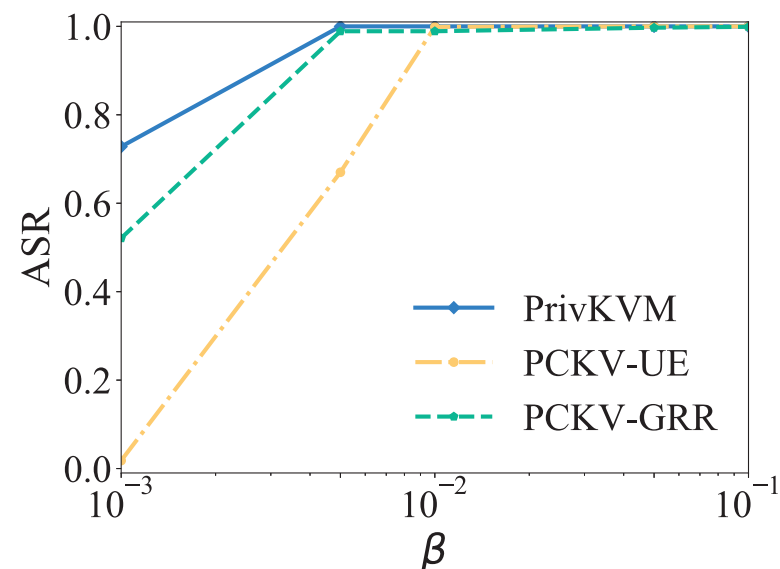
Promoting one target key in a rating dataset with PCKV-UE protocol



β : fraction of fake users

Takeaway: huge frequency and mean gains, even with a small β

Empirical Evaluation – RecSys



Promoting 10 target items in a recommender system

Takeaway: even with a small β , recommendation result is greatly compromised

ASR: success rate (fraction of the 10 target items that are among the top-20 after attack)

Defenses - detect fake users

- One-class classifier
- Anomaly score

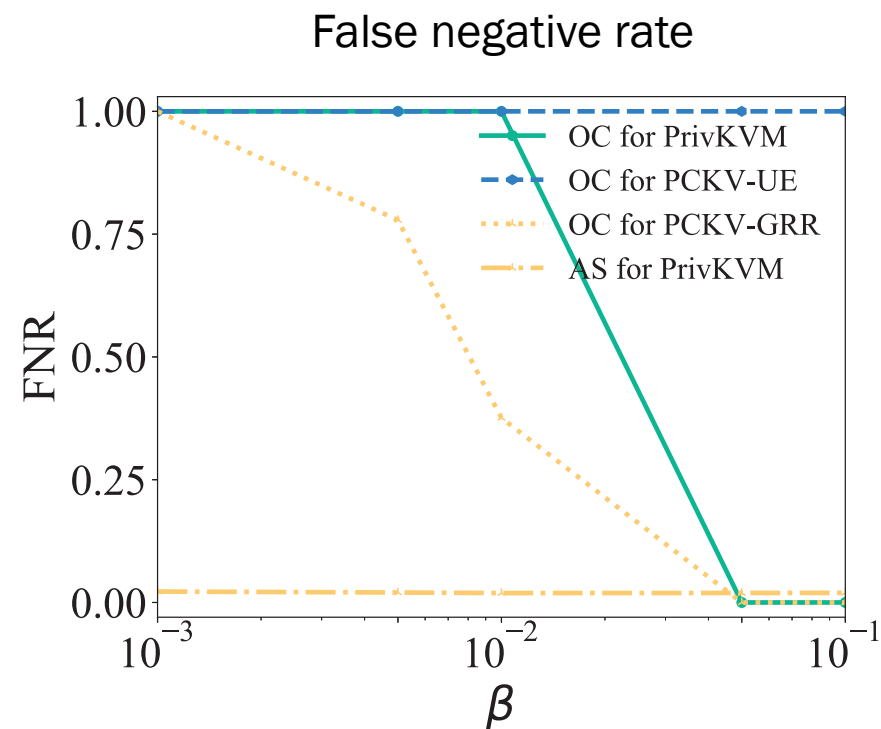
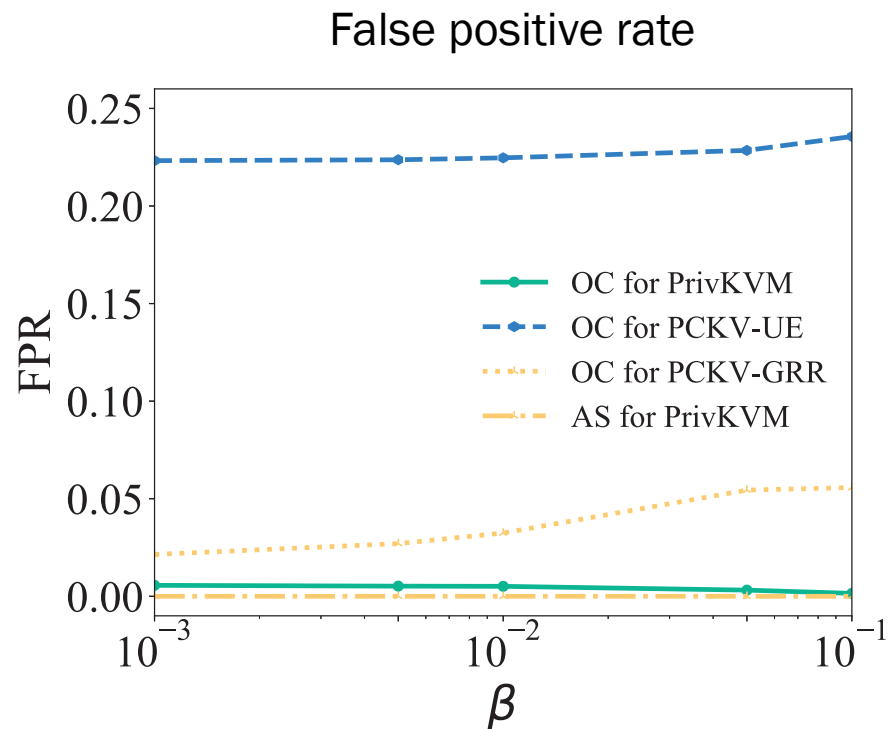
One-class classifier

- Treat each user's message as its features.
- Assumption
 - Server knows a fraction of genuine users

Anomaly score

- Multiple rounds of communications are conducted in PrivKVM
- We can then check consistency of messages from a user across multiple rounds
- We assign an anomaly score to each user
- If the score is greater than anomaly threshold η , consider the user to be fake

Defense results



OC: One-class classifier
AS: Anomaly score

Takeaway: our defenses are effective in some scenarios, but still limited in other cases.

Conclusion

- Key-value LDP protocols are vulnerable to poisoning attacks
- An attacker can promote frequency / mean of any target items
- We highlight the need for strong defenses against such attacks
 - Our defenses help to a degree, but there is more work to do