Fine-grained Poisoning Attack to LDP Protocols for Mean and Variance Estimation

Xiaoguang Li, Ninghui Li, **Wenhai Sun**, Neil Zhenqiang Gong, Hui Li

Background

- Companies are collecting more and more data
- Mean and variance of numerical data are widely-used in:

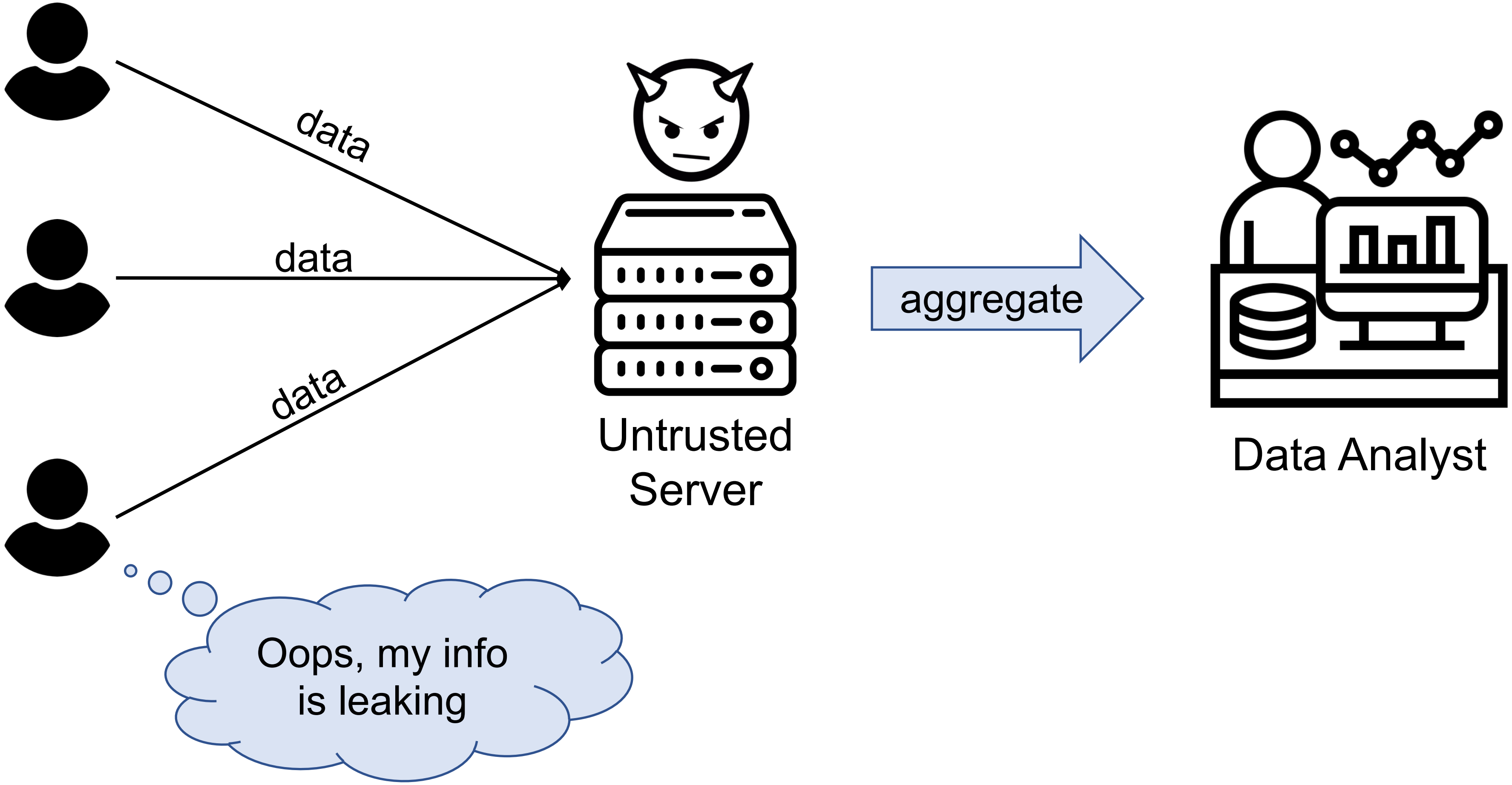
Market Survey

Healthcare Insurance

Real Estate

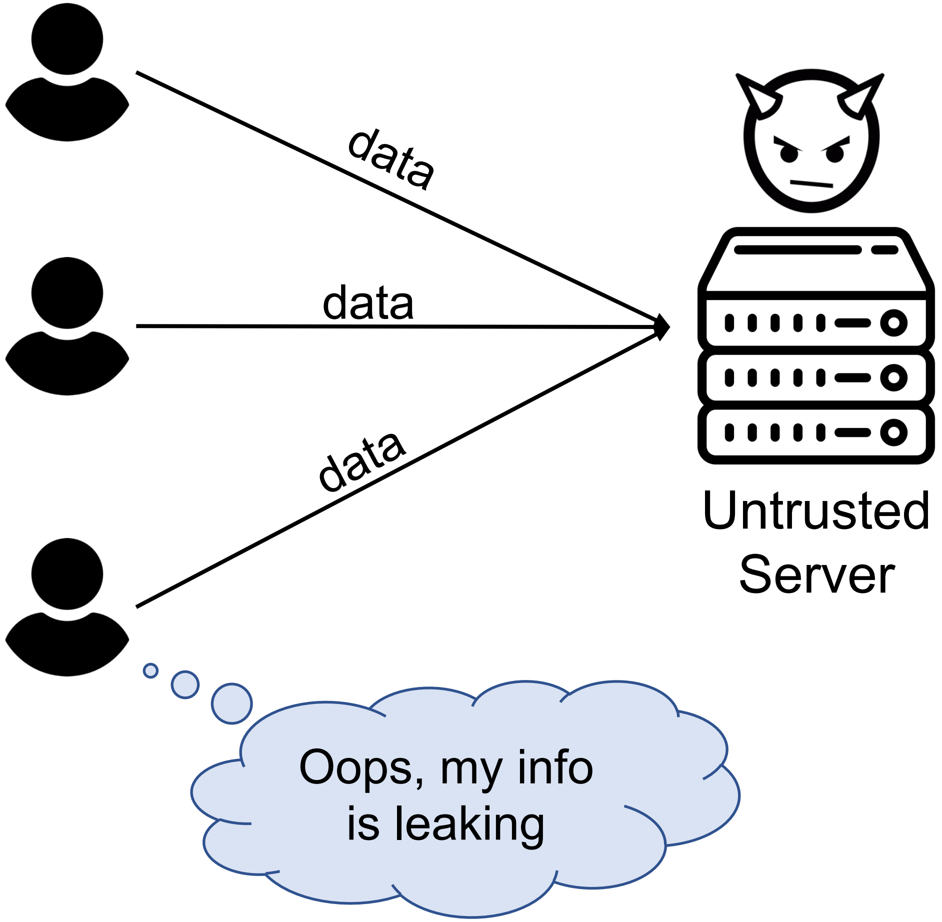
Untrusted Data Collection

In many cases the server is untrusted



Untrusted Data Collection

In many cases the server is untrusted



Moneylife

Congratulations! Your Privacy Is Officially Compromised with No Remedy

Congratulations on your new start up!" "Congratulations!! Few More Things For Your New Venture!!" Isn't this a wonderfully welcoming way to receive an...

4 days ago

New Electronics

The impact of IP address leaks on your privacy

Promoted content: Online privacy has grown to be an important concern in today's interconnected world. The leakage of IP addresses is one of the main...

2 weeks ago

The Quint

'Real-Time' Governance in AP: How Data Collection Is Raising Privacy Concerns

Personal information collected from every household by village volunteers has reportedly been susceptible to leaks. Srinivas Kodali. Published: 13 Jul 2023,...

2 weeks ago

The Times of India

Here's what Realme has to say on personal user data collection concerns

Realme was recently accused of collecting sensitive user data such as call logs, SMS, and location information via the "Enhanced Intelligent Services"...

1 month ago

Newslaundry

CoWIN data leak: Global data protection norms and what's at stake in India

Such breaches not only compromise individuals' privacy but also erode public trust, potentially hampering vaccination efforts.

1 month ago

Scroll.in

CoWIN breach: 'No government in a developed country would have survived a data leak of this scale'

Anivar Aravind, a public interest technologist, explains why you should be worried about the CoWIN data leak and why the government should take...

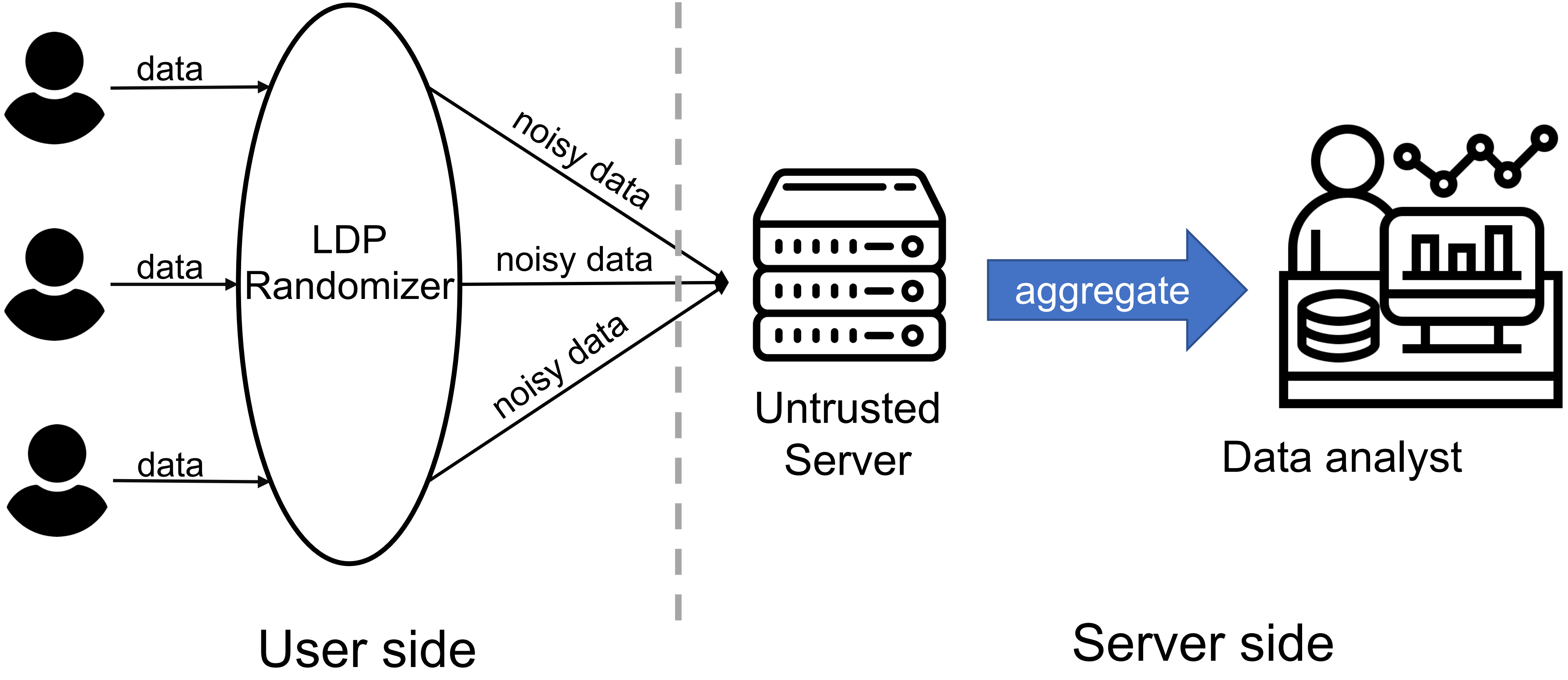
1 month ago

Local Differential Privacy

Local Differential Privacy [Duchi *et al*. FOCS'13]: A randomized algorithm M is ϵ-LDP if and only if

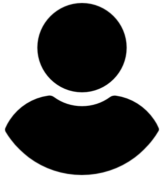
$$\Pr[M(x_1) = t] \leq e^{\epsilon} \Pr[M(x_2) = t]$$

where x_1 and x_2 are any pair of inputs in the domain.



Mean and Variance Estimation

- Stochastic Rounding (SR) [Duchi *et al.* JASA'18]



Encode: $v \in [-1,1]$

$$\Pr[M(v) = v'] = \begin{cases} q + \dfrac{(p-q)(1-v)}{2}, if\ v' = -1 \\ q + \dfrac{(p-q)(1+v)}{2}, if\ v' = 1 \end{cases}$$

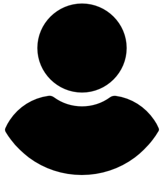
Perturb

v'

$$\hat{\mu} = \frac{\sum_{i=1}^{n} v_i'}{(p-q)n}$$

Aggregate

- Piecewise Mechanism (PM) [Wang *et al.* ICDE'19]



Encode: $v \in [-1,1]$

$$\Pr[M(v) = v'] = \begin{cases} \dfrac{e^{\epsilon/2}(e^{\epsilon/2}-1)}{2(e^{\epsilon/2}+1)}, if\ v' \in [l(v), r(v)] \\ \dfrac{e^{\epsilon/2}-1}{2(e^{\epsilon/2}+e^{\epsilon})}, \quad otherwise \end{cases}$$

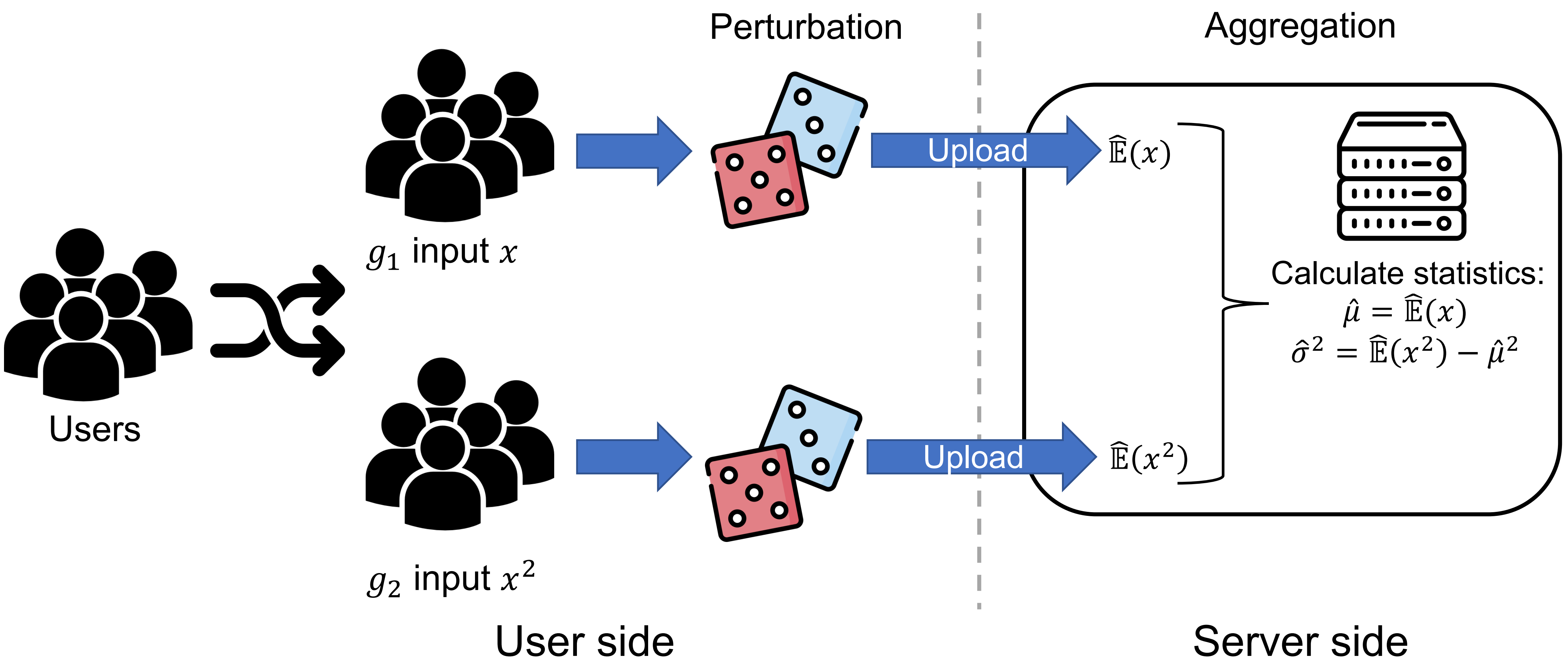
Perturb

v'

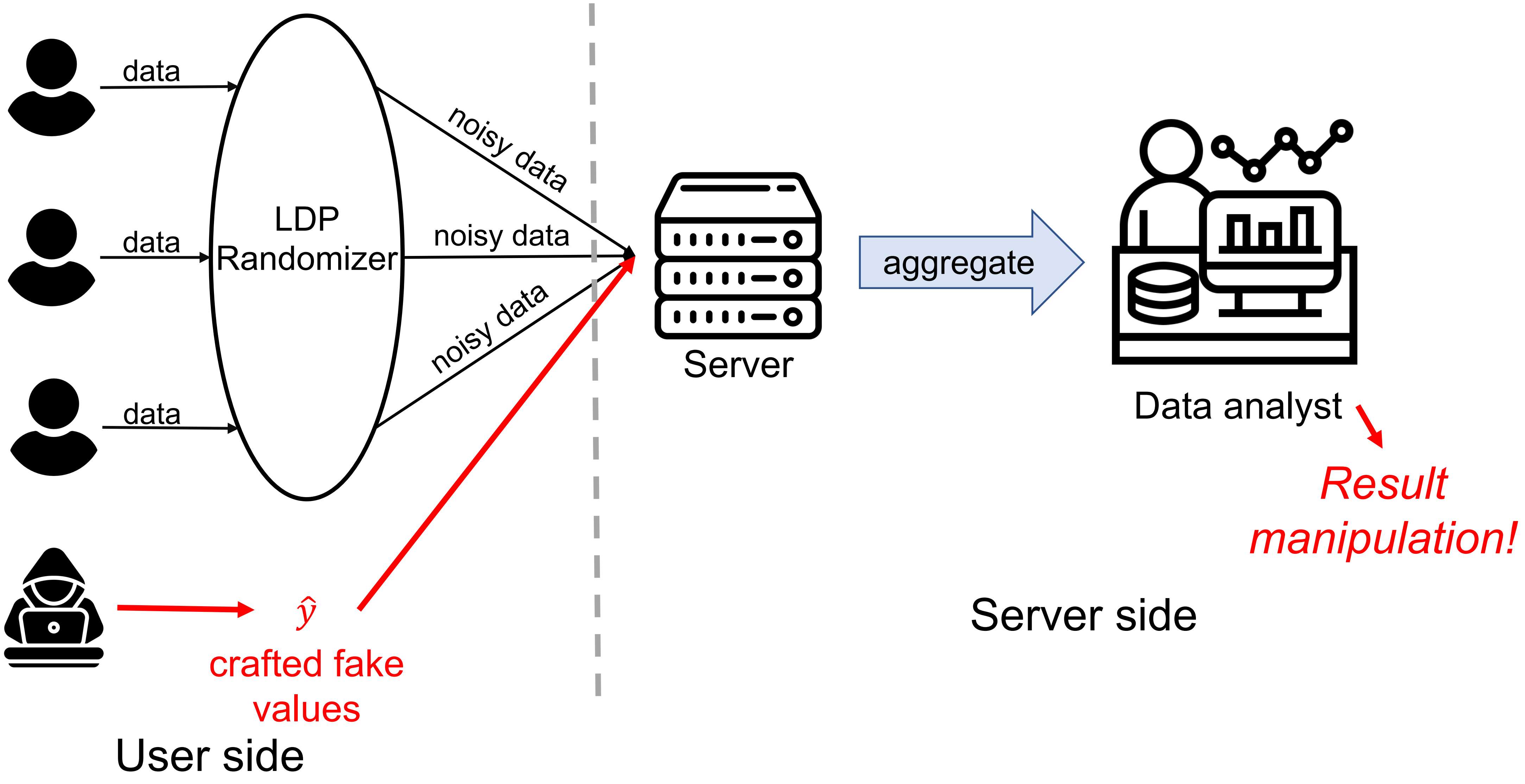
$$\hat{\mu} = \frac{\sum_{i=1}^{n} v_i'}{n}$$

Aggregate

Workflow

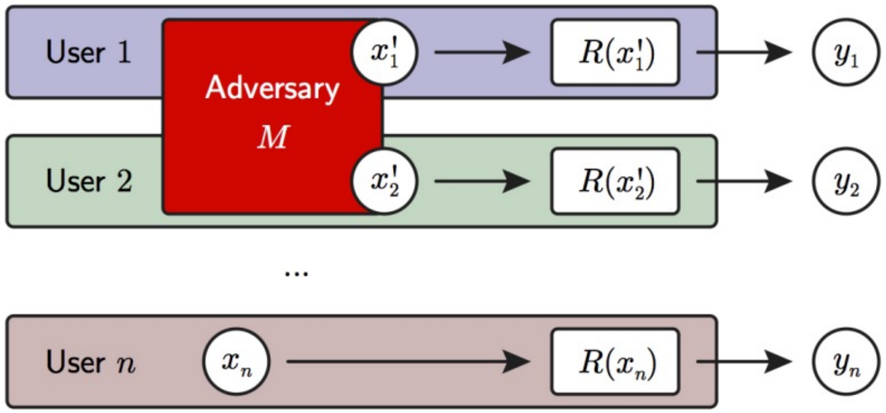

Perturbation
Aggregation
Users
g_1 input x
g_2 input x^2
Upload
Upload
$\widehat{\mathbb{E}}(x)$
$\widehat{\mathbb{E}}(x^2)$
Calculate statistics:
$\hat{\mu} = \widehat{\mathbb{E}}(x)$
$\hat{\sigma}^2 = \widehat{\mathbb{E}}(x^2) - \hat{\mu}^2$
User side
Server side

Data Poisoning Attack



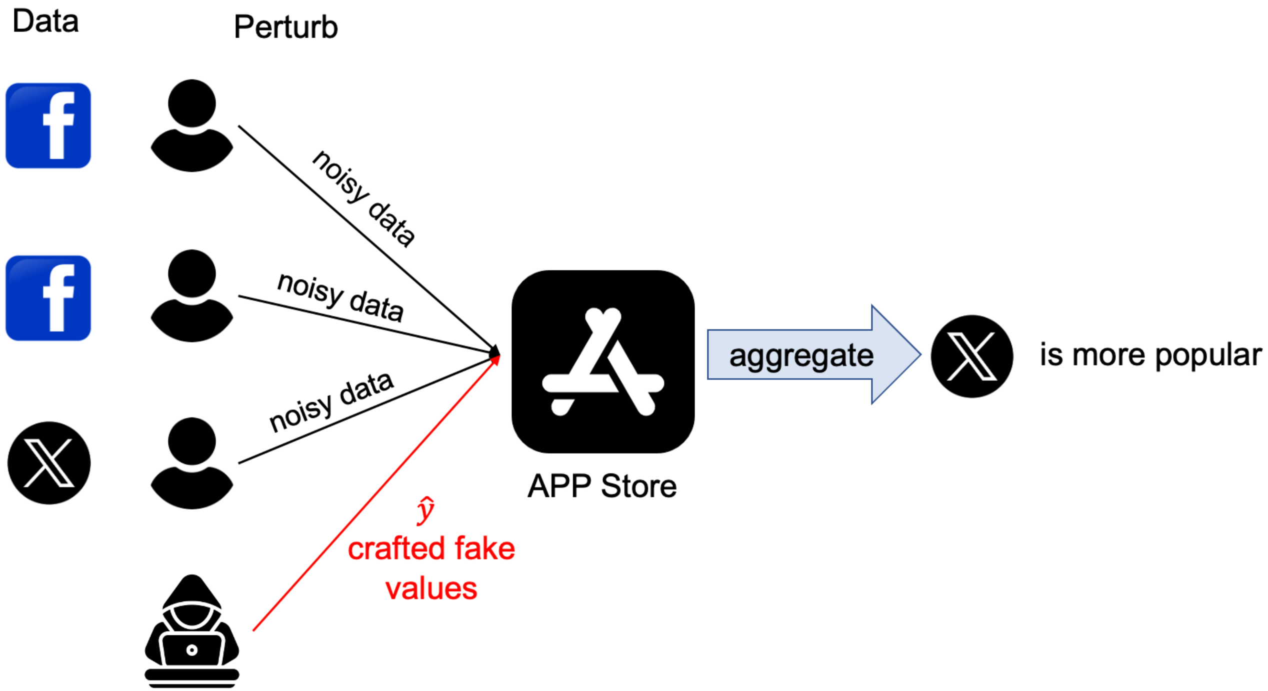
Existing Attacks

[Cheu *et al.* *IEEE* S&P'21]



Goal: Degrade estimation accuracy

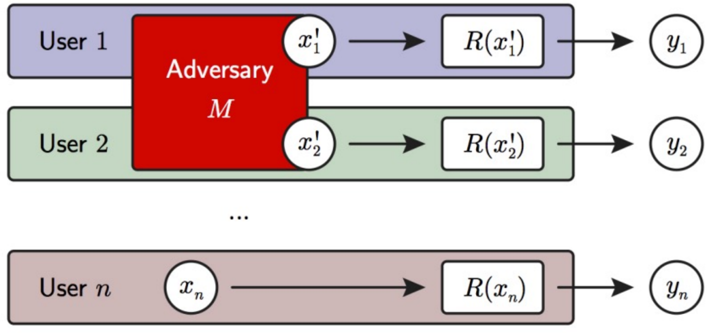
[Cao *et al.* USENIX Security'21; Wu *et al.* USENIX Security'22]



Goal: Promote targeted items by maximizing their associated statistics

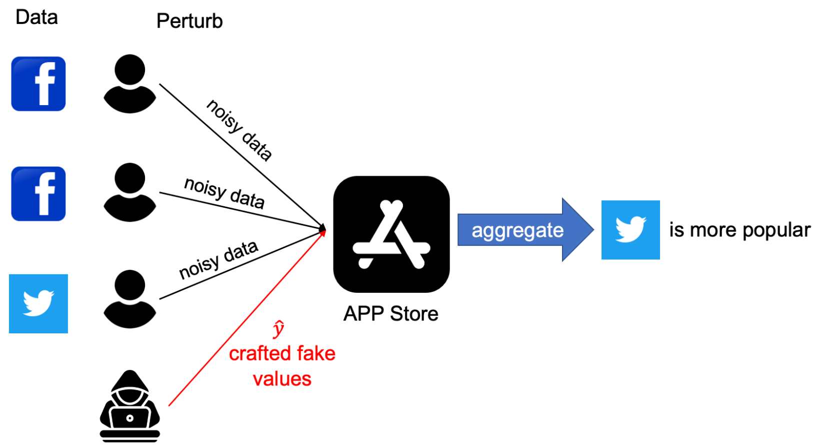
Existing Attacks

[Cheu *et al.* *IEEE* S&P'21]



Goal: Degrade estimation accuracy.

[Cao *et al.* USENIX Security'21; Wu *et al.* USENIX Security'22]



Goal: Promote targeted items by maximizing their associated statistics

Our *fine-grained* attack: manipulate the statistics to an intended value

Threat Model

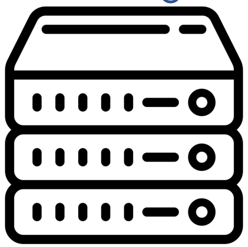
Attack goal: Simultaneously modify the estimated mean $\hat{\mu}$ and variance $\hat{\sigma}^2$ through LDP protocols to target values $\hat{\mu}_t$ and $\hat{\sigma}_t^2$.

Attacker's capabilities:

1. Estimate related statistics
   - The number of users.
   - The sum of users' value
   - The sum of squared users' values
2. Inject fake users into LDP protocols
3. Manipulate input/output of LDP perturbation

Attack Example

I want to do a
market survey for
mean/variance of
users' income



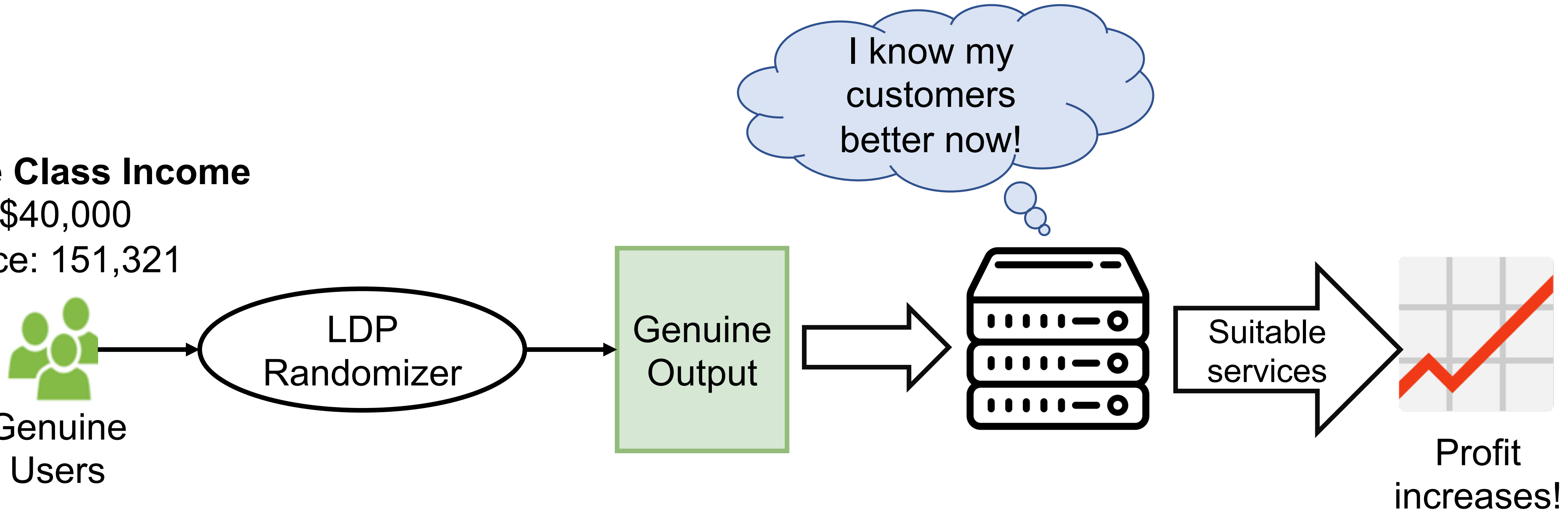
Target Group Income
Mean: ???
Variance: ???

Genuine
Users

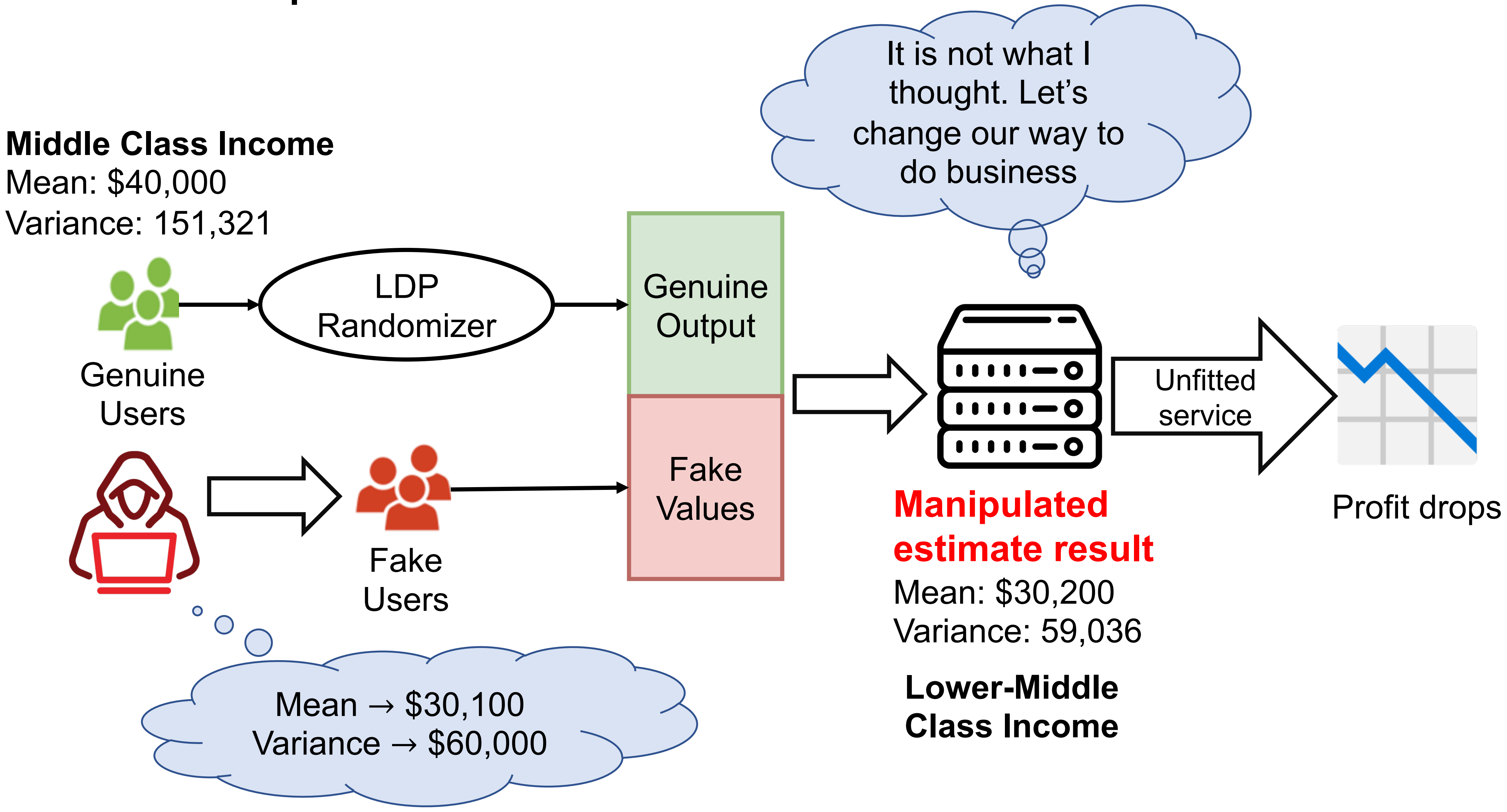
I do not want
to sacrifice
privacy

Attack Example

Middle Class Income
Mean: $40,000
Variance: 151,321

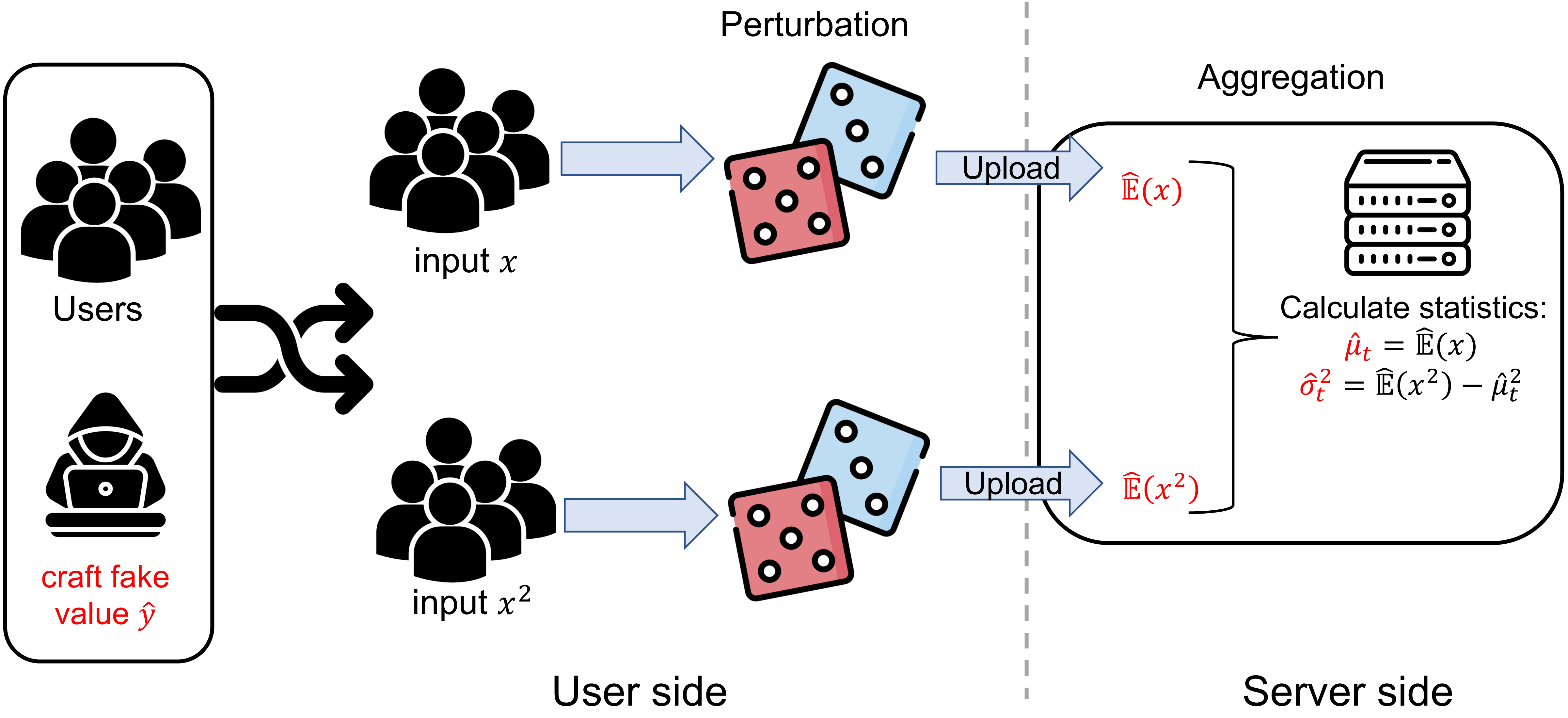


Attack Example



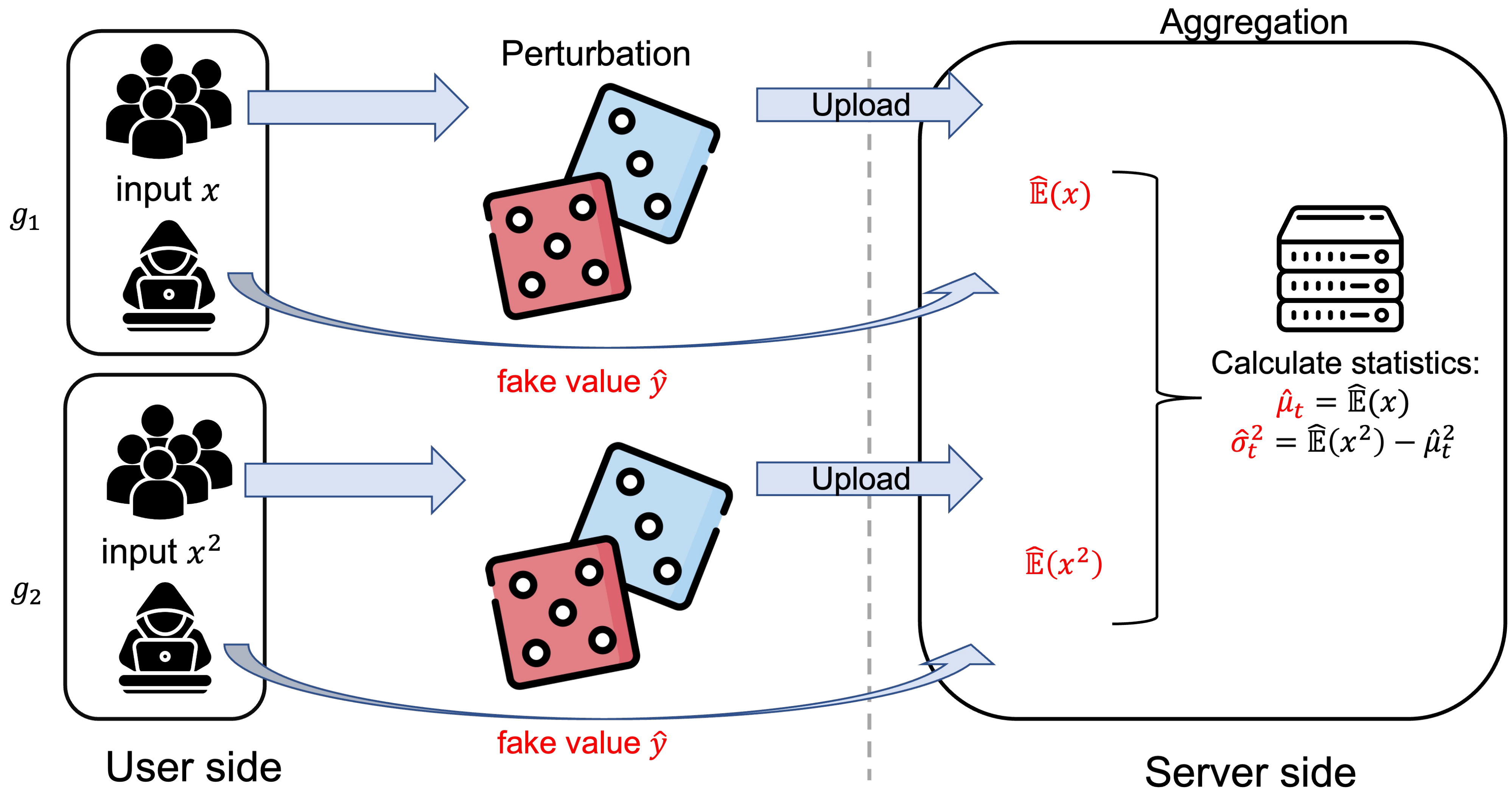
Our Attack

Baseline—Input Poisoning Attack (IPA)



Our Attack

Output Poisoning Attack (OPA) — Manipulate perturbation output directly



Error Analysis

Analyze attack error: $\mathbb{E}[(\hat{\mu}_t - \mu_t)^2]$ and $\mathbb{E}[(\hat{\sigma}_t^2 - \sigma_t^2)^2]$

	Baseline (IPA)	OPA
$\mathrm{Err}(\hat{\mu}_t)$ in SR	$\mathcal{P} + \frac{2}{(m+n)(p-q)^2} - Q$	$\frac{\left(2n-2(p-q)^2 S^{(2)}\right)}{(m+n)^2(p-q)^2} + \frac{S^{(2)}}{(m+n)^2} + \mathcal{P}$
$\mathrm{Err}(\hat{\sigma}_t^2)$ in SR	$\leq \frac{2}{(m+n)(p-q)^2} - \frac{S^{(4)}}{(m+n)^2} + \mathcal{T}_{\mathrm{SR}}^{IPA} + 1$	$\leq \frac{2n-2(p-q)^2 S^{(4)}}{(m+n)^2(p-q)^2} + \frac{S^{(4)}}{(m+n)^2} + \mathcal{T}_{\mathrm{SR}}^{OPA} + 1$
$\mathrm{Err}(\hat{\mu}_t)$ in PM	$\frac{2(e^{\varepsilon/2}+3)}{3(n+m)(e^{\varepsilon/2}-1)^2} + \mathcal{P} + Q + \frac{2Q}{(e^{\varepsilon/2}-1)}$	$\mathcal{P} + \frac{2n(e^{\varepsilon/2}+3)}{3(m+n)^2(e^{\varepsilon/2}-1)^2} + \frac{(1+e^{\varepsilon/2})S^{(2)}}{(m+n)^2(e^{\varepsilon/2}-1)}$
$\mathrm{Err}(\hat{\sigma}_t^2)$ in PM	$\leq \frac{2(e^{\varepsilon/2}+3)}{3(n+m)(e^{\varepsilon/2}-1)^2} + \frac{2(S^{(4)}+\mathcal{Y}_u^{(4)})}{(n+m)^2(e^{\varepsilon/2}-1)} + \frac{(S^{(4)}+\mathcal{Y}_u^{(4)})}{(m+n)^2} + \mathcal{T}_{\mathrm{PM}}^{IPA} + 1$	$\leq \frac{2n(e^{\varepsilon/2}+3)}{3(m+n)^2(e^{\varepsilon/2}-1)^2} + \frac{(1+e^{\varepsilon/2})S^{(4)}}{(m+n)^2(e^{\varepsilon/2}-1)} + \mathcal{T}_{\mathrm{PM}}^{OPA} + 1$

$\hat{\mu}_t, \hat{\sigma}_t^2$: The final estimate of mean and variance

μ_t, σ_t^2 : The target values set by the attacker

$\mathcal{P}, Q, \mathcal{T}_{SR}^{IPA}, \mathcal{T}_{PM}^{IPA}, \mathcal{T}_{SR}^{OPA}, \mathcal{T}_{PM}^{OPA}$: Intermediate constant

Error Analysis

Analyze attack error: $\mathbb{E}[(\hat{\mu}_t - \mu_t)^2]$ and $\mathbb{E}[(\hat{\sigma}_t^2 - \sigma_t^2)^2]$

	Baseline (IPA)	OPA
$\mathrm{Err}(\hat{\mu}_t)$ in SR	$\mathcal{P} + \frac{2}{(m+n)(p-q)^2} - Q$	$\frac{\left(2n-2(p-q)^2 S^{(2)}\right)}{(m+n)^2(p-q)^2} + \frac{S^{(2)}}{(m+n)^2} + \mathcal{P}$
$\mathrm{Err}(\hat{\sigma}_t^2)$ in SR	$\leq \frac{2}{(m+n)(p-q)^2} - \frac{S^{(4)}}{(m+n)^2} + \mathcal{T}_{\mathrm{SR}}^{IPA} + 1$	$\leq \frac{2n-2(p-q)^2 S^{(4)}}{(m+n)^2(p-q)^2} + \frac{S^{(4)}}{(m+n)^2} + \mathcal{T}_{\mathrm{SR}}^{OPA} + 1$
$\mathrm{Err}(\hat{\mu}_t)$ in PM	$\frac{2(e^{\varepsilon/2}+3)}{3(n+m)(e^{\varepsilon/2}-1)^2} + \mathcal{P} + Q + \frac{2Q}{(e^{\varepsilon/2}-1)}$	$\mathcal{P} + \frac{2n(e^{\varepsilon/2}+3)}{3(m+n)^2(e^{\varepsilon/2}-1)^2} + \frac{(1+e^{\varepsilon/2})S^{(2)}}{(m+n)^2(e^{\varepsilon/2}-1)}$
$\mathrm{Err}(\hat{\sigma}_t^2)$ in PM	$\leq \frac{2(e^{\varepsilon/2}+3)}{3(n+m)(e^{\varepsilon/2}-1)^2} + \frac{2(S^{(4)}+\mathcal{Y}_u^{(4)})}{(n+m)^2(e^{\varepsilon/2}-1)} + \frac{(S^{(4)}+\mathcal{Y}_u^{(4)})}{(m+n)^2} + \mathcal{T}_{\mathrm{PM}}^{IPA} + 1$	$\leq \frac{2n(e^{\varepsilon/2}+3)}{3(m+n)^2(e^{\varepsilon/2}-1)^2} + \frac{(1+e^{\varepsilon/2})S^{(4)}}{(m+n)^2(e^{\varepsilon/2}-1)} + \mathcal{T}_{\mathrm{PM}}^{OPA} + 1$

How do our attacks perform under different LDP protocols?

Error Analysis

Analyze attack error: $\mathbb{E}[(\hat{\mu}_t - \mu_t)^2]$ and $\mathbb{E}[(\hat{\sigma}_t^2 - \sigma_t^2)^2]$

	Baseline (IPA)	OPA
$\mathrm{Err}(\hat{\mu}_t)$ in SR	$\mathcal{P} + \frac{2}{(m+n)(p-q)^2} - Q$	$\frac{\left(2n-2(p-q)^2 S^{(2)}\right)}{(m+n)^2(p-q)^2} + \frac{S^{(2)}}{(m+n)^2} + \mathcal{P}$
$\mathrm{Err}(\hat{\sigma}_t^2)$ in SR	$\leq \frac{2}{(m+n)(p-q)^2} - \frac{S^{(4)}}{(m+n)^2} + \mathcal{T}_{\mathrm{SR}}^{IPA} + 1$	$\leq \frac{2n-2(p-q)^2 S^{(4)}}{(m+n)^2(p-q)^2} + \frac{S^{(4)}}{(m+n)^2} + \mathcal{T}_{\mathrm{SR}}^{OPA} + 1$
$\mathrm{Err}(\hat{\mu}_t)$ in PM	$\frac{2(e^{\varepsilon/2}+3)}{3(n+m)(e^{\varepsilon/2}-1)^2} + \mathcal{P} + Q + \frac{2Q}{(e^{\varepsilon/2}-1)}$	$\mathcal{P} + \frac{2n(e^{\varepsilon/2}+3)}{3(m+n)^2(e^{\varepsilon/2}-1)^2} + \frac{(1+e^{\varepsilon/2})S^{(2)}}{(m+n)^2(e^{\varepsilon/2}-1)}$
$\mathrm{Err}(\hat{\sigma}_t^2)$ in PM	$\leq \frac{2(e^{\varepsilon/2}+3)}{3(n+m)(e^{\varepsilon/2}-1)^2} + \frac{2(S^{(4)}+\mathcal{Y}_u^{(4)})}{(n+m)^2(e^{\varepsilon/2}-1)} + \frac{(S^{(4)}+\mathcal{Y}_u^{(4)})}{(m+n)^2} + \mathcal{T}_{\mathrm{PM}}^{IPA} + 1$	$\leq \frac{2n(e^{\varepsilon/2}+3)}{3(m+n)^2(e^{\varepsilon/2}-1)^2} + \frac{(1+e^{\varepsilon/2})S^{(4)}}{(m+n)^2(e^{\varepsilon/2}-1)} + \mathcal{T}_{\mathrm{PM}}^{OPA} + 1$

How do our attacks perform under different LDP protocols?

- When ϵ is small (large), it is easier to manipulate estimates in SR (PM) with small attack error.

Error Analysis

Analyze attack error: $\mathbb{E}[(\hat{\mu}_t - \mu_t)^2]$ and $\mathbb{E}[(\hat{\sigma}_t^2 - \sigma_t^2)^2]$

	Baseline (IPA)	OPA
$\mathrm{Err}(\hat{\mu}_t)$ in SR	$\mathcal{P} + \frac{2}{(m+n)(p-q)^2} - \mathcal{Q}$	$\frac{\left(2n-2(p-q)^2 S^{(2)}\right)}{(m+n)^2(p-q)^2} + \frac{S^{(2)}}{(m+n)^2} + \mathcal{P}$
$\mathrm{Err}(\hat{\sigma}_t^2)$ in SR	$\leq \frac{2}{(m+n)(p-q)^2} - \frac{S^{(4)}}{(m+n)^2} + \mathcal{T}_{\mathrm{SR}}^{IPA} + 1$	$\leq \frac{2n-2(p-q)^2 S^{(4)}}{(m+n)^2(p-q)^2} + \frac{S^{(4)}}{(m+n)^2} + \mathcal{T}_{\mathrm{SR}}^{OPA} + 1$
$\mathrm{Err}(\hat{\mu}_t)$ in PM	$\frac{2(e^{\varepsilon/2}+3)}{3(n+m)(e^{\varepsilon/2}-1)^2} + \mathcal{P} + \mathcal{Q} + \frac{2\mathcal{Q}}{(e^{\varepsilon/2}-1)}$	$\mathcal{P} + \frac{2n(e^{\varepsilon/2}+3)}{3(m+n)^2(e^{\varepsilon/2}-1)^2} + \frac{(1+e^{\varepsilon/2})S^{(2)}}{(m+n)^2(e^{\varepsilon/2}-1)}$
$\mathrm{Err}(\hat{\sigma}_t^2)$ in PM	$\leq \frac{2(e^{\varepsilon/2}+3)}{3(n+m)(e^{\varepsilon/2}-1)^2} + \frac{2(S^{(4)}+\mathcal{Y}_u^{(4)})}{(n+m)^2(e^{\varepsilon/2}-1)} + \frac{(S^{(4)}+\mathcal{Y}_u^{(4)})}{(m+n)^2} + \mathcal{T}_{\mathrm{PM}}^{IPA} + 1$	$\leq \frac{2n(e^{\varepsilon/2}+3)}{3(m+n)^2(e^{\varepsilon/2}-1)^2} + \frac{(1+e^{\varepsilon/2})S^{(4)}}{(m+n)^2(e^{\varepsilon/2}-1)} + \mathcal{T}_{\mathrm{PM}}^{OPA} + 1$

Does our OPA attack outperform the baseline by leveraging LDP characteristics?

Error Analysis

Analyze attack error: $\mathbb{E}[(\hat{\mu}_t - \mu_t)^2]$ and $\mathbb{E}[(\hat{\sigma}_t^2 - \sigma_t^2)^2]$

	Baseline (IPA)	OPA
$\mathrm{Err}(\hat{\mu}_t)$ in SR	$\mathcal{P} + \frac{2}{(m+n)(p-q)^2} - \mathcal{Q}$	$\frac{\left(2n-2(p-q)^2 S^{(2)}\right)}{(m+n)^2(p-q)^2} + \frac{S^{(2)}}{(m+n)^2} + \mathcal{P}$
$\mathrm{Err}(\hat{\sigma}_t^2)$ in SR	$\leq \frac{2}{(m+n)(p-q)^2} - \frac{S^{(4)}}{(m+n)^2} + \mathcal{T}_{\mathrm{SR}}^{IPA} + 1$	$\leq \frac{2n-2(p-q)^2 S^{(4)}}{(m+n)^2(p-q)^2} + \frac{S^{(4)}}{(m+n)^2} + \mathcal{T}_{\mathrm{SR}}^{OPA} + 1$
$\mathrm{Err}(\hat{\mu}_t)$ in PM	$\frac{2(e^{\varepsilon/2}+3)}{3(n+m)(e^{\varepsilon/2}-1)^2} + \mathcal{P} + \mathcal{Q} + \frac{2\mathcal{Q}}{(e^{\varepsilon/2}-1)}$	$\mathcal{P} + \frac{2n(e^{\varepsilon/2}+3)}{3(m+n)^2(e^{\varepsilon/2}-1)^2} + \frac{(1+e^{\varepsilon/2})S^{(2)}}{(m+n)^2(e^{\varepsilon/2}-1)}$
$\mathrm{Err}(\hat{\sigma}_t^2)$ in PM	$\leq \frac{2(e^{\varepsilon/2}+3)}{3(n+m)(e^{\varepsilon/2}-1)^2} + \frac{2(S^{(4)}+\mathcal{Y}_u^{(4)})}{(n+m)^2(e^{\varepsilon/2}-1)} + \frac{(S^{(4)}+\mathcal{Y}_u^{(4)})}{(m+n)^2} + \mathcal{T}_{\mathrm{PM}}^{IPA} + 1$	$\leq \frac{2n(e^{\varepsilon/2}+3)}{3(m+n)^2(e^{\varepsilon/2}-1)^2} + \frac{(1+e^{\varepsilon/2})S^{(4)}}{(m+n)^2(e^{\varepsilon/2}-1)} + \mathcal{T}_{\mathrm{PM}}^{OPA} + 1$

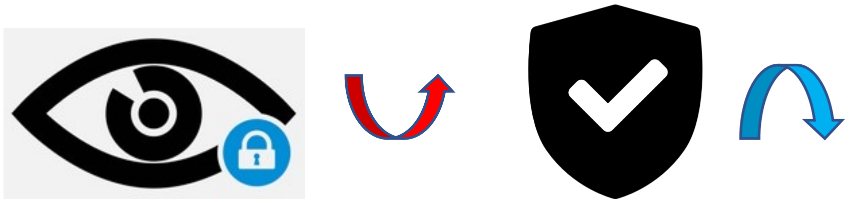
Does our OPA attack outperform the baseline by leveraging LDP characteristics?

- **OPA is more effective** with small attack error

Privacy-security Relationship

Prior attacks

Privacy-security tradeoff: Higher privacy (smaller ϵ), lower security (better attack result).

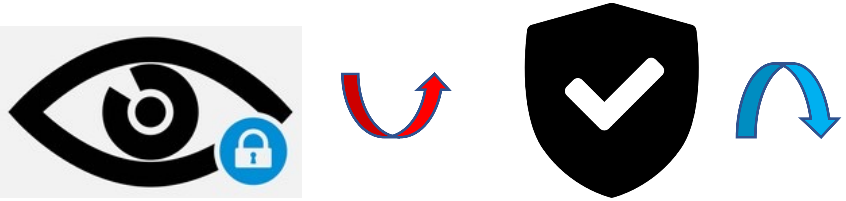


Strong privacy Weaker Security

Privacy-security Relationship

Prior attacks

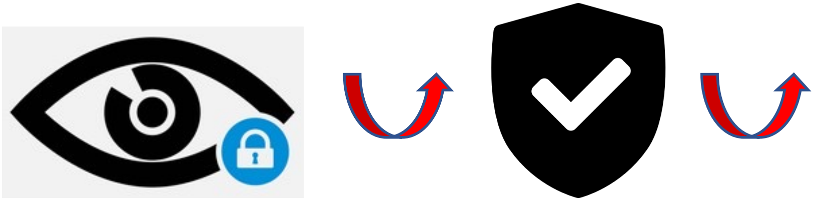
Privacy-security tradeoff: Higher privacy (smaller ϵ), lower security (better attack result).



Strong privacy Weaker Security

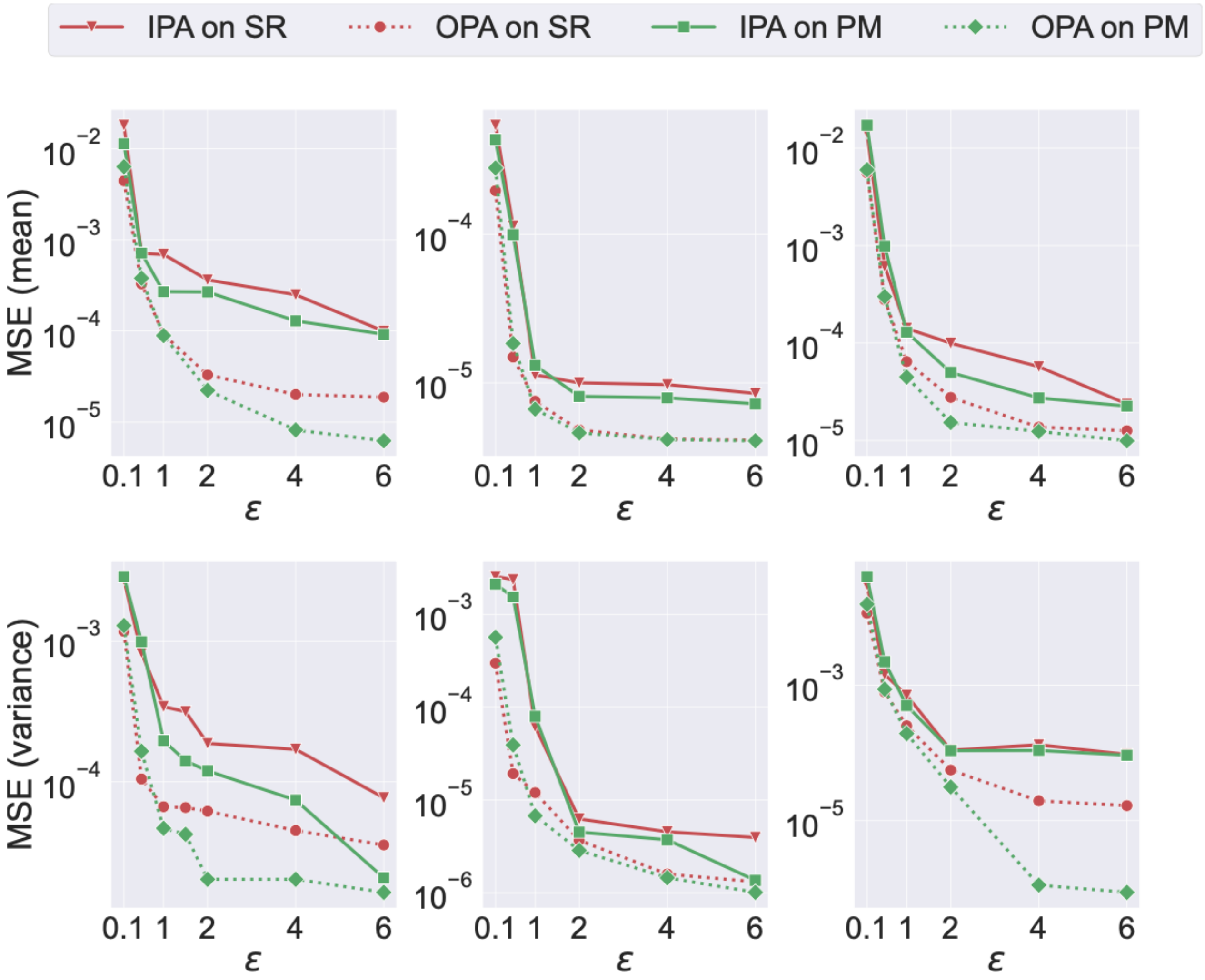
Our attack

Privacy-security consistency: Higher privacy (smaller ϵ), higher security (worse attack result).



Strong privacy Strong Security

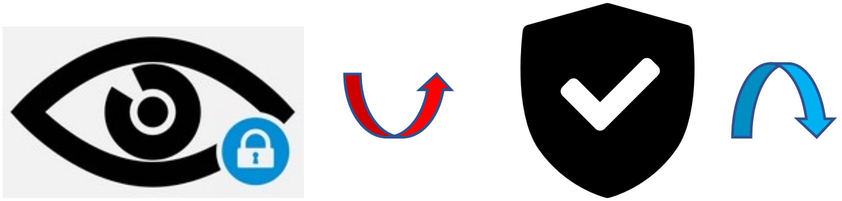
Privacy-security Consistency



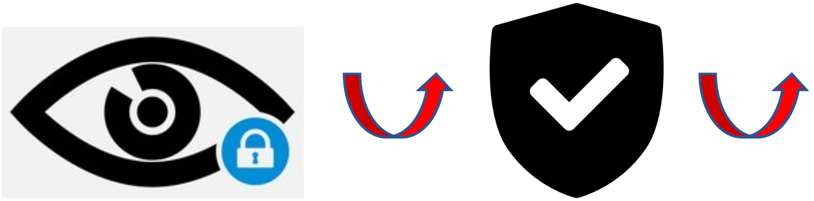
(a) Taxi (b) Income (c) Retirement

Which is true?

Privacy-security tradeoff



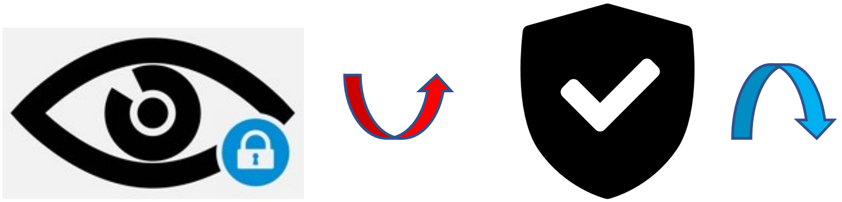
Privacy-security consistency

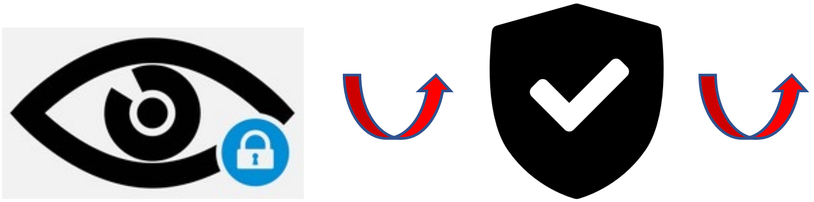


Which is true?

Both are correct!

The relationship depends on how you perform the attack and attack goal.





Prior Attack

Intuition:

- Higher privacy facilitates attack [Cheu *et al. IEEE* S&P'21];
- A smaller ϵ allows attacker to contribute more to the estimates [Cao *et al.* USENIX Security'21; Wu *et al.* USENIX Security'22]

Our Attack

Intuition: Difficult to precisely manipulate the LDP estimates under large noise (small ϵ)

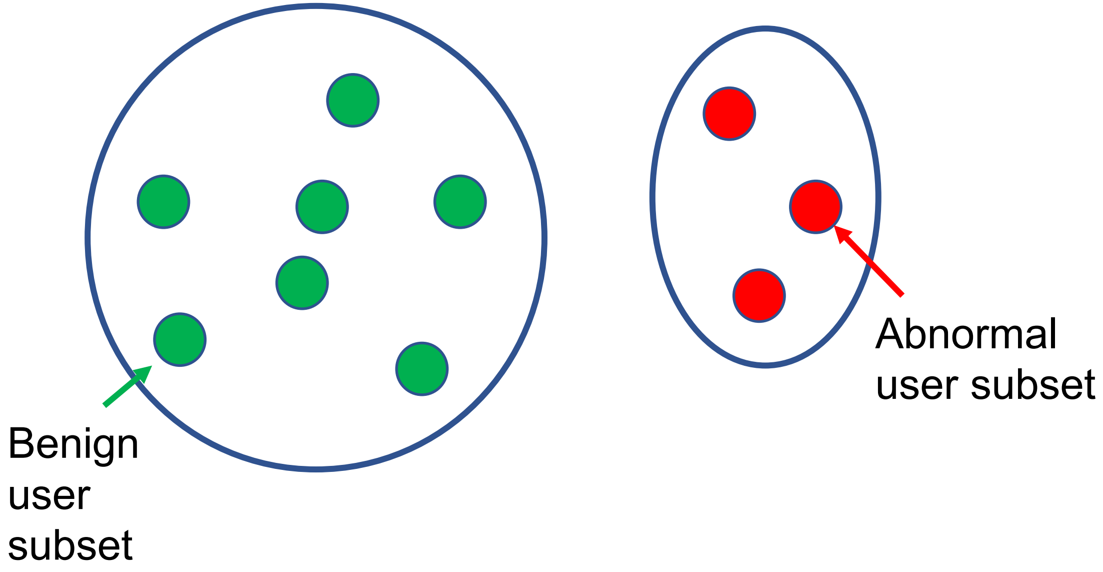
Defense Exploration

Clustering-based mitigation

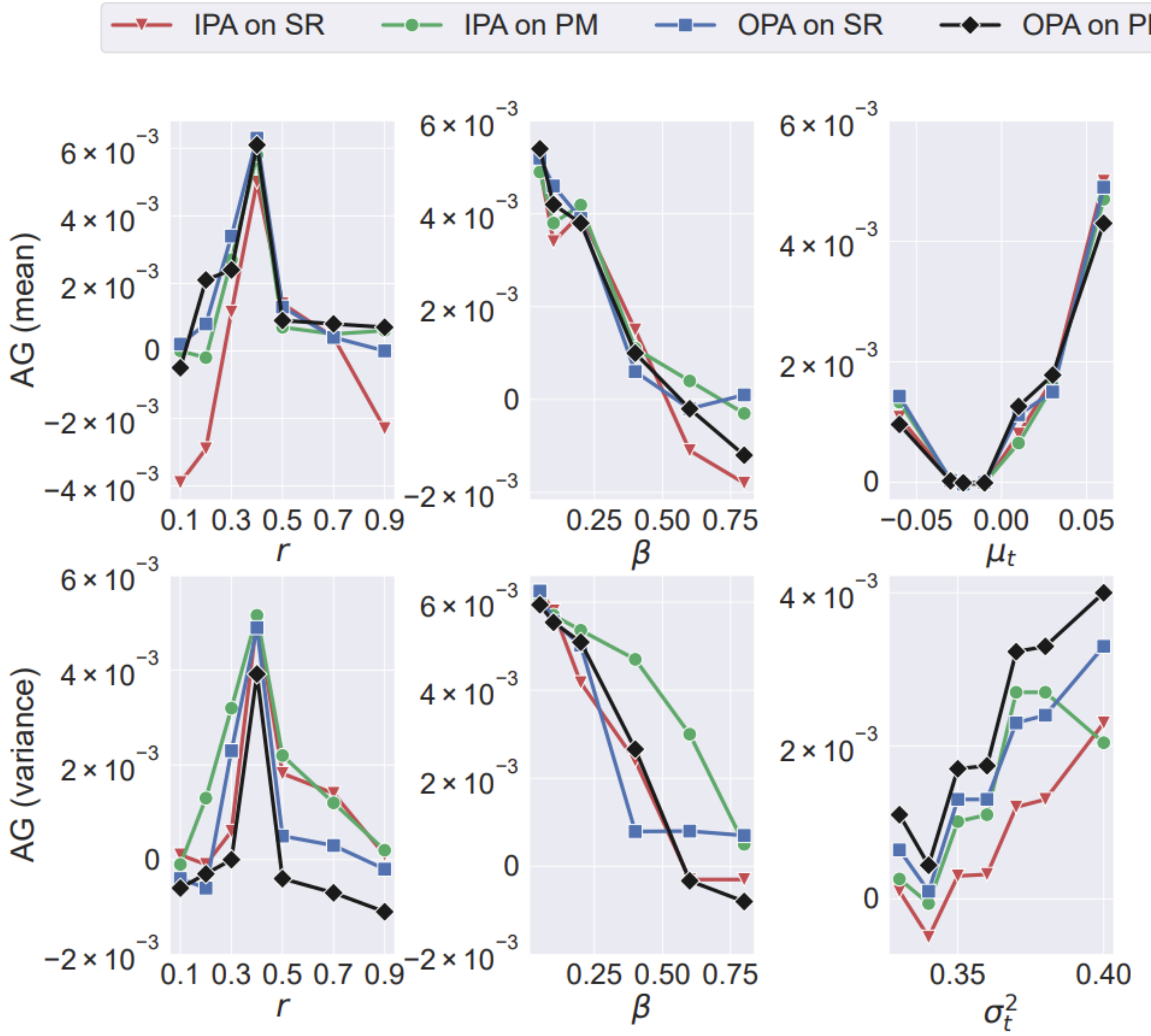
- The majority of users are benign
- Sample multiple subsets of users (sampling rate r)
- Cluster containing most subsets used for estimation

Metric

- Accuracy Gain (AG): $MSE_{before} - MSE_{after}$.
- Larger AG means better defense result



Mitigation Evaluation



- Sampling rate r has an impact on defense;
- Fewer fake users makes mitigation easier
- The mitigation is more effective when the target values are farther away from the true values.

More Research Needed

Protocol Robustness Analysis

- Robustness of different LDP protocols under poisoning attacks
- Provides insights into future design

Defense Design

- Attack detection for fake values and fake users
- Fault tolerance

Conclusion

- We propose fine-grained poisoning attacks for LDP protocols
- A disturbing fact for secure LDP setup: both privacy-security tradeoff and consistency are true
- We propose the mitigation and highlight the urgent needs for
  - Robust LDP design
  - More effective defenses

Thank you!