Guarding Multiple Secrets: Enhanced Summary Statistic Privacy for Data Sharing

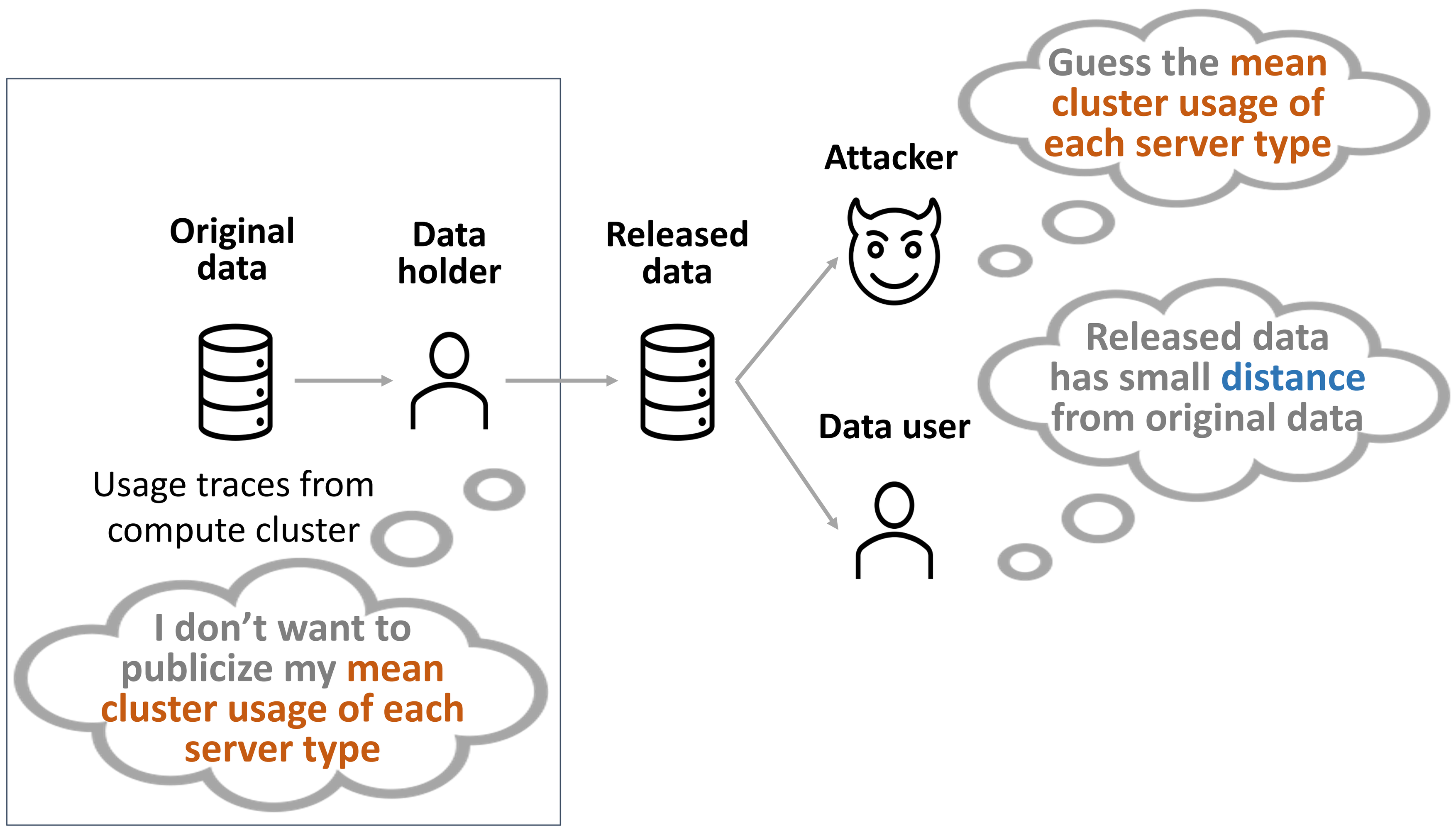
Shuaiqi Wang, Rongzhe Wei, Mohsen Ghassemi, Eleonora Kreacic, Vamsi K. Potluru

We propose a framework to define, analyze, and protect multi-secret summary statistics privacy in data sharing. Given an attacker's objective spanning from inferring a subset to the entirety of summary statistic secrets, we systematically design and analyze tailored privacy metrics. We analyze the tradeoff between privacy and distortion.

Data Sharing in Practice

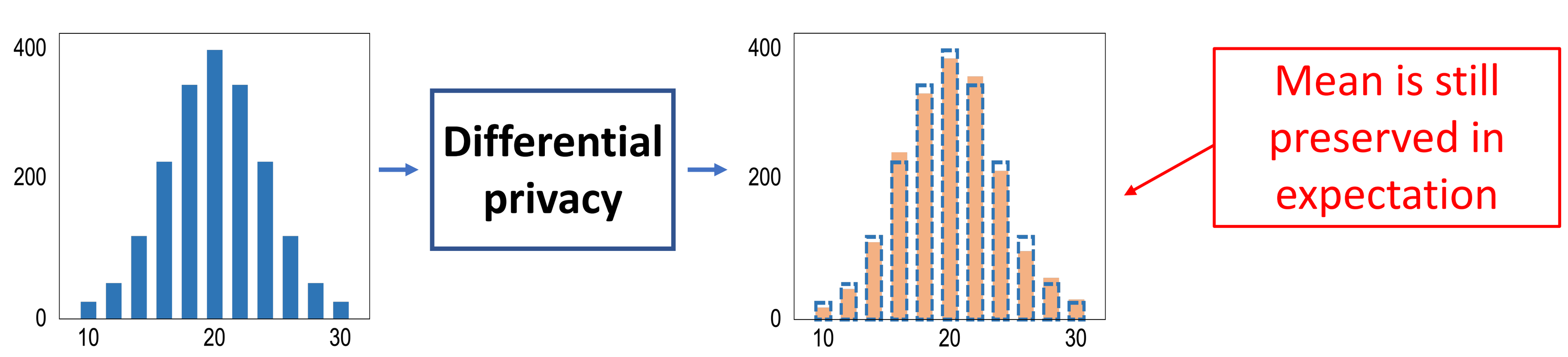
- ***Motivating Scenario***

Data holder produces released cluster usage traces for the data user.



- ***Differential Privacy Doesn't Work***

Differential privacy is designed to *preserve* the underlying data distribution, while protecting individual-level privacy.



Problem Formulation

Data holder

Distributional secrets to protect :

mathematically defined as functions of the data distribution, e.g., secrets g = means in the motivating scenario

Attacker

Privacy metric Π_ϵ:

given the attacker's objective, probability of guessing the secrets within tolerance ϵ by the best attack strategy

Data user

Distortion metric Δ:

worst-case distance between the original distribution and the released distribution

Objective:

$$\min \Pi_\epsilon \qquad \text{subject to } \Delta \le T$$

Privacy Metric Design

Within diverse data sharing paradigms, we design tailored privacy metrics. (see paper for the detailed definitions)

Union Privacy: *prevents attackers guessing any secret correctly*

probability of the attacker guessing ***any*** secret to within a tolerance range, ϵ_i for secret g_i

Intersection privacy: *secrets are compromised only when the attacker guesses all of them simultaneously*

probability of the attacker guessing ***all*** secrets to within a tolerance range, ϵ_i for secret g_i

Group privacy: *secrets are compromised when the attacker guesses a certain group of them*

probability of the attacker guessing ***any certain group*** of secrets to within a tolerance range, ϵ_i for secret g_i

l_p norm privacy: *ensures a significant separation between original and attacker-guessed secret vectors*

probability of l_p norm distance between original and attacker-guessed secret being within a tolerance ϵ

Privacy-Distortion Tradeoff

Theorem (Union Privacy)

(see paper for the detailed statements of each privacy metric)

- For any $T > 0$, when $\Pi_\epsilon \le T$, we have

$$\Delta > 2\gamma \left[\frac{1}{1-(1-T)^{1/d}} - 1\right] \cdot \left(\prod_{i\in[d]} \epsilon_i\right)^{1/d},$$

where d is the secret number and

$$\gamma = \min \frac{\text{distance between two potential distributions of original data}}{\text{difference between two potential secrets of original data}}$$

Future Work

- Develop mechanisms tailored to various data distributions and secret types that achieves (near) optimal privacy-distortion tradeoffs.
- Measure privacy empirically for arbitrary secrets and data distributions.

Shuaiqi Wang: shuaiqiw@andrew.cmu.edu
Rongzhe Wei: rongzhe.wei@gatech.edu
Mohsen Ghassemi: mohsen.ghassemi@ jpmchase.com
Eleonora Kreacic: eleonora.kreacic@ jpmchase.com
Vamsi K. Potluru: vamsi.k.potluru@jpmchase.com