Unlocking the Value of Privacy: Trading Aggregate Statistics over Private Correlated Data

Chaoyue Niu, Zhenzhe Zheng, Fan Wu, Shaojie Tang†, Xiaofeng Gao, and Guihai Chen
Shanghai Key Laboratory of Scalable Computing and Systems, Shanghai Jiao Tong University, China
†Department of Information Systems, University of Texas at Dallas, USA
Email:{rvince, zhengzhenzhe, wu-fan}@sjtu.edu.cn; tangshaojie@gmail.com; {gao-xf, gchen}@cs.sjtu.edu.cn

Introduction

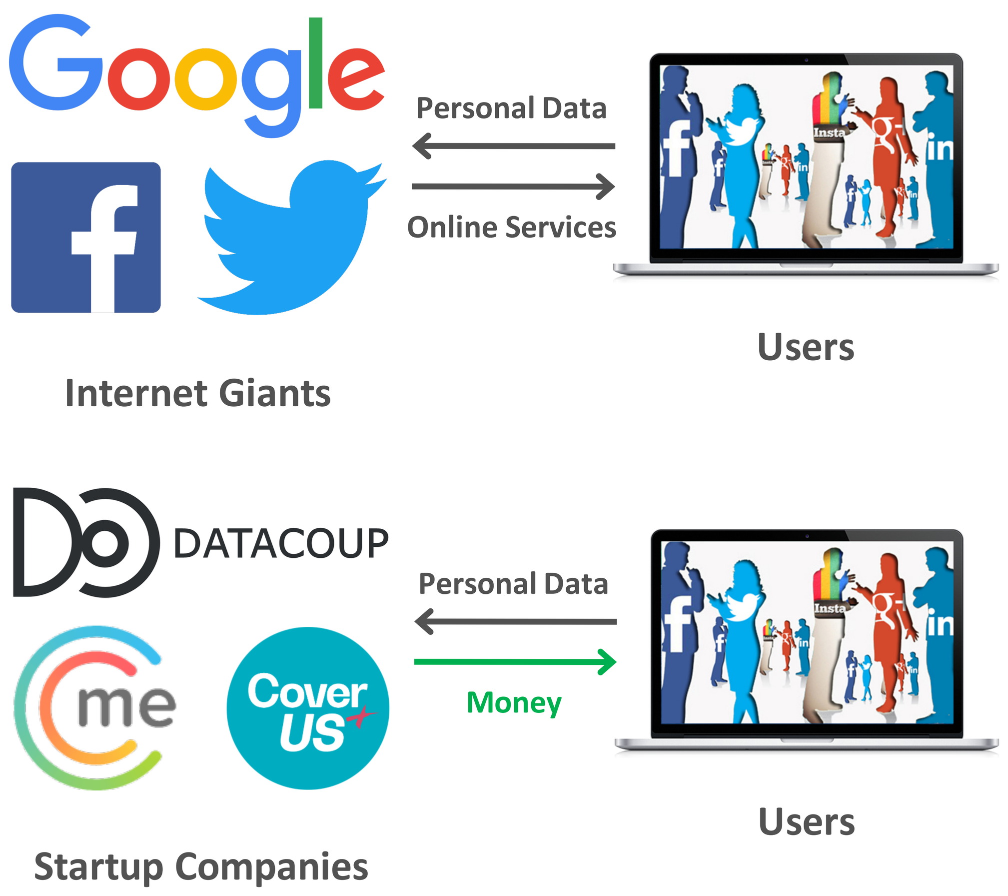


Fig. 1. Private Data Circulations in Real Life.

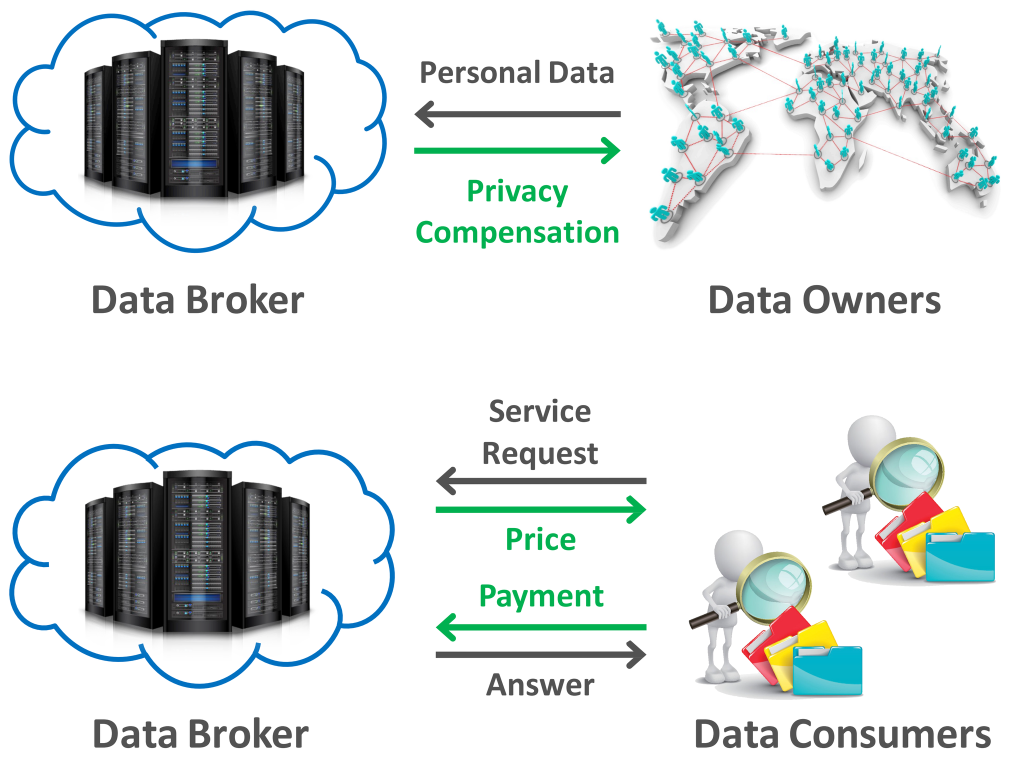


Fig. 2. A General System Model of Data Markets.

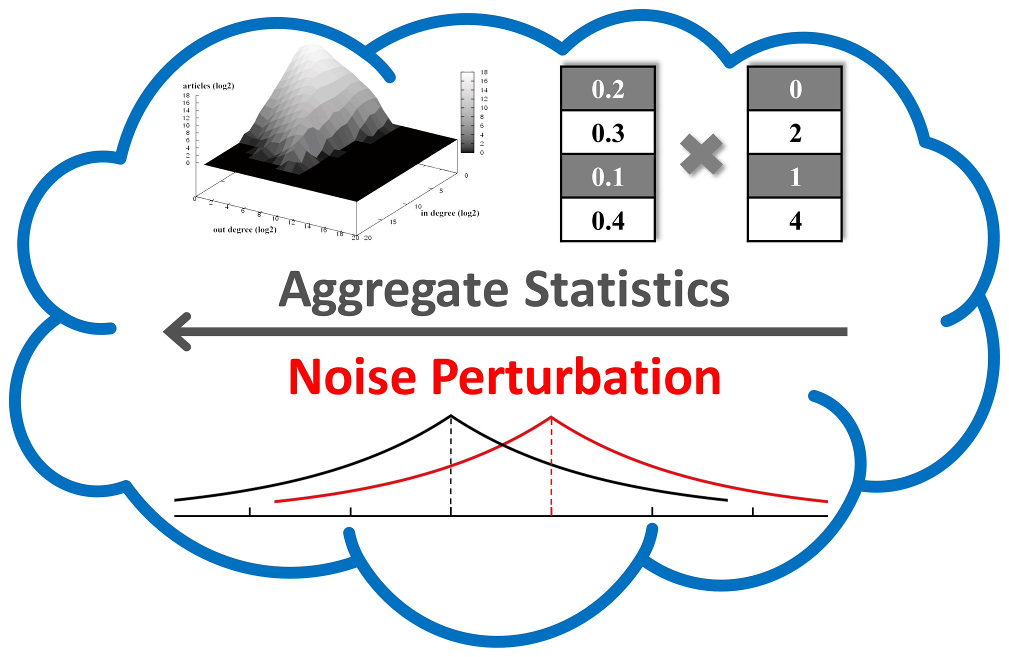
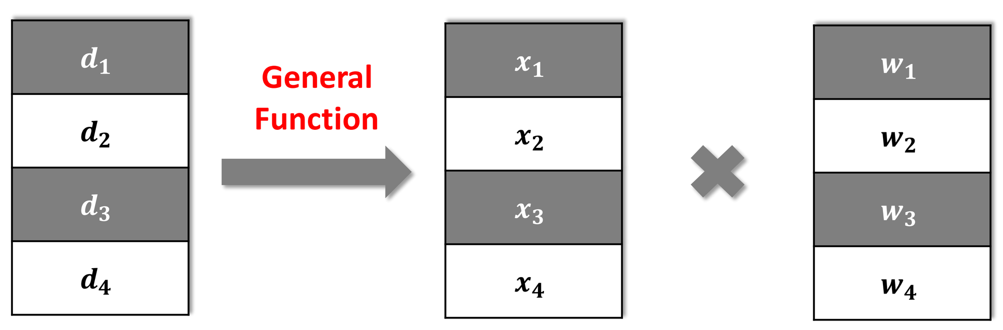


Fig. 3. Service Requests from Data Consumers.



Originates from Practical Computation over Encrypted Data Using Homomorphic Encryption (Popa et al. 2011, Shi et al. 2011).

Fig. 4. The Elementary Dot Product Operation Underlying Common Aggregate Statistics.

Service Pricing

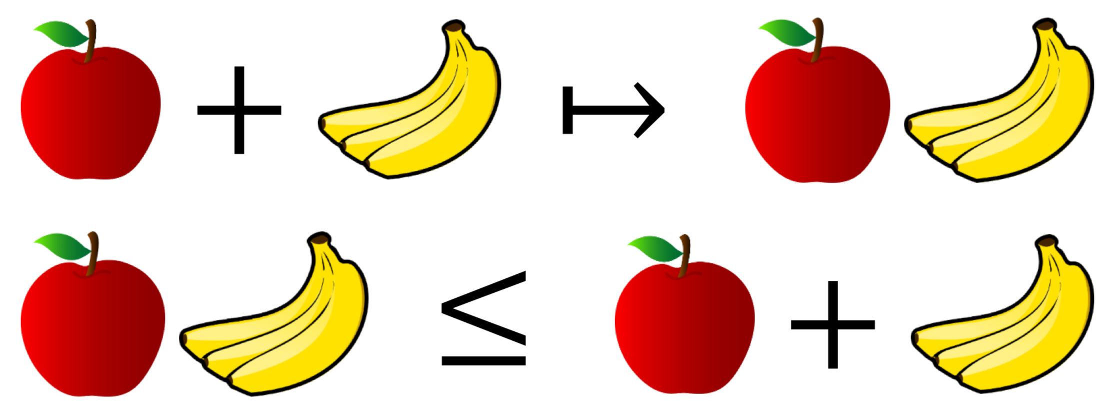


Fig. 5. Service Determinacy and Arbitrage Freeness.

Incorporating Variance of Noise

Motivating Example. Get $(\mathbf{w}, v)$ with a lower price?

$$\{(\mathbf{w}, v_1)\ldots(\mathbf{w}, v_m)\} \mapsto \left(m\mathbf{w}, \sum_{j=1}^{m} v_j\right) \mapsto \left(\mathbf{w}, \frac{1}{m^2}\sum_{j=1}^{m} v_j\right).$$

Theorem 1. For any arbitrage-free pricing function $\pi(\mathbf{w}, v)$, with two independent parts: the weight vector $\mathbf{w}$ and the variance of noise v, it **cannot decrease faster than** $1/v$.

Incorporating Weight Vector

Theorem 2 (Basic Arbitrage-free Pricing Functions). Let $\pi(\mathbf{w}, v) = g(\mathbf{w})^2/v$ be the pricing function for some positive function $g(\mathbf{w})$ that only depends on $\mathbf{w}$. Then, $\pi(\mathbf{w}, v)$ is **arbitrage free iff** $g(\mathbf{w})$ is a **semi-norm**.

Theorem 3 (Composite Arbitrage-free Pricing Functions). Let $\Gamma : \mathbb{R}^\phi \to \mathbb{R}$ be a **nondecreasing** and **subadditive** function. For any set of arbitrage-free pricing functions $\{\pi_1(S), \ldots, \pi_\phi(S)\}$, the composite pricing function $\pi(S) = \Gamma(\pi_1(S), \ldots, \pi_\phi(S))$ is also arbitrage free.

Privacy Compensation

Definition 1 (Individual Privacy Loss). The privacy loss of the data owner i is defined as:

$$\epsilon_i(\mathcal{M}) = \sup_{\mathbf{x}, O} \left| \log \frac{P\left(\mathcal{M}\left(\mathbf{x}\left(L, R\right)\right) = O\right)}{P\left(\mathcal{M}\left(\mathbf{x}^{(i)}\left(L, R\right)\right) = O\right)} \right|.$$

Here, $\mathbf{x}(L, R)$ and $\mathbf{x}^{(i)}(L, R)$ **initially** differ in x_i.

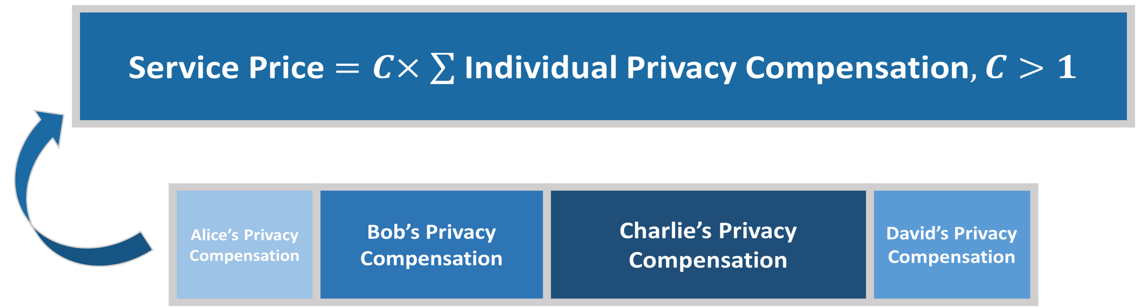


Fig. 6. Bottom-up Design of Privacy Compensation.

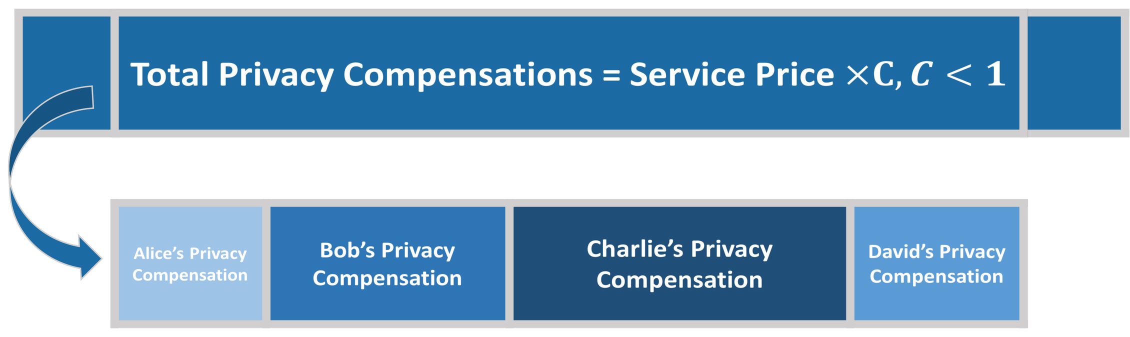


Fig. 7. Top-down Design of Privacy Compensation.

Evaluation Results

- **MovieLens** 1M dataset: **1,000,209 ratings** of approximately **3900 movies** made by **6040 users**; **Displayed ratings** as **target variables** in supervised learning.
- 2009 **Residential Energy Consumption Survey** (RECS) dataset: Released by U.S. **EIA** in Jan. 2013; Diverse energy usages in **12,083 U.S. homes**.
- Two social network datasets from **Stanford Network Analysis Platform** (SNAP): ego-Twitter: **81,306 nodes** and **1,768,149 edges** from **Twitter**; ego-Google+: **107,614 nodes** and **13,673,453 edges** from **Google+**.

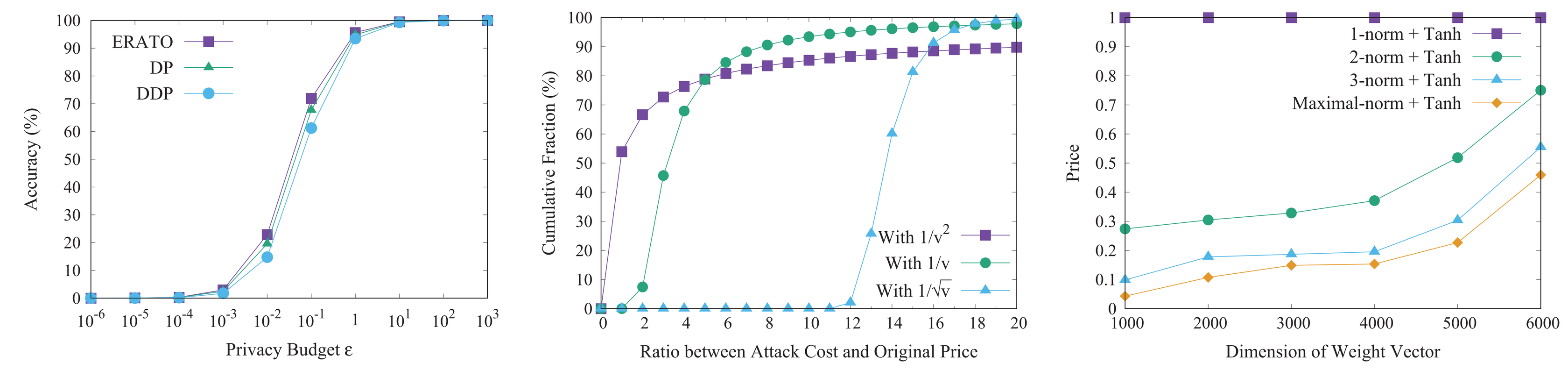


Fig. 8. Privacy vs. Utility and Arbitrage Freeness in Weighted Sum.

Fig. 8 reveals that ERATO can balance privacy and utility better than differential privacy (DP) and dependent differential privacy (DDP), and avoid arbitrage in service pricing.

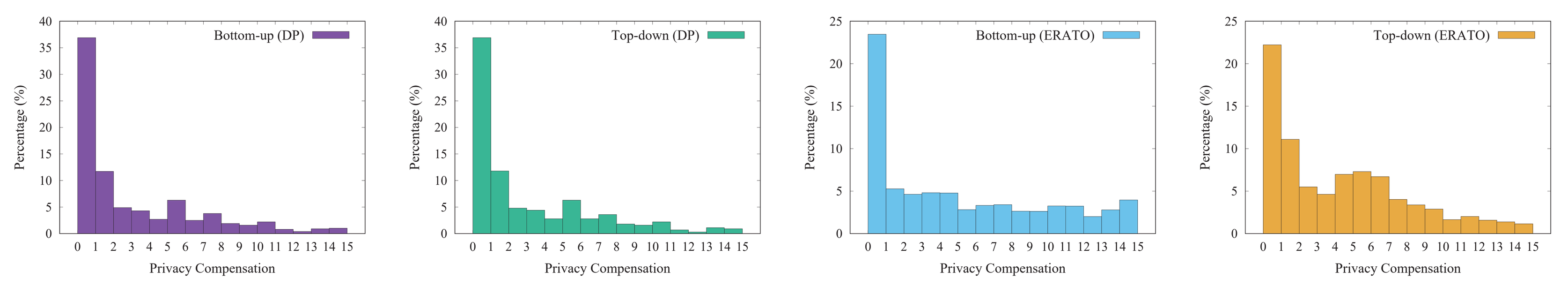


Fig. 9. DP and ERATO based Privacy Compensations in Weighted Sum.

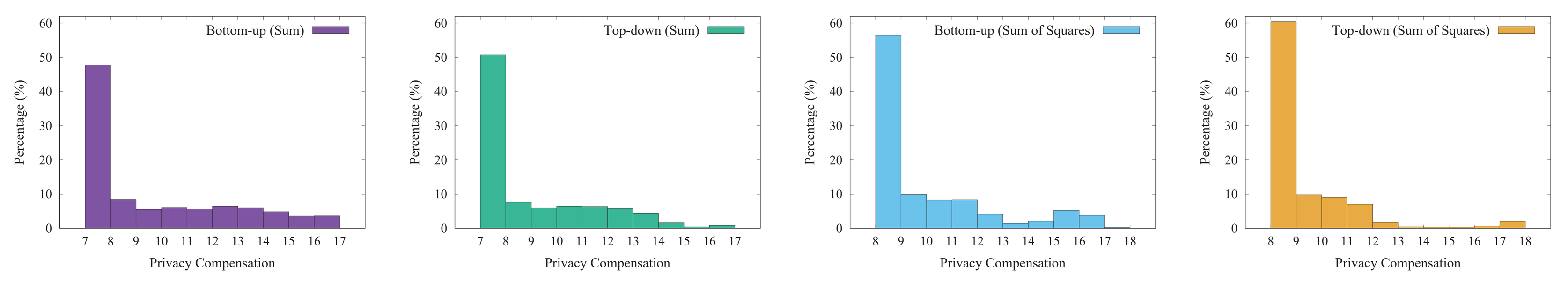


Fig. 10. ERATO based Privacy Compensation in Gaussian Distribution Fitting.

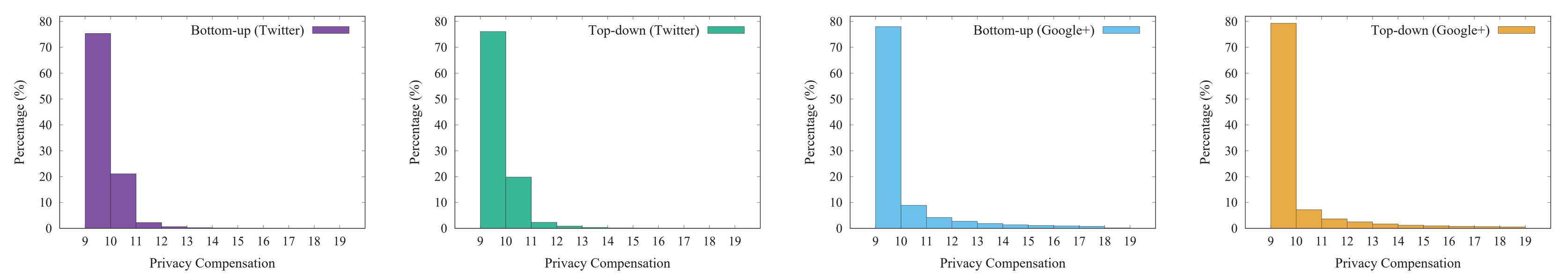


Fig. 11. ERATO based Privacy Compensation in Twitter/Google+ Degree Distribution.

These figures reveal that in the context of common aggregate statistics, the bottom-up and top-down designs of privacy compensation in ERATO can compensate the data owners for their privacy losses more fairly than DP based approaches.

Conclusions

- Have considered how to trade noisy aggregate statistics over private correlated data from the perspective of a data broker in data markets, and thus proposed ERATO.
- Have applied ERATO to three different aggregate statistics, and extensively evaluate their performances on four real-world datasets.
- Evaluation results have demonstrated the feasibility of ERATO, from privacy and utility guarantees, arbitrage-free pricing functions, and fine-grained privacy compensations.

上海交通大学先进网络实验室
Advanced Network Laboratory