ERATO: Trading Noisy Aggregate Statistics over Private Correlated Data

Chaoyue Niu, *Student Member, IEEE,* Zhenzhe Zheng, *Student Member, IEEE,* Fan Wu, *Member, IEEE,* Shaojie Tang, *Member, IEEE,* Xiaofeng Gao, *Member, IEEE,* and Guihai Chen, *Senior Member, IEEE*

Abstract—With the commoditization of personal privacy, pricing private data has become an intriguing problem. In this paper, we study noisy aggregate statistics trading from the perspective of a data broker in data markets. We thus propose ERATO, which enables aggrEgate statistics pRicing over privATe cOrrelated data. On one hand, ERATO guarantees arbitrage freeness against cunning data consumers. On the other hand, ERATO compensates data owners for their privacy losses using both bottom-up and top-down designs. We further apply ERATO to three practical aggregate statistics, namely weighted sum, probability distribution fitting, and degree distribution, and extensively evaluate their performances on MovieLens dataset, 2009 RECS dataset, and two SNAP large social network datasets, respectively. Our analysis and evaluation results reveal that ERATO well balances utility and privacy, achieves arbitrage freeness, and compensates data owners more fairly than differential privacy based approaches.

Index Terms—Data trading, data privacy, data correlation

1 Introduction

In today's big data economy, a common practice for Internet giants, like Google, Facebook, and Twitter, is to provide free online services in exchange for private information [1]. Nevertheless, when data owners become more aware of the economic values of personal data and the potential consequences of privacy disclosure, they would have stronger motivations to receive monetary compensations in return [2]. In particular, a study by JPMorgan Chase found that each unique user is worth roughly $4 to Facebook and $24 to Google [3]. Furthermore, several startup companies, including Datacoup [4], CitizenMe [5], and CoverUS [6], have already paid data owners for access to their private data. In a nutshell, data privacy has become a commodity to be bought and sold in practice.

To facilitate private data circulation, many open information platforms have emerged to bridge the gap between data owners and data consumers. For example, according to an FTC's survey on the nine typical data markets [7], Acxiom, which is the largest data broker, collects personal data from about 700 million users worldwide, and then sells aggregate statistics to top companies, such as Microsoft, Oracle, AT&T, etc. However, as further investigated by CBS News [8], such a multibillion-dollar industry has raised great attention together with serious doubt. One critical concern is that the data brokers make huge profits from private information, whereas they do not properly compensate data owners for their privacy losses. This criticism prompts the intermediate data brokers to devise a feasible privacy compensation mechanism for the data owners. In addition, the pricing strategy for the data consumers, which initially neither respects privacy nor provides economic guarantee [9], also requires new design.

To design a pricing framework for practical data markets trading aggregate statistics over private data, there are three major challenges. The first and the thorniest challenge is to rigorously quantify privacy loss. Markets for sensitive personal data significantly differ from those for ordinary information goods in privacy compensation. To compensate each data owner properly, it is necessary to quantify her privacy loss during the usage of her data. In the context of aggregate statistics, differential privacy [10], [11] has a natural utility-theoretic interpretation, which makes it a compelling measure to quantify individual privacy loss [12]. However, if the ubiquitous data correlations are further taken into account, there are two striking differences: (1) Due to data correlations, data owners, who are not involved in an aggregate statistic, may still suffer privacy losses. For example, if Alice is not but one of her friends is involved in the counting statistic about how many people have infected a contagious disease, Alice's status can still be leaked to an attacker who knows her social network [13], [14]. (2) Data owners with different sets of correlated data owners, or even the same set but with different correlation coefficients, can have distinct privacy losses. For example, in degree distribution, if Bob's degree is larger than Charlie's, which implies that Bob has more social connections, Bob thus can suffer a higher risk of privacy leakage [15]. If differential privacy is adopted for privacy loss quantification in two cases, the privacy loss of Alice is zero, and the privacy losses of Bob and Charlie are the same, which are both unreasonable in practice.

Yet, another challenge comes from the rich and complex formulas of common aggregate statistics. The data consumers in data markets are normally permitted to purchase multiple statistics. As a consequence, a critical concern is that they may circumvent the advertised price of a statistic through buying a bundle of cheaper ones. This economic practice is called arbitrage, while desirable pricings should be arbitrage free. Besides, the key issue in investigating arbitrage freeness is to determine whether a certain statistic can be derived from others. Such a concept of the determinacy relation has been well studied in queries/views answering from the database community [1], [16], [17]. Nevertheless, aggregate statistics tend to take different and even more complicated forms, such as linear polynomial in weighted sum [18], quadratic polynomial in Gaussian

- *C. Niu, Z. Zheng, F. Wu, X. Gao, and G. Chen are with the Shanghai Key Laboratory of Scalable Computing and Systems, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. E-mail: {rvince, zhengzhenzhe}@sjtu.edu.cn; {fwu, gao-xf, gchen}@cs.sjtu.edu.cn.*
- *S. Tang is with the Department of Information Systems, University of Texas at Dallas, Richardson, TX 75080. E-mail: tangshaojie@gmail.com.*
- *F. Wu is the corresponding author.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2019.2934100, IEEE Transactions on Knowledge and Data Engineering

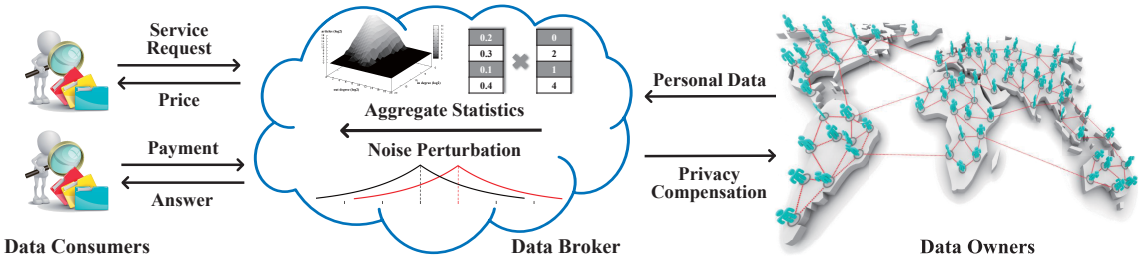


Fig. 1. A general system model for aggregate statistics based data markets.

distribution [19], [20], and nonlinear comparison in degree distribution [21]. Hence, it is highly nontrivial to design universal pricing functions for diverse aggregate statistics.

Last but not least challenge is to avoid the arbitrage opportunities in varying degrees of perturbation. For the sake of privacy issues, *e.g.*, the successive Facebook data scandals [22], [23], it is necessary for the data broker to sell noisy answers of aggregate statistics. Besides, to allow different prices for the same statistic but with diverse accuracies, the data consumer can specify her customized noise level, *e.g.*, the variance of noise used in [24]. In particular, if more noise is added to the true answer, the price should be lower. However, this setting makes reasoning about arbitrage freeness even harder. For example, a hidden arbitrage attack is that a clever data consumer is interested in an aggregate statistic with low variance of noise, while she is reluctant to pay its full price. She may instead turn to buying the same statistic multiple times but with diverse high variances. She can reduce the variance by averaging the returned answers. Therefore, economically-robust data markets have to rule out such arbitrage opportunities.

In this paper, by jointly considering above three challenges, we propose ERATO, which is an aggrEgate statistics pRicing framework over privATe cOrrelated data. ERATO consists of a service pricing mechanism (Section 3) and a privacy compensation mechanism (Section 4). For service pricing, ERATO first models common aggregate statistics as a set of dot product operations, where the dot product is between a weight vector and a data vector. ERATO then ensures arbitrage freeness with respect to both the variance of noise and the weight vector. On one hand, by combating the arbitrage attack as mentioned above, ERATO finds that arbitrage-free pricing functions cannot decrease faster than linearly with the variance of noise. On the other hand, ERATO establishes the equivalency between basic arbitrage-free pricing functions and semi-norms of the weight vector. Besides, ERATO constructs new composite pricing functions by means of subadditive and nondecreasing functions. In particular, activation functions from neural networks are introduced to allow high but finite prices for unperturbed answers. For balanced privacy compensation, ERATO offers both bottom-up and top-down designs. In the bottom-up design, the broker first needs to compensate each data owner for her privacy loss at bottom, and then to determine the price of a service request at top, by scaling up the total privacy compensation. Conversely, in the top-down design, the broker first determines the service price charged from the data consumer, and then spares some fraction of the payment for privacy compensation. Moreover, ERATO borrows key principles from dependent differential privacy to quantify individual privacy loss over correlated data, and further tightens its upper bound by distinguishing negative or positive weights and correlations. At last, ERATO extends the conventional fairness to a general dependent fairness, which clarifies the counterintuitive problem that a data owner, who is not involved in the service, can still receive privacy compensation, if at least one of her correlated data owners is involved.

We summarize our key contributions as follows.

- To the best of our knowledge, ERATO is the first pricing framework for trading aggregate statistics over private correlated data from the perspective of a data broker.
- ERATO features the properties of norms and activation functions to avoid arbitrage in pricings. Considering pervasive data correlations, ERATO quantifies privacy losses with dependent differential privacy, and compensates data owners in either a bottom-up or top-down manner.
- We instructively instantiate ERATO with three different kinds of aggregate statistics. Besides, we extensively evaluate their performances on four practical datasets (Section 5). Our analysis and evaluation results demonstrate that ERATO improves the utility of aggregate statistics, guarantees arbitrage freeness, and compensates data owners in a fairer way than the classical differential privacy based approaches. Specifically, when the privacy budget is 0.01 and the dimension of weight vector is 1000, ERATO improves 10.67% and 4.20% of accuracies than dependent differential privacy and differential privacy based approaches, respectively. Besides, when the pricing functions decrease quadratically with the variance of noise, there exist arbitrage opportunities with probability 53.91%. Moreover, compared with differential privacy based approaches, the number of data owners with no privacy compensation decreases by 17.7% for weighted sum; the data owners receive distinct privacy compensations rather than the same compensation for Gaussian distribution fitting and degree distribution.

2 Problem Formulation

In this section, we present system model and technical preliminaries for data markets providing aggregate statistics. For clarity, we list the frequently used notations in Table 1.

2.1 System Model

As shown in Fig. 1, we consider a general system model for data markets. The model has a data acquisition layer and a data trading layer. There are three major kinds of entities, including data owners, a data broker, and data consumers.

In the data acquisition layer, the data broker procures massive personal data, denoted by $\mathbf{d} = (d_1, \ldots, d_n)$, from n distinct data owners. Typical examples of personal data include product ratings, electrical usages, social media data, location data, and health records. Due to social, behavioral, and genetic interactions in practice [25], there exist correlations among the collected data items.

TABLE 1
FREQUENTLY USED NOTATIONS.

Notation	Remark
$\mathbf{d} = \{d_1, \ldots, d_n\}$	Original database contributed by n data owners
$S = (f, v)$	Data consumer's service request, including a concrete statistic and a tolerable variance of noise
ϵ_i	Data owner i's privacy loss due to S
$\psi_i(S)$	Data owner i's privacy compensation in S
$\pi(S)$	Price of S
$\mathbf{x}$	A data vector by preprocessing $\mathbf{d}$
$\mathbf{w}$	A weight vector over $\mathbf{x}$ to specify f
$\mathcal{M}$	A randomized mechanism
$\mapsto$	Service determinacy relation
L	Dependent size of $\mathbf{x}$
R	Probabilistic dependence relationship over $\mathbf{x}$
ϵ	A privacy budget
$\mathbb{C}_i$	Index set of x_i's correlated data items
$\rho_{ij} \in [0, 1]$	Dependence coefficient of x_j on x_i
DS_i^f	Dependent sensitivity of f at x_i
Δf_j	Sensitivity of f at x_j
$Lap(\lambda)$	Laplace distribution centered at 0 with scale λ
g	A semi-norm, *e.g.*, ℓ_p norm
Γ	A non-decreasing and subadditive function
$\mathbf{x}, \mathbf{x}^{(i)}$	A pair of dependent neighboring databases, initially differing in x_i
$\underline{\beta}_i, \bar{\beta}_i$	Infimum and supremum of x_i's domain
ξ_i	A contract function between data broker and data owner i with respect to privacy compensation
B	Total privacy compensation
$\sigma_{ij} = -1$ or 1	x_i, x_j are negatively or positively correlated

In the data trading layer, we consider that the data broker tends to trade aggregate statistics, *e.g.*, histogram count, weighted sum, mean, standard deviation, and probability distribution fitting, rather than directly offering sensitive raw data to the data consumers. Besides, each data consumer can request her customized service $S = (f, v)$, where f is a concrete statistic, and v denotes a tolerable variance of noise added to the true answer. We note that the self-defined variance of noise allows the data consumer to adjust the statistic's accuracy with a certain confidence based on Chebyshev's inequality. Formally, we let $\bar{O}$ denote the true answer, and let O denote the returned answer, then we have $P(|O - \bar{O}| \leq t\sqrt{v}) \geq 1 - \frac{1}{t^2}$, *i.e.*, the returned answer has at least $1 - \frac{1}{t^2}$ probability to be no more than $t\sqrt{v}$ away from the true answer.

Depending on the service $S = (f, v)$, on one hand, the data broker charges the data consumer with the price $\pi(S)$; on the other hand, the data broker compensates the data owner i with $\psi_i(S)$ for her privacy leakage ϵ_i. Specifically, if the variance of perturbing noise v is higher, the returned answer is less accurate, the price $\pi(S)$ should be lower, the privacy loss ϵ_i is smaller, and thus the privacy compensation $\psi_i(S)$ would be lower. Furthermore, a pricing framework is balanced if the utility of the data broker is no less than zero, *i.e.*, the price is sufficient to cover all the privacy compensations, namely $\pi(S) \geq \sum_{i=1}^{n} \psi_i(S)$.

2.2 Technical Preliminaries

In this section, we introduce the underlying mathematical operation of common aggregate statistics and the fundamental economic property of the pricing framework, namely dot product and arbitrage freeness, respectively. Besides, we briefly review dependent differential privacy.

Dot Product: We first identify the elementary mathematical operation underlying common aggregate statistics. Without loss of generality, we consider the following three practical aggregate statistics in detail.

Example 1. *A commercial company wants to capture the popularity of its product among customers. Besides, it assigns a weight w_i to each customer's rating d_i. The final score takes the form of a weighted sum $\sum_{i=1}^{n} w_i d_i$ [18].*

Example 2. *A researcher would like to learn the Gaussian distribution over U.S. residential energy consumptions. The key parameters are mean and variance. It suffices to compute the sum $\sum_{i=1}^{n} d_i$ and the sum of squares $\sum_{i=1}^{n} {d_i}^2$ [19], [20].*

Example 3. *A traffic analyst intends to count the drivers exceeding a certain speed limit δ. She needs to compare d_i with δ, and then do summation $\sum_{i=1}^{n} 1\{d_i \leq \delta\}$ [26].*

Given the above three application scenarios, we model common aggregate statistics as a set of dot product operations. In particular, the dot product operation is conducted between a weight vector $\mathbf{w}$ and a data vector $\mathbf{x}$, namely $\mathbf{w}^T\mathbf{x} = \sum_{i=1}^{n} w_i x_i$. Here, x_i represents any general function of the original data d_i, *e.g.*, quadratic polynomial in Example 2 and nonlinear comparison in Example 3. Besides, the weight w_i, set by the data consumer, indicates her preference/importance over x_i. Moreover, the purpose of introducing an interfaced database $\mathbf{x}$ by preprocessing the original database $\mathbf{d}$ is to simplify and unify statistic models. This concept originates from practical aggregate statistics over encrypted data, where homomorphic encryption can be applied over a general function of the original data, mainly to reduce time-consuming homomorphic multiplications [19], [20], [26]. Furthermore, the preprocessing from $\mathbf{d}$ to $\mathbf{x}$ can be viewed as a sort of feature mapping in learning theory [27]. In addition to manual engineering utilized by the above three examples, the feature mapping can also be realized by kernel tricks, deep learning, and so on. In the following context of a clear service type, for brevity, we use the weight vector $\mathbf{w}$ to specify the data consumer's requested statistic f, *i.e.*, $S = (\mathbf{w}, v)$.

Arbitrage Freeness: We next introduce a fundamental and desirable property of pricing functions, namely arbitrage freeness. Before investigating arbitrage freeness, we first establish the key concept of service determinacy. A similar concept has been studied in randomized query/view answering from the database community [1]. Under our data market model, the noisy answers can still be regarded as random variables. In particular, given a service request $S = (\mathbf{w}, v)$ over the database $\mathbf{x}$, the data broker answers using a randomized mechanism $\mathcal{M}$, and returns the result $\mathcal{M}(\mathbf{x})$, where its expectation is $\mathbf{w}^T\mathbf{x}$, and its variance is no more than v. We give the formal definition of service determinacy as follows.

Definition 1. *The service determinacy relation is between a service $S = (\mathbf{w}, v)$ and a multiset of services $\mathbf{Q} = \{S_1, \ldots, S_m\}$. We say that $\mathbf{Q}$ determines S, denoted as $\mathbf{Q} \mapsto S$, if the following rules are satisfied:*

- *Summation:*

$$\{(\mathbf{w}_1, v_1), \ldots, (\mathbf{w}_m, v_m)\} \mapsto \left(\sum_{j=1}^{m} \mathbf{w}_j, \sum_{j=1}^{m} v_j\right).$$

- *Scalar Multiplication:* $\forall c \in \mathbb{R}, (\mathbf{w}, v) \mapsto (c\mathbf{w}, c^2 v)$.
- *Relaxation:* $\forall v \geq v', (\mathbf{w}, v') \mapsto (\mathbf{w}, v)$.
- *Transitivity:*

If $\mathbf{Q}_1 \mapsto S_1, \ldots, \mathbf{Q}_m \mapsto S_m$ *and* $\{S_1, \ldots, S_m\} \mapsto S$, *then* $\bigcup_{j=1}^{m} \mathbf{Q}_j \mapsto S$.

We now explain the key intuitions behind Definition 1. (1) The rules of summation and scalar multiplication inherit basic mathematical operations over random variables. For example, a data consumer requests two services $S_1 = (\mathbf{w}_1, v_1), S_2 = (\mathbf{w}_2, v_2)$, and obtains noisy answers O_1, O_2. Here, O_1 (resp., O_2) can be viewed as a random variable with mean ${\mathbf{w}_1}^T\mathbf{x}$ (resp., ${\mathbf{w}_2}^T\mathbf{x}$) and variance v_1 (resp., v_2). Besides, if the data consumer adds O_1 to O_2, she can obtain another random variable O_3 with mean $(\mathbf{w}_1 + \mathbf{w}_2)^T\mathbf{x}$ and variance $v_1 + v_2$, which is in fact the answer of another service $S_3 = (\mathbf{w}_1 + \mathbf{w}_2, v_1 + v_2)$. Moreover, if the data consumer multiplies O_1 by $1/2$, she can obtain a random variable O_4 with mean ${\mathbf{w}_1}^T\mathbf{x}/2$ and variance $v_1/4$, which is the answer of the service $S_4 = (\mathbf{w_1}/2, v_1/4)$. Therefore, S_1, S_2 can determine S_3, and S_1 can determine S_4. Here, "determine" is kind of "derive". (2) We clarify the relaxation rule from expected accuracy. When answering the same statistic, if less noise is added to the true answer, the returned answer will be more accurate in expectation. Here, "determine" is kind of "more accurate than". (3) Transitivity is an important rule of both partial order relations and equivalence relations [28], and has been widely used in defining the determinacy relation among database queries [29], [30].

Based on the service determinacy relation, we define arbitrage freeness in a formal way.

Definition 2 (Arbitrage Freeness)**.** *A pricing function $\pi(\cdot)$ is arbitrage free, if $\forall m \geq 1, \{S_1, \ldots, S_m\} \mapsto S$ implies:*

$$\pi(S) \leq \sum_{j=1}^{m} \pi(S_j). \tag{1}$$

The intuition behind the above definition is that if there exists arbitrage in the pricing function $\pi(\cdot)$, *e.g.*, $\pi(S) > \sum_{j=1}^{m} \pi(S_j)$, then the data consumer would never pay the full price of the service S. Instead, she would turn to buying a cheaper set of services $\{S_1, \ldots, S_m\}$ to answer S.

Dependent Differential Privacy: We now introduce dependent differential privacy [31] from the privacy preservation perspective, *i.e.*, we focus on the randomized mechanism $\mathcal{M}$ itself. Yet, some of its disciplines will be used to mathematically quantify the privacy losses of data owners.

Dependent differential privacy is essentially a variant of the celebrated differential privacy [10]. In particular, differential privacy imposes a bound on the maximum ratio between the probabilities of returning a certain aggregate result with and without any individual's record, and thus limits the adversary's ability to infer private information. As an enhanced version, dependent differential privacy further considers data correlations. We introduce its technical notations as follows.

Given the statistical database $\mathbf{x} = (x_1, \ldots, x_n)$, if any data item in $\mathbf{x}$ is dependent on at most $L-1$ other items, the dependent size of $\mathbf{x}$ is defined to be L. Besides, the probabilistic dependence relationship over the whole database $\mathbf{x}$ is denoted as R. For example, to capture social, temporal, and spatial correlations, R can be some probabilistic graphical models, such as Bayesian networks and Markov chains. In addition, the existence of R may be due to a certain data generation process, or some other social, behavioral, and genetic relationships. For example, as illustrated in [31], R in the Gowalla location dataset was introduced from its relevant social network dataset [25]. Moreover, a pair of dependent neighboring databases is defined as follows.

Definition 3 (Dependent Neighboring Databases)**.** *Two databases $\mathbf{x}(L, R), \mathbf{x}'(L, R)$ are dependent neighboring databases, if the modification of one data item in $\mathbf{x}(L, R)$ (e.g., x_i changes to x_i') causes changes in at most $L-1$ other data items in $\mathbf{x}'(L, R)$ due to the probabilistic dependence relationship R.*

For the sake of brevity, when the dependent/correlated context is clear, we omit the parameters L, R, and write $\mathbf{x}, \mathbf{x}'$ instead. Based on dependent neighboring databases, the definition of dependent differential privacy is formalized as:

Definition 4 (ϵ-Dependent Differential Privacy)**.** *A randomized algorithm $\mathcal{M}$ provides ϵ-dependent differential privacy, if for any pair of dependent neighboring databases $\mathbf{x}$ and $\mathbf{x}'$ and any possible output O, we have:*

$$\exp(-\epsilon) \leq \max_{\mathbf{x}, \mathbf{x}'} \frac{P(\mathcal{M}(\mathbf{x}) = O)}{P(\mathcal{M}(\mathbf{x}') = O)} \leq \exp(\epsilon), \tag{2}$$

where ϵ is the privacy budget. Smaller ϵ provides better privacy and worse utility guarantees.

To achieve ϵ-dependent differential privacy, a matching dependent perturbation mechanism was proposed in [31]. The key idea is to carefully add Laplace noise by introducing fine-grained dependence coefficients between data items. In particular, ρ_{ij} denotes the dependence relationship between x_i and x_j, which quantifies the dependence of x_j on the modification of x_i. With the help of ρ_{ij}'s, the dependent sensitivity of a numeric function f over the database $\mathbf{x}$ caused by the modification of x_i can be expressed as:

$$DS_i^f = \sum_{j \in \mathbb{C}_i} \rho_{ij} \Delta f_j, \tag{3}$$

where $\mathbb{C}_i$ denotes the index set of the data items that are correlated with x_i. Besides, $\mathbb{C}_i$ contains i itself, and the dependence coefficient $\rho_{ii} = 1$. Moreover, Δf_j denotes the sensitivity of f with respect to the modification of x_j itself, *i.e.*, $\Delta f_j = \max_{x_{j_1}, x_{j_2}} \| f(\ldots, x_{j_1}, \ldots) - f(\ldots, x_{j_2}, \ldots) \|_1$. Furthermore, when focusing on the individual data item x_i, the dependent size L and the probabilistic dependence relationship R, which outline the dependent structure of the whole database $\mathbf{x}$ as mentioned earlier, are now reflected in two concrete parameters $\mathbb{C}_i$ and ρ_{ij}. Specifically, the cardinality of $\mathbb{C}_i$ is no more than L, while ρ_{ij} measures the dependence relationship between two data items, and can be computed from R. We finally present the dependent perturbation mechanism as follows.

Theorem 1 (Dependent Perturbation Mechanism)**.** *The randomized mechanism*

$$\mathcal{M}(\mathbf{x}) = f(\mathbf{x}) + Lap\left(\frac{\max_i DS_i^f}{\epsilon}\right) \tag{4}$$

guarantees ϵ-dependent differential privacy.

3 Service Pricing

In this section, we consider the first component of ERA-TO, namely the pricing mechanism for common aggregate statistics. It should be arbitrage free not only to the statistic $\mathbf{w}$ itself but also to the variance of perturbing noise v.

3.1 Design Rationale

Given service determinacy and arbitrage freeness in Definition 1 and Definition 2, respectively, we first list some intuitive properties that any arbitrage-free pricing function $\pi(\mathbf{w}, v)$ should satisfy: (1) The service with zero weight vector is free: $\pi(\mathbf{0}, v) = 0$; (2) The service with higher

variance is cheaper: $\forall v \geq v', \pi(\mathbf{w}, v) \leq \pi(\mathbf{w}, v')$; (3) The service with zero variance is the most expensive: $\forall v > 0, \pi(\mathbf{w}, 0) > \pi(\mathbf{w}, v)$; (4) The service with infinite noise is free: if $\pi(\cdot)$ is continuous at $\mathbf{w} = \mathbf{0}$, then $\pi(\mathbf{w}, +\infty) = 0$.

Proof. For (1), by the summation rule of Definition 1, when $m = 0$, $\emptyset \mapsto (\mathbf{0}, 0)$, and further by the relaxation rule, $(\mathbf{0}, 0) \mapsto (\mathbf{0}, v)$. Thus, by Definition 2, $0 \leq \pi(\mathbf{0}, v) \leq \pi(\mathbf{0}, 0) \leq 0$, which implies $\pi(\mathbf{0}, v) = 0$. For (2), first by the relaxation rule, $(\mathbf{w}, v') \mapsto (\mathbf{w}, v)$, and then by Definition 2, $\pi(\mathbf{w}, v) \leq \pi(\mathbf{w}, v')$. For (3), it directly follows from (2). For (4), by the scalar multiplication rule, $(1/c \cdot \mathbf{w}, 1) \mapsto (\mathbf{w}, c^2)$, then if c is towards positive infinity, we have: $\pi(\mathbf{w}, +\infty) = \lim_{c \to +\infty} \pi(\mathbf{w}, c^2) \leq \lim_{c \to +\infty} \pi(1/c \cdot \mathbf{w}, 1) = \pi(\mathbf{0}, 1) = 0$. Here, the first inequality follows from Definition 2, and the last equality follows from (1). □

We next discuss the existence of arbitrage-free pricing functions. First, we give a trivial example of zero-price function, *i.e.*, $\forall \pi(\mathbf{w}, v) = 0$. This function is arbitrage free. Second, we give a nontrivial example of widely used constant-price function, *i.e.*, $\forall \pi(\mathbf{w}, v) = c$ for some $c > 0$. There exists arbitrage in this function. A simple counter example is $\pi(\mathbf{0}, v) = 0$. Third, the general construction of non-trivial arbitrage-free pricing functions has been proven to be a hard problem [1]. Therefore, we turn to exploring sufficient conditions for arbitrage-free pricing functions.

We further divide an arbitrage-free pricing function into two parts, namely the variance of noise v and the weight vector $\mathbf{w}$, and conquer each part step by step. On one hand, from the above properties (2) and (4), we know that any nontrivial, continuous, and arbitrage-free pricing function should monotonically decrease with v, but the thorniest problem is how fast it can decrease with v. We determine the boundary function $1/v$ by thwarting the arbitrage attack as illustrated in Section 1. On the other hand, we associate service determinacy with norms of the weight vector $\mathbf{w}$, *e.g.*, ℓ_p norms. Besides, we establish the equivalency between arbitrage-free pricing functions and semi-norms. Moreover, we synthesize new pricing functions by applying subadditive and non-decreasing functions. In particular, to allow a high but finite price for the unperturbed answer, we utilize activation functions from neural networks.

3.2 Detailed Design

Following the guidelines given above, we now introduce the detailed design of arbitrage-free pricing functions.

3.2.1 Incorporating Variance of Noise

We start with the first part of an arbitrage-free pricing function $\pi(\mathbf{w}, v)$ involving the variance of noise v. We formulate the arbitrage attack in a formal way to figure out how $\pi(\mathbf{w}, v)$ can decrease with v:

Example 4. *A data consumer, who wants to obtain the service $(\mathbf{w}, v)$ but with a lower price, may turn to buying m other cheaper services of the same statistic but with higher variances, denoted as $\{(\mathbf{w}, v_j) | j \in \{1, \ldots, m\}, v_j > v\}$. The data consumer first applies summation and then scalar multiplication by $1/m$ in Definition 1, i.e., $\{(\mathbf{w}, v_1), \ldots, (\mathbf{w}, v_m)\} \mapsto (m\mathbf{w}, \sum_{j=1}^m v_j) \mapsto (\mathbf{w}, \frac{1}{m^2} \sum_{j=1}^m v_j)$. In other words, the data consumer computes the average of m answers, and gets an unbiased answer but with a lower variance. If the pricing function $\pi(\cdot)$ is arbitrage free, then the following conditional statement must hold:*

$$\frac{1}{m^2} \sum_{j=1}^m v_j \leq v \Rightarrow \sum_{j=1}^m \pi(\mathbf{w}, v_j) \geq \pi(\mathbf{w}, v). \tag{5}$$

We give Theorem 2 to thwart the above attack.

Theorem 2. *For any arbitrage-free pricing function $\pi(\mathbf{w}, v)$ that depends on two independent parts $\mathbf{w}$ and v, it cannot decrease faster than $1/v$.*

Proof. We first prove $1/v$ is the boundary function, *i.e.*, $\pi(\mathbf{w}, v) = g(\mathbf{w})/v$ is arbitrage free for some positive function $g(\mathbf{w})$ that depends only on $\mathbf{w}$. We utilize the antecedent in Equation (5) to show the correctness of its consequent:

$$\sum_{j=1}^m \pi(\mathbf{w}, v_j) = g(\mathbf{w}) \sum_{j=1}^m \frac{1}{v_j} \geq g(\mathbf{w}) \frac{m^2}{\sum_{j=1}^m v_j} \tag{6}$$

$$\geq g(\mathbf{w}) \frac{m^2}{m^2 v} = \frac{g(\mathbf{w})}{v} = \pi(\mathbf{w}, v). \tag{7}$$

Here, the inequality in Equation (6) follows from that the harmonic mean of a list of non-negative real numbers is less than or equal to the arithmetic mean of the same list, namely

$$\frac{m}{\sum_{j=1}^m \frac{1}{v_j}} \leq \frac{\sum_{j=1}^m v_j}{m} \Rightarrow \sum_{j=1}^m \frac{1}{v_j} \geq \frac{m^2}{\sum_{j=1}^m v_j}. \tag{8}$$

Besides, the inequality in Equation (7) follows from the antecedent in Equation (5). Furthermore, when these two inequalities simultaneously take the equal signs, we can obtain boundary variances $\{v_j = mv | j \in \{1, \ldots, m\}\}$, which implies that requesting the service with the same variance multiple times is the most possible way to obtain an arbitrage.

We next show that if $\pi(\mathbf{w}, v)$ decreases faster than $1/v$, we would derive an arbitrage. We consider a sequence of variances $\{v_j | j \in \{1, \ldots, +\infty\}\}$, such that $\lim_{j \to +\infty} v_j = +\infty$ and $\lim_{j \to +\infty} v_j \pi(\mathbf{w}, v_j) = 0$. Thus, we can find $j_0 > 1$ such that $v_{j_0} \pi(\mathbf{w}, v_{j_0}) < \pi(\mathbf{w}, 1)/2$. Now, we can answer the service $(\mathbf{w}, 1)$, through requesting $\lceil v_{j_0} \rceil$ times the same service $(\mathbf{w}, v_{j_0})$ and averaging their answers. However, for these $\lceil v_{j_0} \rceil$ services, we pay

$$\lceil v_{j_0} \rceil \pi(\mathbf{w}, v_{j_0}) < 2 v_{j_0} \pi(\mathbf{w}, v_{j_0}) < \pi(\mathbf{w}, 1), \tag{9}$$

which yields an arbitrage, and completes the proof. □

In what follows, for simplicity, we fix the part of $\pi(\mathbf{w}, v)$ related to the variance v at $1/v$ by default, while investigate other functions, *e.g.*, $1/\sqrt{v}$, in our evaluation part.

3.2.2 Incorporating Weight Vector

We continue to consider the other part of an arbitrage-free pricing function $\pi(\mathbf{w}, v)$, namely the weight vector $\mathbf{w}$.

By carefully studying the rules of the service determinacy in Definition 1, we find a metric in linear algebra with analogous properties, called norm, more precisely semi-norm. In particular, a norm of a vector $\mathbf{w}$ can be viewed as a measure of its "length". Formally speaking, a norm is any function $g : \mathbb{R}^n \to \mathbb{R}$ that satisfies the following properties:

- *Subadditivity:*

$$\forall \mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^n, g(\mathbf{w}_1 + \mathbf{w}_2) \leq g(\mathbf{w}_1) + g(\mathbf{w}_2).$$

- *Homogeneity:* $\forall c \in \mathbb{R}, \mathbf{w} \in \mathbb{R}^n, g(c\mathbf{w}) = |c| g(\mathbf{w})$.
- *Non-negativity:* $\forall \mathbf{w} \in \mathbb{R}^n, g(\mathbf{w}) \geq 0$.
- *Definiteness:* $\mathbf{w} = \mathbf{0} \Leftrightarrow g(\mathbf{w}) = 0$.

If the last property relaxes to $\mathbf{w} = \mathbf{0} \Rightarrow g(\mathbf{w}) = 0$, we call it semi-norm. Besides, the most commonly used norms in the machine learning and data mining algorithms are a family of ℓ_p norms for some real number $p \geq 1$. Furthermore,

considering that the trivial example of zero-price function is arbitrage free, we utilize semi-norms to devise our basic arbitrage-free pricing functions:

Theorem 3 (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.*

Proof. Due to space limitations, we put the proof into our technical report [32]. □

We next consider how to construct more arbitrage-free pricing functions by combining basic/existing ones. We resort to a general class of nondecreasing and subadditive functions. We recall that a function $\Gamma : \mathbb{R}^\phi \to \mathbb{R}$ over $\forall \mathbf{y}, \mathbf{z} \in \mathbb{R}^\phi$ is nondecreasing, if $\mathbf{y} \leq \mathbf{z}, \Gamma(\mathbf{y}) \leq \Gamma(\mathbf{z})$. Besides, it is subadditive, if $\Gamma(\mathbf{y} + \mathbf{z}) \leq \Gamma(\mathbf{y}) + \Gamma(\mathbf{z})$.

Theorem 4 (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.*

Proof. We consider the general form of service determinacy, *i.e.*, $\{S_1, \ldots, S_m\} \mapsto S$. Since $\pi_1, \ldots, \pi_\phi$ are arbitrage free, we have:

$$\forall k \in \{1, \ldots, \phi\}, \pi_k(S) \leq \sum_{j=1}^{m} \pi_k(S_j). \tag{10}$$

Besides, due to the nondecreasing and subadditive properties of the function Γ, we further have:

$$\begin{aligned}\Gamma(\pi_1(S), \ldots, \pi_\phi(S)) &\leq \Gamma\left(\sum_{j=1}^{m} \pi_1(S_j), \ldots, \sum_{j=1}^{m} \pi_\phi(S_j)\right) \\ &\leq \sum_{j=1}^{m} \Gamma(\pi_1(S_j), \ldots, \pi_\phi(S_j)). \end{aligned} \tag{11}$$

This completes the proof. □

We give some typical examples of composite arbitrage-free pricing functions as follows. If $\pi_1(S), \ldots, \pi_\phi(S)$ are arbitrage free, then

- *Linear Combination:* $\forall c_1, \ldots, c_\phi \geq 0, \sum_{k=1}^{\phi} c_k \pi_k(S)$;
- *Geometric Mean:* $\sqrt{\prod_{k=1}^{\phi} \pi_k(S)}$;
- *Maximum:* $\max(\pi_1(S), \ldots, \pi_\phi(S))$;
- *Power:* $\pi_1(S)^c$ for $0 < c \leq 1$;
- *Logarithmic:* $\log(\pi_1(S) + 1)$;
- *Cut-off:* $\min(\pi_1(S), c)$ for $c \geq 0$;
- *Sigmoid:* $\tanh(\pi_1(S))$, $\arctan(\pi_1(S))$, $\frac{\pi_1(S)}{\sqrt{\pi_1(S)^2+1}}$;

are arbitrage free as well. We note that the basic arbitrage-free pricing functions and the first five composite arbitrage-free pricing functions set an infinite price for the unperturbed answer, *i.e.*, the variance of noise $v = 0$. However, these functions may be impractical in data markets, since the data broker tends to sell unperturbed aggregate statistics for high but finite prices. Nevertheless, we can turn to applying some bounding functions for composition, *e.g.*, cut-off and sigmoid functions. In particular, sigmoid functions are commonly used as activation functions in neural networks [27]. At last, we give a sufficient condition to check whether a function Γ is nondecreasing and subadditive.

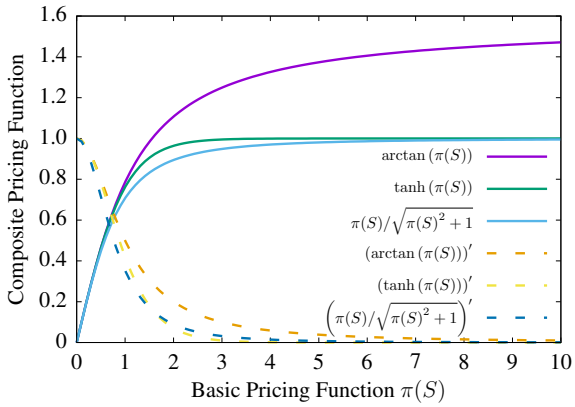


Fig. 2. Three typical sigmoid functions and their first derivatives.

Lemma 1. *Let $\Gamma : \mathbb{R}^\phi \to \mathbb{R}$ be a continuous and twice-differentiable function such that $\Gamma(\mathbf{0}) = 0$. Then, if each element of Γ's gradient (i.e., partial derivative) is no less than zero, Γ is nondecreasing; if each element of Γ's Hessian matrix (i.e., second partial derivative) is no greater than zero, Γ is subadditive.*

To provide an intuitive view of Lemma 1, we plot three typical sigmoid functions and their first derivatives in Fig.2. From Fig. 2, we can see that these functions are increasing and bounded above, while their first derivatives are decreasing. Therefore, according to Theorem 4, they are composite arbitrage-free pricing functions.

4 Privacy Compensation

In this section, we consider the other component of ERATO, *i.e.*, the privacy compensation mechanism for individual privacy loss. We propose both bottom-up and top-down designs. In the bottom-up design, the sum of privacy compensations determines the service price, while this relation is inverse in the top-down design. Besides, another major difference is that the bottom-up design allows each data owner to actively select a privacy compensation function according to her privacy strategy, which is instead not required in the top-down design.

4.1 Privacy Loss for General Function

When the data broker answers aggregate statistics with a randomized mechanism $\mathcal{M}$, some private information of each data owner would be leaked. Based on the disciplines of dependent differential privacy, we formally define the individual privacy loss ϵ_i for an arbitrary real-valued function f, and further give its upper bound related to the dependent sensitivity DS_i^f and the variance of noise v.

We first consider a pair of dependent neighboring databases $\mathbf{x}$ and $\mathbf{x}^{(i)}$, which initially differs in the data item x_i. In fact, $\mathbf{x}$ and $\mathbf{x}^{(i)}$ can simulate the presence and absence of the data owner i. By comparing the output of the randomized mechanism $\mathcal{M}$ over $\mathbf{x}$ and $\mathbf{x}^{(i)}$, we define individual privacy loss as follows.

Definition 5 (Individual Privacy Loss)**.** *The privacy loss of the data owner i in the randomized mechanism $\mathcal{M}$ over the database $\mathbf{x}$ is defined as:*

$$\epsilon_i(\mathcal{M}) = \sup_{\mathbf{x}, O} \left| \log \frac{P(\mathcal{M}(\mathbf{x}) = O)}{P(\mathcal{M}(\mathbf{x}^{(i)}) = O)} \right|, \tag{12}$$

where $\mathbf{x}$ *ranges over all possible database instances, and* O *ranges over all possible outputs.*

We discuss the relationship between the individual privacy loss $\epsilon_i(\mathcal{M})$ and the privacy budget ϵ in Definition 4. (1) The privacy budget ϵ, normally preset by the data broker over the randomized mechanism $\mathcal{M}$, applies to all the data owners. In other words, each data owner is promised to suffer privacy loss no more than ϵ, *i.e.*, $\epsilon = \max_i \epsilon_i(\mathcal{M})$. By comparison, in the context of privacy compensation, we turn this around. Now, the randomized mechanism $\mathcal{M}$, given the variance of perturbing noise v from the data consumer, has already breached privacy, and Definition 5 quantifies the privacy loss for each data owner. (2) From $\epsilon = \max_i \epsilon_i(\mathcal{M})$, we can see that if the randomized mechanism $\mathcal{M}$ is ϵ-dependent differentially private for a tiny ϵ, then the individual privacy loss $\epsilon_i(\mathcal{M})$ would be very small as well.

We further give an upper bound of the individual privacy loss $\epsilon_i(\mathcal{M})$, when the randomized mechanism $\mathcal{M}$ is known to be the dependent perturbation mechanism defined in Theorem 1. In particular, this upper bound depends on the variance of Laplace noise v and the dependent sensitivity of f at x_i.

Theorem 5. *Let* $\mathcal{M}$ *be dependent perturbation mechanism,* f *be any numeric function,* DS_i^f *be the dependent sensitivity of* f *at* x_i*, and* v *be the variance of Laplace noise. The privacy loss of the data owner* i *is bounded above by:*

$$\epsilon_i(\mathcal{M}) \leq \frac{DS_i^f}{\sqrt{v/2}}. \tag{13}$$

Proof. In the dependent perturbation mechanism, the noise η is drawn from the Laplace distribution $Lap(\lambda)$. Here, the scaling factor λ can be computed from the variance v, namely $\lambda = \sqrt{v/2}$. We then derive that:

$$\begin{aligned}
\epsilon_i(\mathcal{M}) &= \max_{O,\mathbf{x}} \left| \log \frac{P(\mathcal{M}(\mathbf{x}) = O)}{P(\mathcal{M}(\mathbf{x}^{(i)}) = O)} \right| \\
&= \max_{O,\mathbf{x}} \left| \log \frac{P\left(f\left(\mathbf{x}\right) + \eta = O\right)}{P\left(f\left(\mathbf{x}^{(i)}\right) + \eta = O\right)} \right| \\
&= \max_{O,\mathbf{x}} \left| \log \frac{\exp\left(-\left|O - f\left(\mathbf{x}\right)\right| / \lambda\right)}{\exp\left(-\left|O - f\left(\mathbf{x}^{(i)}\right)\right| / \lambda\right)} \right| \qquad (14) \\
&= \max_{O,\mathbf{x}} \left| \frac{\left|O - f\left(\mathbf{x}\right)\right| - \left|O - f\left(\mathbf{x}^{(i)}\right)\right|}{\lambda} \right| \\
&\leq \max_{\mathbf{x}} \frac{\left|f\left(\mathbf{x}\right) - f\left(\mathbf{x}^{(i)}\right)\right|}{\lambda} \leq \frac{DS_i^f}{\lambda} = \frac{DS_i^f}{\sqrt{v/2}}. \qquad (15)
\end{aligned}$$

Here, Equation (14) follows from the probability density function of $Lap(\lambda)$. Besides, in Equation (15), the first inequality follows from the triangle inequality, and the second inequality follows from the definition of the dependent sensitivity of f at x_i [31]. □

4.2 Bottom-up Design

In this section, we consider the bottom-up design of privacy compensation. The data broker first needs to satisfy each individual privacy compensation $\psi_i(S)$, and then determine the price $\pi(S)$ for the data consumer. For example, to guarantee the property of balance, the relationship between the service price and the sum of individual privacy compensation can be $\pi(S) = c\sum_{i=1}^{n} \psi_i(S)$ for some $c > 1$.

First, the individual privacy compensation $\psi_i(S)$ should hinge on the individual privacy loss $\epsilon_i(\mathcal{M})$. Besides, the data broker needs to evaluate/approximate $\epsilon_i(\mathcal{M})$ from the service S itself, including the weight vector $\mathbf{w}$ and the variance of noise v. However, the original $\epsilon_i(\mathcal{M})$ in Definition 5 not only depends on the actual randomized mechanism $\mathcal{M}$, but also needs to consider all the database instances and all the possible outputs, which can be infeasible to compute in practice [1], [33]. Therefore, we turn to focusing on the specific dependent perturbation mechanism in Theorem 1, and utilize the upper bound of privacy loss in Theorem 5 to do compensation. We note that the bounded privacy loss in Theorem 5 is given as a function of the variance v and the dependent sensitivity DS_i^f. Here, we can compute DS_i^f in the context of aggregate statistics. We let $\underline{\beta}_i, \bar{\beta}_i \in \mathbb{R}$ denote the infimum and supremum of the data item x_i's domain, respectively. Then, according to Equation (3), we can get:

$$DS_i^f = \sum_{j \in \mathbb{C}_i} \rho_{ij} \left|w_j\right| \left(\bar{\beta}_j - \underline{\beta}_j\right). \tag{16}$$

Suppose we ignore data correlations by setting $\rho_{ij} = 0$ for all $j \in \mathbb{C}_i \backslash i$. The dependent sensitivity in Equation (16) will degenerate to the sensitivity defined in the classical differential privacy [10]:

$$DS_i^f = \left|w_i\right| \left(\bar{\beta}_i - \underline{\beta}_i\right). \tag{17}$$

After quantifying the individual privacy loss in aggregate statistics, we now consider how to compensate each data owner in an appropriate manner. We first identify two desirable properties in the bottom-up design:

Definition 6 (Bottom-up Privacy Compensation)**.** *Let* $\psi_i(S)$ *be a privacy compensation function over the service* $S = (\mathbf{w}, v)$ *in the bottom-up design.* $\psi_i(S)$ *should satisfy:*

- *Dependent Fairness:* $\forall j \in \mathbb{C}_i, w_j = 0 \Rightarrow \psi_i(S) = 0$.
- *Micro Arbitrage Freeness:* $\psi_i(S)$ *is arbitrage free.*

We give some comments on these two properties as follows. (1) Dependent fairness is an extension of fairness defined in the conventional query-based pricing [33] by further incorporating data correlations. The original fairness says that the data owner whose data is not queried should not expect reward. In contrast, our dependent fairness says that only if the data owner and her correlated data owners are not involved in the service, she will receive no privacy compensation. Although the case, where a data owner who is not involved in the service but may still be compensated, seems counterintuitive, it makes sense from the perspective of privacy loss due to data correlations. (2) Micro arbitrage freeness is a necessity in the bottom-up design. The reason is that the service price at top hinges on the total privacy compensations at bottom. Therefore, the data consumer may have strong motivations to circumvent the due privacy compensations, and thus the payment, by asking other alternative services. Besides, the definition of micro arbitrage freeness is identical to that of arbitrage freeness, but the former needs to be verified over the whole data owners.

In a similar way to service pricing, we design basic bottom-up privacy compensation functions directly from the privacy losses, which set infinite compensations for unperturbed answers. This kind of functions are suitable for the data owner, who values her privacy highly, and would never accept full disclosure of personal data.[1]

1. By querying two consecutive unperturbed aggregate statistics with and without a data owner's data item, the data consumer can have full knowledge of the data owner's data item.

Theorem 6. *The privacy compensation functions*

$$\psi_i(S) = c_i \frac{DS_i^f}{\sqrt{v/2}} = c_i \frac{\sum_{j\in\mathbb{C}_i} \rho_{ij} |w_j| \left(\bar{\beta}_j - \underline{\beta}_j\right)}{\sqrt{v/2}} \tag{18}$$

for some constant $c_i > 0$ and for all $i \in \{1, \ldots, n\}$, are basic bottom-up privacy compensation functions.

Proof. First, we prove dependent fairness. We can check that $\forall j \in \mathbb{C}_i, w_j = 0 \Rightarrow \psi_i(S) = 0$. Second, we prove micro arbitrage freeness. We view $\psi_i(S)$ as a linear combination of $\{|w_j|/\sqrt{v/2} | j \in \mathbb{C}_i\}$, where the corresponding coefficients are $\{c_i\rho_{ij}(\bar{\beta}_j - \underline{\beta}_j) | j \in \mathbb{C}_i\}$. By Theorem 4 (Linear Combination), to prove the micro arbitrage freeness of $\psi_i(S)$, it suffices to prove that $|w_j|/\sqrt{v/2}$ is arbitrage free. By Theorem 4 (Geometric Mean), it further suffices to prove the arbitrage freeness of ${w_j}^2/v$. Now, by using the weighted ℓ_2 norm and setting those weights, whose indexes are not j, to be zeros, it completes the proof. □

Analogous to Theorem 4, we can construct new bottom-up privacy compensation functions from basic ones by applying any nondecreasing and subadditive function. In particular, to allow the data owner, who is less concerned about her privacy, to reveal her personal data at some high but finite price, we can make use of sigmoid functions.

Theorem 7. *The privacy compensation functions*

$$\psi_i(S) = b_i \tanh\left(c_i \frac{DS_i^f}{\sqrt{v/2}}\right) \tag{19}$$

for constants $b_i, c_i > 0$ and for all $i \in \{1, \ldots, n\}$, are bounded bottom-up privacy compensation functions.

Proof. First, we can check that $\forall j \in \mathbb{C}_i, w_j = 0 \Rightarrow \psi_i(S) = 0$. Second, in Theorem 6, we have proved the arbitrage freeness of $c_i DS_i^f/\sqrt{v/2}$. Then, by Theorem 4 (Sigmoid and Linear Combination), $\psi_i(S)$ is micro arbitrage free. □

Considering the diversity of individuals' privacy strategies, we demonstrate how the data broker can select customized privacy compensation functions for different kinds of data owners. We introduce a nondecreasing contract function $\xi_i(\epsilon_i)$ between the data broker and each data owner i, *i.e.*, in the event of privacy loss ϵ_i, i should be compensated with at least $\xi_i(\epsilon_i)$. In fact, the contract function $\xi_i(\cdot)$ depends on i's valuation over private information. For example, if i values her privacy highly, and would never accept full disclosure of her personal data, then she may choose a linear contract function $\xi_i(\epsilon_i) = c_i\epsilon_i$ for some $c_i > 0$. In contrast, another data owner j is less concerned about her privacy, and is willing to sell her private data at some high price. Then, she may select the bounded sigmoid contract function $\xi_j(\epsilon_j) = b_j \tanh(c_j\epsilon_j)$ for some $b_j, c_j > 0$. Additionally, the data broker would define the corresponding bottom-up privacy compensation functions $\psi_i(S)$ and $\psi_j(S)$ using Equation (18) and Equation (19) for the data owners i and j, respectively. In particular, the privacy compensation functions are satisfying for both i and j, since $\xi_i(\epsilon_i) \le \psi_i(S)$ and $\xi_j(\epsilon_j) \le \psi_j(S)$ due to Theorem 5 and the fact that the tanh function is nondecreasing.

At last, the data broker can determine the service price $\pi(S)$. Take $\pi(S) = c\sum_{i=1}^n \psi_i(S), c > 0$ for example. We note that if every privacy compensation function $\psi_i(S)$ is micro arbitrage free, then the pricing function $\pi(S)$, which can be viewed as a linear combination of $\psi_i(S)$'s, is factually arbitrage free. Of course, $\pi(S)$ can be any other composite functions under Theorem 4.

4.3 Top-down Design

In this section, we consider a different top-down privacy compensation design, where the data broker first determines the service price $\pi(S)$ using the pricing mechanism in Section 3, and then spares some fraction of the payment for privacy compensation, *i.e.*, $\sum_{i=1}^n \psi_i(S) = c\pi(S)$ for some $0 < c < 1$. If we regard $c\pi(S)$ as a budget B, we can convert the privacy compensation problem to a budget allocation problem, where each data owner i's share in B should be roughly proportional to her privacy loss $\epsilon_i(\mathcal{M})$.

Specific to the dot product operation in common aggregate statistics, we shall tighten the upper bound of the individual privacy loss $\epsilon_i(\mathcal{M})$, by computing the dependent sensitivity DS_i^f more accurately. We first give our motivating examples as follows.

Example 5. *A database $\mathbf{x}$ consists of two entries x_1, x_2, such that $x_2 = 0.5x_1$ and $x_1 \in [0, 1]$. Here, the dependence coefficient $\rho_{12} = 1$, since x_2 is completely dependent on x_1. We then consider two statistics $f = x_1 + x_2$ and $g = x_1 - x_2$, which differ in the sign of x_2's weight. By Equation (16), we compute the dependent sensitivities of f and g at x_1:*

$$DS_1^f = \Delta f_1 + \rho_{12}\Delta f_2 = 1 + 1 \times 0.5 = 1.5, \tag{20}$$
$$DS_1^g = \Delta g_1 + \rho_{12}\Delta g_2 = 1 + 1 \times 0.5 = 1.5, \tag{21}$$

respectively. We can see that the dependent sensitivities of f and g at x_1 are the same. However, g is essentially $g^* = 0.5x_1$, and its dependent sensitivity at x_1 should be:

$$DS_1^{g^*} = \Delta g_1^* = 0.5 < DS_1^g = 1.5. \tag{22}$$

Example 6. *We continue to consider the database $\mathbf{x}$, but now x_1 and x_2 are negatively correlated rather than positively correlated, i.e., $x_2 = -0.5x_1$. According to the definition of nonnegative dependence coefficients in [31], $\rho_{12} = 1$ remains unchanged. Thus, the dependent sensitivities of f and g at x_1 are still 1.5. However, f is essentially $f^* = 0.5x_1$, and thus its sensitivity at x_1 is 0.5, which is less than DS_1^f.*

Given the two examples above, we can observe that the definition and the mechanism of the dependent differential privacy proposed in [31] aim to be applicable for general functions and general positive/negative correlations, which implies that the general dependent sensitivity can be just a loose upper bound in the context of a specific function. Such a key observation enables us to tighten the dependent sensitivity and thus the individual privacy loss by considering two extra factors: whether the weight is negative or positive, and whether the correlation is negative or positive. In our following calculation, we will maintain the original forms of weights rather than utilizing their absolute values as in the dependent differential privacy, namely Equation (16). Additionally, we introduce $\sigma_{ij} = -1$ and $\sigma_{ij} = 1$ to represent the cases that x_i, x_j are negatively and positively correlated, respectively. We thus get:

Lemma 2. *The tight dependent sensitivity of $f = \mathbf{w}^T\mathbf{x}$ at x_i over the database $\mathbf{x}$ is given as:*

$$DS_i^f = \left|\sum_{j\in\mathbb{C}_i} \sigma_{ij}\rho_{ij}w_j \left(\bar{\beta}_j - \underline{\beta}_j\right)\right|. \tag{23}$$

Proof. Due to the linearity of the dot product operation, the dependent sensitivity of f at x_i can occur in two cases:

Case 1 ($x_i : \underline{\beta}_i \rightarrow \bar{\beta}_i$): We first consider the expected dependent modification of f over x_j caused by the modification of x_i, denoted as DS_{ij}^f. There are two additional cases: If x_j is positively correlated with x_i, DS_{ij}^f will occur in the direction from $\underline{\beta}_j$ to $\bar{\beta}_j$, otherwise it will occur in the reverse direction, *i.e.*,

$$DS_{ij}^f = \begin{cases} \rho_{ij} w_j \left(\bar{\beta}_j - \underline{\beta}_j\right) & \text{if } \sigma_{ij} = 1, \\ \rho_{ij} w_j \left(\underline{\beta}_j - \bar{\beta}_j\right) & \text{otherwise,} \end{cases} \tag{24}$$

or equivalently,

$$DS_{ij}^f = \sigma_{ij} \rho_{ij} w_j \left(\bar{\beta}_j - \underline{\beta}_j\right). \tag{25}$$

By summing all the dependent modifications and then taking absolute value, we can obtain the dependent sensitive at x_i:

$$DS_i^f = \left| \sum_{j=\mathbb{C}_i} \sigma_{ij} \rho_{ij} w_j \left(\bar{\beta}_j - \underline{\beta}_j\right) \right|. \tag{26}$$

Case 2 ($x_i : \bar{\beta}_i \rightarrow \underline{\beta}_i$): Similar to Case 1, we can derive:

$$DS_{ij}^f = \begin{cases} \rho_{ij} w_j \left(\underline{\beta}_j - \bar{\beta}_j\right) & \text{if } \sigma_{ij} = 1, \\ \rho_{ij} w_j \left(\bar{\beta}_j - \underline{\beta}_j\right) & \text{otherwise,} \end{cases} \tag{27}$$

or equivalently,

$$DS_{ij}^f = -\sigma_{ij} \rho_{ij} w_j \left(\bar{\beta}_j - \underline{\beta}_j\right). \tag{28}$$

Besides, the final form of DS_i^f is the same as that in Case 1. This completes the proof. □

After obtaining the tight upper bound of individual privacy loss, we can utilize it to compute each data owner's share in the total privacy compensations B. Before this, we note that in the top-down design, the privacy compensation function $\psi_i(S)$ should still guarantee dependent fairness, but no longer needs to ensure micro arbitrage freeness. The reason is that the data consumer has paid the arbitrage-free service price $\pi(S)$, and she is not involved in the separate process of privacy compensation. Thus, it is infeasible for the data consumer, as an attacker, to gain arbitrage.

Theorem 8. *The privacy compensation functions*

$$\begin{aligned} \psi_i(S) &= B \frac{DS_i^f / \sqrt{v/2}}{\sum_{i=1}^n DS_i^f / \sqrt{v/2}} \\ &= B \frac{\left| \sum_{j \in \mathbb{C}_i} \sigma_{ij} \rho_{ij} w_j \left(\bar{\beta}_j - \underline{\beta}_j\right) \right|}{\sum_{i=1}^n \left| \sum_{j \in \mathbb{C}_i} \sigma_{ij} \rho_{ij} w_j \left(\bar{\beta}_j - \underline{\beta}_j\right) \right|} \end{aligned} \tag{29}$$

for the total privacy compensations B and for all $i \in \{1, \ldots, n\}$ are top-down privacy compensation functions.

Proof. We prove the dependent fairness by checking that $\forall j \in \mathbb{C}_i, w_j = 0 \Rightarrow \psi_i(S) = 0$. □

In conclusion, the top-down design divides an integrated pricing framework into two independent parts: service pricing and privacy compensation. Different from the bottom-up design, on one hand, the data broker here just needs to ensure the arbitrage freeness of service pricing rather than the micro arbitrage freeness of privacy compensation; on the other hand, the top-down design is essentially the budget allocation problem according to individual privacy loss at the data broker. Hence, the top-down design can execute without the online participation of data owners, and is applicable to any general aggregate statistic.

5 Evaluation Results

In this section, we present the evaluation results in terms of privacy and utility guarantees, arbitrage-free pricing functions, and fine-grained privacy compensations.

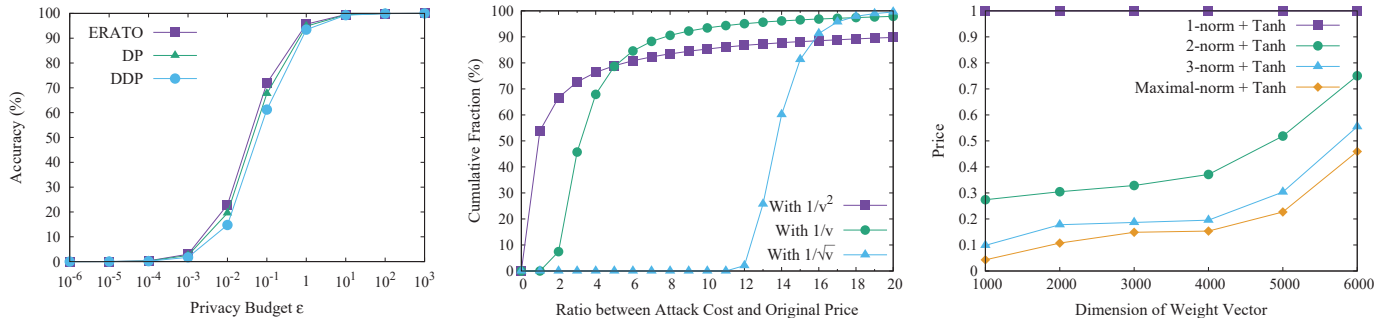
Datasets: We use four real-world datasets, *i.e.*, MovieLens 1M dataset [34], 2009 Residential Energy Consumption Survey (RECS) dataset [35], and two large-scale social network datasets from Stanford Network Analysis Platform (SNAP) [36], for three aggregate statistics, namely weighted sum, probability distribution fitting, and degree distribution, respectively. First, the MovieLens dataset contains 1,000,209 ratings of approximately 3900 movies made by 6040 anonymous users. Besides, we extracted the displayed ratings from MovieLens, which function as target variables in supervised learning. Second, the RECS dataset, which was released by U.S. Energy Information Administration (EIA) in January 2013, provides diverse energy usages in 12,083 U.S. homes. Third, two SNAP datasets are named ego-Twitter and ego-Gplus: ego-Twitter comprises 81,306 nodes and 1,768,149 edges from Twitter, while e-Gplus contains 107,614 nodes and 13,673,453 edges from Google+.

Profiles: To compute the dependence coefficient ρ_{ij} by means of the method developed in [31], we also need to acquire each data owner's profile as auxiliary data. The above four datasets all provide this kind of information: The MovieLens dataset comprises some attributes of users, *e.g.*, gender, age, and occupation; The RECS dataset contains several attributes of each household, such as heating degree days, cooling degree days, total number of rooms, etc; The two SNAP datasets include node features, *e.g.*, gender, institution, and job title. Just as [31], we set the similarity threshold between the profiles of two data owners to be 0.8, and only consider positive correlation, *i.e.*, $\sigma_{ij} = 1$. In contrast, the weight w_j can be either negative or positive in our evaluations, which helps to verify the effect of negative correlation, since $\sigma_{ij} w_j$ in Lemma 2 is in the product form.

Statistics: For weighted sum, we apply linear regression to the ratings of different movies from distinct numbers of users, and can learn different weight vectors with distinct dimensions. For Gaussian distribution fitting, we draw the univariate Gaussian distribution of a certain type of energy consumption, *e.g.*, space heating, air conditioning, or refrigerators. For degree distribution, we count both in and out degrees of every user in Twitter and Google+ networks.

5.1 Privacy and Utility Guarantees

Before investigating economic properties, we first show how ERATO can improve the utility of aggregate statistics, by calculating the dependent sensitivity more accurately for the dependent perturbation mechanism in Theorem 1. Fig. 3a depicts the accuracies of weighted sum under the Laplace perturbation mechanism [10] in the conventional differential privacy (DP), and the dependent perturbation mechanisms in the dependent differential privacy (DDP) and our ERATO, where the privacy budget ϵ varies from 10^{-6} to 10^3 by exponential growth. Here, we select the movie ratings from 1000 users for training, and thus derive



(a) Accuracy (b) Variance of Noise (c) Weight Vector

Fig. 3. Privacy vs. utility and arbitrage freeness in weighted sum.

1000-dimensional weight vectors. Besides, we define the accuracy as $1 - \frac{|\bar{O}-O|}{|\bar{O}+O|}$ [31], where $\bar{O}$ and O are the true and perturbed results, respectively.

One key observation from Fig. 3a is that more accuracy is achieved as the privacy budget ϵ increases, especially when ϵ changes from 0.01 to 0.1. We explain the reason through the formula between the variance of Laplace noise v and the privacy budget ϵ:

$$\epsilon = \frac{\max_i DS_i^f}{\sqrt{v/2}} \Rightarrow v = 2\left(\frac{\max_i DS_i^f}{\epsilon}\right)^2, \qquad (30)$$

which follows from Theorem 1. Here, the function sensitivity $\max_i DS_i^f$ in the numerator remains unchanged for a certain perturbation mechanism. When ϵ becomes larger, less noise is added, which implies a more accurate statistic. Besides, $\max_i DS_i^f$'s for all three perturbation mechanisms are in the magnitude of 0.1. Hence, the accuracy is significantly improved at $\epsilon = 0.1$. Furthermore, when ϵ is too small or too large, the perturbation or the true result completely dominates, and the differences among the accuracies of three perturbation mechanisms are tiny.

The second key observation is derived by comparing the accuracies of three perturbation mechanisms at a fixed privacy budget ϵ, *i.e.*, the denominator in Equation (30) keeps the same. ERATO is more accurate than DDP or even DP. In particular, when $\epsilon = 0.01$, ERATO improves 10.67% and 4.20% of accuracies than DDP and DP, respectively. On one hand, due to the triangle inequality, each individual dependent sensitivity DS_i^f in ERATO, namely Equation (23), is no greater than that in DDP, namely Equation (16), which implies the same relation for $\max_i DS_i^f$. Thus, the true result in ERATO is perturbed with less noise than that in DDP. On the other hand, DP can be regarded as a special case of DDP or ERATO, where the correlated data items are ignored when evaluating DS_i^f, namely Equation (17). Although DS_i^f in DP is always no greater than that in DDP, there exist negative weights here. Besides, the negative part can have more effect on some DS_i^f's than the positive part (not including i itself). Under such circumstance, the final function sensitivity $\max_i DS_i^f$ in ERATO can be less than that in DP, which means higher accuracy.

In conclusion, ERATO can better balance privacy and utility in the aggregate statistics than DP and DDP.

5.2 Arbitrage-free Pricing Functions

In this section, we carry on with the weighted sum application, and further explore arbitrage freeness.

Variance of Noise: We first evaluate the variance of noise v in an arbitrage-free pricing function by simulating the attack illustrated in Example 4. We recall that the data consumer, as an attacker, wants to obtain the service $(\mathbf{w}, v)$ by averaging m the same statistic but with diverse higher variances, namely $\{(\mathbf{w}, v_j)|j \in \{1, \ldots, m\}, v_j > v\}$. We simulate such an arbitrage attack by randomly generating v_j's with the fixed sum $m^2 v$ from the open interval v to $(m^2-m+1)v$. Besides, we set v to be 1 and m to be 100. After simulating 10000 samples, we plot the cumulative fraction of the ratio between the attack cost $\sum_{j=1}^m \pi(\mathbf{w}, v_j)$ and the original price $\pi(\mathbf{w}, v)$ in Fig. 3b, where the pricing function $\pi(\cdot)$ decreases with the variance v from $1/v^2$, to $1/v$, and to $1/\sqrt{v}$. We note that the cumulative fraction here differs from the common cumulative distribution function in that it does not include the endpoint. For example, when the ratio takes 1, the cumulative fraction denotes the fraction of the samples, where the attack cost is strictly less than the original price, *i.e.*, $\sum_{j=1}^m \pi(\mathbf{w}, v_j) < \pi(\mathbf{w}, v)$. More specifically, the cumulative fraction at the ratio of 1 can generally embody the success ratio of finding arbitrage.

By observing the cumulative fractions at the ratio of 1 in Fig. 3b, we can see that there exists arbitrage in $1/v^2$, while the other two pricing functions are arbitrage free, since in $1/v^2$, the cumulative fraction at the ratio of 1 is greater than 0. In particular, the probability that the attacker can find arbitrage in $1/v^2$ is 53.91%. This coincides with our theoretical analysis that arbitrage-free pricing functions cannot decrease faster than $1/v$, namely Theorem 2. From Fig. 3b, we can also observe that an attempt of finding arbitrage in $1/\sqrt{v}$ is expected to be more costly than that in $1/v$, which can be roughly captured by the areas above these two function curves. For instance, to launch an arbitrage attack in $1/\sqrt{v}$, the attacker is most likely to spend 13 to 14 times the original price with probability 34.40%. In contrast, the most possible case in $1/v$ is to pay 2 to 3 times the original price with probability 38.27%. Therefore, in the sense of defending against arbitrage, the pricing function, which decreases slower with the variance v, *e.g.*, $1/\sqrt{v}$ vs. $1/v$, can be more robust. Nevertheless, those legal data consumers may need to pay higher prices when their variances are greater than 1.

Weight Vector: We continue to examine the other part of an arbitrage-free pricing function, namely weight vector. We choose the movie ratings from different numbers of users, and obtain diverse dimensions of weight vectors. Fig. 3c plots four composite pricing functions, when the dimension n increases from 1000 to 6000 with a step of 1000. In particular, the composite pricing functions are derived by

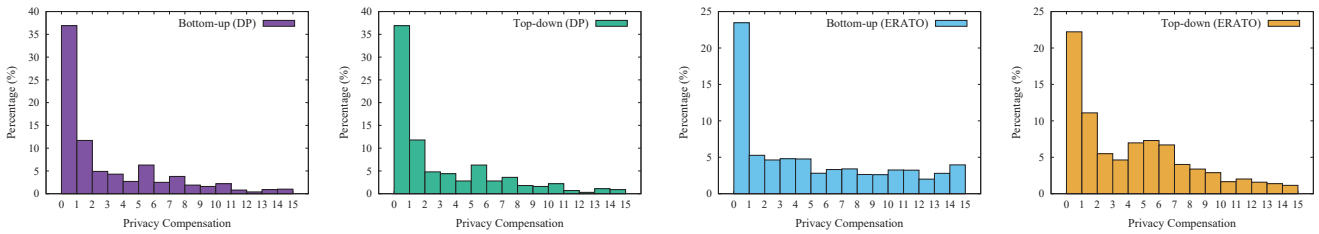


Fig. 4. Differential privacy (DP) and ERATO based privacy compensations in weighted sum.

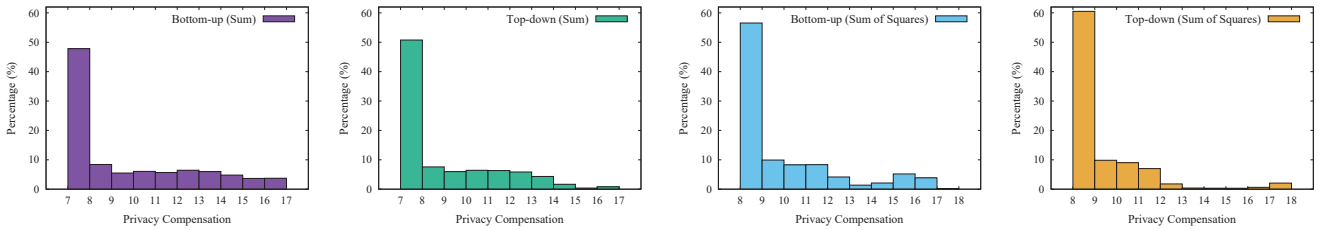


Fig. 5. ERATO based privacy compensation in Gaussian distribution fitting.

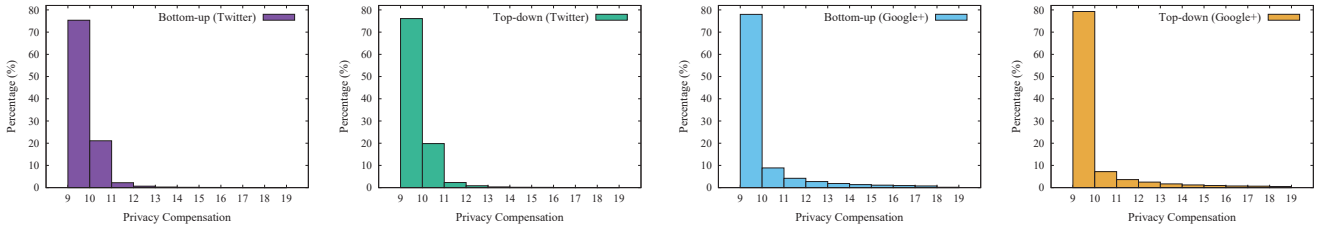


Fig. 6. ERATO based privacy compensation in Twitter and Google+ degree distributions.

first applying $\ell_1, \ell_2, \ell_3, \ell_\infty$ norms and then tanh. Besides, the variance of noise v is set to be 0.1, which gives an error of 1 with 90% confidence by Chebyshev's inequality.

From Fig. 3c, we can see that the composite pricing function using ℓ_1 norm remains almost unchanged at 1, while the other ones increase with the dimension of weight vector n. The reason lies in the characteristics of the bounded tanh function. When $n = 1000$, the pricing function using ℓ_1 norm has already approximated tanh's upper bound 1, and is insensitive to later changes. Besides, the absolute value of each weight is less than 1 here. Thus, as depicted in Fig. 3c, when n is fixed, the price becomes lower for the pricing function using ℓ_p norm with a larger p.

These evaluation results demonstrate that arbitrage freeness is a strong economic property. If not guaranteed, *e.g.*, in the case of $1/v^2$, it is effortless for the data consumer to game the data market. Besides, the data broker can develop her customized pricing strategy by carefully applying Theorem 3 and Theorem 4.

5.3 Fine-grained Privacy Compensations

In this section, we show the privacy compensations in three different aggregate statistics, including weighted sum, Gaussian distribution fitting, and degree distribution. For clarity in presentation and comparison, we fix the total privacy compensations such that one data owner is rewarded with 10 units in average, *i.e.*, $B = 10n$. Besides, we choose the same bounded privacy compensation function in Theorem 7 for each data owner in the bottom-up design.

Before introducing the concrete evaluation results, we first analyze the major differences among three aggregate statistics: (1) From mathematical formula, there exist both positive and negative weights in weighted sum, while the weights in the other two statistics are all constant 1's. Besides, the domain of each data item keeps the same in a certain statistic; (2) From privacy compensation, suppose that we employ the DP framework, which ignores data correlations and compensates the data owner roughly proportional to the absolute value of her weight. Each data owner would be compensated with the average 10 units in Gaussian distribution fitting and degree distribution. Therefore, we only compare DP with ERATO in weighted sum, and directly show ERATO-based evaluation results in the other two statistics.

Weighted Sum: We start with weighted sum, where the dimension of weight vector is fixed at 1000, and the variance of noise v is set to be 0.1 as in Section 5.2. Fig. 4 plots the bottom-up and top-down privacy compensations under DP and ERATO. We note that any pair of neighboring x-axis ticks in Fig. 4 denotes a half-closed interval, *e.g.*, the hist from "9" to "10" stands for the privacy compensations between 9 and 10 excluding 10.

We first compare DP with ERATO in a certain design of privacy compensation. As depicted in Fig. 4, compared with DP, more privacy compensations fall into the center region under ERATO. In particular, 325 data owners receive no privacy compensation in both bottom-up and top-down designs under DP, whereas this number decreases to 148 under ERATO. Such an outcome truly reflects the difference between the properties of fairness and dependent fairness.

We next compare the bottom-up and top-down designs under a certain framework. From Fig. 4, we can see that these two designs of privacy compensation appear identical for DP, but look a slightly different for ERATO.

First, DP does not consider data correlations by setting $\forall j \in \mathbb{C}_i \backslash i, \rho_{ij} = 0$. Thus, a specific data owner i's privacy losses measured by two designs are the same. Besides, when the total privacy compensations are fixed, each data owner's share is proportional to her privacy loss in the top-down design, while is proportional to the tanh value of her privacy loss in the bottom-up design. Moreover, most of the privacy losses ϵ_i's under DP are within 0.1. We further note that tanh has the following property: $0 \leq \epsilon_i \leq 0.1, \tanh(\epsilon_i) \approx \epsilon_i$. Hence, the privacy compensations in two designs look almost the same under DP. In contrast, under ERATO, the dependent sensitivity in the top-down design utilizes a more accurate calculation than that in the bottom-up design, by considering whether the weight is negative or positive. This implies distinct privacy losses and thus distinct privacy compensations under ERATO.

Gaussian Distribution: We show the privacy compensations of Gaussian distribution fitting under ERATO. We recall that the Gaussian distribution can be answered by sum and sum of squares. Here, we set the number of data owners to be 10000. Besides, in the bottom-up design, we set the variance of noise v to be 100, which gives an error of 50 with 96% confidence by Chebyshev's inequality. Moreover, we scale the values inside the tanh function into the range 0 to 5 to better show the differences among privacy compensations in the bottom-up design. We plot the major privacy compensations and their corresponding percentages in Fig. 5, where the results are derived by averaging 10 kinds of energy consumptions.

First, we can see from Fig. 5 that different data owners may obtain distinct privacy compensations rather than the uniform 10 units under DP, although their weights and data domains are the same. The reason is that each data owner has a distinct set of correlated data owners or even the same set but with different strength of correlations, which implies a distinct privacy loss. Second, by comparing privacy compensations in a specific design for two statistics, we can see that they are different from each other, because the dependence coefficient between the same pair of correlated data items in the sum changes in the sum of squares. Third, we compare the privacy compensations in two designs for a certain statistic, and find them consistent in general. This is because when both correlations and weights are positive, the privacy losses measured by two designs are the same. In addition, when the total privacy compensations are fixed, the difference between two designs is that the bottom-up design further applies tanh to the privacy losses. Hence, two designs of privacy compensation are consistent in general.

Degree Distribution: We now investigate how privacy compensations are allocated in large-scale social networks. Fig. 6 depicts the evaluation results of the degree distributions in Twitter and Google+. We set the variance of noise v to be 10 in the bottom-up design. From Fig. 6, we can see that most of privacy compensations fall in the central interval between 9 units and 11 units in both Twitter and Google+. This outcome stems from the fact that the degree distribution of Twitter/Google+ social network asymptotically follows a power law. In particular, 37.17% and 45.49% of Twitter and Google+ users have degrees no more than 5, respectively. Besides, the number of a data owner's degrees has a positive correlation with her privacy loss [15]. Therefore, most of the data owners are compensated around the average 10 units.

We finally give some comments on the ERATO and DP based privacy compensations holistically. First, under DP, the data owners with zero weights receive no compensation in weighted sum. Besides, each data owner is compensated with the indiscriminate 10 units in the other two aggregate statistics. However, such a DP-based allocation scheme is unfair/unreasonable in terms of privacy loss: For weighted sum, a zero-weight data owner can still suffer privacy loss, if her correlated data owners are involved in the service; For the other two statistics, different data owners may have distinct sets of correlated data owners, or even the same set but with different correlation coefficients, which indicates that privacy losses can be different from each other. Specifically, for degree distribution, a higher degree the data owner has, the more social connections she keeps, and the richer private information can be leaked. In short, DP-based privacy compensation is actually another kind of unfairness. In contrast, our ERATO, which discriminates a data owner's privacy compensation with regard to her dependent privacy loss, and introduces the novel property of dependent fairness, has proven to be fairer in practice.

The above evaluation and analysis results demonstrate that two designs of privacy compensation in ERATO can indeed compensate the data owners for their privacy losses in a fairer and more fine-grained way.

6 Related Work

In this section, we briefly review related work.

6.1 Data Market Design

In recent years, data market design has gained increasing attention, especially from the database community. The researchers in this field mainly focus on query-based pricing [29], [30]. Koutris *et al.* [37] showed that the prices of a large class of SQL queries can be computed using ILP solvers. Lin and Kifer [9] designed arbitrage-free pricing functions for arbitrary query formats. Deep and Koutris [16] characterized the structure of arbitrage-free pricing functions in both answer-dependent and instance-independent settings. Based on this work, they also implemented a scalable pricing framework for more relational queries [17]. Specific to private data, Ghosh and Roth [12] considered differential privacy as a commodity, and proposed to selling privacy at auction for single counting query. The follow-up works by Li *et al.* [1], [33], [38] further extend to multiple linear queries by introducing arbitrage freeness. Different from these data trading works, Wang *et al.* [39] focused on the data collection process, where the data broker is untrusted, and each data owner tends to report a noisy version of her private data. They thus established a game-theoretic model to measure the value of privacy.

However, none of above works has taken data correlations into account, and further considered service pricing and privacy compensation in practical aggregate statistics.

6.2 Privacy Preserving Aggregate Statistics

An explosive demand of mining private data from a variety of sources contributes to growing interest in privacy preserving aggregate statistics, where untrusted data analysts can study patterns or statistics over a population while maintaining individual privacy. Shi *et al.* [19] considered the sum statistic for time-series data, *e.g.*, electrical usage and medical telemetry data. Their design is based on distributed differential privacy and additively homomorphic encryption. Popa *et al.* [26] developed a practical system, called PrivStats, to support common aggregate statistics

over location data. PrivStats guarantees privacy and accountability by exploiting additively homomorphic encryption and zero-knowledge proof of knowledge. In particular, to facilitate efficient oblivious evaluation, PrivStats requires the data owner to upload a general function value of her raw data d_i. Such a practice inspires us to introduce an interfaced database $\mathbf{x}$ in the data market setting, and to further model common aggregate statistics as a set of dot product operations. In contrast to the above works, Corrigan-Gibbs and Boneh [40] introduced multiple data analysts to collaboratively compute aggregate statistics in a private, robust, and scalable fashion. Their system mainly integrates secret-shared non-interactive proofs (SNIPs) with affine-aggregatable encodings (AFEs).

Unfortunately, the original intention of these works is preserving privacy against untrusted data analysts rather than pricing noisy aggregate statistics for data consumers, and quantifying and compensating privacy losses for data owners, which are instead the major focuses of our work.

6.3 Differential Privacy over Correlated Data

The classical differential privacy framework, proposed by Dwork *et al.* [10], [11], adopts a different security assumption that the data analyst can be trusted. Under this assumption, the data analyst adds appropriate noises to aggregate results before releasing them, which can protect an individual's private information. However, as pointed by Kifer and Machanavajjhala [13], when there exist correlations among the data items, the perturbation in differential privacy can be inadequate. They thus proposed a generalized version of differential privacy, called Pufferfish privacy [41]. Many follow-up research works have been going on around this particular issue. In addition to the dependent differential privacy [31] utilized in this work, Yang *et al.* [15] focused on the correlation structure modeled by Gaussian Markov random fields. Xiao *et al.* [42] considered how to protect a user's consecutive locations, and employed Markov chains to model temporal correlations. Cao *et al.* [43] quantified the risk of differential privacy under the continuous aggregate release over multiple users' locations. Song *et al.* [14] proposed a Wasserstein mechanism for any general Pufferfish instantiation, together with a computationally efficient Markov quilt mechanism for Bayesian networks.

However, the above works still aim at privacy preservation but now against external attackers, *e.g.*, data consumers in data markets. Yet, some of their principles can be borrowed to quantify fine-grained privacy losses for a wider range of aggregate statistics.

7 Conclusion and Future Work

In this paper, we have proposed the first pricing framework ERATO for data markets, which provide common aggregate statistics over private correlated data. In ERATO, the data consumer has to faithfully request the desired service rather than gaming the system through buying a bundle of cheaper services. Besides, the data owners can be compensated for their dependent privacy losses in a more fine-grained way. Furthermore, we have instantiated ERATO with three different kinds of aggregate statistics, and extensively evaluated their performances on four practical datasets. Evaluation results have demonstrated the feasibility of ERATO from the improvement of statistic utility, the arbitrage freeness of service pricing, and the fairness of privacy compensation.

As for future work, one interesting direction is to investigate how to trade more kinds of personal data in practice, *e.g.*, health records, physical activities, and driving trajectories. Specific to a concrete kind of data, we should first determine an appropriate trading format, and further rule out arbitrage opportunities when pricing different trading settings. For example, in the case of trading time-series data, the data consumer may be allowed to designate a pair of starting and ending points together with a sampling period. In addition to the trading format, we also need to consider the underlying data characteristics, especially when quantifying privacy loss, *e.g.*, social, temporal, and spatial correlations among the multiple data owners' driving trajectories. Yet, another potential research direction is to balance pros and cons brought by relaxing the arbitrage freeness requirement. Here, pros are for the data broker, and cons are from cunning data consumers. In essence, arbitrage freeness implies the computational infeasibility of arbitrage attack, and requires the pricing functions to preserve strict mathematical properties. Suppose the data broker relaxes arbitrage freeness, *e.g.*, by abandoning some rules in the determinacy relation. She can choose a wider range of pricing functions, and support more aggregate statistics. However, the arbitrage attack now becomes computationally feasible. If attack cost is no more than revenue, the data consumers are well-motivated to launch arbitrage attacks.

Acknowledgments

This work was supported in part by the National Key R&D Program of China 2018YFB1004703, in part by China NSF grant 61672348, 61672353, and 61872238, in part by the Open Project Program of the State Key Laboratory of Mathematical Engineering and Advanced Computing 2018A09, in part by the State Key Laboratory of Air Traffic Management System and Technology SKLATM20180X, in part by the Shanghai Science and Technology Fund 17510740200, in part by the Huawei Innovation Research Program HO2018085286, in part by Alibaba Group through Alibaba Innovation Research Program, and in part by the Tencent Social Ads Rhino-Bird Focused Research Program. The opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies or the government.

References

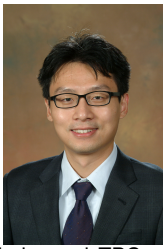
[1] C. Li, D. Y. Li, G. Miklau, and D. Suciu, "A theory of pricing private data," *Communications of the ACM*, vol. 60, no. 12, pp. 79–86, 2017.
[2] "Personal data: The emergence of a new asset class," https://www.weforum.org/reports/personal-data-emergence-new-asset-class, 2011.
[3] J. Brustein, "Start-ups seek to help users put a price on their personal data," *The New York Times*, Feb. 2012.
[4] "Datacoup," https://datacoup.com/, 2012.
[5] "CitizenMe," https://www.citizenme.com/, 2013.
[6] "CoverUS," https://www.coverus.io/, 2018.
[7] Federal Trade Commission (FTC), "Data brokers: A call for transparency and accountability," https://www.ftc.gov/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014, 2014.
[8] "The data brokers: Selling your personal information," https://www.cbsnews.com/news/data-brokers-selling-personal-information-60-minutes/, 2014.
[9] B. Lin and D. Kifer, "On arbitrage-free pricing for general data queries," *PVLDB*, vol. 7, no. 9, pp. 757–768, 2014.
[10] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.
[11] C. Dwork, F. McSherry, K. Nissim, and A. D. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. of TCC*, 2006, pp. 265–284.

[12] A. Ghosh and A. Roth, "Selling privacy at auction," in *Proc. of EC*, 2011, pp. 199–208.
[13] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proc. of SIGMOD*, 2011, pp. 193–204.
[14] S. Song, Y. Wang, and K. Chaudhuri, "Pufferfish privacy mechanisms for correlated data," in *Proc. of SIGMOD*, 2017, pp. 1291–1306.
[15] B. Yang, I. Sato, and H. Nakagawa, "Bayesian differential privacy on correlated data," in *Proc. of SIGMOD*, 2015, pp. 747–762.
[16] S. Deep and P. Koutris, "The design of arbitrage-free data pricing schemes," in *Proc. of ICDT*, 2017, pp. 12:1–12:18.
[17] ——, "QIRANA: A framework for scalable query pricing," in *Proc. of SIGMOD*, 2017, pp. 699–713.
[18] A. Vanderveld, A. Pandey, A. Han, and R. Parekh, "An engagement-based customer lifetime value system for e-commerce," in *Proc. of KDD*, 2016, pp. 293–302.
[19] E. Shi, T. H. Chan, E. Rieffel, R. Chow, and D. Song, "Privacy-preserving aggregation of time-series data," in *Proc. of NDSS*, 2011.
[20] C. Niu, Z. Zheng, F. Wu, X. Gao, and G. Chen, "Achieving data truthfulness and privacy preservation in data markets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 1, pp. 105–119, 2019.
[21] N. Eikmeier and D. F. Gleich, "Revisiting power-law distributions in spectra of real world networks," in *Proc. of KDD*, 2017, pp. 817–826.
[22] The New York Times, "Facebook is not the problem. lax privacy rules are." https://www.nytimes.com/2018/04/01/opinion/facebook-lax-privacy-rules.html, 2018.
[23] ——, "Facebook Hack Included Search History and Location Data of Millions," https://www.nytimes.com/2018/10/12/technology/facebook-hack-investigation.html, 2018.
[24] R. Cummings, K. Ligett, A. Roth, Z. S. Wu, and J. Ziani, "Accuracy for sale: Aggregating data with a variance constraint," in *Proc. of ITCS*, 2015, pp. 317–324.
[25] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proc. of KDD*, 2011, pp. 1082–1090.
[26] R. A. Popa, A. J. Blumberg, H. Balakrishnan, and F. H. Li, "Privacy and accountability for location-based aggregate statistics," in *Proc. of CCS*, 2011, pp. 653–666.
[27] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.
[28] K. H. Rosen, *Discrete mathematics and its applications*, 7th ed. McGraw-Hill, 2011.
[29] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," in *Proc. of PODS*, 2012, pp. 167–178.
[30] ——, "Query-based data pricing," *Journal of the ACM*, vol. 62, no. 5, pp. 43:1–43:44, 2015.
[31] C. Liu, S. Chakraborty, and P. Mittal, "Dependence makes you vulnerable: Differential privacy under dependent tuples," in *Proc. of NDSS*, 2016.
[32] "Technical Report for ERATO," https://www.dropbox.com/s/zokshkynh2e8wf8/, Jan. 2019.
[33] C. Li, D. Y. Li, G. Miklau, and D. Suciu, "A theory of pricing private data," in *Proc. of ICDT*, 2013, pp. 33–44.
[34] "MovieLens 1M Dataset," https://grouplens.org/datasets/movielens/1m/, 2003.
[35] "2009 RECS Dataset," https://www.eia.gov/consumption/residential/data/2009/index.php?view=microdata, 2013.
[36] "SNAP Datasets: Stanford Large Network Dataset Collection," http://snap.stanford.edu/data, 2014.
[37] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Toward practical query pricing with querymarket," in *Proc. of SIGMOD*, 2013, pp. 613–624.
[38] C. Li, D. Y. Li, G. Miklau, and D. Suciu, "A theory of pricing private data," *ACM Transactions on Database Systems*, vol. 39, no. 4, pp. 34:1–34:28, 2014.
[39] W. Wang, L. Ying, and J. Zhang, "The value of privacy: Strategic data subjects, incentive mechanisms and fundamental limits," in *Proc. of SIGMETRICS*, 2016, pp. 249–260.
[40] H. Corrigan-Gibbs and D. Boneh, "Prio: Private, robust, and scalable computation of aggregate statistics," in *Proc. of NSDI*, 2017, pp. 259–282.
[41] D. Kifer and A. Machanavajjhala, "A rigorous and customizable framework for privacy," in *Proc. of PODS*, 2012, pp. 77–88.
[42] Y. Xiao and L. Xiong, "Protecting locations with differential privacy under temporal correlations," in *Proc. of CCS*, 2015, pp. 1298–1309.
[43] Y. Cao, M. Yoshikawa, Y. Xiao, and L. Xiong, "Quantifying differential privacy under temporal correlations," in *Proc. of ICDE*, 2017, pp. 821–832.

Chaoyue Niu is working toward the PhD degree in the Department of Computer Science and Engineering, Shanghai Jiao Tong University, P. R. China. His research interests include privacy preservation and verifiable computation in data management. He is a student member of the ACM and IEEE.

Zhenzhe Zheng is now a Post Doc in the University of Illinois at Urbana-Champaign (UIUC). He received the PhD degree in Computer Science from Shanghai Jiao Tong University, in 2018. His research interests include algorithmic game theory, resource management in wireless networking and data center. He is a student member of the ACM, IEEE, and CCF.

Fan Wu is a professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University. He received the BS degree in Computer Science from Nanjing University, in 2004, and the PhD degree in Computer Science and Engineering from the State University of New York at Buffalo, in 2009. He has visited the University of Illinois at Urbana-Champaign (UIUC) as a Post Doc Research Associate. His research interests include wireless networking and mobile computing, algorithmic game theory and its applications, and privacy preservation. He has published more than 100 peer-reviewed papers in technical journals and conference proceedings. He is a recipient of the first class prize for Natural Science Award of China Ministry of Education, NSFC Excellent Young Scholars Program, ACM China Rising Star Award, CCF-Tencent "Rhinoceros bird" Outstanding Award, CCF-Intel Young Faculty Researcher Program Award, and Pujiang Scholar. He has served as the chair of CCF Y-OCSEF Shanghai, on the editorial board of Elsevier Computer Communications, and as the member of technical program committees of more than 60 academic conferences. For more information, please visit http://www.cs.sjtu.edu.cn/~fwu/.



Shaojie Tang is currently an assistant professor of Naveen Jindal School of Management at University of Texas at Dallas. He received the PhD degree in computer science from Illinois Institute of Technology, in 2012. His research interests include social networks, mobile commerce, game theory, e-business, and optimization. He received the Best Paper Awards in ACM MobiHoc 2014 and IEEE MASS 2013. He also received the ACM SIGMobile service award in 2014. Dr. Tang served in various positions (as chairs and TPC members) at numerous conferences, including ACM MobiHoc and IEEE ICNP. He is an editor for International Journal of Distributed Sensor Networks.

Xiaofeng Gao received the B.S. degree in information and computational science from Nankai University, China, in 2004; the M.S. degree in operations research and control theory from Tsinghua University, China, in 2006; and the Ph.D. degree in computer science from The University of Texas at Dallas, USA, in 2010. She is currently a professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. Her research interests include distributed system, wireless communications, data engineering, and combinatorial optimizations. She has published more than 160 peer-reviewed papers in the related area, including well-archived international journals such as IEEE TC, TKDE, TMC, TPDS, JSAC, and also in well-known conference proceedings such as WWW, SIGKDD, INFOCOM, ICDCS, etc. She has served on the editorial board of Discrete Mathematics, Algorithms and Applications, and as the PCs and peer reviewers for a number of international conferences and journals.

Guihai Chen earned the BS degree from Nanjing University, in 1984, the ME degree from Southeast University, in 1987, and the PhD degree from the University of Hong Kong, in 1997. He is a distinguished professor of Shanghai Jiaotong University, China. He had been invited as a visiting professor by many universities including Kyushu Institute of Technology, Japan, in 1998, University of Queensland, Australia, in 2000, and Wayne State University, USA during September 2001 to August 2003. He has a wide range of research interests with focus on sensor network, peer-to-peer computing, high-performance computer architecture and combinatorics. He has published more than 200 peer-reviewed papers, and more than 120 of them are in well-archived international journals such as IEEE Transactions on Parallel and Distributed Systems, Journal of Parallel and Distributed Computing, Wireless Network, The Computer Journal, International Journal of Foundations of Computer Science, and Performance Evaluation, and also in well-known conference proceedings such as HPCA, MOBIHOC, INFOCOM, ICNP, ICPP, IPDPS, and ICDCS.