Making Big Money from Small Sensors: Trading Time-Series Data under Pufferfish Privacy

Chaoyue Niu, Zhenzhe Zheng, Shaojie Tang†, Xiaofeng Gao, and Fan Wu
Shanghai Key Laboratory of Scalable Computing and Systems, Shanghai Jiao Tong University, China
†Department of Information Systems, University of Texas at Dallas, USA
Email: {rvincency, zhengzhenzhe220, tangshaojie}@gmail.com; {gao-xf, fwu}@cs.sjtu.edu.cn

***Abstract*—With the commoditization of personal data, pricing privacy has become an intriguing topic. In this paper, we study time-series data trading from the perspective of a data broker in data markets. We thus propose HORAE, which is a PufferfisH privacy based framewOrk for tRAding timE-series data. HORAE first employs Pufferfish privacy to quantify privacy losses under temporal correlations, and compensates data owners with distinct privacy strategies in a satisfying way. Besides, HORAE not only guarantees good profitability at the data broker, but also ensures arbitrage freeness against cunning data consumers. We further apply HORAE to physical activity monitoring, and extensively evaluate its performance on the real-world Activity Recognition with Ambient Sensing (ARAS) dataset. Our analysis and evaluation results reveal that HORAE compensates data owners in a more fine-grained manner than entry/group differential privacy based approaches, well controls the profit ratio of the data broker, and thwarts arbitrage attacks launched by data consumers.**

***Index Terms*—Data Trading; Data Privacy; Time-Series Data**

I. Introduction

The past few years have witnessed the proliferations of smart devices and Internet of Things (IoTs) in people's daily lives. Tremendous volumes of data are collected periodically by the embedded sensors to monitor human activities. Typical examples of monitoring data in time series include breathing volumes [1], heartbeats [2], physical activities [3], and residential energy consumptions [4]. However, for privacy concerns, most of data owners are reluctant to share their data, resulting in a number of isolated data islands. To promote private data circulation, many data brokers [5]–[8] have emerged to bridge the gap between data owners and data consumers. On one hand, a data broker needs to offer monetary rewards to incentivize the data owners to contribute sensitive data. On the other hand, the data broker charges the data consumers for their queries over the collected dataset.

In this paper, we study a novel time-series data trading problem from the data broker's point of view in data markets. We summarize three major design challenges. The first and the thorniest challenge is to rigorously quantify privacy loss at timestamp level. Markets for sensitive personal data significantly differ from those for ordinary information goods in privacy compensation [9]. To compensate each data owner properly, it is necessary to quantify her privacy loss during the usage of her data. Existing works on private data trading [9]–[11] mainly considered data from multiple data owners, and utilized differential privacy [12], [13] to measure individual privacy loss. However, these works cannot directly apply to our context, since we intend to investigate a sequence of data from a certain data owner, and to quantify her privacy loss at each timestamp. Although two modified versions of differential privacy, called entry differential privacy [14] and group differential privacy [15], may be applicable, they mishandle temporal correlations, where all states are assumed to be independent (resp., correlated) in the entry (resp., group) differential privacy. In other words, they treat each state equally, which is unreasonable in practice. For example, Alice's states at 8:00am and 8:00pm have different sets of correlated states, and thus can suffer distinct privacy losses.

Yet, another challenge comes from the query format over time-series data. In data markets, each data consumer should be permitted to purchase data analysis over her interested data items rather than the whole dataset [16]. For time-series data, we consider that the data consumer can designate a pair of starting and ending points together with a sampling period. Nevertheless, this query format has two striking impacts on the entire framework: On one hand, different query settings can induce distinct structures of temporal correlations, and thus can affect the quantification and compensation of privacy loss as mentioned above; On the other hand, the data broker needs to set appropriate prices for different query settings. In particular, a desirable pricing mechanism should be balanced and arbitrage free. Balance enforces a constraint that the price of a query is sufficient to cover the total privacy compensation, while arbitrage freeness requires that the data consumer cannot circumvent the advertised price of a query through buying other cheaper ones. Considering the naive zero-price function is arbitrage free but not balanced, it is highly nontrivial to guarantee these two economic properties simultaneously.

Last but not least challenge is to avoid arbitrage opportunities in varying degrees of perturbation. For the sake of privacy issues, *e.g.*, the successive Facebook data scandals [17], [18], it is necessary for the data broker to sell noisy answers. Besides, to allow different prices for the same data analysis

This work was supported in part by the National Key R&D Program of China 2018YFB1004703, in part by China NSF grants 61672348, 61672353, 61472252, and 61872238, in part by the Huawei Innovation Research Program (HO2018085286), and in part by the State Key Laboratory of Air Traffic Management System and Technology (SKLATM20180X). 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.

F. Wu is the corresponding author.

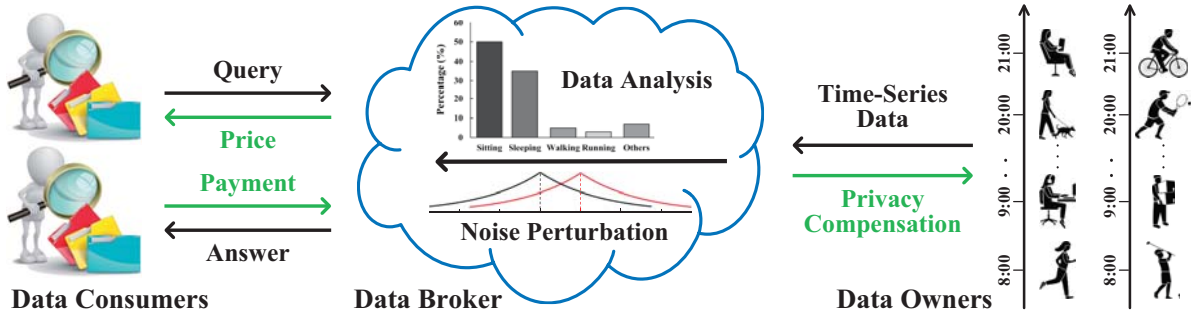


Fig. 1. A General System Model of Data Markets Trading Time-Series Data.

but with diverse accuracies, the data consumer can specify her customized noise level, *e.g.*, the variance of noise used in [10], [11]. In particular, if less noise is added to the true answer, the price of the query should be higher. However, this setting makes reasoning about arbitrage freeness even harder. A hidden arbitrage attack is that a savvy data consumer may obtain a certain data analysis with low variance of noise at a lower price, particularly through averaging multiple the same analysis but with diverse high variances. Furthermore, if the variance of noise is not independent in the pricing of query, it can make thwarting the above attack more challenging.

In this paper, by jointly considering the above three challenges, we propose HORAE, which is a PufferfisH privacy based framewOrk for tRADing timE-series data. HORAE adopts a bottom-up design holistically, and considers privacy loss quantification, privacy compensation, and pricing of query in sequence. HORAE first employs Markov chains to model private data with temporal correlations. HORAE then borrows key principles from Pufferfish privacy to define the privacy loss of a data owner at each timestamp, and further gives its upper bound in the context of Markov quilt mechanism. Based on this upper bound, HORAE devises customized privacy compensation functions, which are satisfying for data owners with different privacy strategies. Besides, when determining the price of a query, HORAE gracefully enlarges the total privacy compensation, such that it can not only guarantee non-negative utility at the data broker, but also ensure arbitrage freeness against the data consumer with respect to both the queried states and the variance of noise.

We summarize our key contributions as follows.

- To the best of our knowledge, we are the first to study trading private data with temporal correlations from the perspective of a data broker in data markets.
- Our proposed framework HORAE features the properties of Pufferfish privacy and Markov quilt mechanism to quantify privacy losses, and compensates diverse data owners in a satisfying manner. Besides, when pricing queries from data consumers, HORAE guarantees balance and avoids arbitrage.
- We instructively instantiate HORAE with the physical activity monitoring application, and extensively evaluate its performance on the real-world ARAS dataset. Our analysis and evaluation results demonstrate that HORAE compensates data owners in a more fine-grained way than entry/group differential privacy based approaches, well controls the profit ratio at the data broker, and discourages data consumers from launching arbitrage attacks. Specifically, a certain data owner receives distinct rather than the same privacy compensations for her privacy losses at different timestamps. Besides, the data broker can control the lowest point on the convex curve of profit ratio, which in turn can guide consumption or maximize her expected profit ratio. Moreover, to launch an arbitrage attack in the advanced, unbounded, and valid pricing function, the data consumer, as an attacker, has to pay the amount of 40 to 41 times the original price with 45.35% probability.

II. Preliminaries

In this section, we introduce system model, Pufferfish privacy, and Markov quilt mechanism.

A. System Model

As shown in Fig. 1, we consider a general system model for data markets trading time-series data. There are three kinds of entities: data owners, a data broker, and data consumers.

The data broker first procures time-series data, such as physical activities, heart rates, and electrical usages, from the data owners. We use a sequence of variables $X = X_1 \to X_2 \to \ldots \to X_T$ to describe the data from a specific data owner, where X_t denotes her state at time t. In addition, we use $X_t = a$ to represent the event that X_t takes the concrete value a from a state space A. For example, A stands for a set of all possible human activities, and $X_t = a$ denotes an event that Alice jogged at 8:00am.

We consider that each data consumer can make her customized query $Q = (f, v)$. Here, f is a general function over X, and is assumed to be ℓ-Lipschitz with respect to L_1 norm, *i.e.*, the change of any state in X (with all the other states fixed) can vary the L_1 norm of f's output by at most ℓ. Typical instantiations of f include common data analysis methods, *e.g.*, histogram count, weighted sum, and probability distribution fitting. Besides, v denotes a tolerable variance of noise added to the true result $f(X)$. In other words, the data broker answers the query Q with a randomized mechanism $\mathcal{M}$, and returns the answer $\mathcal{M}(X)$, where its expectation is $f(X)$, and its variance is no more than v.

Depending on the query Q, the data broker compensates the data owner with $\rho(Q)$ for her privacy leakage, and charges the data consumer a price $\pi(Q)$. Specifically, if the variance of noise v is higher, the privacy loss becomes smaller, the privacy compensation $\rho(Q)$ would be lower, the returned answer is less accurate, and the price $\pi(Q)$ should be lower.

B. Pufferfish Privacy

The privacy framework of our choice is Pufferfish privacy [19], which is an elegant generalization of the celebrated differential privacy [12], [15], by incorporating general data correlations. We introduce its technical details from the privacy preservation perspective, *i.e.*, we focus on the randomized mechanism $\mathcal{M}$ itself. Yet, its key principles will be used to mathematically quantify privacy loss in Section III-A.

A Pufferfish framework is instantiated by three parameters: a set of secrets S that we wish to hide, a set of secret pairs $SP \subseteq S \times S$ that we want to be indistinguishable, and a class of probability distributions Θ that can plausibly generate the data. For time-series data X, $S = \{X_t = a | a \in A, t \in [T]\}$ indicates that the concrete event a at each time t is a secret; $SP = \{(X_t = a, X_t = b) | a \neq b \in A, t \in [T]\}$ means that whether the data owner is engaging in event a or b at any time t cannot be distinguished; Θ can be a set of Markov chains that capture how the data owner switches between states.

Definition 1. *A randomized mechanism $\mathcal{M}$ is said to be ϵ-Pufferfish private, if for any $\theta \in \Theta$ with $X \sim \theta$, any secret pair $(X_t = a, X_t = b) \in SP$, and any possible output O:*

$$e^{-\epsilon} \leq \frac{P(\mathcal{M}(X) = O | X_t = a, \theta)}{P(\mathcal{M}(X) = O | X_t = b, \theta)} \leq e^{\epsilon},$$

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

We note that the definition of Pufferfish privacy is not only with respect to the randomized mechanism $\mathcal{M}$, like in differential privacy, but also to the distribution class Θ, where Θ can well control the amount and nature of data correlations.

C. Markov Quilt Mechanism

To achieve ϵ-Pufferfish privacy for any general Bayesian network G, an efficient Markov quilt mechanism was proposed in [20]. For consistency in notations, we use "state" to represent "node" in the network G considered here, and use X to denote the vertex set of G. Besides, we hereinafter let $\mathsf{card}(\cdot)$ denote the cardinality of a set.

The main insight behind Markov quilt mechanism is that if two states X_t and X_r are "far apart" in G, then X_r is largely independent of X_t. Thus, to obscure the effect of X_t on the result of a function, it is sufficient to add noise proportional to the number of states that are "close" to X_t plus a correction term accounting for the effect of "distant" states. Now, two key problems arise: one is how to separate "close" states from "distant" states, and the other is how to calculate the correction term for "distant" states.

For the first problem, a novel concept of Markov quilt is established, which generalizes the standard Markov blanket in probabilistic graphical models. We recall that the Markov blanket of a state X_t in a Bayesian network consists of its parents, its children, and the other parents of its children. Additionally, the rest states in the network are independent of X_t conditioned on its Markov blanket. We present a generalization of Markov blanket, Markov quilt, as follows.

Algorithm 1: Markov Quilt Mechanism

Input: A dataset X, an ℓ-Lipschitz function f, a privacy budget ϵ, Markov quilt set $S_{M,t}$ of each state X_t, and Laplace distribution $Lap(\cdot)$ centered at 0.
Output: ϵ-Pufferfish private perturbation mechanism $\mathcal{M}(X)$.

```
foreach X_t ∈ X do
    foreach X_M ∈ S_{M,t} (with X_C, X_R) do
        if ψ_Θ(X_M|X_t) < ε then
            σ(X_M) = ℓ·card(X_C) / (ε − ψ_Θ(X_M|X_t))
        else
            σ(X_M) = +∞
    σ_t = min_{X_M ∈ S_{M,t}} σ(X_M)
return M(X) = f(X) + Lap(max_t σ_t)
```

Definition 2. *A set of states X_M is a Markov quilt of a state X_t in a Bayesian network G, if the following conditions hold:*

- *Deleting X_M partitions G into two parts X_C and X_R, such that $X = X_C \bigcup X_M \bigcup X_R$ and $X_t \in X_C$.*
- *X_R is independent of X_t conditioned on X_M.*

Intuitively, X_C is a set of "close" states, X_R is a set of "remote" states, and they are separated by the Markov quilt X_M. For example, $X_M = \varnothing$ (with $X_C = X, X_R = \varnothing$) is a trivial Markov quilt. Besides, different from the uniqueness of Markov blanket, a state can have multiple Markov quilts.

Regarding the second problem, we note that "distant" states contain both the Markov quilt and "remote" states, *i.e.*, $X_M \bigcup X_R$. Nevertheless, since X_R is independent of X_t given X_M, to quantify the effect of X_t on $X_M \bigcup X_R$, it suffices to measure the effect of X_t on X_M. For this purpose, the max-influence between a variable X_t and a set of variables X_M is defined, which quantifies how much changing the value of X_t can affect X_M, where their probabilistic dependence is described by a distribution class Θ.

Definition 3. *The max-influence of a variable X_t on a set of variables X_M under a distribution class Θ is defined as:*

$$\psi_\Theta(X_M | X_t) = \sup_{\theta \in \Theta} \max_{a,b \in A} \max_{x_M \in A^{\mathsf{card}(X_M)}} \log \frac{P(X_M = x_M | X_t = a, \theta)}{P(X_M = x_M | X_t = b, \theta)}.$$

The max-influence $\psi_\Theta(X_M | X_t)$ is essentially the maximum max-divergence between two distributions $X_M | X_t = a, \theta$ and $X_M | X_t = b, \theta$, where the maximum is taken over any $a, b \in A, \theta \in \Theta$. Thus, the mathematical forms of max-influence here and the privacy budget ϵ in Definition 1 are consistent in nature [21]. From this perspective, $\psi_\Theta(X_M | X_t)$ can function as a share of the total privacy budget ϵ allocated

to "distant" states $X_M \bigcup X_R$, and $\epsilon - \psi_\Theta(X_M|X_t)$ is the complementary share allocated to "close" states X_C.

After solving two problems, we present Markov quilt mechanism in Algorithm 1. If we want to release the result of an ℓ-Lipschitz function f while protecting X_t, and if we find a Markov quilt X_M of X_t, it is sufficient to add Laplace noise with scale $\ell \cdot \mathsf{card}(X_C)/(\epsilon - \psi_\Theta(X_M|X_t))$ (Line 4). We search over the Markov quilt set $S_{M,t}$ of X_t (Line 2), and pick the one which requires the least amount of noise (Line 7). We then iterate over all X_t's (Line 1), and add to $f(X)$ the maximum amount of noise needed to protect the whole X (Line 8). We finally note that if the output of f is of multiple dimensions, we just need to add noise drawn from the same Laplace distribution to each dimension. Hence, for brevity, we hereinafter focus on a specific dimension of the output.

III. Design of HORAE

In this section, we propose HORAE. HORAE takes a bottom-up design, where the data broker first needs to compensate the privacy loss of the data owner at bottom, and then determines the price of the query for the data consumer at top.

A. Privacy Loss

When the data broker answers the data consumer's query Q with the randomized mechanism $\mathcal{M}$, some private information of the data owner can be leaked. Based on the principles of Pufferfish privacy, we formally define privacy loss.

We consider a pair of time-series data instances $X \sim \theta|X_t = a$ and $X \sim \theta|X_t = b$ for any $a, b \in A, \theta \in \Theta$, which initially differ in the state X_t at time t. In fact, they can simulate every possible change of X_t, together with ripple effects on the other states due to the data correlations modeled by Markov chains Θ. By comparing the output of the randomized mechanism $\mathcal{M}$ over these two data instances, we define the privacy loss ξ_t at time t.

Definition 4. *The privacy loss of the data owner at time t in the randomized mechanism $\mathcal{M}$ over the time-series data X is:*

$$\xi_t = \sup_{a,b \in A, \theta \in \Theta, O} \log \left| \frac{P(\mathcal{M}(X) = O | X_t = a, \theta)}{P(\mathcal{M}(X) = O | X_t = b, \theta)} \right|. \quad (1)$$

We further give an upper bound of the privacy loss ϵ_t, when the randomized mechanism is known to be the Markov quilt mechanism in Algorithm 1.

Theorem 1. *Let $\mathcal{M}$ be Markov quilt mechanism, f be an ℓ-Lipschitz function, $S_{M,t}$ be the Markov quilt set of X_t, and v be the variance of Laplace noise. The privacy loss of the data owner at time t is bounded above by:*

$$\xi_t \leq \min_{X_M \in S_{M,t}} \left(\frac{\ell \cdot \mathsf{card}(X_C)}{\sqrt{v/2}} + \psi_\Theta(X_M|X_t) \right).$$

When computing the upper bound of privacy loss given in Theorem 1, specific to the topological structure of Markov chains, we limit the size of Markov quilt set $S_{M,t}$ of each state X_t, rather than traversing all its exponential number of Markov quilts like in general Bayesian networks [20].

Lemma 1. *To find the upper bound of privacy loss, it suffices to search the Markov quilt set: $S_{M,t} = \{\{X_{t-j}, X_{t+k}\}, \{X_{t-j}\}, \{X_{t+k}\}, \varnothing | 1 \leq j \leq t-1, 1 \leq k \leq T-t\}$.*

Proof. We give a proof sketch. We first prove that any $X_M \in S_{M,t}$ is a Markov quilt of X_t by Definition 2 and d-separation. We next prove that for any Markov quilt $X_{M'} \notin S_{M,t}$, there exists $X_M \in S_{M,t}$ such that the upper bound of privacy loss over X_M is no more than that over $X_{M'}$. □

Interested readers can refer to our technical report [22] for detailed proofs of Theorem 1 and Lemma 1.

B. Privacy Compensation

After quantifying the privacy loss ξ_t, we consider the second component of HORAE, namely the privacy compensation mechanism for the data owner.

Just as [10], we first introduce a nondecreasing contract function $\omega(\xi_t)$ between the data broker and the data owner, such that $\omega(0) = 0$. This is to ensure that the data owner will be compensated with at least $\omega(\xi_t)$ in the event of a privacy loss ξ_t. Based on the contract function, we define the valid privacy compensation function in a formal way.

Definition 5. *Let $\omega(\cdot)$ be a contract function between the data broker and the data owner. Let $\rho(\cdot)$ be a privacy compensation function. If $\forall t \in [T], \rho(\xi_t) \geq \omega(\xi_t)$, then $\rho(\cdot)$ is valid.*

Intuitively, a privacy compensation function is valid with respect to the contract function, if it is satisfactory for the data owner for her privacy loss at any time t.

We next demonstrate how the data owner can select a customized contract function, and how the data broker can construct a valid privacy compensation function accordingly. In fact, the contract function hinges on the data owner's privacy strategy. For example, if the data owner values her privacy highly, and would never accept full disclosure of personal data, she may choose a linear contract function, which sets infinite compensations for unperturbed answers, *i.e.*, the variance of noise $v = 0$. Correspondingly, the data broker can utilize the upper bound of privacy loss in Theorem 1 to devise a valid privacy compensation function.

Theorem 2. *Let $\omega(\xi_t) = c\xi_t$ for $c > 0, t \in [T]$. Then, the privacy compensation function*

$$\rho(\xi_t) = c \min_{X_M \in S_{M,t}} \left(\frac{\ell \cdot \mathsf{card}(X_C)}{\sqrt{v/2}} + \psi_\Theta(X_M|X_t) \right)$$

for $t \in [T]$ is unbounded and valid.

Proof. We prove unboundedness by checking that $v = 0 \Rightarrow \rho(\xi_t) = \infty$. We further prove validness by checking that $\forall t \in [T], \rho(\xi_t) \geq \omega(\xi_t)$, which follows from Theorem 1. □

However, this kind of contract function may be unsuitable for the data owner, who is less concerned about her privacy, and is willing to sell her private data at some high but finite price. Nevertheless, she can turn to selecting some bounded contract functions, *e.g.*, cut-off and sigmoid functions.

Theorem 3. *Let $\omega(\xi_t) = c\tanh(d\xi_t)$ for $c, d > 0, t \in [T]$. Then, the privacy compensation function*

$$\rho(\xi_t) = c\tanh\left(d \min_{X_M \in S_{M,t}} \left(\frac{\ell \cdot \mathsf{card}\,(X_C)}{\sqrt{v/2}} + \psi_\Theta\,(X_M|X_t)\right)\right)$$

for $t \in [T]$ is bounded and valid.

Proof. We prove boundedness by showing that $0 \leq \rho(\xi_t) \leq c$. We further prove validness by checking that $\forall t \in [T], \rho(\xi_t) \geq \omega(\xi_t)$, which follows from Theorem 1 and the fact that the tanh function is nondecreasing. □

We finally take a closer look at the format of the query Q, and define the total privacy compensation of the data owner. In addition to a tolerable variance of noise v, Q also contains a general function f. Here, we consider that f should explicitly express its input $X_Q \subset X$ rather than the entire time-series data X. For example, X_Q can be further designated by a pair of starting and ending points together with a sampling period. In particular, such a query setting can change the initial distributions and transition matrixes of Markov chains, and thus affect the privacy compensation functions defined in Theorem 2 and Theorem 3. By incorporating the queried states X_Q, we give the definition of total privacy compensation.

Definition 6. *Let X_Q be the set of states in the query Q. The total privacy compensation of the data owner in Q is:*

$$\rho\,(Q) = \sum_{X_t \in X_Q} \rho(\xi_t),$$

where $\rho(\xi_t)$ is the unbounded (resp., bounded) and valid privacy compensation at time t in Theorem 2 (resp., Theorem 3).

C. Pricing of Query

We present the last component of HORAE, namely the pricing mechanism for the data consumer.

We first identify three desirable economic properties.

Definition 7. *Let $\pi(Q)$ be a valid pricing function for the query Q. $\pi(Q)$ should satisfy:*

- *Fairness: If $X_Q = \varnothing$, $\pi(Q) = 0$.*
- *Balance: $\pi(Q) \geq \rho(Q)$.*
- *Arbitrage freeness: If Q determines Q', $\pi(Q) \geq \pi(Q')$.*

We give some comments on these properties: (1) Fairness says that if no states are queried, the data broker should charge zero price; (2) Balance is to guarantee non-negative utility at the data broker; (3) The intuition behind arbitrage freeness is that if there exists arbitrage in $\pi(\cdot)$, *e.g.*, $\pi(Q) < \pi(Q')$, then the data consumer would never pay the full price of Q'. Instead, she would buy a cheaper query Q to answer Q'.

Given the fact that the *basic pricing function*, *i.e.*, setting the query price $\pi(Q)$ to be the total privacy compensation $\rho(Q)$, can guarantee the first two properties in Definition 7, we shall focus on arbitrage freeness. To investigate arbitrage freeness, the key issue is to determine whether a query can be derived from others. Such a concept of determinacy relation has been established in randomized query/statistic answering [10], [11], where the function f can have different Lipschitz parameters, and applies to the whole dataset X. Complementary to these works, we here consider f with a certain Lipschitz parameter ℓ, but over any subset X_Q of X. For brevity, we use the queried states X_Q to specify the data consumer's requested function f, *i.e.*, $Q = (X_Q, v)$. Additionally, we give the the formal definition of the determinacy relation in our new context.

Definition 8. *The determinacy relation is between $Q = (X_Q, v)$ and $Q' = (X_{Q'}, v')$. We say that Q determines Q', if either of the following conditions holds:*

- *Less noise: $X_Q = X_{Q'}, v \leq v'$.*
- *More states: $X_{Q'} \subset X_Q, v = v'$.*

Based on the determinacy relation, we consider how to guarantee arbitrage freeness in a pricing function $\pi(Q)$. First is with respect to the queried states X_Q. A trivial example is the basic pricing function $\pi(Q) = \rho(Q)$ raised above. If we regard the elementary part of $\pi(Q)$, namely the privacy compensation $\rho(\xi_t)$ for a state $X_t \in X_Q$, as the price of an item, we find that the basic pricing function belongs to item pricing [23], and can inherently guarantee arbitrage freeness with respect to X_Q. Second is about the variance of noise v. Intuitively, $\pi(Q)$ should monotonically decrease with v, but the thorniest challenge is how fast it can decrease with v. To figure out the boundary function, we formulate the arbitrage attack raised in Section I as our motivating example.

Example 1. *A data consumer, who wants to obtain the query (X_Q, v) with a lower price, may turn to buying n other cheaper queries of the same states X_Q but with higher variances, denoted as $\{(X_Q, v_i)|i \in [n], v_i > v\}$. Afterwards, the data consumer computes the average of n answers, and gets an unbiased result $f(X_Q)$, but with a lower variance $\frac{1}{n^2}\sum_{i=1}^{n} v_i$. If the pricing function $\pi(\cdot)$ is arbitrage free, then the following conditional statement must hold: $\frac{1}{n^2}\sum_{i=1}^{n} v_i \leq v \Rightarrow \sum_{i=1}^{n} \pi\,(X_Q, v_i) \geq \pi\,(X_Q, v)$.*

We further give Lemma 2 to thwart the above attack, and put its detailed proof into our technical report [22].

Lemma 2. *For any arbitrage-free pricing function $\pi(Q)$ that depends on two independent parts X_Q and v, it decreases not faster than $1/v$.*

According to Lemma 2, we revise the basic pricing function $\pi(Q) = \rho(Q) = \sum_{X_t \in X_Q} \rho(\xi_t)$. We recall that the valid privacy compensation function $\rho(\xi_t)$ hinges on the upper bound of privacy loss, namely $\min_{X_M \in S_{M,t}}((\ell \cdot \mathsf{card}(X_C))/\sqrt{v/2} + \psi_\Theta(X_M|X_t))$. Here, we can observe that for a certain state $X_t \in X_Q$, if the variance of noise v changes, this upper bound may select a different Markov quilt $X_M \in S_{M,t}$ to achieve the minimum value. In other words, the change of v in the basic pricing function is not independent, which may break the less noise rule in Definition 8, and thus make reasoning about arbitrage freeness with respect to v extremely hard. To handle this problem, we fix the Markov quilt $X_{M,t}$ (with $X_{C,t}, X_{R,t}$) for each state X_t in our advanced pricing function. Besides,

we provide two strategies for the data broker to determine a sequence of stable Markov quilts:

- *Randomized strategy:* She randomly fixes a sequence of Markov quilts, such that the max-influence of each state on its Markov quilt is finite.
- *Leading strategy:* She searches for a leading query setting, and uses its sequence of Markov quilts for all the other query settings.

Our proposed solution above has three nice properties: (1) The assumption of Lemma 2 is satisfied, and can be used to definitely rule out arbitrage opportunities involving the variance of noise v; (2) Due to the fact that any element in a set is no less than the minimum element in the same set, the advanced pricing function $\pi(Q)$, now regarded as an upper bound of the total privacy compensation $\rho(Q)$, is balanced or even profitable. Specifically, the data broker can well control her profit ratio with the leading strategy, where it is 0 at the leading setting, and increases when the data consumer's concrete query setting deviates from the leading setting. We shall elaborate more on this in Section IV-B; (3) Just as the basic pricing function, the advanced pricing function, falling into the category of item pricing, can still ensure arbitrage freeness involving the queried states X_Q.

Specific to two kinds of valid privacy compensation functions in Theorem 2 and Theorem 3, we present the corresponding advanced pricing functions in Theorem 4 and Theorem 5, respectively. In particular, the first advanced pricing function sets an infinite price for the unperturbed answer.

Theorem 4. *For the privacy compensation function in Theorem 2, the following advanced pricing function*

$$\pi(Q) = c \sum_{X_t \in X_Q} \left(\frac{\ell \cdot \mathsf{card}\left(X_{C,t}\right)}{\sqrt{v/2}} + \psi_\Theta\left(X_{M,t} | X_t\right) \right)$$

is unbounded and valid.

Proof. First, we prove fairness by checking that $X_Q = \varnothing \Rightarrow \pi(Q) = 0$. Second, we prove balance by checking that $\pi(Q) \geq \rho(Q)$, which follows from the minimum function. Third, we prove arbitrage freeness. For v, we prove by Lemma 2, *i.e.*, $\pi(Q)$ decreases with $1/\sqrt{v}$, which is slower than $1/v$. For X_Q, we prove by Definition 8. We can check that $X_{Q'} \subset X_Q, v = v' \Rightarrow \pi(Q) \geq \pi(Q')$, which follows from: $X_Q \backslash X_{Q'} \neq \varnothing \Rightarrow \pi(Q) - \pi(Q') = c \sum_{X_t \in X_Q \backslash X_{Q'}}(\cdot) \geq 0$. □

Yet, with the following advanced pricing function, the data broker can sell the unperturbed answer at some finite price.

Theorem 5. *For the privacy compensation function in Theorem 3, the following advanced pricing function*

$$\pi(Q) = c \sum_{X_t \in X_Q} \tanh\left(d \left(\frac{\ell \cdot \mathsf{card}\left(X_{C,t}\right)}{\sqrt{v/2}} + \psi_\Theta\left(X_{M,t} | X_t\right) \right) \right)$$

is bounded and valid.

Proof. The proof differs from that of Theorem 4 in that $\pi(Q)$ decreases with $\tanh(1/\sqrt{v})$, which is slower than $1/v$. □

TABLE I
UPPER BOUND OF PRIVACY LOSS PER TIMESTAMP.

Period	Data Owner 1	Data Owner 2	Data Owner 3	Data Owner 4
1 second	*59.391 (8.070)	62.588 (7.815)	68.329 (8.485)	66.141 (8.801)
1 minute	30.958 (2.763)	31.032 (2.697)	30.026 (2.899)	31.826 (2.946)
1 hour	6.972 (0.295)	6.919 (0.265)	8.203 (0.341)	10.360 (0.528)

*Each entry is stored in the format: mean (standard deviation).

IV. EVALUATION RESULTS

In this section, we focus on the physical activity monitoring scenario, and present the evaluation results of HORAE in terms of fine-grained privacy compensation and economically-robust pricing of query, respectively.

Dataset: We use the public Activity Recognition with Ambient Sensing (ARAS) dataset [24]. This dataset contains 30-day real-world monitoring data, including sensor readings and activity labels, from 2 houses with 2 residents per house. In addition, the original sampling period is 1 second, and the size of the whole activity occurrences is 10,368,000. Moreover, 27 kinds of daily activities were recorded, *e.g.*, sleeping, studying, shaving, having dinner, and listening to music.

Setups: We treat 4 residents as 4 different data owners, and index them from 1 to 4, where the first/last two data owners live in the same house. Besides, we regard 27 kinds of activities as the state space A. Moreover, we consider Markov chains with different parameters, by varying the size and sampling period of a certain data owner's activity data. For a concrete pair of size and sampling period, we cover all possible starting points, which can help to verify the effect of different ending points. The implementation code in Matlab is online available from [25].

A. Fine-grained Privacy Compensation

In this section, we show privacy loss and privacy compensation in physical activity monitoring.

1) Privacy Loss: Before investigating privacy compensation, we first show the upper bounds of the privacy losses of four data owners per timestamp. Table I lists their mean values and standard deviations, where the sampling period ranges from 1 second, to 1 minute, and to 1 hour. Besides, all Markov chains take the fixed length of 720, which is the maximum size of a certain data owner's activity data at the sampling period of 1 hour. Furthermore, the Lipschitz parameter ℓ is set to be 1, and the variance of noise v is fixed at 10, which gives an error of 10 with 90% confidence by Chebyshev's inequality.

From Table I, we can see that when the sampling period becomes longer, the upper bound of a certain data owner's privacy loss decreases. We explain the reason from data correlations. A longer sampling period tends to imply weaker data correlations, and thus less privacy leakage. We can also see from Table I that the upper bounds of the first/last two data owners' privacy losses are consistent in general. This is because either pair of data owners, living in the same house, can share similar activity patterns and thus privacy losses.

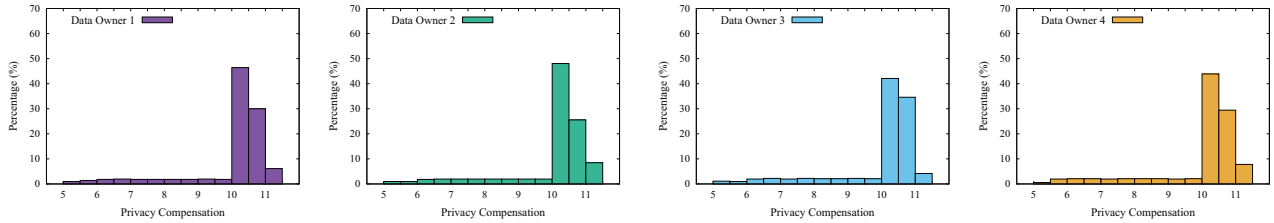


Fig. 2. Unbounded and Valid Privacy Compensation in Physical Activity Monitoring.

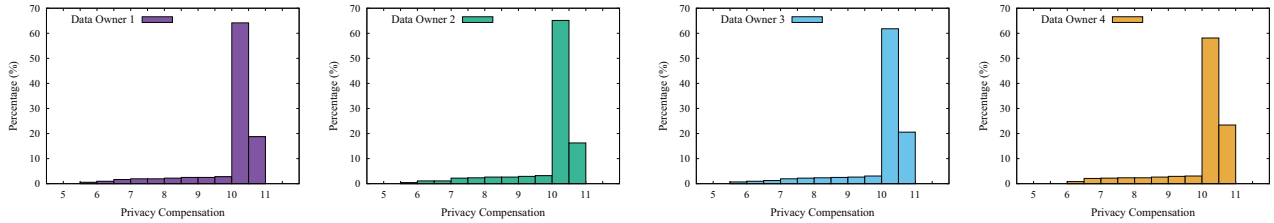


Fig. 3. Bounded and Valid Privacy Compensation in Physical Activity Monitoring.

2) Privacy Compensation: We now focus on the original sampling period, and further explore how privacy compensations are allocated at different timestamps. For clarity in presentation and comparison, we fix the total privacy compensation of each data owner, such that her privacy loss at one timestamp is compensated with 10 units in average, *i.e.*, $\rho(Q) = 10 \times \mathsf{card}(X_Q)$. Besides, we consider both bounded and unbounded privacy compensation functions. For the bounded one, we choose the parameter d for each data owner, such that the terms within tanh are scaled down to 1 or below. Fig. 2 and Fig. 3 plot the evaluation results, where a pair of neighboring x-axis ticks denote a half-closed interval, *e.g.*, the bin from "10" to "10.5" stands for the privacy compensations between 10 and 10.5 excluding 10.5.

First, we can see from Fig. 2 and Fig. 3 that each data owner may obtain distinct privacy compensations for her privacy losses at different timestamps, rather than the uniform 10 units under entry/group differential privacy. The reason is that by Lemma 1, the states of a data owner at different timestamps have distinct sets of Markov quilts, and may have different max-influences on the Markov quilts, which jointly imply distinct privacy compensations. Second, we compare two kinds of valid privacy compensations for a certain data owner, and find that more privacy compensations fall into the center region between 10 and 10.5 under the bounded privacy compensation. This outcome truly reflects the difference between the linear and tanh functions utilized in the unbounded and bounded privacy compensations, respectively. The tanh function is less sensitive to the changes of its variable than the linear function. Third, we compare the privacy compensations of the first/last two data owners in Fig. 2 or Fig. 3, and find them consistent in general. Besides, compared with the first two data owners, more privacy compensations of the last two data owners deviate from the center region. These results conform to the standard deviations at the original sampling period in Table I.

The above evaluation results demonstrate that HORAE can indeed compensate the data owners for their privacy losses at different timestamps in a more fine-grained way.

B. Economically-robust Pricing of Query

In this section, we show the pricing of query from queried states, variance of noise, and arbitrage freeness, respectively. For brevity, the pricing functions hereinafter refer to the advanced ones by default. Nevertheless, when comparing with the basic ones, we shall reserve "advanced" for clarity.

To quantify the profitability of the data broker, we first introduce a common metric from economics, called profit ratio, which is defined as the ratio between revenue minus cost and cost. Under our data market model, we regard the payment $\pi(Q)$ from the data consumer as revenue, and view the total privacy compensation $\rho(Q)$ allocated to the data owner as cost. Therefore, the profit ratio here is equal to $\pi(Q)/\rho(Q) - 1$.

We next note that compared with the unbounded pricing of query in Theorem 4, the bounded one in Theorem 5 further applies a scaling factor d and a tanh function, which can mitigate the differences among the evaluation results of four data owners. This characteristic has been depicted in the above privacy compensation. Therefore, in what follows, we only present the evaluation results of unbounded pricing of query.

1) Queried States: We start with the queried states X_Q in a pricing function, by varying its size and sampling period separately. In particular, when evaluating the size of queried states, we use the original sampling period, and take the randomized strategy to determine the sequence of stable Markov quilts for the pricing function. In contrast, when evaluating the sampling period, we fix the size of queried states at 250, which is roughly the maximum length of Markov chains at the sampling period of 10000 seconds. Besides, we take the leading strategy at the original sampling period, where the leading variance of noise is 10. Furthermore, for consistency with privacy compensation, we use the same values of the other parameters in Section IV-A. The first two subfigures in Fig. 4 plot the evaluation results.

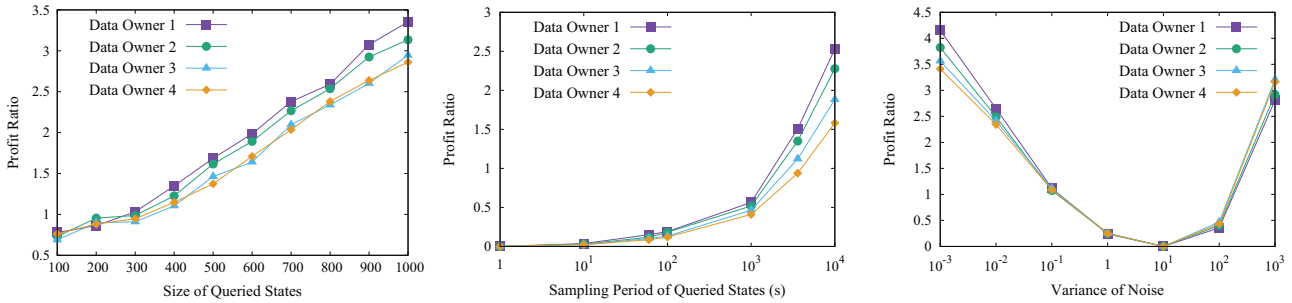


Fig. 4. Advanced, Unbounded, and Valid Pricing of Query in Physical Activity Monitoring.

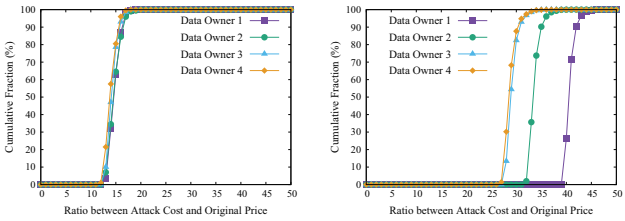
From the leftmost subfigure in Fig. 4, we can see that the profit ratio increases with the size of queried states. This is because by Theorem 2, with more queried states, a larger size of Markov quilt set $S_{M,t}$ can be searched to obtain a lower privacy compensation $\rho(\xi_t)$ for each state X_t. In contrast, when determining the query price $\pi(Q)$, we fix the Markov quilt $X_{M,t}$ for X_t, which indicates that the part of $\pi(Q)$ involving X_t is almost irrelevant with the size of queried states. Therefore, the profit ratio, depending on $\pi(Q)/\rho(Q)$, grows with the size of queried states.

From the middle subfigure in Fig. 4, one key observation is that the profit ratio is 0 at the original sampling period. This outcome stems from that the leading strategy is taken here, and thus the sequence of Markov quilts in the pricing of query is the same as that in the privacy compensation, which further implies that the query price $\pi(Q)$ is equal to the total privacy compensation $\rho(Q)$. The second key observation is that the profit ratio increases with the sampling period. The reason is that when a sampling period is farther away from the leading sampling period of 1 second, the difference between the sequence of stable Markov quilts in the pricing of query and the sequence of Markov quilts in the privacy compensation becomes larger, which implies a higher profit ratio.

2) Variance of Noise: We continue to examine the other part of a pricing function, namely the variance of noise v. The rightmost subfigure in Fig. 4 plots the evaluation results. Here, we take the original sampling period, and keep the other settings the same as those in evaluating the sampling period, *i.e.*, we take the leading strategy with leading variance of 10.

From the rightmost subfigure in Fig. 4, we can observe that the four curves of profit ratio convex downward. In particular, the minimum profit ratio is 0 at $v = 10$. This coincides with the leading strategy adopted here. Besides, when v deviates from 10, the profit ratio increases. The reason is analogous to that in evaluating the sampling period, where the difference is that the changes of v here can be bi-directional.

We now give some comments on how the data broker can exploit the convexities of the profit ratio curves. On one hand, if the data broker releases her leading strategy to the public, she can guide the data consumers to buy the query with the leading setting, *e.g.*, a certain pair of sampling period and variance of noise. On the other hand, if the leading strategy is hidden, the data broker can carefully craft the strategy, *e.g.*, through learning the query histories of the data consumers, such that her expected profit ratio is maximized.



(a) Basic (b) Advanced

Fig. 5. Arbitrage Freeness in Unbounded and Valid Pricing of Query.

3) Arbitrage Freeness: We finally investigate arbitrage freeness in the pricing of query by simulating the attack in Example 1. After simulating 10000 samples, we plot the cumulative fraction of the ratio between the attack cost $\sum_{i=1}^{n} \pi(X_Q, v_i)$ and the original price $\pi(X_Q, v)$ in Fig. 5, where $\pi(\cdot)$ can be basic or advanced unbounded pricing function. In addition, we note that the cumulative fraction 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. 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. 5, we can see that both basic and advanced pricing functions are arbitrage free. We can also observe that an attempt of finding arbitrage in the advanced pricing function is expected to be more costly than that in the basic one, which can be roughly captured by the areas above a certain data owner's two function curves in Fig. 5(b) and Fig. 5(a), respectively. Let's examine the case of the first data owner in detail. To launch an arbitrage attack in the advanced pricing function, the attacker is most likely (with 45.35% probability) to pay the amount of 40 to 41 times the original price. In contrast, the most possible case in the basic pricing function is to pay the amount of 14 to 15 times the original price with 31.14% probability. Therefore, in the sense of defending against arbitrage, the advanced pricing function can be more robust than the basic one. This coincides with our theoretic analysis in Section III-C.

The above evaluation results illustrate that the pricing mechanism in HORAE can preserve good profitability and arbitrage freeness. In particular, the data broker can well control her profit ratio, by developing the leading strategy. Besides, it is infeasible for data consumers to game the data market.

V. Related Work

In this section, we briefly review related work.

A. Data Market Design

An explosive demand for sharing data contributes to growing interest in data market design. The researchers from the database community mainly focus on arbitrage freeness in pricing queries over relational databases [16], [26]. Specific to personal data trading, Ghosh and Roth [9] considered differential privacy as a commodity, and proposed to selling privacy at auction for single counting query. The follow-up work by Li *et al.* [10] further extended to multiple linear queries by introducing arbitrage freeness. We proposed to trade noisy aggregate statistics over private data with correlations at the individual level [11]. Different from these data trading works, Wang *et al.* [27] focused on the data collection process, and measured the value of privacy through a game-theoretic model. Zhang *et al.* [28] considered crowdsourcing based image collection, and studied the ownership and privacy issues.

However, none of the above works has taken temporal correlations into account, and further considered privacy compensation and pricing of query for time-series data.

B. Pufferfish Privacy over Correlated Data

The classical differential privacy framework, proposed by Dwork *et al.* [12], [15], allows trusted data curators to add appropriate noises to aggregate results before releasing them, which can protect an individual's private information. However, as pointed by Kifer and Machanavajjhala [29], when there exist correlations among data items, the perturbation in differential privacy can be inadequate. They thus proposed a generalized version of differential privacy, called Pufferfish privacy [19]. Many follow-up research works have been going on around this particular issue. In addition to the Markov quilt mechanism utilized in this work, Liu *et al.* [13] proposed dependent differential privacy for correlations between individuals. Yang *et al.* [30] focused on the correlation structure modeled by Gaussian Markov random fields. Xiao *et al.* [31] considered continuous location sharing, and employed Markov chains to model temporal correlations.

The original intention of these works is preserving privacy rather than pricing privacy, which is our major focus.

VI. Conclusion

In this paper, we have proposed the first framework HORAE for trading private data with temporal correlations. In HORAE, the data owners can be compensated for their privacy losses in a satisfying and fine-grained way. Besides, the data broker can well control her profit ratio. Moreover, the data consumers have to faithfully purchase their desired queries rather than gaming the system. We have instantiated HORAE with physical activity monitoring, and extensively evaluated its performance on the ARAS dataset. Evaluation and analysis results have demonstrated the feasibility of HORAE.

References

[1] P. Nguyen, X. Zhang, A. C. Halbower, and T. Vu, "Continuous and fine-grained breathing volume monitoring from afar using wireless signals," in *Proc. of INFOCOM*, 2016.

[2] K. Qian, C. Wu, F. Xiao, Y. Zheng, Y. Zhang, Z. Yang, and Y. Liu, "Acousticcardiogram: Monitoring heartbeats using acoustic signals on smart devices," in *Proc. of INFOCOM*, 2018, pp. 1574–1582.

[3] J. Ryoo, Y. Karimi, A. Athalye, M. Stanacevic, S. R. Das, and P. M. Djuric, "BARNET: towards activity recognition using passive backscattering tag-to-tag network," in *Proc. of MobiSys*, 2018, pp. 414–427.

[4] S. Tang, Q. Huang, X.-Y. Li, and D. Wu, "Smoothing the energy consumption: Peak demand reduction in smart grid," in *Proc. of INFOCOM*, 2013, pp. 1133–1141.

[5] "ThingSpeak," https://thingspeak.com/, 2010.

[6] "Datacoup," https://datacoup.com/, 2012.

[7] FTC, "Data brokers: A call for transparency and accountability," 2014.

[8] 60 Minutes on CBS News, "The data brokers: Selling your personal information," 2014.

[9] A. Ghosh and A. Roth, "Selling privacy at auction," in *Proc. of EC*, 2011, pp. 199–208.

[10] 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.

[11] C. Niu, Z. Zheng, F. Wu, S. Tang, X. Gao, and G. Chen, "Unlocking the value of privacy: Trading aggregate statistics over private correlated data," in *Proc. of KDD*, 2018, pp. 2031–2040.

[12] 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.

[13] C. Liu, S. Chakraborty, and P. Mittal, "Dependence makes you vulnberable: Differential privacy under dependent tuples," in *NDSS*, 2016.

[14] M. Hardt and A. Roth, "Beyond worst-case analysis in private singular vector computation," in *Proc. of STOC*, 2013, pp. 331–340.

[15] 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.

[16] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," in *Proc. of PODS*, 2012, pp. 167–178.

[17] The New York Times, "Facebook is not the problem. lax privacy rules are." 2018.

[18] ——, "Facebook Hack Included Search History and Location Data of Millions," 2018.

[19] D. Kifer and A. Machanavajjhala, "A rigorous and customizable framework for privacy," in *Proc. of PODS*, 2012, pp. 77–88.

[20] S. Song, Y. Wang, and K. Chaudhuri, "Pufferfish privacy mechanisms for correlated data," in *Proc. of SIGMOD*, 2017, pp. 1291–1306.

[21] C. Dwork, G. N. Rothblum, and S. P. Vadhan, "Boosting and differential privacy," in *Proc. of FOCS*, 2010, pp. 51–60.

[22] "Technical Report for HORAE," https://www.dropbox.com/s/dkx6a4c2vw901pe/, 2018.

[23] M. Balcan, A. Blum, and Y. Mansour, "Item pricing for revenue maximization," in *Proc. of EC*, 2008, pp. 50–59.

[24] "ARAS Dataset," https://www.cmpe.boun.edu.tr/aras/, 2013.

[25] "Source Code for HORAE," https://github.com/NiuChaoyue/INFOCOM-2019-HORAE, 2018.

[26] S. Deep and P. Koutris, "QIRANA: A framework for scalable query pricing," in *Proc. of SIGMOD*, 2017, pp. 699–713.

[27] 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.

[28] L. Zhang, Y. Li, X. Xiao, X.-Y. Li, J. Wang, A. Zhou, and Q. Li, "CrowdBuy: privacy-friendly image dataset purchasing via crowdsourcing," in *Proc. of INFOCOM*, 2018, pp. 2735–2743.

[29] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proc. of SIGMOD*, 2011, pp. 193–204.

[30] B. Yang, I. Sato, and H. Nakagawa, "Bayesian differential privacy on correlated data," in *Proc. of SIGMOD*, 2015, pp. 747–762.

[31] Y. Xiao and L. Xiong, "Protecting locations with differential privacy under temporal correlations," in *Proc. of CCS*, 2015, pp. 1298–1309.