Privacy-Aware Data Acquisition under Data Similarity in Regression Markets

Shashi Raj Pandey, *IEEE Member*, Pierre Pinson, *IEEE Fellow*, and Petar Popovski, *IEEE Fellow*

Abstract

Data markets facilitate decentralized data exchange for applications such as prediction, learning, or inference. The design of these markets is challenged by varying privacy preferences as well as data similarity among data owners. Related works have often overlooked how data similarity impacts pricing and data value through statistical information leakage. We demonstrate that data similarity and privacy preferences are integral to market design and propose a query-response protocol using local differential privacy for a two-party data acquisition mechanism. In our regression data market model, we analyze strategic interactions between privacy-aware owners and the learner as a Stackelberg game over the asked price and privacy factor. Finally, we numerically evaluate how data similarity affects market participation and traded data value.

Index Terms

information leakage, regression markets, collaborative learning, mechanism design, Stackelberg game

I. Introduction

A. Context and Motivation

In recent years, there has been a surge in Internet of Things (IoT) devices with sensing and computing capabilities, leading to an abundance of IoT data. Massively distributed heterogeneous

Shashi Raj Pandey and Petar Popovski are with the Connectivity Section, Department of Electronic Systems, Aalborg University, Denmark. Email: {srp, petarp}@es.aau.dk.

Pierre Pinson has primary affiliation with Dyson School of Design Engineering, Imperial College London, UK. Email: {p.pinson}@imperial.ac.uk. He is also affiliated to the Technical University of Denmark, Department of Technology, Management and Economics, as well as with Halfspace

This work was supported by the Villum Investigator Grant “WATER” from the Velux Foundation, Denmark.

data fuels various emerging applications, such as forecasting and analytics [1], learning models [2], [3], and in diverse industry verticals [4], [5]. It is thus critical to investigate the ways in which this data becomes available, respecting the privacy constraints and/or offering incentives to the data owners. *Data markets* act as platforms that facilitate the collection, exchange, and utilization of both personal and IoT data. In this study, the focus is on *regression data markets* [1], [6], which addresses the regression problem in which data is distributed across multiple agents. We motivate the use and operation of regression data markets through two examples.

Example 1: Consider a labour market where agents Alice and Bob are generating privacy-sensitive distinct explanatory features $\mathbf{x}$ that may explain a common target variable y. Say, Alice wants to learn a regression model that quantifies the target value y, e.g., *hiring salary* using features $\mathbf{x}$ (such as age, gender, academic qualifications, etc.). This is effectively done through learning the regression mapping function that involves features obtained from Bob as input. However, Bob's disclosure of features in a truthful manner leads to privacy loss and requires reasonable monetary compensation from Alice. Another agent, Carol, enters the market with an extensive history of employee hiring experiences, gathering similar features held by Bob to explain y. However, she is constrained by her company's intellectual property (IP) regulations, preventing *direct* (unaltered) sharing of features with Alice and instead employing privacy guarantees and appropriate compensation. Bob's features are correlated with Carol's, and his willingness to share affects Carol's decisions, while Bob and Carol have different privacy preferences.

Example 2: Consider a startup company offering a location-based ride-sharing service. There exist publicly accessible cameras at bus stations to capture passenger activities, owned by online platforms such as "NAVER" [7] and "Kakao" [8]. Arrangements are made to uphold passenger privacy and confidentiality. The data derived from these cameras are of poor quality and are intentionally non-real-time due to security considerations. The startup wants to attain accuracy in predicting the volume of ride-sharing requests. Prediction accuracy hinges upon diverse passenger features, such as age, gender, number of passengers, and more. The setting also includes security companies, and owners of cameras that monitor both these bus stops and nearby taxi stands. The startup can incentivize the camera owners to retrieve data from these sources and enhance its predictions. However, the data streams from these disparate cameras, although resembling each other, necessitate measures to protect passenger privacy, which constrains the attainable prediction accuracy.

A regression data market allows data-driven analysis using a conditional model that explores the dependency between the feature vector and the target value of interest. In Example 1, Alice acts as a *learner* and can initiate a collaboration protocol to build a regression model, soliciting features from Bob and Carol. Alice can employ any regression mapping functions to the distributed features, subsequently assessing the collaboration's merit using a chosen convex loss function, such as mean square error (MSE). Both Bob and Carol share heterogeneous privacy preferences and offer data of different qualities, such that Alice must incentivize both of them, employing tailored pricing signals. In Example 2, companies owning different cameras observe similar data and are market competitors. These companies may agree to provide their data to the startup in a manner that is equitable and ensures data privacy. Their aim is to limit the exposure of data similarities and the ensuing ramifications, such as data value depression [9], [10], while simultaneously meeting passenger privacy mandates. However, privacy breaches can arise from collaboratively computed target values or direct data exchanges among agents. This includes questions about who controls IoT devices and maintains control over the collected data as well as whether strategic agents can even exert influence over the outcome of computations.

In both examples, the agents are sensitive towards information leakage due to data similarity. Consequently, they might opt for strategies like secure data computation methods (e.g., MPC [11], [12]) or leverage differential privacy (DP) techniques [13], [14] to execute statistical analysis while preserving privacy. Nevertheless, the employment of privacy mechanisms, such as DP, invariably leads to a trade-off between data privacy and the accuracy of learned statistical models. The application of rigorous data privacy techniques during data exchange can potentially compromise the model's efficacy, with diminishing accuracy, particularly in situations characterized by increased data similarity.

Thus, the learner is challenged with the optimization of query signals to extract distributed features. Two critical aspects should be considered: (i) the value brought forth by each agent's features in addressing the regression problem and (ii) the impact of correlated features. These considerations are bounded by constraints pertaining to pricing budgets and the inherent diversity in agent privacy preferences. This brings us to the main question addressed by the present study: *"How to find the optimal trade-off between data privacy and agent's utility under data similarity in a regression data market?"*

B. Related Work

Existing literature [1], [6], [9], [15] on data markets explore various interactions among agents, formulating algorithmic solutions. The interplay between data privacy and ownership significantly influences the broader data trading process and the quality of offered services provided through data exchanged. This duality, often termed the *privacy-utility* trade-off [13], [16], can be addressed through incentives that balance data privacy considerations without compromising utility. In the context of a regression data market, an effective incentive structure becomes pivotal in aligning the strategies of distributed data sources towards a shared goal, like training a learning model [2], [3], [17]. The overall operation is challenged by the data characteristics, competition [9], [10], privacy requirements and the computing capabilities of agents.

Recent works on IoT data markets focus on data acquisition models leading to efficient incentive mechanism designs. The objective is to improve participation for various applications, such as training a learning model [3], [18] and data utility maximization. Auction-based designs are studied in the data market design, leading to truthful mechanisms for data exchanges. Numerous privacy-preserving distributed model training approach exists where updates on local raw data are shared rather than the data itself, as done in Federated Learning (FL) [2], [17]. Alternatively, variants of DP methods are implemented [13], [19], where the agents' supply features they have for training at a central entity after adding noise proportional to privacy sensitivity. Therein, it is required to quantify the value of an individual agent's contribution in lowering the prediction error of the trained regression model. However, strategic interaction between agents having heterogeneity in local privacy preferences and developing a participation mechanism that respects the impact of the unknown degree of information leakage due to data similarity (as hinted in Example 2) has not been explored yet in a regression market setting. In [10], the authors proposed a game-theoretic setup that enables distributed coalition amongst devices with similar data properties to minimize information leakage to prevent adversarial under-pricing and data rivalry issues in the data market. However, they ignore the influence of heterogeneity in individual privacy preferences. In [20], the authors introduced a *Bayesian regression markets* design, which is compatible when considering a more general class of regression tasks. The work builds the mechanism by adopting a Bayesian framework that offers a fair allocation of market revenue for data trading. Regarding the literature gap, the existing works [1], [10] mostly focus on incentive design, participation, and the value of data trading. Here, instead, we look into the optimal participation strategy of privacy-sensitive data holders in order to realize a regression data market under information leakage due to data similarity.

C. Contributions and Paper Structure

We model a data acquisition method where we jointly analyze the impact of data similarity, particularly correlation, on statistical information leakage in a privacy-aware finite player regression data market. We develop incentive strategies that enforce loss minimization objectives during collaborative data trading for solving the regression problem. This is done by taming agents to provide high-quality data through incentives. We propose an incentive design that offers control of the heterogeneous privacy factors on the distributed data and elicits them to the *learner*. For the offered pricing and privacy budget, we summarize the strategic interaction between the learner and the agents following a single-leader, multiple-followers Stackelberg game structure, with the learner acting as a leader and the devices as followers. We show the Nash best response strategies of the agents lead to a unique equilibrium in the non-cooperative game amongst followers. With devices employing the local DP technique to trade data, we show there exists a trade-off between the available data privacy budget and the value of device participation. Finally, we show extensive numerical evaluations to verify these observations.

The paper is structured as follows. Section II introduces the system model and problem setup. Section III develops the interaction framework between agents in the regression data market as a two-stage Stackelberg game. This involves designing utility models and the mechanism design with a first-order, low-complexity iterative solution to the posed problem. Section IV provides a performance evaluation of the proposed algorithm with extensive numerical results and shows comparative analysis with the intuitive baselines. Finally, Section V concludes this work with discussions on the future outlook.

II. System Model and Problem Setup

We consider a network comprised of agents accumulating distinct features that, in principle, contribute to learning the parameters of a regression model [1]. These agents can communicate with each other through a platform that hosts a data market for training a regression model: a *regression data market*. Then, any one of the interested agents can position herself as the "learner" and initiate the exchange of the feature observations. In particular, the learner can send a query request to incorporate high-quality features from other agents to improve the overall prediction accuracy of her model. Meanwhile, as agents are privacy-aware, they execute differential privacy methods (details on DP in Definition 1, Section III-A) to avoid potential privacy loss during data trading with the learner. Therefore, to maintain the quality of the collected features and the utility they bring, the learner specifically asks agents to abide by a common privacy measure (i.e., the *privacy factor*) in sharing their features for an offered pricing. However, in practice, these devices exhibit heterogeneity in their data privacy preference, *unknown* to the learner, as opposed to the commonly asked privacy factor. This suggests the strategic interaction of agents with the learner. Further, the features they hold might be correlated, but to an *unknown* degree, resulting in possible information leakage during features trading. In principle, correlated data contribute less to improving the performance of the trained model. Therefore, not all agents but a subset of available agents in the platform participate strategically in training the regression model by supplying high-quality features. In the following, we formalize the problem setup.

Agents. Like [1], we consider a regression market consisting of a set of n agents $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$. One of these agents i, called *learner* a_i, is trying to fit a regression model based on its own features as well as features bought from other agents. The learner is initiating interaction with the other agents through the platform. Following the naming convention in [1], we call the learner a_i the *central agent* and the remaining agents $a_j, j \neq i$ the *support agents*. The term "data samples" means "features" in the context of regression tasks and will be used interchangeably throughout this work. Without loss of generality, the learner is a_1.

We assume that data samples are collected by agents at discrete points in time $t \in \{1, 2, \ldots, \tau\}$. The central agent has a response target variable $Y_t \in \mathcal{Y}$, denoted for each time instance, and is trying to train a regression model that allows it to understand some statistics of $\{Y_t\}$, collected over τ time instances. Then, a time series collection $\{y_t\}$ is obtained, which denotes the realization of target variable $\{Y_t\}$, one for each time instance. The regression model relies on several explanatory variables (commonly called input features) indexed by a set Ω. Consider K explanatory variables such that $\Omega = \{x_k, k = 1, \ldots, K\}$ and denote $x_{k,t} \in \mathbb{R}, \forall k, t$, be the x_k feature observed[1] at time instance t. Correspondingly, $\mathbf{y} = [y_1, \ldots, y_\tau]^\top$ is the collected target variable. Following our assumption that the features are observed at each time instance, we denote $\mathbf{x}_k = [x_{k,1}, \ldots, x_{k,\tau}]^\top$ be the vector of all observed values for the feature x_k and

[1]For simplicity, we restrict $x_{k,t} \in \mathbb{R}$.

$\mathbf{x}_t = [x_{1,t}, \ldots, x_{K,t}]^\top$ be the vector of all features observed at time instance t. Each agent in $\mathcal{A}$ holds a subset of relevant explanatory variables $\omega \subset \Omega$ and $\mathbf{x}_{\omega,t}$ the vector of features at time instance t. We assume that the data is complete and non-redundant, i.e., for each explanatory variable at each time step, a unique agent is holding the corresponding feature. However, we will not assume these explanatory variables are uncorrelated with each other, i.e., there is a possibility of having data similarity between samples of different agents. If we closely look at this setting, it is a setup in the vertical model training setting of federated optimization [21], where the features are distributed across agents. This requires the acquisition of data distributed amongst several privacy-aware agents to identify the explanatory variable over time, which is our focus in this work. We denote $X_\omega \in \mathbb{R}^{\tau \times |\omega|}$ as the design matrix, whose column denotes the features observed, and the t'th row is $\mathbf{x}_{\omega,t}^\top$.

Learning Model. The central agent a_i holds a set of features $\omega_i \subset \Omega$ and the target variable y. In a general setting, each agent $j \in \mathcal{A} \backslash a_i$ has a set of features $\omega_j \subset \Omega$, such that $|\omega_i| + \sum_j |\omega_j| = K$. As the features of potential relevance are distributed amongst the supporting agents j, at a regular time interval or at every time instance, the central agent aims to obtain those features to maximize its prediction ability on the target variable. Then, the regression problem of the central agent is to describe a mapping function f between the set of explanatory variables $\omega \subset \Omega$ and the target variable y, i.e., $f : \mathbf{x}_{\omega,t} \in \mathbb{R}^{|\omega|} \to y_t \in \mathbb{R}$. As in the usual regression setup, the structure of f defines the set of explanatory variables. For simplicity, we consider a linear regression mapping that can be described fully with the set of parameters $\boldsymbol{\theta}_\omega = [\theta_0, \theta_1, \ldots, \theta_d]^\top$, where $d = |\omega|$. Then, the mapping can be described as $y_t = \theta_0 + \sum_{k|x_k \in \omega} \theta_k x_{k,t} + n_t, \forall t$, with n_t defining a Gaussian noise with zero mean and unit variance. The learning objective is to find the optimal set of parameters $\hat{\boldsymbol{\theta}}$ that minimizes the chosen loss function l, commonly taken as a quadratic function of prediction errors (residuals) denoted $e_t = y_t - \boldsymbol{\theta}_\omega^\top \tilde{\mathbf{x}}_t$, in expectation, as

$$\hat{\boldsymbol{\theta}}_\omega = \arg\min_{\boldsymbol{\theta}_\omega} \mathbb{E}[l(e_t)], \tag{1}$$

where $\tilde{\mathbf{x}}_t = [1, \mathbf{x}_t]^\top$ with the first element a unit value to incorporate the bias term θ_0 during computation. Here, $\boldsymbol{\theta}_\omega^\top \tilde{\mathbf{x}}_t$ results in the prediction $\tilde{y}_t$ with the parameter $\boldsymbol{\theta}_\omega^\top$. In some settings, the central agent might have additional budget constraints, and the optimization of ℓ will happen subject to these constraints in this case[2].

[2]As an example, consider the central agent incentivize supporting agents to find the best estimate of $\hat{\boldsymbol{\theta}}_\omega$. Then, the budget constraint at the central agent can be defined in terms of the total available monetary value to offer.

The map f is commonly described in a matrix form $\tilde{\mathbf{y}} = \tilde{X}_\omega \boldsymbol{\theta}_\omega$ with the t'th row of $\tilde{X}_\omega$ is $\tilde{\mathbf{x}}_{\omega,t}$. The model is then completely determined by the design matrix $\tilde{X}_\omega$, where $\tilde{X}_\omega$ is constructed executing the data acquisition protocol. For example, the loss function at any time instance can be given by

$$\ell(\mathbf{y}; \boldsymbol{\theta}_\omega) = \|\mathbf{y} - \tilde{X}_\omega \boldsymbol{\theta}_{\boldsymbol{\omega}}\|^2, \quad (2)$$

where the central agent aims to obtain the best parameter $\boldsymbol{\theta}_\omega$ that solves (1). In the considered learning setup, we define $\tilde{\boldsymbol{\theta}}_{\omega,t}$ as the optimal set of parameters minimizing the expected loss l over time and $\tilde{L}_{\omega,t}$ as the time-varying estimator of the loss function. Note that, in an online learning setting, the optimization problem for updating the parametric information follows a recursive approach, where new information is a function of the latest residuals (c.f. [1], Eq. (16) and Eq. (17)). To that end, we consider $q_n \in [0,1], \forall n \in \mathcal{A} \setminus a_i$ is the participation probability to define the involvement of privacy-aware supporting agents in trading explanatory data samples. This means the construction of X_ω at the learner relies on the randomly perturbed features (see Definition 1, Sec. III-A) from the supporting agents, particularly as per their individual data privacy preference. Intuitively, as in [22], such noisy, lower-quality contributing features impact the learner's ability to correctly map input features to the target variable, following larger parameter estimation errors. Furthermore, with $q_n = 0$, the agent n opts out of data trading in the market.

Problem Definition. At any time instance, the ultimate goal of the central agent is to find the optimal mapping parameters [3] $\hat{\boldsymbol{\theta}}$ executing the data acquisition protocol with the supporting agents. The central agent aims to maximize the regression model performance by influencing supporting agents with appropriate pricing signals to exchange high-quality data, i.e., with asked perturbations as the privacy factor in DP terms. The supporting agents tune their responses as per individual privacy preferences while considering the impact of data similarity and the announced pricing. In the following, we present the details of the interaction framework.

III. Interaction Framework in the Regression Market

In this section, we show the strategic interaction between the central agent and the supporting agents in the linear regression market. We will model the two-stage game and derive strategies

[3] In principle, the optimal mapping parameters can be static and time-varying as per the regression problem setting. In any case, this will not influence the overall analysis made hereafter.

of the supporting agents for participation given the asked pricing and privacy factor under a potential data similarity situation from the central agent at equilibrium, respectively. After the mechanism design, we analyze the properties of the derived solution through a low-complexity, first-order backward induction method.

A. Strategies of agents

In this subsection, we explore the strategic behaviour of agents in the regression market. Specifically, we characterize the privacy preferences of individual agents and the impact of data similarity on their participation strategy.

Each agent $a_n \in \mathcal{A}$ has a preference on data privacy ϵ_n *unknown* to the central agent. While the value of privacy preferences differs amongst agents, we assume the agents shared information about its distribution. Let F_ε be the cumulative distribution function (cdf) capturing the realizations ϵ_n from the random variable (RV) $\varepsilon \sim U[\epsilon_{\mathrm{l}}, \epsilon_{\mathrm{u}}]$. A higher value of ϵ_n implies a lower sensitivity towards data privacy, while a lower value means the agents prefer injecting more noise on the traded data - generating perturbed statistics on explanatory features- to meet tighter privacy requirements. Intuitively, smaller $\epsilon \sim f_\varepsilon$ corresponds to lower information leakage in the data market, where f_ε is the probability density function (pdf) of ε. Then, the agent can be called $\epsilon-$*type* to characterize their privacy preference, and the corresponding participation strategy is referred to as $q_n(\epsilon_n) \in [0, 1]$; we use a shorthand q_n hereafter. Therein, we formalize the privacy preference with the following definition of local differential privacy.

Definition 1 (ϵ-Local Differential Privacy [13])**.** *A mechanism $M(\cdot)$ satisfies $\epsilon-$local differential privacy ($\epsilon - LDP$) for $\epsilon \geq 0$, if an only if, given any input data sample $x, x' \in Dom(M(\mathcal{X}))$, we have*

$$\forall y \in \boldsymbol{Dom}(M) : \mathbb{P}[M(x) = y] \leq e^{\epsilon}\mathbb{P}[M(x') = y], \tag{3}$$

where $M(\mathcal{X})$ is a mapping to discrete values denoting the set of all possible outcomes of M.

As outlined, the direct consequence of data similarity, particularly the correlation between traded data, is the undervaluation of data and price allocation mismatch amongst the agents - leading to market distrust and, eventually, a dropout scenario. To mitigate information disclosure, the agents employ a local DP strategy that equivalently limits their participation contribution; as such, the offered reward gets lowered. Hence, the supporting agents are reluctant to participate beyond their privacy budget, as defined by the realizations of ε.

Because the exact privacy preference of each agent is private information, the disclosure of such information induces additional privacy costs for the supporting agents [23]; hence, it is often *unknown* to the central agent. Instead, the central agent plays around with various pricing signals and the asked privacy factors to align the strategies of supporting agents in improving the trained model performance. On the other hand, to counter potential information leakage, the supporting agents compete non-cooperatively and align strategies – *as per their type* – of injecting structured noise into the trading data. We assume any rational agent joins the data market with the pricing compatible with their incurred costs, both in terms of data privacy and information leakage due to adjustments in the privacy budget; however, they opt out from the market if the asked noise level is out of their individual privacy preference.

Remark 1. *We define the statistical information leakage as the ability of the central agent to decode the true type of the supporting agents precisely, manipulating the pricing and data valuation. The supporting agents opt out of the market if the information leakage exceeds their privacy preference.*

We observe the interaction between the central agent and the supporting agents, therefore, can be realized as a single leader multiple followers Stackelberg game, where supporting agents are stimulated by the central agent (the leader) to align their strategies on participation and the adjustment of privacy factor as DP noise during data trading in a non-cooperative manner.

Next, we model the utility functions that characterize such interactions between the central agent and the supporting agents. Based on the utility models, we develop the two-stage game and show the existence of Stackelberg equilibrium in such interactions.

B. Utility Models

We use p to define the offered pricing by the central agent for the asked privacy factor ϵ. Then, for a given p, we model the valuation of the central agent as a monotonically increasing concave function of the privacy factor: $U(\epsilon) = \ln[\alpha\epsilon p + 1]^{-\beta}$, where $\alpha > 0, \beta \in (-1, 0)$ are system parameters. Intuitively, the proposed valuation function captures the improvement in the central agent's utility when obtaining high-quality data samples. Therein, we can define the central agent's utility as follows.

Central Agent's Utility: Making the standard assumption of the concavity of the utility function [24], [25], we propose the learner utility function based on the valuation $U(\epsilon)$ with the

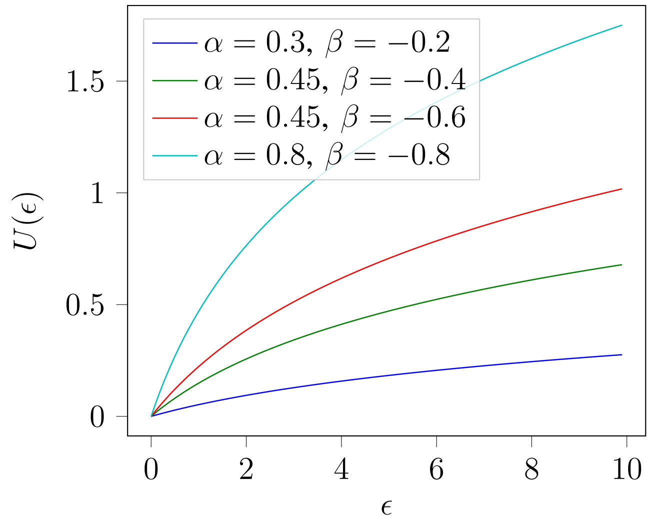

Figure 1: An illustration of learner's valuation $U(\epsilon)$ for asked data privacy factor ϵ.

following definition.

Definition 2. *Considering the improvement in model prediction accuracy $L(\zeta) = 1/(|\tilde{L}_{\omega_i} - \tilde{L}_{\Omega}|)$, using features obtained from supporting agents through participation, the central agent's utility is defined as*

$$S(p;\epsilon) = L(\zeta)\frac{1}{\ln[\alpha\epsilon p + 1]^{\beta}} - p\sum\nolimits_{n\in\mathcal{A}} \mathbb{1}_{\epsilon_n(q_n)>\epsilon}, \tag{4}$$

where $\zeta < \zeta_{ref}$ is the relative accuracy of the trained regression model for a reference requirement of ζ_{ref} in the regression market, p is the offering pricing of the central agent to stimulate individual participation with no worse than ϵ privacy factor on the available data samples.

Following Definition 2, given the participation of all agents, we have $p \geq \sum_n p_{a_n}$. And for a known participation that quantifies $L(\zeta)$, $S(p;\epsilon)$ is decreasing in α and β, and is concave with the privacy factor ϵ. In Fig. 1, we illustrate the influence of the asked privacy factor on the valuation of the central agent for different system parameters. We observe a larger β offers flexibility in data privacy factor with fair compromise on the utility. It is of interest for the central agent to solicit high-quality data with a relaxed privacy factor for the available pricing budget. Then, the central agent adopts an ex-ante differentiated pricing scheme to incentivize supporting agents in terms of the valuation of their data as a model contribution (c.f., Definition 3).

Remark 2. *Participation of supporting agents with quality data in the data trading improves the*

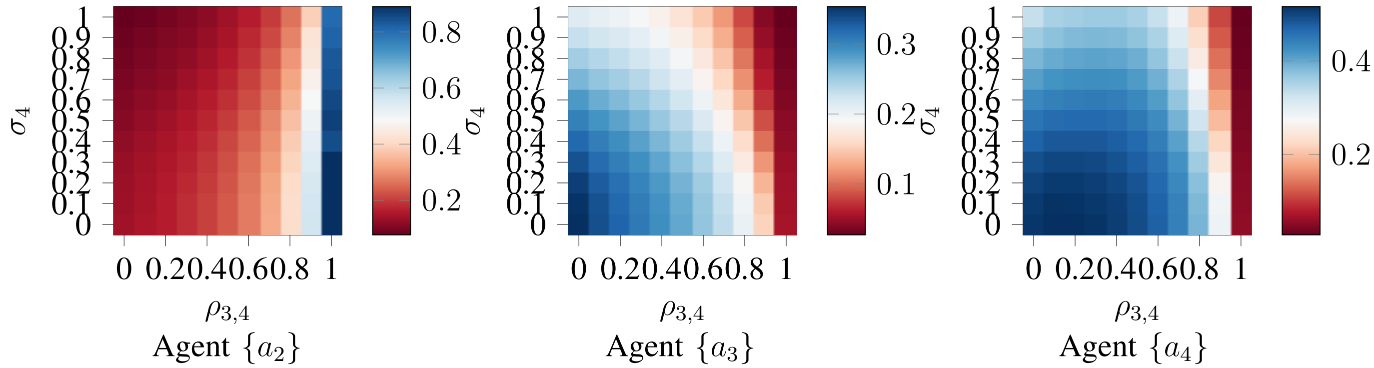

Figure 2: Example scenario: heatmap represents the impact on the normalized contribution of each agent given information leakage due to data correlation $\rho_{3,4}$ between $\{a_3\} - \{a_4\}$ and the noise injection σ_4 by $\{a_4\}$.

performance of the loss estimates. The model contribution is then defined and evaluated as the marginal contribution of individual participation, where $L_n(\zeta)$ is calculated using the standard Shapley Value [26], following Definition 3.

We assume the central agent updates the privacy factor during interactions with the supporting agents.

At each time instance, the central agent chooses the pricing signal p that maximizes its utility defined as a composite function of the performance improvement, expressed in terms of relative accuracy ζ, and the information leakage due to the injection of the statistical uncertainty (i.e., the noise) ϵ.

Definition 3. *The contribution of supporting agents $a_n \in \mathcal{A}$ in each iteration of interaction as*

$$L_n(\zeta) = \frac{1}{|\mathcal{A}|!} \sum\nolimits_{\pi \in \Pi(q_n)} [V(\mathcal{A}^{\pi}_{a_n} \cup \{a_n\}) - V(\mathcal{A}^{\pi}_{a_n})], \tag{5}$$

where $V(\cdot)$ is the standard valuation of traded data [27] – commonly known as performance score – contributing to improving the model accuracy or lowering the loss function.

We then have three particular interpretations following the utility function of the central agent, given all supporting agents trade their data as per their privacy budget, as follows.

C-I. When $\epsilon = 0$, we have $S(p;\epsilon) < 0$, considering the definition of the utility function of the central agent $S(p;\epsilon)$. Furthermore, due to privacy restrictions, for $\epsilon = 0$, we also have $L(\zeta) \approx 0$, i.e., no contribution of the supporting agents.

C-II. When $\exists n : \epsilon_n \sim f_\varepsilon$, and $\epsilon \sim \mathrm{U}[\epsilon_\mathrm{l}, \epsilon_\mathrm{u}]$, we have $S(p;\epsilon) > 0$, leaving the central agent to solve the following optimization problem:

$$\begin{aligned} \mathbf{P:} \ \underset{p}{\text{maximize}} \quad & S(p;\epsilon) && \text{(6a)} \\ \text{subject to} \quad & \sum\nolimits_n p_{a_n} \leq p, && \text{(6b)} \\ & p > 0, && \text{(6c)} \\ & L(\zeta) > \zeta_{\text{ref}}. && \text{(6d)} \end{aligned}$$

Problem **P** is, therefore, an integral structure of the mechanism design problem where the central agent plays with its pricing signal for arbitrary privacy restriction on the supporting agents to ensure a level of performance ζ_{ref}.

C-III. Following **C-I**, we have $S(p;\epsilon) \leq 0$ when $\epsilon \notin [\epsilon_\mathrm{l}, \epsilon_\mathrm{u}]$. Conversely, this leads to a similar scenario where $\epsilon = 0$, i.e., no participation; hence, $S(p;\epsilon) = 0$.

Supporting Agent's Utility Each agent a_n of type ϵ_n responds strategically over the offered reward p_{a_n} to minimize costs on data privacy, for the privacy budget ϵ asked by the central agent, and the information leakage due to data similarity. To simplify our analysis, we use a linear-cost model to define the agent's privacy cost, i.e., $c_n(\epsilon)$, where $c_n(\cdot)$ models the cost of participation employing a privacy factor of ϵ. We use the shorthand c_n, hereafter. The individual data owner aims to minimize the information leakage during data trading and tune privacy factor ϵ_n over the offered pricing p_{a_n} for maximal benefit of participation in the data market:

$$u_n(\epsilon_n, p_{a_n}) = \gamma V_n(q_n, p_{a_n}) - \epsilon_n \mathbb{E}[\mathcal{I}(q_n, \rho_n;\epsilon)] - \psi_n c_n, \tag{7}$$

where $V_n(q_n, p_{a_n})$ is the valuation[4] of agent n on participation for the offered pricing p_n, $\gamma > 0$ captures the participation preference of supporting agents, wherein a larger γ implies a higher valuation on participation, and $\mathbb{E}[\mathcal{I}(q_n, \rho_n)]$ is a strictly increasing[5] that captures information leakage due to data similarity with added privacy cost on participation c_n; we model it with the number of active agents $N(q_n)$ as $\varphi_n \log(1 + N(q_n))$ for $q_n = 1$, and φ_n and ψ_n are weight parameters on information leakage and local privacy cost for agent n, respectively. We further

[4]For ease, we model it as a linear function of asked privacy factor for the offered pricing.

[5]The degree of information leakage is increasing with the number of high-quality data samples in the regression market, as discussed in the data acquisition models [28], [29].

exemplify information leakage with the following numerical example and discuss the utility model design.

Example 3: In Fig. 2, we provide an example scenario to assess the impact of information leakage due to data similarity on the valuation of the supporting agents' data. We set supporting agents $\mathcal{A} \setminus \{a_1\} = \{\{a_2\}, \{a_3\}, \{a_4\}\}$ aiding the linear regression market initiated by the central agent $\{a_1\}$. The numerical evaluation follows the settings of [1]. In addition, we consider agents $\{a_2\}$ and $\{a_3\}$ have correlated data samples denoted as $\rho_{3,4}$ while agent $\{a_4\}$ is injecting noise σ_4 of different magnitude. The central agent is solving the regression problem, as defined in (2), while the contribution is normalized contribution of each supporting agent is evaluated following Definition 3. In the corresponding heatmap scales, we observe the information leakage due to data similarity between agents $\{a_3\}$ and $\{a_4\}$ impacts the normalized contribution of both agents and to the extent, where the high degree of noise injection in the shared data and perfect data correlation would leave to having agent $\{a_2\}$ as the only contributor in solving the regression problem.

Next, we formalize our proposed mechanism design that allows privacy-aware data acquisition under data similarity in the regression market.

C. Mechanism Design

In this subsection, we derive the optimal strategies for the participation of the supporting agents through their sequential interaction with the central agent over pricing. We build on the utility models designed in Sec. III-B and formulate the two-stage game model of interaction under data similarity.

Recall the utility models defined in (4) and (7). While it is true that agent n accrues a utility $u_n(\epsilon_n, p_{a_n})$ only for a positive pricing signal, i.e., $p_{a_n} > 0$, such that

$$u_n(\epsilon_n, p_{a_n}) = \begin{cases} u_n(\epsilon_n, p_{a_n}), \text{if } \epsilon_n \in \varepsilon \\ -\infty, \quad \text{otherwise.} \end{cases} \tag{8}$$

We remark $p_{a_n} > 0$ is a necessary but not sufficient condition for agent n to participate in the data trading.

We have two participation scenarios to characterize the utility function of the individual agent, as follows:

S-I. We have $q_n = 1$ when $\epsilon_n(q_n) > \epsilon$, where

$$\epsilon_n(q_n) = \frac{\gamma V_n(q_n, p_{a_n}) - \psi_n c_n}{\mathbb{E}[\mathcal{I}(q_n, \sigma_n; \epsilon)]}, \tag{9}$$

and $q_n = 0$, otherwise.

S-II. The requirements of $\epsilon_n(q_n) > \epsilon$ restrict some of the supporting agents from participation in the regression market, primarily due to individual privacy budgets.

Then, considering scenarios S-I and S-II results in the following derivations of the optimal participation response of the supporting agents with necessary definitions.

Definition 4 (Feasibility). *The mechanism is feasible for the offered pricing signal p if $\exists n \in \mathcal{A} : u_n(\epsilon_n, p_{a_n}) > 0$. Feasibility criteria can be satisfied as the central agent is aware of the distribution on privacy preference profiles F_ε and stimulates interaction for the exchange of data samples considering the agent with the highest privacy budget ϵ_u and set $\epsilon = \epsilon_u$.*

Following Definition 4, and the utility profile of the supporting agents, however, we cannot guarantee that the criterion for joining the regression market will be fulfilled for all available agents $\epsilon_n(q_n) > \epsilon$, i.e., we have

$$q_n = \int_0^\infty q(\epsilon) dF_\varepsilon(\epsilon) \tag{10}$$

that quantifies the joining fraction of supporting agents, in probability, in the market. This leads to

$$q_n = \int_0^{\epsilon_n(q_n)} dF_\varepsilon(\epsilon) = F_\varepsilon(\epsilon_n(q_n)). \tag{11}$$

Therefore, we can formalize the participation probability of the supporting agent as follows.

Definition 5. *Given the following condition satisfies, as*

$$q_n^* = F_\varepsilon(\epsilon_n(q_n^*)), \tag{12}$$

we define q_n^* as a Nash equilibrium of the supporting agent.

Lemma 1. *Given the incurred cost of data exchanges c_n in the regression market, with a shared value of instantaneous information leakage, there exists a unique Nash equilibrium q_n^* defining the probability of supporting agents joining the collaborative training in the regression market.*

Proof. We begin the proof for the uniqueness of the solution by defining a variable $\xi(q_n) := F_\varepsilon(\epsilon_n(q_n)) - q_n$. As $\mathbb{E}[\mathcal{I}(q_n, \rho_n)]$ is a strictly increasing in (9), consequently, we have $\epsilon_n(q_n)$

a strictly decreasing on its domain, and leading to $F_\varepsilon(\cdot)$ as an increasing function [30]. Then, following the Definition 5, $\xi(q_n)$ should have a unique solution, i.e., a root, to guarantee q_n^* is an equilibrium at the best response. Then,

1) if $V_n(q_n, p_{a_n}) \leq \left(\frac{\psi_n}{\gamma}\right)c_n$, we have $q_n^* = 0$, resulting in unique root of $\xi(q_n)$.
2) if $V_n(q_n, p_{a_n}) \geq \frac{1}{\gamma}\left[\psi_n c_n + \epsilon_n \mathbb{E}[\mathcal{I}(q_n, \rho_n)\right]$ for any $\gamma > 0$, we have $q_n^* = 1$, resulting unique root of $\xi(q_n)$.
3) otherwise, if we have a region between (1) and (2), i.e., $c_n < V_n(q_n, p_{a_n}) < \left(\frac{\gamma}{\psi_n}\right)\epsilon_n \mathbb{E}[\mathcal{I}(q_n, \rho_n)]$, there exists a unique root $q_n^* \in (q_n', 1)$. This can be concluded based on the following observations. We also drop the normalizing constants hereafter for simplifying the analysis, as it won't influence the conclusion made. Choose arbitrary $q_n' \in (0, 1)$, then there exists $c_n < V_n(q_n, p_{a_n}) = c_n + \epsilon_n \mathbb{E}[\mathcal{I}(q_n, \rho_n)] < c_n + \epsilon_n \mathbb{E}[\mathcal{I}(1, \rho_n)]$ as $\mathbb{E}[\mathcal{I}(1, \rho_n)]$ is strictly increasing. Then, using the definition of $\xi(q_n)$, which is a continuous, decreasing function, we have $\xi(q_n) = 1 - q_n > 0, \forall q_n \in [0, q_n']$ and $\xi(1) = F_\varepsilon(\epsilon_n(1)) - 1 < 0$.

Next, we show that q_n^* is the Nash equilibrium with the following observations. As such, the optimal strategy of supporting agents is to adopt their true privacy preference for incentives during data trading. From conditions in (9), this is straightforward as we have,

1) the expected utility $\tilde{V}_n$ of supporting agent n is $\tilde{V}_n := V_n(q_n^*, p_{a_n}) - \frac{1}{\gamma}\left[\psi_n c_n + \epsilon_n \mathbb{E}[\mathcal{I}(q_n^*, \rho_n)\right] > 0$, if $q_n = 1$. For any other strategy $\tilde{q}_n \in [0, 1)$, the expected utility is $\tilde{q}_n\Bigg(V_n(q_n^*, p_{a_n}) - \frac{1}{\gamma}\left[\psi_n c_n + \epsilon_n \mathbb{E}[\mathcal{I}(q_n^*, \rho_n)]\right]\Bigg) < \tilde{V}_n$.
2) The converse is true, as deviating from optimal strategy $q_n = 0$ only lowers the expected utility of the supporting agent $\tilde{V}_n$, when $V_n(q_n^*, p_{a_n}) < \frac{1}{\gamma}\left[\psi_n c_n + \epsilon_n \mathbb{E}[\mathcal{I}(q_n^*, \rho_n)\right]$.
3) Finally, when $V_n(q_n^*, p_{a_n}) = \frac{1}{\gamma}\left[\psi_n c_n + \epsilon_n \mathbb{E}[\mathcal{I}(q_n^*, \rho_n)\right]$, any deviation in the supporting agent's strategy does not improve its expected utility, i.e., $u_n(\epsilon_n, p_{a_n}) = 0$; hence, the supporting agent has no incentive to deviate.

This completes the proof. □

Following Lemma 1, after meeting the participation criteria, the supporting agents derive the optimal response for maximizing their overall valuation with the pricing to the query made by the central agent by solving the following optimization problem.

$$\textbf{P1:} \underset{q_n(\epsilon)}{\text{maximize}} \quad u_n(\epsilon, p_{a_n}|p) \tag{13a}$$

$$\text{subject to} \quad V_n \geq \frac{1}{\gamma}\bigg[\psi_n c_n + \epsilon\mathbb{E}[\mathcal{I}(q_n, \rho_n)]\bigg], \tag{13b}$$

$$\epsilon_n(q_n) > \epsilon, \tag{13c}$$

$$p_n > 0, \tag{13d}$$

where (13b) ensures a positive return on participation and constraint (13c) satisfies the individual privacy budget. Consider $q_n(\tilde{\epsilon})$ is the solution of **P1**, then, we have $q_n^* = \min\{q_n(\tilde{\epsilon}), q_n^*\}$.

Following the solution to **P1**, i.e., the strategic participation, without loss of generality, we then formulate the overall regression market problem as follows:

$$\textbf{P2:} \underset{p,\,\epsilon}{\text{maximize}} \quad S(p; \epsilon|\mathbf{q}^*) \tag{14a}$$

$$\text{subject to} \quad \sum\nolimits_n p_{a_n} \leq p(\epsilon), \tag{14b}$$

$$\epsilon_n(q_n^*) > \epsilon > \epsilon_{\text{ref}}, \forall n \in \mathcal{A}, \tag{14c}$$

$$p(\epsilon) > 0, \tag{14d}$$

$$L(\zeta) > \zeta_{\text{ref}}, \tag{14e}$$

where $\mathbf{q}^*$ is a vector with the best response strategies of the supporting agents over the offered pricing signal and the privacy requirements set by the regression market; constraint (14c) restricts the asked privacy guarantees to a reference ϵ_{ref} value. While it is in the best interest of the central agent to keep the value of ϵ as large to improve participation with less perturbed explanatory data samples in the regression market, it is mostly impractical. This is due to the influence of heterogeneous privacy preferences. In principle, ϵ_{ref} is set as $\max\{\epsilon_n\}, \forall n \in \mathcal{A}$ in each iteration of interaction to satisfy participation constraint (14c), while satisfying the price allocation under budget constraints (14b).

Lemma 2. *The optimal solution for **P2** ϵ^* for the known pricing budget at the central agent can be derived as $\max\{\epsilon_n(q_n^*), \epsilon_{ref}\}$, where the asked privacy guarantees are set by the central agent to a reference ϵ_{ref} value.*

Proof. The proof follows the characteristics of the constraints, leading the optimal solution to be the boundary conditions for a fixed offered pricing $p(\epsilon) > 0$. Note that the differential price allocation constraint on the available monetary budget, $\sum_n p_{a_n} \leq p(\epsilon)$ is evaluated as per the proportional contribution measure of the individual supporting agent, as in [10], [31]. □

Algorithm 1 First-order Iterative Backward Induction

```
Start with random sample ϵ_ref < ϵ ∼ f_ε, offered pricing signal p = p_max, normalizing parameters γ, {ψ_n, φ_n}, ∀n set to 1, ζ_ref = 0.9.
P = {}, A ← {n|q_n = 1, ∀n};
repeat
    R = A;
    Evaluate the performance improvement L(ζ);
    Solve the following optimization problem:
```

$$\underset{p>\tilde{p}}{\text{maximize}}\; L(\zeta)\frac{1}{\ln[\alpha\epsilon^* p(\epsilon^*)+1]^{\beta}} - p(\epsilon^*)\sum\nolimits_{n\in\mathcal{A}} q_n(\epsilon^*);$$

```
    for all agents n ∈ A do
        Invoke proportionally fair price allocation p_{a_n} with the marginal contribution (5);
        P ← P + {p_{a_n}}
        R = A \ n;
    end for
    Set ϵ = max{ϵ_n(q*_n), ϵ_ref}, p̃ = sup P;
until A = {∅};
```

For the pricing signal p and the asked privacy budget ϵ, following **P1** and problem **P2**, leads the regression market problem a two-staged leader-follower game, where the market aims to receive high-quality explanatory data samples for the feasible pricing signal to the agents who strategically response to the query made for participation in a non-cooperative setting.

Following Lemma 2 results in the following overall utility maximization problem:

$$\textbf{P3:}\ \underset{p>0,\mathbf{q}}{\text{maximize}}\quad \frac{L(\zeta)}{\ln[\alpha\epsilon^* p(\epsilon^*)+1]^{\beta}} - \ p(\epsilon^*)\sum\nolimits_{n\in\mathcal{A}} q_n\epsilon^*(q_n) \tag{15a}$$

$$\text{subject to}\quad L(\zeta) > \zeta_{\text{ref}} \tag{15b}$$

$$q_n \in \{0,1\}, \forall n \in \mathcal{A} \tag{15c}$$

The solution to the optimization problem **P3** is non-trivial, first, given the participation constraints following response to feasibility (Definition 4), the influence of pricing on participation, and its overall consequence on the performance improvement factor $L(\zeta)$. Second, with a possible $2^{|\mathcal{A}|}$ configurations, it might require exponential-complexity effort to solve the problem with an exhaustive search solution. We propose a low-complexity solution to address this.

We develop a first-order iterative solution (Algorithm 1) that aims to satisfy the conditions for the Nash solution and builds on top of the backward induction method to reach the Stackelberg equilibrium. The proposed method satisfies the following economic properties.

Incentive Compatibility: The mechanism is incentive compatible given if all supporting agents behave rationally according to their local privacy preference, i.e., their true type, such that $\mathbb{E}[\epsilon_n, u_n(p_{a_n})] \geq \mathbb{E}[\epsilon, u_n(p_{a_n})], \forall \epsilon, \forall n \in \mathcal{A}$. *Individual Rationality*: For each supporting agent, there exists a non-negative utility $u_n(\epsilon_n, p_{a_n}) \geq 0, \forall n \in \mathcal{A}$, if they respond with their true type. This is the participation constraint, which is observed following Lemma 1. The supporting agents always opt for their true data type, ensuring a positive utility.

Besides the fundamental properties of our mechanism, its computational properties are covered through the following theorem.

Theorem 1. *The first-order Algorithm 1 solves the overall utility maximization problem* ***P3*** *with linear complexity.*

Proof. The proof follows the convexity property of the maximization problem in (line 6) with the responses of supporting agents to the asked privacy factor $\epsilon = \max\{\epsilon_n(q_n^*), \epsilon_{\text{ref}}\}$ (line 12), that eventually reduces the problem into a single variable optimization with the initial maximum offered pricing. □

With this, we now have all the ingredients to characterize the *Stackelberg equilibrium*. Obtaining the solution of (15) $p^*(\epsilon)$, involving best-responses of supporting agents with Definition 5, we have the following proposition.

Proposition 1. *For any values of p and ϵ, we have the Stackelberg equilibrium if the following conditions are satisfied:*

$$S(p^*, \boldsymbol{\epsilon}^*) \geq S(p, \boldsymbol{\epsilon}^*) \tag{16}$$

$$u_n(\epsilon_n(q_n^*), p^*) \geq u_n(\epsilon_n(q_n), p^*), \forall n \in \mathcal{A}. \tag{17}$$

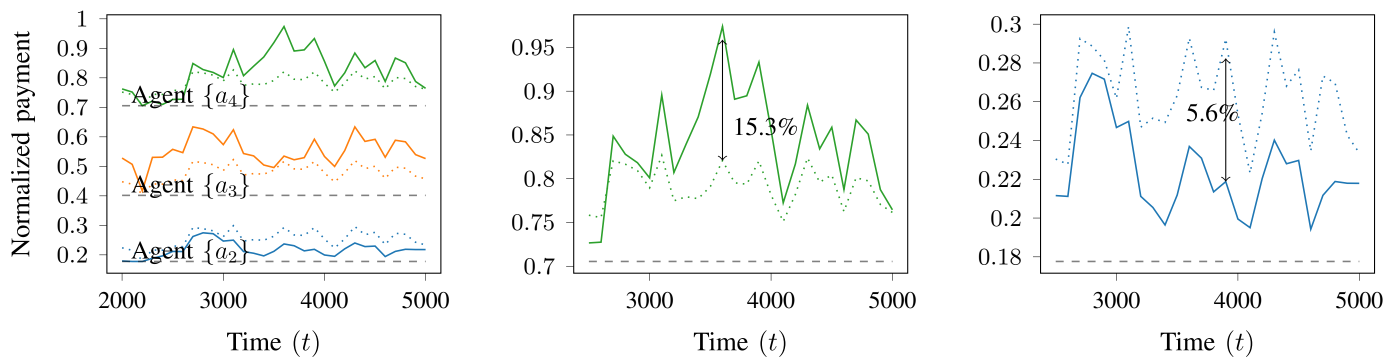

Figure 3: (Left – Right) Evolution of normalized payment to agents $\{a_2, a_3, a_4\}$ during training in the online regression market by the agent $\{a_1\}$.

IV. Numerical Results

This section introduces the evaluation results for the proposed framework. We begin with the evaluation setup, where we show the underlying model used in the regression task. Then, we provide a performance evaluation following an analysis of the pricing, strategic participation, and comparison under the impact of data similarity.

Setup: We consider a plain regression learning problem. For this, we generate the data to match the setup in which four agents, where $\{a_1\}$ is a central agent posing the online regression problem, as in [1], and $\{a_2, a_3, a_4\}$ are the supporting ones supplying contributing features. The agents use distinct features sampled from the Gaussian distribution with unit variance. Central agent $\{a_1\}$ own feature x_1, while the supporting agents $\{a_2, a_3, a_4\}$ hold relevant features x_2, x_3, and x_4, respectively, at each time step. We particularly consider a single-order regression model such as $y_t = \theta_0 + \theta_{1,t}x_{1,t} + \theta_{2,t}x_{2,t} + \theta_{3,t}x_{3,t} + \theta_{4,t}x_{4,t} + \beta_t$, where the last term β_t is Gaussian noise with zero mean and a finite variance of 0.3, with quadratic loss, as in (2). Furthermore, to demonstrate the impact of participation due to data similarity and, further, the information leakage, we model features as correlated with each other through linear models, as in [10]. We first simulate the process using the true parameters as $\theta^\top = [0.2 \ 0.4 \ -0.3 \ -0.6 \ 0.2]$. We follow the batch estimation process [see Section 2.3.2 in [1], equation (15)], gathering features for $\tau = 10000$ times, and later use this initialization for the online regression. The privacy budget is set as $\epsilon_{\text{ref}} = \ln 10$. and the agents strategically employ $\epsilon-$LDP following Definition 1, while the utility model uses parameters $\alpha = 0.45$ and $\beta = -0.4$.

Analysis on offered pricing and loss estimates: For evaluation, we consider the following two

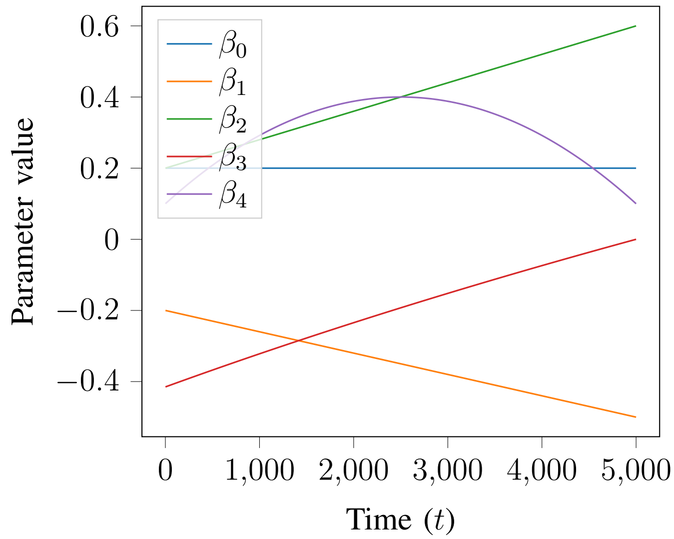

Figure 4: Temporal evolution of the parameters over the period.

intuitive baselines: (i) **Case I**, which ignores the strategic participation of the agent allocates pricing as per the contribution made by the agents, (ii) **Case II**, which considers strategic participation and allocates pricing as per the contribution made by the agents. Our method **Alg. 1** considers the strategic participation of devices with a proportional price allocation scheme as per their data contribution. In Fig.3, we show the evolution of normalized payment to three agents with contributing features in the regression market established by agent $\{a_1\}$, where agent $\{a_2\}$ is sharing the poorest quality data (i.e., implementing extreme privacy measures) and agents $\{a_1\}, \{a_2\}$ have correlated data. We observe that in both cases, the agents are offered payments more than the asked price, with a variability of 5.6% to 15.3%. This situation ensures the participation of the agents; however, they do not align their privacy budget accordingly as asked by the learner - leading to unintended consequences in the payment for the learner and its utility. As compared to Case I, in Case II, the payment is shared amongst agents as per their strategic participation that influences individuation contributions.

In Fig. 4, we see the temporal evolution of the model parameters as per the underlying model we have adopted. Correspondingly, Fig. 5 validates the performance of the learned regression model for the online regression task with four agents. We consider three scenarios for the evaluation of loss estimates: (i) Central Info, where only the feature available at agent $\{a_1\}$ is provided; (ii) Partial Info, where only features of agents $\{a_1, a_2, a_3\}$ are solicited by the agent $\{a_1\}$; and (iii) Full Info, where all features are provided, i.e., full participation. It is intuitive

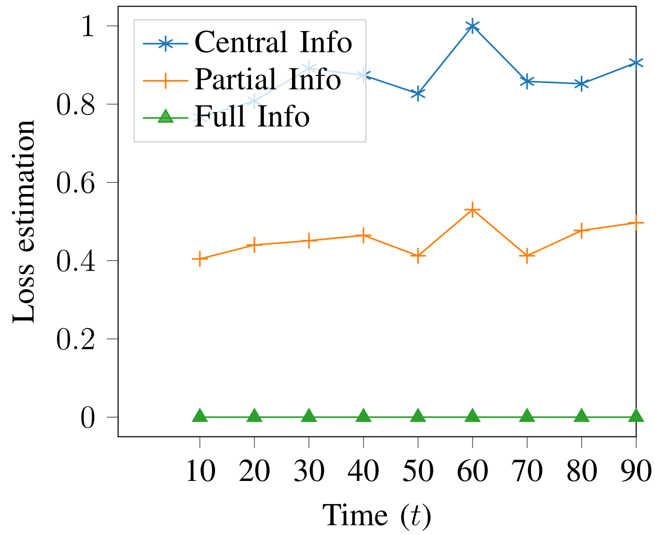

Figure 5: Impact of participation on the normalized loss estimates for different collaborative online learning scenarios with four agents: (i) Central Info, with only agent $\{a_1\}$, (ii) Partial Info, with agents $\{a_1, a_2, a_3\}$, and (iii) Full Info, with all agents.

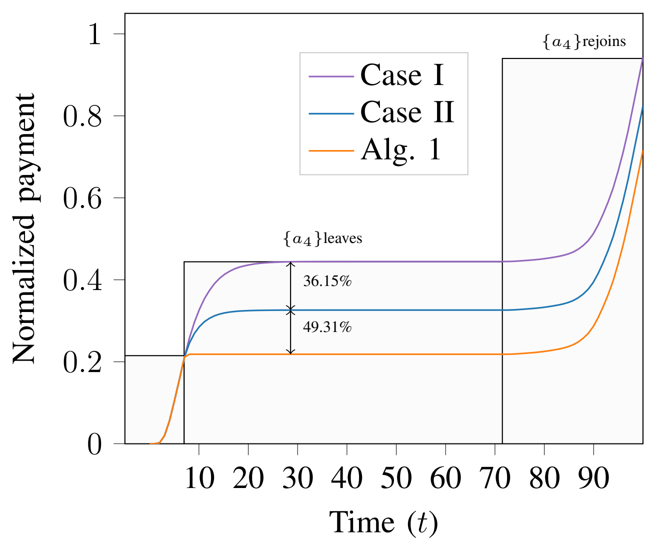

Figure 6: Impact of dynamic participation of agent $\{a_4\}$ on normalized payment: Case I, Case II, Alg. 1.

that the temporal evolution of loss estimates with Partial Info is better than the Central Info, where the central agent is missing relevant features in keeping track of estimating the true model parameters. This results in poor loss estimates, which is intuitive.

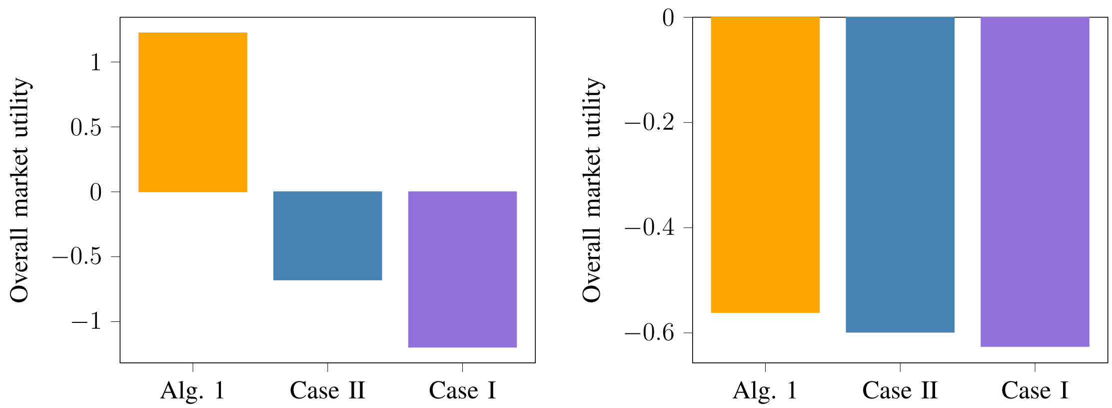

Figure 7: Performance comparison of regression market's utility defined for the central agent while varying reference privacy factor: (left) $\epsilon_{\text{ref}} = \ln 10$, and (right) $\epsilon_{\text{ref}} = \ln 60$.

Impact of data similarity and privacy factor: In Fig.6, we analyze the influence of strategic participation on data valuation and normalized payments in the regression market. Agent $\{a_4\}$ has intermittent connectivity and leaves the regression market after a few iterations. Following **Case I**, the Shapley valuation is obtained based on all estimation subsets and allocated to the agents. This results in higher payment (up to 49%) for even those agents, i.e., $\{a_4\}$, particularly without making a contribution, as compared with **Case II**. **Case II** accounts for strategic participation but still allocates payment following the recursive nature of the contribution evaluation procedure. Alg. 1 account for the participation variable in evaluating the performance improvement, followed by the price allocation. Hence, the contribution made by the remaining agents is only considered, offering a gain of two factors.

Fig. 7 shows the performance gain in terms of the overall market utility as compared with the baselines for different asked privacy budgets. We evaluate the mechanism for two extreme privacy budgets: $\epsilon_{\text{ref}} = \{\ln 10, \ln 60\}$, which allows a comparative performance evaluation of the proposed method against baselines participation strategies. First, the central agent relaxing privacy budget enforces supporting agents with better privacy preferences to opt out from the market; hence, we observe a reduction in the market utility for $\epsilon_{\text{ref}} = \ln 60$ in Fig. 7. This follows our analysis for the drop in the loss estimates under such a scenario, as in Fig. 5. Second, we observed the naive scenario of Case I performs the worst among all, where the agents are not strategic in participation but are offered higher pricing. Our approach considers the heterogeneous privacy preferences of the supporting agents and offers pricing to ensure the

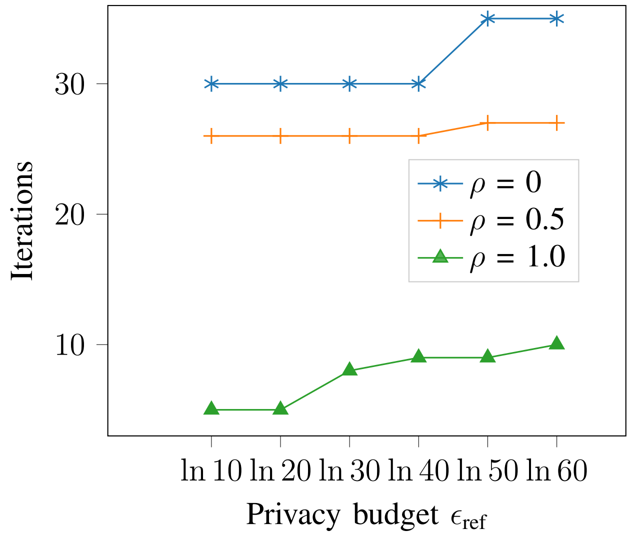

Figure 8: Convergence analysis (in terms of number of required iterations) at different privacy budgets under data similarity.

supply of quality data. Consequently, this improves the overall market utility, as seen in Fig. 7 for $\epsilon_{\text{ref}} = \ln 10$. This follows improved contributions of the supporting agents, considering the adjustment in privacy factor, as defined in Alg. 1, line 12.

Finally, in Fig. 8, we analyze the number of iterations required for the convergence of the proposed approach under the influence of different degrees of data similarity and against different privacy budgets. Following the original underlying model (as in Fig. 2), we vary the data similarity measure ρ between the agents $\{a_3, a_4\}$. We observe with the increase in the privacy budget, i.e., lower sensitivity towards data privacy; the algorithm takes more iterations to converge following the tight privacy preference used for the agents. Furthermore, involving the participation of all agents, the results show the adversarial influence of data similarity on the convergence iterations. Statistical information leakage due to data similarity lowers the number of iterations where the correlated feature is less, contributing to the loss estimates. This supports the analysis made in Fig. 2 regarding data contribution in solving the regression problem under data similarity.

V. Conclusion

In this work, we analyzed the interplay between data pricing, privacy, and learning. In particular, we showed the impact of statistical information leakage (for instance, due to data similarity, e.g., data correlation) on the offered pricing and its influence on the value of traded data in a regression data market setting. We proposed a holistic market design where we account for such dependencies. Therein, we developed a query-response strategy for a leader-followers data acquisition mechanism that enjoys a local differential privacy technique, where participation in the trading of data samples happens between a number of data owners and the learner in an elastic fashion. We modelled the strategic interactions between the privacy-aware data owners and the learner as a Stackelberg game and evaluated the consequences of data similarity in terms of participation and the value of traded data on the online regression data market setup.

REFERENCES

[1] P. Pinson, L. Han, and J. Kazempour, "Regression markets and application to energy forecasting," *Top*, vol. 30, no. 3, pp. 533–573, 2022.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, pp. 1273–1282, PMLR, 2017.

[3] S. R. Pandey, N. H. Tran, M. Bennis, Y. K. Tun, A. Manzoor, and C. S. Hong, "A crowdsourcing framework for on-device federated learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3241–3256, 2020.

[4] P. Popovski, F. Chiariotti, V. Croisfelt, A. E. Kalør, I. Leyva-Mayorga, L. Marchegiani, S. R. Pandey, and B. Soret, "Internet of things (iot) connectivity in 6g: An interplay of time, space, intelligence, and value," *arXiv preprint arXiv:2111.05811*, 2021.

[5] Y. M. Saputra *et al.*, "Energy demand prediction with federated learning for electric vehicle networks," in *2019 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, IEEE, 2019.

[6] O. Dekel, F. Fischer, and A. D. Procaccia, "Incentive compatible regression learning," *Journal of Computer and System Sciences*, vol. 76, no. 8, pp. 759–777, 2010.

[7] N. Corporation, "Naver map."

[8] K. Corporation, "Kakaomap."

[9] D. Acemoglu, A. Makhdoumi, A. Malekian, and A. Ozdaglar, "Too much data: Prices and inefficiencies in data markets," *American Economic Journal: Microeconomics*, vol. 14, no. 4, pp. 218–256, 2022.

[10] S. R. Pandey, P. Pinson, and P. Popovski, "Strategic coalition for data pricing in iot data markets," *IEEE Internet of Things Journal*, 2023.

[11] R. Cramer, I. B. Damgård, *et al.*, *Secure multiparty computation.* Cambridge University Press, 2015.

[12] R. B. Christensen, S. R. Pandey, and P. Popovski, "Semi-private computation of data similarity with controlled leakage," *arXiv preprint arXiv:2206.06650*, 2022.

[13] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*, pp. 1–19, Springer, 2008.

[14] A. Fallah, A. Makhdoumi, A. Malekian, and A. Ozdaglar, "Optimal and differentially private data acquisition: Central and local mechanisms," *arXiv preprint arXiv:2201.03968*, 2022.

[15] A. Agarwal, M. Dahleh, and T. Sarkar, "A marketplace for data: An algorithmic solution," in *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 701–726, 2019.

[16] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pp. 265–284, Springer, 2006.

[17] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, “Federated learning: Strategies for improving communication efficiency,” *arXiv preprint arXiv:1610.05492*, 2016.

[18] L. D. Nguyen, S. R. Pandey, S. Beatriz, A. Broering, and P. Popovski, “A marketplace for trading ai models based on blockchain and incentives for iot data,” *arXiv preprint arXiv:2112.02870*, 2021.

[19] C. Dwork, A. Roth, *et al.*, “The algorithmic foundations of differential privacy.,” *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3-4, pp. 211–407, 2014.

[20] T. Falconer, J. Kazempour, and P. Pinson, “Bayesian regression markets,” *arXiv preprint arXiv:2310.14992*, 2023.

[21] Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated machine learning: Concept and applications,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[22] X. Zheng and Z. Cai, “Privacy-preserved data sharing towards multiple parties in industrial iots,” *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 5, pp. 968–979, 2020.

[23] R. E. Verrecchia, “Discretionary disclosure,” *Journal of accounting and economics*, vol. 5, pp. 179–194, 1983.

[24] W. M. Gorman, “The structure of utility functions,” *The Review of Economic Studies*, vol. 35, no. 4, pp. 367–390, 1968.

[25] N. Singh and X. Vives, “Price and quantity competition in a differentiated duopoly,” *The Rand journal of economics*, pp. 546–554, 1984.

[26] L. S. Shapley *et al.*, “A value for n-person games,” 1953.

[27] S. R. Pandey, L. Nguyen, and P. Popovski, “Fedtoken: Tokenized incentives for data contribution in federated learning,” in *Workshop on Federated Learning: Recent Advances and New Challenges (in Conjunction with NeurIPS 2022)*.

[28] R. Cummings, S. Ioannidis, and K. Ligett, “Truthful linear regression,” in *Conference on Learning Theory*, pp. 448–483, PMLR, 2015.

[29] R. Cummings, H. Elzayn, E. Pountourakis, V. Gkatzelis, and J. Ziani, “Optimal data acquisition with privacy-aware agents,” in *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 210–224, IEEE, 2023.

[30] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[31] Y. K. Tun, N. H. Tran, D. T. Ngo, S. R. Pandey, Z. Han, and C. S. Hong, “Wireless network slicing: Generalized kelly mechanism-based resource allocation,” *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 8, pp. 1794–1807, 2019.

Shashi Raj Pandey (Member, IEEE) received his B.E. degree in Electrical and Electronics with a specialization in communication engineering from Kathmandu University, Nepal, and the Ph.D. degree in Computer Science and Engineering from Kyung Hee University, Seoul, South Korea. He is currently an Assistant Professor at the Department of Electronics Systems, Aalborg university. Prior, he was a Postdoctoral Researcher at the Connectivity Section, Aalborg University, from 2021 to 2023. He served as a Network Engineer at Huawei Technologies Nepal Co. Pvt. Ltd, Nepal, from 2013 to 2016. His research interests include network economics, game theory, wireless communications, data markets and distributed machine learning. He was the recipient of the Best Paper Award at several conferences, including IEICE APNOMS 2019. He was a Member at Large at the IEEE Communication Society Young Professionals 2020 – 2021. He currently serves as a Member at Large in the IEEE Communication Society On-Line Content Board and is on the editorial advisory board of IEEE’s Spectrum The Institute 2022 – 2024. He is an affiliated member of the Pioneer Center for AI, Denmark.

Pierre Pinson (Fellow, IEEE) received the M.Sc. degree in applied mathematics from the National Institute of Applied Sciences (INSA), Toulouse, France, in 2002 and the Ph.D. degree in energetics from Ecole des Mines de Paris, France, in 2006. He is the chair of data-centric design engineering at Imperial College London, United Kingdom, Dyson School of Design Engineering. He is also an affiliated professor of operations research and analytics with the Technical University of Denmark and a chief scientist at Halfspace (Denmark). He is the editor-in-chief of the International Journal of Forecasting. His research interests include analytics, forecasting, optimization and game theory, with application to energy systems mostly, but also logistics, weather-driven industries and business analytics. He is a Fellow of the IEEE, an INFORMS member and a director of the International Institute of Forecasters (IIF).

Petar Popovski (Fellow, IEEE) is a Professor at Aalborg University, where he heads the section on Connectivity and a Visiting Excellence Chair at the University of Bremen. He received his Dipl.-Ing and M. Sc. degrees in communication engineering from the University of Sts. Cyril and Methodius in Skopje and the Ph.D. degree from Aalborg University in 2005. He received an ERC Consolidator Grant (2015), the Danish Elite Researcher award (2016), IEEE Fred W. Ellersick prize (2016), IEEE Stephen O. Rice prize (2018), Technical Achievement Award from the IEEE Technical Committee on Smart Grid Communications (2019), the Danish Telecommunication Prize (2020) and Villum Investigator Grant (2021). He was a Member at Large at the Board of Governors in IEEE Communication Society 2019-2021. He is currently an Editor-in-Chief of IEEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. He also serves as a Vice-Chair of the IEEE Communication Theory Technical Committee and the Steering Committee of IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING. Prof. Popovski was the General Chair for IEEE SmartGridComm 2018 and IEEE Communication Theory Workshop 2019. His research interests are in the area of wireless communication and communication theory. He authored the book “Wireless Connectivity: An Intuitive and Fundamental Guide”, published by Wiley in 2020.