Participation and Data Valuation in IoT Data Markets through Distributed Coalitions

Pandey, Shashi Raj; Pinson, Pierre; Popovski, Petar

DOI (link to publication from Publisher):
https://doi.org/10.48550/arXiv.2206.07785

Publication date:
2022

Participation and Data Valuation in IoT Data Markets through Distributed Coalitions

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

***Abstract*—This paper considers a market for Internet of Things (IoT) data that is used to train machine learning models. The data is supplied to the market platform through a network and the price of the data is controlled based on the value it brings to the machine learning model. We explore the correlation property of data in a game-theoretical setting to eventually derive a simplified distributed solution for a data trading mechanism that emphasizes the mutual benefit of devices and the market. The key proposal is an efficient algorithm for markets that jointly addresses the challenges of availability and heterogeneity in participation, as well as the transfer of trust and the economic value of data exchange in IoT networks. The proposed approach establishes the data market by reinforcing collaboration opportunities between device with correlated data to avoid information leakage. Therein, we develop a network-wide optimization problem that maximizes the social value of coalition among the IoT devices of similar data types; at the same time, it minimizes the cost due to network externalities, i.e., the impact of information leakage due to data correlation, as well as the opportunity costs. Finally, we reveal the structure of the formulated problem as a distributed coalition game and solve it following the simplified split-and-merge algorithm. Simulation results show the efficacy of our proposed mechanism design toward a trusted IoT data market, with up to 32.72% gain in the average payoff for each seller.**

***Index Terms*—Internet of things (IoT), IoT Data market, data trading, incentive mechanism, information leakage, coalition game.**

I. Introduction

A. Context and Motivation

The massive volume of Internet of Things (IoT) devices and services lead to exponential growth of IoT data [1]. Various networked cyber physical systems (CPSs) are accumulating and processing data at a large scale, often contributing to the training of some learning model or carrying out an inference. For instance, massively distributed data when integrated with Machine Learning (ML) tools stimulate both *real time* and *non-real time* decision-making services that create a value of data in the IoT networks [2]. This brings the question of economic opportunities in IoT data markets, where data and its value to the services can be traded or exchanged. It is thus relevant to study the IoT data markets in terms of mechanisms for attaining the desired economic properties in offering learning services, such as prediction, detection, classification, forecasting, and similar. Furthermore, it is necessary to investigate strategies involved in the execution of such distributed cooperation amongst devices having data of value for IoT data markets.

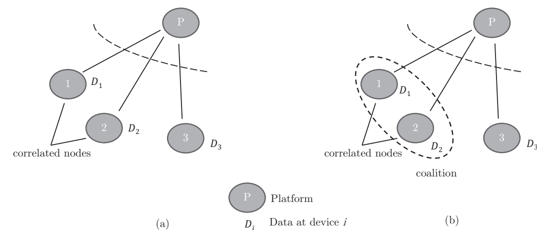

Figure 1: Devices with correlated data and coalition formation when interacting with the platform (learner).

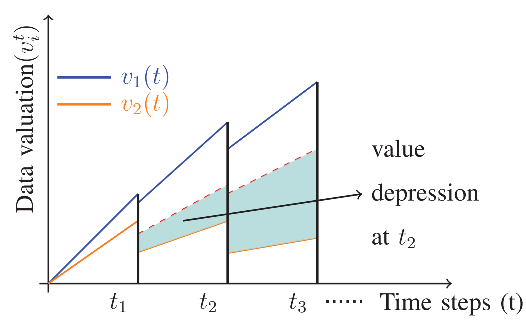

Figure 2: Value depression due to leakage of correlated information in a two seller $i \in \{1, 2\}$ one buyer scenario in a linear pricing scheme.

The two fundamental aspects of an IoT data markets are: (1) offered *pricing*[1], and (2) device participation in the data trading process. An IoT device should be stimulated by the network to participate and share data. The stimulation is achieved by pricing signals that compensate the IoT device based on the data valuation and the cost of data privacy, with additional computational and communication costs. The statistical properties of traded data over IoT networks raise fundamental challenges on the scope/impact of exploiting data trading mechanisms to realize IoT data market, which has been overlooked in the recent literature [3]–[6].

The seminal work [7] models the data market by relating the correlation among devices' data to the price depression. To illustrate the main ideas of [7] in a data streaming setting, consider an IoT data market, as in Fig. 1(a), featuring devices with correlated data that interact and trade data with the platform. Note that in this case the platform acts as a learner, such that we will use the terms learner and platform interchangeably. Assume that the device $\{1\}$ shares its dataset

This work has been in part supported by the Villum Investigator Grant "WATER" from the Velux Foundation, Denmark.

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 is with the Department of Technology, Management and Economics, Technical University of Denmark, Denmark. Email: {ppin}@dtu.dk.

[1]Pricing indicates monetary reward or incentives of any form in general, such as discount vouchers.

$\mathcal{D}_1$ with the platform, after which the device $\{2\}$ does the same. If the two datasets are correlated and the market learns the total variation distance between $\mathcal{D}_1$ and $\mathcal{D}_2$, then it can prioritize pricing for the earliest traded data, and drop offered price rate for the latter. Consider Fig. 2, which shows the evolution of data valuation over time $t \geq 0$, assuming a linear pricing scheme. The market defines $v_i(t)$ as the valuation of the data of seller i at time t. We observe that the valuation of data $v_2(t)$ for seller $\{2\}$ drops in subsequent interactions with the market because device $\{1\}$ leaks information about the data of device $\{2\}$.

Motivated by this observation, the key contribution of this work is a method by which the devices can challenge the market to limit price depression: forming a coalition within devices with correlated data, as shown in Fig. 1(b), and bargaining as a group instead of individually. To elaborate, take the IoT data market scenario in Fig. 3. A number of IoT devices connected to a platform collaboratively train a learning model and create a value; e.g., this could be a predictor in the Federated Learning (FL) setting [4], [5], [8], [9]. In this regard, the authors in [10] discuss a marketplace for data where a robust Shapley technique is developed to capture replicable properties of exchanged data and ways to capitalize the value when sharing them. In general, such market offers incentives to the devices in a way that (i) stimulates their participation [11]–[13], and (ii) strikes a balance between the data privacy concerns, the trustworthiness of the data market, and the cost of data trading [7], [14]. As explained before about [7], the market may also leverage the correlated information or *information leakage*, and other statistical properties of data between sellers [10], [15]–[17], to unilaterally steer the pricing signals for self-benefit. This leads to depression of the value of the data that would not contribute much to the model trading process, leading to uncontrolled competition in data sharing, particularly, due to data rivalry. As the market exploits more data, it cause devices to drop their participation out of mistrust, or negligible pricing. This network externality creates a *loop of mistrust*, by which the platform can manipulate the data market causing and affect device participation in the data trading process.

Our solution to counter the loop of mistrust consists of two steps: (1) form a coalition among a group of devices to tighten the information leakage within the group; (2) challenge the platform to execute the data trading mechanism in a trusted setting. Overall, this brings value in collaboration with improved pricing offers. Coalition formation limits data rivalry amongst sellers, lowers the impact of information leakage due to uncontrolled competitive data trading on pricing, and fosters availability of devices to establish data markets. However, forming coalition to realize a data market is not straightforward, since the devices need to: (i) learn correlated statistical properties of data of the other devices, and without revealing it through the market, (ii) characterize and formalize relevant utility models that identify conditions for coalition formation and price determination amongst devices within coalition, and (iii) handle time-complexity and efficiency of coalition formation at scale.

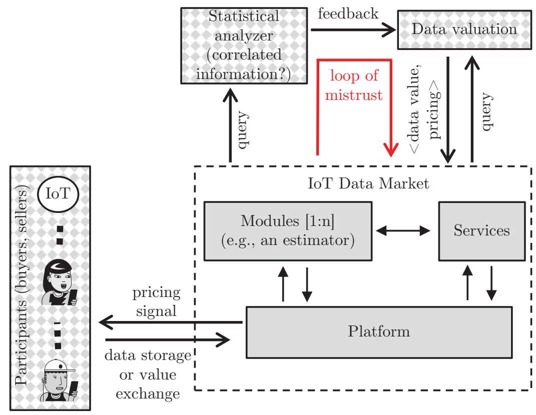

Figure 3: A schematic framework: loop of mistrust in the IoT data market.

B. *Challenges and Contributions*

Based on the previous discussion, we have identified the challenges expressed through the following questions:

- **Q1**: How do IoT devices protect correlated information of their data without allowing the learner to manipulate pricing in the IoT data market?
- **Q2**: How can the platform infer when the data from devices have equal marginal values or valuations?
- **Q3**: What is the impact of device availability in the data trading process?

Addressing **Q1** means preventing the market to identify possible correlation between different data type of devices and further, monopolize pricing and data trading strategies of devices. To eliminate this, we devise distributed coalitions of devices with similar data types, which enforces the platform to derive the marginal contribution of the coalition, or simply, the coalition value, instead of the individual interactions. Addressing **Q2** positions us to develop reasonable utility models for the IoT market, that equally benefits the platform, without hurting participation of devices in coalition due to information leakage and unreliable connectivity in the IoT networks[2]. Then, we show this eventually leads to the formation of different coalition structures that balances individual payoffs and stability of coalition. Addressing **Q3** positions us to evaluate the value of participation and set us to develop a holistic framework that jointly incorporates **Q1** and **Q2** in the IoT data market design.

As a main contribution, we develop a *novel cooperation protocol*, termed multi-agent joint policy (MAJP), in an IoT data market that enables devices to maximize their value of participation in coalition. We explore the characteristics of the formulated MAJP problem, which is intractable due to binary constraints and coupling of variables. Therein, we devise a game-theoretic mechanism that offers a distributed solution of MAJP problem where the proposed approach reinforces data sellers into collaboration for data trading with the objective

[2]In this work, we realize unreliable connectivity in terms of participation. Particularly, the connectivity is considered as an uncontrollable factor in the IoT network, wherein we reflect the unreliable connectivity with the availability of market players, i.e., the devices for the data trading.

to minimize the information leakage in a distrusted IoT data market. The proposed mechanism of distributed coalition game captures the properties of information leakage, value of collaboration and the opportunity costs during the coalition formation. In doing so, we first highlight the *tension between data sellers and the data market over offered pricing and issues due to information leakage.* Then, we offer a distributed solution to overcome the presented challenges. To that end, we derive stability conditions of the coalition game following the developed distributed coalition formation mechanism based on the merge-and-split algorithm [18] to realize trusted IoT data market. Finally, through a sequence of numerical evaluations, we show the efficacy of our proposed mechanism. The rest of the paper is organized as follows. Section II provides preliminaries, introduces truthful IoT data market model and defines device data type. Section III develops the underlying utility models, and presents the problem formulation as an optimization problem to obtain a multi-agent joint policy (MAJP) in a distributed coalition setting. Section IV develops the market design and proposed a coalition game solution to the MAJP problem with complexity analysis. Section V provides the performance evaluation of the proposed approach and shows comparative analysis with the competitive baselines. Finally, Section VI concludes this work.

II. Preliminaries

A. Network Setup for IoT Data Market

A typical IoT data market considers the interaction between the buyer and the sellers IoT devices to trade data or services (e.g., training learning models) with pricing signals [19]–[21]. Consider a network with a finite set of IoT devices $\mathcal{M}$ as $|\mathcal{M}| = M$ training a global learner (e.g., a predictor[3]). The learner acts as an intermediary (for simplicity, we consider it as the buyer, or equivalently, the platform) purchasing data from the distributed devices (sellers); thus, forming a marketplace where strategic data sellers get incentives for their contribution in improving model at the global learner. Indeed, to later explore the strategic interaction between the devices and the learner, the following remark is useful.

Remark 1. *In a game-theoretic setting, the said set of $\mathcal{M}$ participating devices are often referred as "agents". These finite set of agents hold explanatory data samples for specific learning tasks, and aim to exchange it with the learner, fully or in a privacy-preserving manner, e.g., in FL [9], for training learning models of interest to the learner. Then, any rational agent is willing to participate in data trading as such the offered pricing compensates their cost of participation.*

Each device $m \in \mathcal{M}$ stores the data samples at time $t \in \mathcal{T} = \{1, 2, \dots, T-1\}$, defined as the local data set $\mathcal{D}_m(t)$ of size $D_m(t)$. Note that, in a typical distributed learning mechanism under the synchronous settings, the observation time t is a single round of global interaction between the platform and the devices [9], [11]. Then, the collective data sample size at time t across the network is $D(t) = \sum_{m=1}^{M} D_m(t)$. In a supervised learning setting, $\mathcal{D}_m(t)$ is a collection set of data samples at device m defined as $\{x_i, y_i\}_{i=1}^{D_m}$ with $x_i \in \mathbb{R}^d$ corresponding label $y_i \in \mathbb{R}$. The data samples are informative about the learning model; hence, brings a value at the learner in terms of their contribution in improving the learning performance. We refer to it as *data type*. Following this intuition, we associate the type of data samples available at the devices as a realization of random variable Φ_i (explained in Definition 2). This setting can be extended to a more generalized form where each device share a stream of data samples with features accounting for the time instance, such as in the time series prediction.

[3]The proposed model is generic in a sense that it works well for training learning models, such as in FL, or an estimator minimizing the mean square error (MSE).

Consider $n \in \mathcal{N}(t)$ data samples available in the network for trading such that $D_m(t) \leq |\mathcal{N}(t)|$. Then, the goal of a supervised learner is to learn a single model defined in Definition 1.

Definition 1. *We define a supervised learner interested in minimizing the empirical risk with respect to parameter $w \in \mathbb{R}^d$ on all distributed data samples $D_m(t)$ as the finite-sum objective of the form*

$$\min_{w \in \mathbb{R}^d} J(w,t) \quad \textit{where} \quad J(w,t) := \sum\nolimits_{m=1}^{M} \frac{D_m(t)}{D(t)} \cdot J_m(w,t). \tag{1}$$

Then, the data market particularly looks at the contribution of each device m in solving (1), which is expressed as the empirical risk with respect to the improvement in $w \in \mathbb{R}^d$ on their local data set $\mathcal{D}_m(t)$ as

$$J_m(w,t) := \frac{1}{D_m(t)} \sum\nolimits_{i=1}^{D_m(t)} f_i(w). \tag{2}$$

For simplicity and without loss of generality, we make a common assumption: $f_i(w)$ is either a $(1/\gamma)$-smooth function or a L-Lipschitz continuous function (cf. [22]); hence, ensuring convergence of the solution. We note the network topology is not restrictive towards changes, i.e., the devices can perform data trading with each other via an arbitrator (a central learner) because they are connected through a network. Therefore, for the performance analysis of the proposed approach later on, we provide a scenario-based statistical analysis.

B. Data Type and Market Model

We make a common assumption that the market is interested in data exchange, and therefore, stimulates the devices with pricing signals based on the value of the traded data in improving the learning performance. In our setting, this translates to finding the type of data each device has. Then, for the offered pricing $p_m > 0$, every rational device m determines its strategy for participation $a_m \in \{0, 1\}$ in the data trading process so as to maximize their individual benefits; such strategies are captured in terms of the defined utility function. For this, we first define the type of data samples $\mathcal{D}_m$ of device m by a random variable Φ_m. To this end, we have the following definition.

Definition 2. *We define the data type $\Phi_m, \forall m \in \mathcal{M}$ as a composite measurement obtained following the preference* *profile ξ_m of the devices to share their available data, fully or partially, and the average marginal contribution value of supporting mini-batch of data points $z \in \mathcal{D}_m$ defined as $\theta_m(z, \mathcal{D}_m)$ that brings to the learner.*

Following Definition 2, we formally define the composite mapping $\xi_m(\theta_m(z, \mathcal{D}_m))$, where $\theta_m(z, \mathcal{D}_m)$ can be evaluated following a modified distributed Shapley [23] value for a known potential function of the global loss[4] $J(\cdot)$ (as defined in (1)) such that

$$\theta_m(z, \mathcal{D}_m) = \mathop{\mathbb{E}}_{\substack{i \sim [\bar{m}] \\ \tilde{D} \sim \mathcal{D}_m^{i-1}}} \left[\tilde{J}(\tilde{D} \cup \{z\}) - \tilde{J}(\tilde{D}) \right], \tag{3}$$

where, respectively, $\bar{m}$ is the mini-batch size, $\tilde{D}$ is the i.i.d. samples drawn from the data distribution $\mathcal{D}_m$ supported on $\mathcal{Z}$ with mini-batch of data point $z \in \mathcal{Z}$, and the potential function $\tilde{J} : \mathcal{Z} \to [0, 1]$ defined by the output $0 \leq \epsilon \leq 1$ such that $|\nabla \tilde{J}(w^{(t)})| \leq \epsilon |\nabla \tilde{J}(w^{(t-1)})|$.

More precisely, following (3), the mapping $\xi_m(\theta_m(z, \mathcal{D}_m)) \in [0, 1]$ quantifies data type as the expected value of data and the device's preference profile, i.e., the willingness to trade data with the learner, to offer that value in the data market. For simplicity, we use shorthand ξ_m for $\xi_m(\theta_m(z, \mathcal{D}_m))$, with $\xi_m = 0$ when the device m reserves no privacy concern on the shared data. In practice, we observe heterogeneity in ξ_m, which is an important metric that captures the function of individual preference on sharing data (i.e., data privacy), and thereof, each device may not reveal their true data type, or perform optimal local computation, as expected by the market, for the offered pricing scheme to participate in the data trading process. Hence, the learner face consequences of the partial knowledge in the state of information exchanged[5] in a setting where payments for traded data are provided after collecting them. We settle the aforementioned analysis with the formalization of an efficient trading mechanism in the proposed market model as below.

Definition 3. *The proposed data trading mechanism is a tuple $(\Pi, \Omega, \mathbf{p}, \mathbf{a})$, where Π is the coalition set following data types, Ω is the outcome space capturing the final learning performance, with $\Omega : \Phi \times \mathbf{a} \to [0, 1]$, $\mathbf{p} = (p_m(\Phi_m))_{m \in \mathcal{S}_\Pi}$ is the pricing vectors defined for coalition $\mathcal{S}_\Pi$, with $a_{m \in \mathcal{S}_\Pi} = 1$, and $\mathbf{a} = (a_m(\Phi_m))_{m \in \mathcal{S}_\Pi}$ is the vector of participation to tighten the information leakage due to data correlation.*

Definition 3 hints the underlying game-theoretic interaction between the learner and the devices for the mechanism design, summarized as the following. The learner (i) evaluates the receive data (including device's importance value towards privacy), and aims at (ii) quantifying the type of device's data so as to lower the offered pricing. Particularly, the traded data is evaluated for its contributing in improving the performance of the learning model, i.e., with $\Omega : \Phi \times \mathbf{a} \to [0, 1]$. Whereas, the devices $m \in \mathcal{M}$ with correlated data samples form a coalition $m \in \mathcal{S}_\Pi$ to challenge the learner in hiding their own data type Φ_m, or adopt sharing data in bundles to mitigate the information leakage and price depression. Following Definition 3, the mechanism aims to foster improved participation for training learning models while addressing impacts of data correlation on the offered pricing.

With these preliminaries, next, we formally start to tackle the research problems **Q1**, **Q2**, and **Q3**, raised in Section I-B, with the considered simple setting. In the following, we present an overview of the problem formulation about data trading in the IoT data market and formalize the data valuation procedure as per the data properties, resulting in specific utility models.

III. Problem Formulation

A. A basic setup

We revisit Definition 2 and make an assumption that the vector of random variables $\Phi = [\Phi_1, \Phi_2, \ldots, \Phi_m]$ follows a joint normal distribution $\mathcal{N}(\mu_\Phi, \Sigma)$, where $\Sigma \in \mathbb{R}^{m \times m}$ is the covariance matrix. This setup provides convenience in further analysis; we simply assume this to reflect the presence of devices with correlated data type. However, the developed framework is not limited to this assumption, as in the case otherwise, the problem eventually boils down to the deconstruction of the data type and our approach follows. Consider $a_m(t)$ as a binary decision variable for device m to join the data market such that

$$a_m(t) = \begin{cases} 1, & \text{if device } m \text{ joins the market at time } t, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Then, in every round of interaction with the learner for the offered pricing $p_m(t), \forall m$, the interested device (if in the agreement to participate) trade their data as the mixture of their data type and the learning parameters such that $S_m = f(\xi_m) + N_m$, where $N_m \sim \mathcal{N}(0, 1)$ is the Gaussian noise. In this regard, as shown in [7], the learner can have an estimate of ξ_m with the traded data D_m with a solution to minimization of the estimation error of the data type. In doing so, the learner can employ both convex/non-convex loss function in (3) that defines the data type of a device. This means, the learner can efficiently reconstruct the mapping function $\hat{\phi} =< g(\xi_m)|_{m \in \mathcal{M}} >$ to derive $\xi_m, \forall m$ precisely by solving the squared-error minimization problem as

$$\operatorname*{arg\,min}_{g(\xi_m)} \mathbb{E}\left[(\phi_m - g(\xi_m | D_m, a_m(t), p_m(t)))^2 \right], \forall m \in \mathcal{M}. \tag{5}$$

In this regard, we outline the following three cases:

- **Case I:** When $\xi_m = 0$, i.e., $f^{-1}(\phi_m) = 0, \forall m$, the learner adopts the following economic properties.
  - (i) **Monotonicity:** If we have $\mathcal{D}_i \subseteq \mathcal{D}_j$ for any pair of devices $i, j \in \mathcal{M}$, then following the standard assumption of monotonicity in the valuation on data defined as $v(\cdot)$, we have $v(D_i) \leq v(D_j)$.
  - (ii) **Additive:** If we have $\mathcal{D} = \mathcal{D}_i \cap \mathcal{D}_j$ for any pair of devices $i, j \in \mathcal{M}$, then we have $v(D) \leq v(D_i) + v(D_j)$.

[4] Potential function reflects the performance metric in terms of learner's model accuracy.

[5] This is often termed as *information asymmetry*.

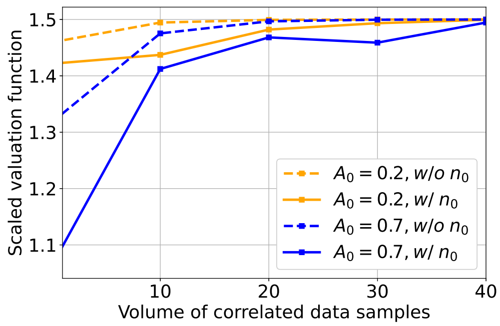

Figure 4: **Case study** on variability in scaled valuation function at the learner in terms of model precision: a two sellers one buyer scenario.

Considering these two properties, if the learner already received data $\mathcal{D}_i$ such that $\mathcal{D}_i, \mathcal{D}_j \subseteq \mathcal{D}$, then it formalizes the valuation function for the data $\mathcal{D}_j$ as

$$v^{D_i}(D_j) = \gamma \cdot v(\mathcal{D}_i \cap \mathcal{D}_j) + v(\mathcal{D}_j \setminus \mathcal{D}_i | \mathcal{D}_i \cap \mathcal{D}_j), \forall i, j \in \mathcal{M} \quad (6)$$

where $\gamma \geq 1$ is a design parameter quantifying the effect of available data samples at the learner. The right hand part characterizes the marginal contribution of the remaining data samples. In fact, the valuation can be explicitly defined in terms of its contribution in improving the learning performance. As an example, we discuss the following case study.

Case Study: Take $v(\mathcal{D}_i \cap \mathcal{D}_j)$ as a log-concave valuation function, defined according to the experimental results in [24], of the learning precision (or accuracy) ζ such as $J(\zeta)$, where $\zeta = 1 - A_0 e^{-2|\mathcal{D}_i \cap \mathcal{D}_j|(1-n_0)}$ for a known A_0 defined as per the learning problem (1) and n_0 is the noise factor sampled from $\mathcal{N}(0.5, 1)$. To put in a context, the noise factor n_0 simply captures the notion of unreliable connectivity. In this regard, Fig. 4 identifies the variability in the scaled valuation function, measured from the buyer's perspective, in terms of model precision for a scenario with two sellers having correlated information and a buyer acting as the learner. We observe the addition of a random noise factor lowers the valuation function, i.e., a negative impact of unreliable connectivity on data trading, which is quite intuitive and straightforward. However, we also see a positive contribution of obtained information on the volume of correlated data samples that in return maximizes the valuation function of learner. □

- **Case II:** When $\xi_m > 0, \forall m$, the learner only has access to a subset of device's data (in the best case scenario), or just partial data (for example, the learning parameters). In this later scenario, the learner can use several distance measures, such a L2-norm, cosine similarity and so on, to figure out correlated learning parameters.
- **Case III:** When $\xi_m \geq 0, \forall m$, i.e., a particular case of I and II.

Next, in the following, we define the relevant utility function for the learner and devices to characterize these properties in the IoT data market.

B. Utility formulations

The learner exploits the solution obtained from solving (5) to maximize its advantage of knowing the data types of devices for a best-suited pricing scheme. In particular, the learner aims at maximizing the following problem:

$$\sum\nolimits_{m \in \mathcal{M}} V(g(\xi_m) | D_m, a_m(t), p_m(t)) - a_m p_m, \quad (7)$$

where $V(\cdot)$ is a non-decreasing, concave valuation function evaluated at the learner knowing data types to lower down the pricing. We later discuss the details of it. On the contrary, with the given pricing, the devices intend to lower the risk of exposing their data types to the learner. An approach to address this concern while to have a method that captures the concerns raised in **Q1**, **Q2** and **Q3** is the following maximization problem for each device:

$$\left[\sum\nolimits_{i \in \mathcal{M} \setminus m} V_i(\cdot) + a_m p_m \right] - \delta_m V_m(D_m, a_m(t), p_m(t)), \quad (8)$$

where $V_{i \in \mathcal{M} \setminus m}(\cdot)$ is the value of added data in the data market, $\delta_m \geq 0$ is the sensitivity of the data market in optimization (5) over the revealed data type for device m. However, solving (8) exerts additional communication overhead to calculate the valuation of all participating devices in the market, and therefore, is inefficient in deriving low-cost, distributed solution to meet the research objectives. Therefore, we need to redesign the interaction scenario between devices and the learner in the data market. To this end, we develop a composite objective that stimulates participation of the devices and brings value from the exchange of data between the devices and the learner in the data market. Our focus is to realize a trusted data market setting that brings participation of the devices with correlated data properties in a group, without uncontrolled competition and probable leakage of information about each other's data properties; hence, the market cannot unilaterally depress pricing.

C. Multi-Agent Joint Policy (MAJP): A distributed coalition strategy

We start by developing the underlying distributed coalition game structure of the posed problem statement as an optimization problem. We recast the interaction between devices and the learner considering the possibility of leakage of correlation information as a multi-agent cooperative game where the payoff during coalition is allocated amongst the devices for tightening correlated information. This fundamentally means the devices with corelated information negotiate to derive an equilibrium solution, where both of them benefit from the data market. Recall the data trading mechanism in Definition 3, this also means the coalition strategy works best of everyone's interest bringing higher value of data, pricing, privacy and learning. Herein, we also drop the notion of time t and evaluate the system for each round of interaction between the devices and the data market, which is a valid assumption to make.

Let $\mathcal{M}$ denotes a grand coalition and $\mathcal{S} \subseteq \mathcal{M}$ is a set of devices in coalition to protect their corelated data. In particular,

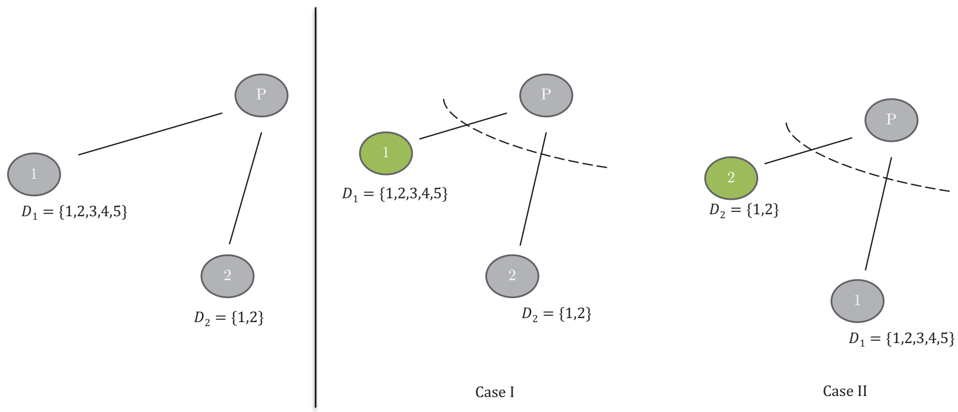

Figure 5: An illustrative experimental setup for data trading and value exchange.

a set of devices $\mathcal{S}$ agree to act as an single entity to negotiate with the value of their collective data $\mathcal{D}_\mathcal{S}$ with the platform during data trading. The value of coalition is therefore related to the pricing rate $p(\mathcal{D}_\mathcal{S})$ such that $p : P(\mathcal{M}) \times \mathcal{M} \rightarrow \mathbb{R}_{\geq 0}$, where $P(\mathcal{M})$ is the power set of $\mathcal{M}$. In what follows, we reuse the data type ϕ_m of device m, defined as a function of the importance value ξ_m it allocates for the privacy of data D_m, as a consensus constraint on the coalition property. Consider $\mathcal{A} = a_1 \times a_2 \times \ldots \times a_m$ is the action space defining device's joint agreement in the data trading process. Then, with reference to (8), the added contribution of coalition $\mathcal{S}$ in the system can be defined by a utility function as follows.

Definition 4. *For a given coalition* $\mathcal{S} \subseteq \mathcal{M}$, $f_\mathcal{S}((\phi_j(\mathcal{S}, \mathcal{A})) : \mathbb{N} \rightarrow \mathbb{R} \in [0, 1]$ *is a positive, concave utility function that adds return on investment for having a coalition and tightening the information about data types* $\phi_j, \forall j \in \mathcal{S}$.

Following Definition 4, we can define the coalition value instead of individual utilities as

$$v(\mathcal{S}, \mathcal{A}) = \sum\nolimits_{j \in \mathcal{S}} a_j \Bigg[\bigg(p_j(n_j) + f_\mathcal{S}(\phi_j(\mathcal{S}, \mathcal{A})) \bigg) - \bigg(p_j \cdot c(\Delta\phi_j) + c(\mathcal{S}) \bigg) \Bigg], \tag{9}$$

where $p_j(n_j), \forall j \in \mathcal{S}$ is a proportional pricing for n_j data available at the devices in coalition, $c(\Delta\phi_j)$ is the opportunity cost when leaking data type information to the learner following early trading, i.e., the learner is allowed to optimally minimize $\mathbb{E}\Big[(\phi_m - g(\xi_m | D_m, a_m(t), p_m(t)))^2\Big], \forall m \in \mathcal{S}$, and $c(\mathcal{S})$ is the cost of coalition defined in terms of total power required to exchange information on correlation. More formally, we define $p_j(n_j) = p_s \Big[\frac{n_j}{\bigcup_{j \in \mathcal{S}} n_j}\Big]$ for a defined budget p_s on the coalition $\mathcal{S}$, $c(\Delta\phi_j) = \sum_{i \neq j} g_{ij} a_i a_j, \forall i, j \in \mathcal{S}$, where g_{ij} is the normalized influence of device i to j due to correlation properties in the data.

Theorem 1. *For a single seller case, the optimal coalition value* $v^*(\mathcal{S}, \mathcal{A})$ *is proportional to the offered pricing* $p^*(n)$ *for trading data of samples* n*. In a multiple seller case, given a known cost of coalition* $c(\mathcal{S})$*, the optimal coalition value is proportional to the gain from tightening the information leakage due to data correlation and the availability of data samples itself.*

Proof. The first case is simple to prove. The absence of data rivalry leads to data trading with pricing signal sufficient enough for active participation of the data seller devices, leading $c(\Delta\phi_j)$ and $c(\mathcal{S})$ to zero, i.e., $a_j = 1$ and $a_i = 0, \forall i \in \mathcal{S} \setminus j$. Given $c(\mathcal{S})$ and a linear pricing scheme, maximizing the coalition value corresponds to minimizing the components that captures the impact of data correlation defined as $g_{i,j}, \forall i, j \in \mathcal{S}$ between sellers pair $\{i, j\}$, i.e., $p_j \cdot c(\Delta\phi_j)$; hence, the participation of devices to add value within coalition and maximize $f_\mathcal{S}((\phi(\mathcal{S}, \mathcal{A}))$ with more data samples. □

We note that (9) presents a holistic outlook to the problem that connects data value, pricing, privacy and learning in the IoT data market. In what follows, if we consider a typical learning problem (1) solved by the data market via data trading, it is of particular interest to realize the data value and its impact on the learning performance for a given pricing scheme, as a usual case in the data market. That also poses a feasible approach where the platform feedback the impact of parameter dissimilarity as in the opportunity cost $c(\Delta\phi_j)$ across devices due to their data properties. This can be achieved with the following definition.

Definition 5. *The devices participating in the data market exhibit parameter dissimilarity* $\rho_m \geq 0, \forall m \in \mathcal{M}$ *in terms of gradients on the global and local loss as* $||\nabla J_m(w) - \nabla J(w)|| \leq \rho_m, \forall w$.

Then, we can derive the average data dissimilarity in the coalition $\mathcal{S}$ following the Definition 5 as

$$\rho_\mathcal{S} = \sum\nolimits_{j \in \mathcal{S}} a_j \Bigg[\rho_j \cdot \frac{n_j}{\bigcup_j n_j}\Bigg]. \tag{10}$$

Illustrative analysis of coalition strategy on opportunity costs: As an appetizer, we set the availability of two devices in the trading system. Consider two sellers with data $\{\mathcal{D}_{i|i=1,2}\}$ and the buyer setups the data market for $\{\mathcal{D}_i \cup \mathcal{D}_j \subseteq \mathcal{D} | \mathcal{D}_i \cap$

Sellers	Case I	Case II
$\mathcal{D}_1 : \{1,2,3,4,5\}$	{1,2,3,4,5}	{3,4,5}
$\mathcal{D}_2 : \{1,2\}$	-	{1,2}

Table I: Illustrative example on data trading where Case I indicates seller $\{1\}$ approaching the platform first, and Case II, otherwise.

Cases	Coalition		No coalition	
	$c(\Delta\phi_1)$	$c(\Delta\phi_2)$	$c(\Delta\phi_1)$	$c(\Delta\phi_2)$
Case I:	1	-	0	1
Case II:	-	1	1	0

Table II: Analysis of coalition strategy on linear opportunity costs $c(\Delta\phi_j), j \in \{1,2\}$ for a unit cost per sample with $\mathcal{D}_{j|j=\{1,2,3,4,5\}}$ and $\mathcal{D}_{j|j=2} = \{1,2\}$.

$\mathcal{D}_j \neq \emptyset\}$ data samples, as shown in Fig. 5. Then, if $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively, posses a subset of data samples in the market, i.e., $\mathcal{D}_1 = \{1,2,3,4,5\}$ and $\mathcal{D}_2 = \{1,2\}$, then for a unit monetary value on the data dissimilarity $\rho_j, \forall j \in \{1,2\}$, we have two specific cases for data trading and the involved opportunity costs. Case I: Seller $\mathcal{D}_1$ trading its data first and Case II: Seller $\mathcal{D}_2$ trading its data first. In this regard, Table I shows the trading procedure and Table II evaluates the opportunity cost under two particular scenarios: (i) when a coalition is formed and (ii) when individual trading is performed. To elaborate, let's say Seller $\mathcal{D}_1$ considers for Case I and opt out of coalition to lower its opportunity cost. Then, $\mathcal{D}_2$ can switch for early trading scenario, i.e., Case II to lower its own opportunity cost; consequently, forcing $\mathcal{D}_1$ to form a coalition with seller $\mathcal{D}_2$. Likewise is the narratives on the stability of coalition formation under the Case II. This illustrative analysis establishes motive of devices to self-organize into coalition in the data market, which we later show is stable, to alleviate the impact of data similarity between them on the offered pricing and the aftermaths of data rivalry. $\square$

With respect to the above analysis, the data market intends to solve the following optimization problem in its general form.

$$
\begin{aligned}
\textbf{P:} \quad & \underset{\{a_i, p_i\}_{i\in\mathcal{S}}}{\text{maximize}} && v(\mathcal{S}, \mathcal{A}) && \text{(11a)}\\
& \text{subject to} && \textstyle\sum_{i\in\mathcal{S}} p_i = p_{\mathcal{S}}, && \text{(11b)}\\
& && V_{i\in\mathcal{S}}(D_i, a_i, p_i|\phi_i) \geq 0, \forall i \in \mathcal{S}, && \text{(11c)}\\
& && V_{i\in\mathcal{S}}(D_i, a_i, p_i|\phi_i) \geq \\
& && V_{i\in\mathcal{S}'}(D_i, a_i, p_i|\phi_i), \mathcal{S}' \in \mathcal{S} \setminus i, && \text{(11d)}\\
& && a_i \in \{0,1\}, \forall i \in \mathcal{S}, && \text{(11e)}\\
& && \rho_{\mathcal{S}} \leq \rho, && \text{(11f)}\\
& && c(\mathcal{S}) \text{ is bounded}, && \text{(11g)}\\
& && c(\Delta\phi_i) \leq \phi_i^{\text{th}}, \forall i \in \mathcal{S}, && \text{(11h)}
\end{aligned}
$$

where (11b) is the budget constraint available to distributed amongst the members in coalition; constraints (11c) and (11d) jointly captures the positive valuation in participation; (11e) defines the participation strategy of the devices, (11f) quantifies the measure of average data dissimilarity such the coalition is stable; (11g) is the accepted tolerance on the cost of coalition, and (11h) is the bound on individual opportunity cost of the members in coalition. We notice, the optimization problem **P** is hard to solve and mostly intractable due to (i) binary constraints, (ii) the stability of the mechanism due to the coupling in data types and valuation for the unknown heterogeneity in data distribution, for a large number of devices, and (iii) private cost information. To address the technical challenges defined for solving **P**, in the following, we recast the market design so as to offer a novel cooperation protocol that mitigates the tension between data sellers and the data market using the pricing signal for the exchange of value and data. In particular, we characterize a subset of coalition is formed accordingly to the data types, and then, derive association and pricing scheme based on the properties of devices associated in the particular coalition. Note that, the obtained solution is sub-optimal, but a low-complexity alternative to address the aforementioned challenges.

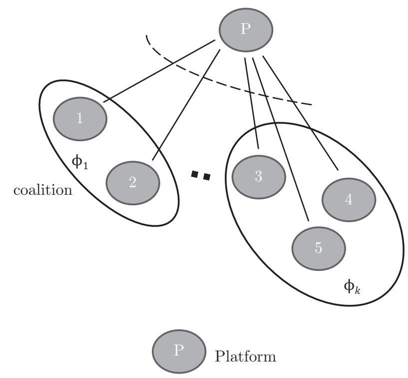

Figure 6: An illustrative snapshot of the distributed coalition game.

IV. Market Design: MAJP as a Coalition Game

In the best interest of the devices ("agents"), we observe, as illustrated in the experimental (subsection III-C), devices having *similar data types form coalition* to maximize their utility and reach stability in the data market offering transferable utility (TU). Hence, we formalize the elegant framework of distributed coalition games, the *Hedonic game* [18], [25] to solve the problem of distributed coalition with the objective of maximizing **P**. In fact, it is intuitive that the devices have individual preferences to form coalition groups with similar data types, which is a common concept for coalition-based games [18]. This captures two necessary conditions to design the Hedonic game: (i) the payoff of the devices is defined only based on the other members in the coalition, and (ii) the coalition structure is the direct consequence of preference profiles of each device in the coalition.

Remark 2. *The interaction between a single learner and a finite set of devices (sellers) in coalition upon the pricing signal and the value of data exchange together form a coalition game to protect their corelated data. Formally, the game is characterized to capture the coalition strategies of devices and the update in pricing signals of the learner towards maximization of coalition during participation in data market.*

Definition 6. *A coalition partition is defined as the set $\Pi = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_K\}$ dividing the total set of devices $\mathcal{M}$ in the system such that $\mathcal{S}_k \subseteq \mathcal{M}$ and $\cup_{k=1}^{K} \mathcal{S}_k = \mathcal{M}$, where $\mathcal{S}_k$ are the coalitions sets based on device type $k \in \{1, 2, \ldots, K\}$. Then, we have the following preference definition for the device participating the coalition [25].*

Definition 7. *The preference profile of any device m is defined by the relation or an order $\succeq_m$ that is a complete, reflexive, and transitive binary relation over the set $\{\mathcal{S}_k \subseteq \mathcal{M} : m \in \mathcal{S}_k\}$.*

Following Definition 7, we have for any pair of coalition sets $\mathcal{S}_1$ and $\mathcal{S}_2$, $\mathcal{S}_1 \succeq_m \mathcal{S}_2$ means device $m \in \mathcal{M}$ prefers coalition S_1 (or least equally prefers both), than S_2.

Then, formally, with the given set of devices $\mathcal{M}$ as players and their preference profiles $\succeq_m, \forall m \in \mathcal{M}$, a Hedonic coalition game can be defined as follows.

Definition 8. *A Hedonic game is defined by the pair of set of players (i.e., the devices) and their preference profiles $(\mathcal{M}, \succ)$.*

Once the coalition $\mathcal{S}_k$, for illustration, as shown in Fig. 6, with TU (9) is agreed upon by the devices, the coalition utility can be divided amongst the devices as the payments, quantified in the form of contract. In our formulation, we define the value of coalition as the coupling between the obtained overall revenue due to participation in the data market and the consequence of limited information leakage due to correlation properties amongst the devices data. Particularly, the payment under *contracts* for device $m \in \mathcal{M}$ is defined as $p_m(\mathcal{S}_k) = p_{\mathcal{S}_k}\left[\frac{n_m}{\bigcup_{m \in \mathcal{S}_k} n_m}\right]$ for a obtained revenue $p_{\mathcal{S}_k}$ due to coalition, where $\sum_{\mathcal{S}_k \in \Pi} p_{\mathcal{S}_k} = p_{\mathcal{S}}$ on the coalition $\mathcal{S}_k$, using the earlier defined concepts.

Next, following Definition 7, we evaluate the preference profile of the devices as follows. Let's define $\mathcal{S}_{\Pi}(m)$ is the coalition set where the device m should belong following its type, i.e., $m \in \mathcal{S}_k$ such that $\mathcal{S}_{\Pi}(m) = \mathcal{S}_k \in \Pi$. Then, as explained in Section II, the preference of devices is defined as $\mathcal{S}_1 \succeq_m \mathcal{S}_2 \Leftrightarrow u_m(\mathcal{S}_1) \succeq_m u_m(\mathcal{S}_2), \forall m \in \mathcal{M}$, where $u_m : 2^{\mathcal{K}} \rightarrow \mathbb{R}$ is the preference function of any device m such that

$$u_m(\mathcal{S}) = \begin{cases} p_m(\mathcal{S}), & \text{if } m \in \mathcal{S}_{\Pi}(m), \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

In doing so, the devices verify the conditions of bound on their individual opportunity costs and the measure of parameter dissimilarity, as defined in Definition 5. Then, it is quite straightforward to have the preference function as (12). The devices in coalition gets benefit from the TU obtained following contracts mechanism, as discussed before, where two specific economic properties (or conditions) are satisfied: (i) *Individual Rationality (IR)*, a condition that captures the motive behind devices undergoing distributed coalition with positive return on investment, (ii) *Incentive Compatibility (IC)*, a condition that ensures devices get to maximize their utilities, as in the form of obtained payments, if they act as per their type.

Algorithm 1 MAJP Solution with Coalition Formation

1: **Initialization:** Partition Π_{initial} with devices in set $\mathcal{M}$ having a total of D data samples at $t \in \mathcal{T}$.
2: **Output:** Stable partition Π_{final}, participation vector $\mathbf{a} = \{a_1, a_2, \ldots, a_m\}, \forall m \in \mathcal{M}$, pricing signal $\mathbf{p} = \{p_{\mathcal{S}_1}, p_{\mathcal{S}_1}, \ldots, p_{\mathcal{S}_k}\}$, $\mathcal{S}_k \in \Pi_{\text{final}}$.
3: **Phase I**:
Private discovery of device types;
Execute Algorithm 2;
4: **Phase II**:
Distributed coalition formation;
5: **repeat**
6: **for all** device $m \in \Pi_{\text{initial}}$ **do**
7: Randomly select two coalitions;
8: Evaluate the preference function for the given coalition with (12) and preference profile (Definition 4);
9: Invoke switch operations between coalition groups, comparing possible payoffs;
10: Add device to the observed coalition sets in Π_{initial};
11: Repeat evaluation of preferences on different coalition groups (line 6) until no further switch operations exists;
12: **end for**
13: **until** Π_{final} is reached;
14: **Phase III**: Computation of coalition value using (9);
15: Evaluate the final pricing signal $\mathbf{p}$;

In Algorithm 1, we develop a solution approach to the optimization problem **P**, where the objective is to derive coalition partitions following device types that uniquely maximizes the overall coalition value. In particular, we adopt the modified merge-and-split algorithm [18], [26], [27] which works in an iterative manner on two coalitions at a time, and design a solution that reaches stable partitions. To that end, the following theorem is useful.

Theorem 2. *While satisfying IR and IC constraints, we show, the developed MAJP framework offering the distributed coalition solution with payment under contracts to realize an IoT data market for model training yields the following desirable properties.*

(i) Budget balance – Following Definition 7, for any device $m \in \mathcal{S}_k$, the sharing of total per group budget allocation $p_{\mathcal{S}}$ between the coalition members is proportional to the value of their participation such that $\sum_{\mathcal{S}_k \in \Pi} p_{\mathcal{S}_k} = p_{\mathcal{S}}$ is satisfied.
(ii) Linearity (within the coalition) – The linearity property, basically implying the revenue allocation for the exchange of data $\mathcal{D}_1$ and $\mathcal{D}_2$ from devices $\{1\}$ and $\{2\}$, respectively, is the same as any one device trading $\mathcal{D}_1 \cup \mathcal{D}_2$, comes as a direct consequence of proportional payoff within the group.
(iii) Truthfulness – The mechanism is truthful such that there exists a non-negative payoff proportional to the data contribution, as defined in the coalition *contracts*, only

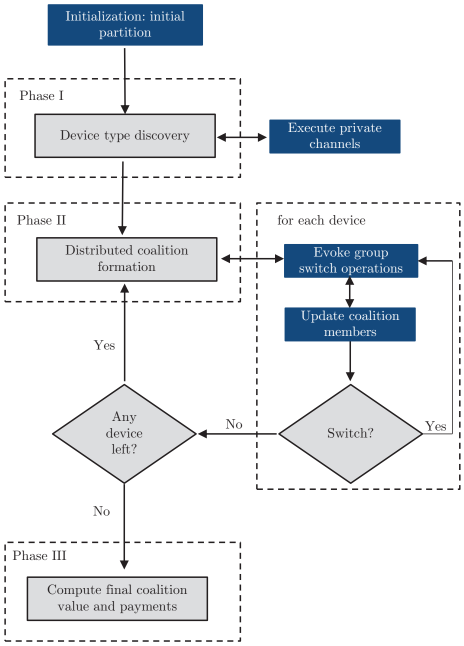

Figure 7: Execution flowchart of MAJP solution.

if the device form coalition as per its data type, and the linearity. It readily follows the definition of preference profile of any device in (12) and the designed corresponding individual utility function, which penalizes untruthful reporting.

(iv) Symmetry – The output of the mechanism is invariant to any permutation of participation during coalition formation. The proof follows for a finite set of devices, given a mechanism satisfying properties (ii) and (iii).

(We rely on the above proof-sketch and omit further details on analytical proof of Theorem 2, as it straightforward with the given explanation.)

Algorithm 1 operates in three phases, as shown in Fig. 7; **Phase I:** device type discovery, **Phase II:** distributed coalition formation, and **Phase III:** computation of final coalition value and payments. In Phase I, the devices exploit a secured MPC channel to compare their type with available proxy device type sets. To allow its successful execution, we assume devices can "ping" each other through broadcast network or using beacons during discovery phase to evaluate individual devices types, similar to [28]. In this regard, the cost of coalition can be explicitly defined as units per round of communication. In Phase II, we employ the execution of split-and-merge algorithm [18], where the devices are allowed to make a switch between two coalition groups at a time, following their preference profile on the device types and utility functions, that maximizes their payoff. Following Definition 5, the bound on individual opportunity costs, and the amount of information leakage, the devices perform coalition switch operations as per their preference profile. This iterative procedure eventually leads to a stable coalition structure, as proven in [18], [27]. Finally, in Phase III, once the stable partition Π_{final} is reached, the coalition value is calculated to further define the final pricing signal for the data trading.

Algorithm 2 Private Discovery of Device Types

1: Begin with initial partition Π_{initial} of K with proxy data type sets in $\mathcal{K}$
2: Permute devices within $\mathcal{S}_k, k \in \{1, 2, \ldots, K\}$ over available private channel to evaluate device data type.
3: Return evaluated device types to individual devices $m \in \mathcal{M}$.

A. *Complexity analysis*

The complexity analysis of the proposed approach is done following the "propose and swap" method of stable matching algorithm with externalities [29]. Particularly, the impact of preference profiles with the number of devices and their interactions during the distributed coalition formation with merge-and-split algorithm is an instance of the propose and swap method, where devices opt to join coalition based on their type so as to maximize formulated individual utility-leading to stability. We begin with two randomly selected coalitions; hence, $\binom{M}{2}$ defines the number of possible switch between coalition groups and $M \times \Pi_{\text{initial}}$ number of possible options to split and merge as a member of coalition group. Following to which, this leads to a sub-linear complexity $\mathcal{O}(M\Pi_{\text{final}} \log(M \times \Pi_{\text{initial}}))$ with the number of devices and coalitions formed, similar to the analysis done in [29]. In the following, we provide simulation results to evaluate the performance analysis of Algorithm 1.

V. Numerical Results

In this section, we evaluate, compare, and validate the performance of the proposed market model with intuitive baselines. To begin with, we conduct statistical analysis to first measure the information leakage due to data correlation in its simplified version. For this, we consider a few number of seller devices available in the market generating explanatory data samples for trading, with the quality of data defined for the model's performance at the learner, as in [24]. Second, we the measure the impact of information leakage on data valuation and, with experimental, show the impact of data rivalry on value depression. Finally, we conduct numerical evaluations of the proposed solution; particularly, we follow a linear pricing scheme to compare the performance of our proposed approach in the IoT data market.

A. *Analysis on data correlation*

In this subsection, we conduct the statistical analysis on information leakage for a three seller $\mathcal{M} = \{1, 2, 3\}$ and one buyer scenario in the developed IoT data market. For simplicity, we consider each device's data provide equal marginal contribution to the learner, and the data samples are revealed in a sequence. As we talked about FL setting, the contribution of data samples is, in fact, reflected in the gradient (or parametric) response provided by the individual sellers. Based on this analysis, we show the impact of data correlation on information leakage. Note that, absence of this simplification won't alter the result, whatsoever, as the framework well-captures individual contribution of data samples in improving

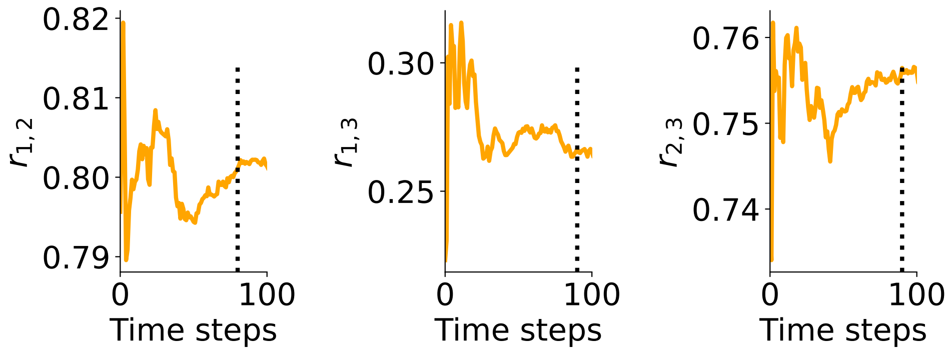

Figure 8: Statistical analysis on information leakage $r_{i,j}, \forall i,j \in \mathcal{M}$ to reflect data rivalry with $\mathcal{M} = \{1,2,3\}$ in a three sellers one buyer scenario with a precision of 5×10^{-4}. We consider each round indicates revelation of a single data sample with the leaner.

the learner's model following (3). We associate each device with a random variable and generate synthetic dataset using Gaussian process priors; the objective is limited to finding the joint probability distribution, and the data type is quantified following a uniform k-level quantization on the model performance, i.e., in the order of contribution in the model improvement, as defined in (3). The definition of average data similarity in (10) is used to perform such quantization, equivalently, satisfying the constraint in **P**. For instance, using in a three sellers setup represented with random variables (RVs) X, Y, and Z, respectively, we model $Z = 0.5X + Y$, with X$\sim \mathcal{N}(\mu_X, \sigma_X)$ and Y$\sim \mathcal{N}(\mu_Y, \sigma_Y)$, where the pairs (μ_X, σ_X) and (μ_Y, σ_Y) defines mean and standard deviation of the corresponding RV. Then, as shown in Fig. 8, the buyer learns data correlation $r_{i,j}, \forall i,j \in \mathcal{M}$ and reach convergence while obtaining device data type information with a precision of 5×10^{-4}. In practice, the sellers can use a batch of data samples; hence, the time step taken is much less than the approach where each sample is revealed. With this, we next evaluate the impact of information leakage on the data valuation.

B. Value depression with information leakage

We use the valuation function following linear pricing models defined as $V = 1/(1+g)\left[\sum_{i=1}^{M}\left(\frac{D_i}{\sum_j D_j}\right)^{1-bp}\right]^{1/b}$ for each device, with information leakage factor $0 \leq g \leq 1$ and a positive weight factor $b > 0$ capturing the characterstics of valuation function. We set $b = 0.1$ and the range of offered pricing in $[1, 10]$ monetary units. We observe the influence of the correlated data of seller $\{2\}$ on the data valuation of seller $\{1\}$ without the execution of proposed coalition solution approach. Fig. 9 shows the data valuation is proportional to the number of data samples, which is intuitive, and also with the offered pricing signal; however, it drops significantly as the learner identifies data properties between the sellers. Interestingly, as shown in Fig. 9, the price depression is prominent for seller $\{2\}$, when seller $\{1\}$ is given the competitive advantage of arrival in the market. This is obvious given the characterization of data properties between $\{2\}$ and $\{1\}$ in Fig. 8 in quantifying their data type.

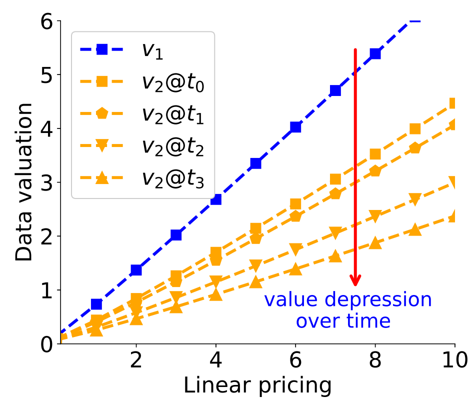

Figure 9: Evaluation of value depression with the offered linear pricing.

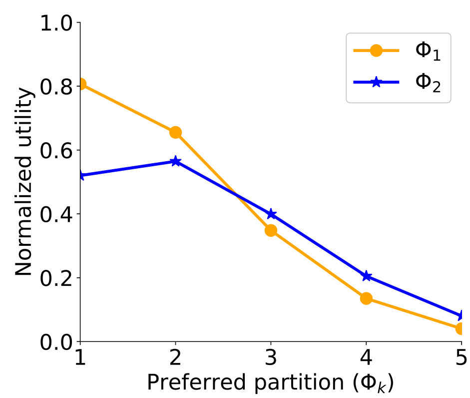

Figure 10: Impact of coalition strategy as per individual data type on normalized utility of devices.

Next, we evaluate individual utilities of the devices following our proposed MAJP solution approach, and provide further analysis of the developed coalition strategy.

C. Analysis of coalition strategy

Fig. 10 shows the impact of switch operation between coalition groups on the normalized utility for each device and the identified data types. As discussed before, while the data

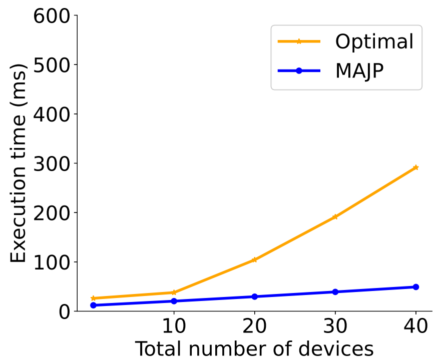

Figure 11: Computational complexity in terms of the execution time (in ms).

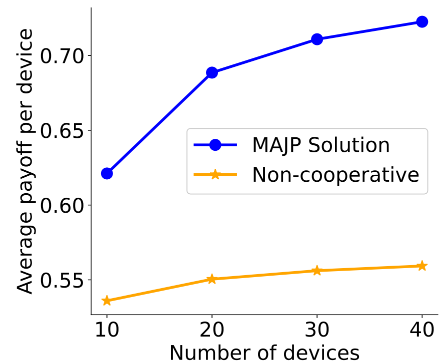

Figure 12: Performance comparison in terms of average payoff per device in coalition.

similarity constraints and the cost of coalition are satisfied, any deviation of the sellers to the group different from their true data type consequently lowers its utility. This hints the seller will opt to join appropriate coalition and undergo data trading in a group so as to maximize their utilities. We also observe, as the data are not perfectly correlated, the impact of information leakage won't penalize normalized utility value to zero, but only lowers it. Interestingly, we have an intuitive observation in Fig. 10 – joining a nearby coalition group is more beneficial for the sellers as the impact of information leakage is higher, otherwise. In Fig. 11, we validate the sub-linear complexity of the MAJP solution. For this, we set $\Pi \in \{1, 2, 3, 4, 5\}$ and include number of devices per group as 1, 10, 20, 30, and 40, respectively. We shuffle the devices and their data type, and compare execution of the MAJP solution with the Optimal [30]. The combinatorial nature of the coalition formation with the increased number of partitions and the number of associated devices results the Optimal solution to be computationally expensive as compared to the proposed MAJP solution.

Next, we consider the following intuitive baseline along to show the gain of adopting coalition strategy than individual interaction with the learner under a scenario with the information leakage.

- Non-cooperative: The learner exploits data properties between the sellers and impose price depression.
- MAJP solution (Cooperative): The pricing allocation follows the proposed solution approach in Algorithm 1.

For this evaluation, we reuse the linear pricing scheme with a log-concave utility on the coalition strategy adopted by devices with similar data type to lower information leakage, as illustrated in Fig. 8. For simplicity, we set $\Pi \in \{1, 2\}$ and include the number of devices per group as 10, 20, 30, and 40. The results are then obtained following Monte-Carlo simulations to check and validate the consistency of the obtained results. In Fig. 12, we observe the proposed MAJP solution provides a gain of up to 32.72% while imposing collaborative interaction between the devices with similar data type. Interestingly, we also observe an almost flat payoff when devices opt a non-cooperative strategy. This is reasonable given the value of information leakage with a fixed similarity in the number of data samples across devices. In this manner, the sellers benefit from coalition to tackle price depression and the information leakage to setup a trusted IoT data market.

VI. Conclusion and Discussion

In this work, we have proposed an approach that establishes a trusted IoT data market by reinforcing collaboration opportunities between devices with correlated data to avoid information leakage. We set out to tackle the challenges posed due to the loop of mistrust in the data market; we jointly study three research questions (as indicated in Q1, Q2, and Q3), where we have shown devices with similar data types can cooperate to deal with the price depression, data rivalry, and uncontrolled participation issues in the data market. We have formalized a network-wide optimization problem that maximizes the social value of coalition between the IoT devices of similar data types while minimizing the overall costs, defined in terms of network externalities, i.e., the impact of information leakage due to data correlation, and the opportunity costs. The formulated problem is intractable due to binary constraint and hard to solve directly given the presence of private information; thereby, we have developed a novel cooperative protocol, namely MAJP, that offered a sub-linear complexity in obtaining the solution using preference-based coalition strategy. To that end, we have shown, via statistical analysis and numerical evaluations, our proposed approach provides benefits (around 32.72% gain) as compared to the non-cooperative baseline, revealing truthful participation of devices without uncontrolled competition due to the information leakage and data rivalry.

We remark that the proposed approach could open additional benefits. For example, knowing others data properties a priori also indicate devices can learn when it is reasonable to collaborate for training learning models, as discussed in [31]. An interesting direction for future work is to consider a more practical network setup with intermittent links and resource constraints a the IoT devices. Another aspect is to better quantify the amount of privacy leakage by using notions of differential privacy [32] or using multi-party computation [31], and develop closer-to-the-real-world utility models. We also foresee challenges in implementing this in a practical scenario given the scale of additional signaling required to identify the data properties. It would be interesting to study the scalability issues in a pure distributed network architecture.

REFERENCES

[1] I. Report, “The growth in connected iot devices is expected to generate 79.4zb of data in 2025, according to a new idc forecast.” [Online]https://www.idc.com, 2019. (Accessed on 12/04/2020).

[2] 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.

[3] M. Marjani, F. Nasaruddin, A. Gani, A. Karim, I. A. T. Hashem, A. Siddiqa, and I. Yaqoob, “Big iot data analytics: architecture, opportunities, and open research challenges,” *ieee access*, vol. 5, pp. 5247–5261, 2017.

[4] 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.

[5] Y. Liu, J. James, J. Kang, D. Niyato, and S. Zhang, “Privacy-preserving traffic flow prediction: A federated learning approach,” *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7751–7763, 2020.

[6] P. Vepakomma, S. N. Pushpita, and R. Raskar, “Private measurement of nonlinear correlations between data hosted across multiple parties,” *arXiv preprint arXiv:2110.09670*, 2021.

[7] D. Acemoglu, A. Makhdoumi, A. Malekian, and A. Ozdaglar, “Too much data: Prices and inefficiencies in data markets,” tech. rep., National Bureau of Economic Research, 2019.

[8] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, “Federated learning for mobile keyboard prediction,” *arXiv preprint arXiv:1811.03604*, 2018.

[9] 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.

[10] 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.

[11] 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.

[12] T. H. T. Le, N. H. Tran, Y. K. Tun, M. N. Nguyen, S. R. Pandey, Z. Han, and C. S. Hong, “An incentive mechanism for federated learning in wireless cellular network: An auction approach,” *IEEE Transactions on Wireless Communications*, 2021.

[13] Y. Jiao, P. Wang, D. Niyato, B. Lin, and D. I. Kim, “Toward an automated auction framework for wireless federated learning services market,” *IEEE Transactions on Mobile Computing*, 2020.

[14] S. N. Ali, G. Lewis, and S. Vasserman, “Voluntary disclosure and personalized pricing,” in *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 537–538, 2020.

[15] S. Gupta, D. C. Jain, and M. S. Sawhney, “Modeling the evolution of markets with indirect network externalities: An application to digital television,” *Marketing Science*, vol. 18, no. 3, pp. 396–416, 1999.

[16] S. Ichihashi, “The economics of data externalities,” *Journal of Economic Theory*, vol. 196, p. 105316, 2021.

[17] D. B. Rubin, “Statistical disclosure limitation,” *Journal of official Statistics*, vol. 9, no. 2, pp. 461–468, 1993.

[18] K. R. Apt and A. Witzel, “A generic approach to coalition formation,” *International game theory review*, vol. 11, no. 03, pp. 347–367, 2009.

[19] D. Niyato, M. A. Alsheikh, P. Wang, D. I. Kim, and Z. Han, “Market model and optimal pricing scheme of big data and internet of things (iot),” in *2016 IEEE international conference on communications (ICC)*, pp. 1–6, IEEE, 2016.

[20] W. Mao, Z. Zheng, and F. Wu, “Pricing for revenue maximization in iot data markets: An information design perspective,” in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pp. 1837–1845, IEEE, 2019.

[21] Q. Li, Z. Li, Z. Zheng, F. Wu, S. Tang, Z. Zhang, and G. Chen, “Capitalize your data: Optimal selling mechanisms for iot data exchange,” *IEEE Transactions on Mobile Computing*, 2021.

[22] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik, “Federated optimization: Distributed machine learning for on-device intelligence,” *arXiv preprint arXiv:1610.02527*, 2016.

[23] R. Jia *et al.*, “Towards efficient data valuation based on the shapley value,” in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1167–1176, PMLR, 2019.

[24] Y. Zhan, P. Li, Z. Qu, D. Zeng, and S. Guo, “A learning-based incentive mechanism for federated learning,” *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6360–6368, 2020.

[25] A. Bogomolnaia and M. O. Jackson, “The stability of hedonic coalition structures,” *Games and Economic Behavior*, vol. 38, no. 2, pp. 201–230, 2002.

[26] K. R. Apt and T. Radzik, “Stable partitions in coalitional games,” *arXiv preprint cs/0605132*, 2006.

[27] W. Saad, Z. Han, M. Debbah, A. Hjorungnes, and T. Basar, “Coalitional game theory for communication networks,” *Ieee signal processing magazine*, vol. 26, no. 5, pp. 77–97, 2009.

[28] I. Clarke, O. Sandberg, M. Toseland, and V. Verendel, “Private communication through a network of trusted connections: The dark freenet,” *Network*, 2010.

[29] A. E. Roth and M. Sotomayor, “Two-sided matching,” *Handbook of game theory with economic applications*, vol. 1, pp. 485–541, 1992.

[30] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1998.

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

[32] 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.