Package Exchange Mechanisms with Quality Constraints for Data Markets

Malcolm Egan, Nir Oren

Package Exchange Mechanisms with Quality Constraints for Data Markets

Malcolm Egan[1*] and Nir Oren[2]

[1*]Univ Lyon, Inria, INSA Lyon, CITI, 6 Avenue des Arts, Villeurbanne, 69621, France.

[2]School of Natural and Computing Sciences, University of Aberdeen, Meston Building, Aberdeen, AB24 3UE, United Kingdom.

*Corresponding author(s). E-mail(s): malcolm.egan@inria.fr;
Contributing authors: n.oren@abdn.ac.uk;

Abstract

In this paper, we consider matching and pricing of buyers and sellers in data exchanges. Buyers are data consumers that wish to obtain data of a sufficient quality to achieve their goals and who may value quality levels differently. Sellers are data providers that desire to properly value the data that they are selling. Our goal in this work is to develop a tractable formulation of the winner determination problem for such a data exchange, and we show that the problem can be solved via bi-level optimization methods. We also examine how different pricing rules are affected when a data provider is able to replicate their data and thus sell it to multiple buyers. We demonstrate that Vickrey and Balanced Winner Contribution rules can introduce inherent disincentives for data replication. Therefore, we introduce a new rule, the modified Balanced Winner Contribution rule, and show that it can provide flexible incentives for data replication in thin markets.

Keywords: Data Market, Package Exchange, Data Quality, Pricing

1 Introduction

Data has become the lifeblood of organizations, and is fundamental to nearly all real-world decision-making processes. The rise of machine learning and AI systems has meant that large volumes of data are needed for decision making, for example when trying to understand consumer preferences, or indeed, when training AI systems. This has led to situations where specialist *data service providers* (DSPs) curate or generate data which is packaged for provision in the form of a *data product*. Furthermore, different DSPs may have access to datasets covering the same domain (e.g., different computer vision data providers may have large corpora of tagged images), and can offer the data at different quality levels (e.g., different resolutions, or levels of tagging). At the same time, those requiring the data, the *data consumers*, may have different quality needs with regards to the data (for example due to cost, storage, transfer, or data processing constraints). Different data consumers may require access to similar data products, and may desire exclusive access (to obtain a competitive advantage) or non-exclusive access to these products. Moreover, data consumers may require data from multiple data providers, such as in the context of collaborative machine learning [1]. Combined, all these requirements suggest the need for a data marketplace where DSPs and consumers are matched, and where appropriate prices for data products can be identified.

A popular framework for matching buyers and sellers of combinations of goods in applications such as spectrum auction design [2] is the family of package exchanges [3]. Package exchanges can be viewed as a generalization of a package auction allowing for multiple buyers. However, there are critical differences between traditional applications of package exchanges and data markets. For one, it may be possible for DSPs to provide data products at multiple different quality levels. Moreover, organizations acting as data consumers may require that data products satisfy some quality constraints. A natural question is then how to develop package exchange mechanisms for data markets which can deal with such quality constraints.

In this paper, we address the problem of mechanism design for data markets with data quality constraints. The main contributions of our work are:

(i) The development of a market mechanism targeting data markets. This market mechanism generalizes the standard package exchange mechanism by incorporating data quality constraints. More specifically, we introduce a bid structure where organizations can specify the data quality of each data product and the amount each agent is prepared to pay is parameterized by the data quality.

(ii) We develop a new variant of the winner determination problem which optimizes over both which agents trade, and the quality of traded data. We demonstrate that the winner determination problem can be reformulated as a bi-level optimization problem.

(iii) We study *value-based pricing*, which determines how the surplus is shared between the agents. This includes Shapley, Vickrey and balanced winner contribution [4] values, as well as a new *modified balanced winner contribution value*.

(iv) We study the impact of each value-based pricing rule on incentives for DSPs to replicate the data products they hold. We focus on two scenarios; namely, when data exclusivity constraints are present, and when it is desirable to incentivize data replication by DSPs (relevant for thin markets with few participants and heterogeneous demands).

The remainder of the paper is structured as follows. In Section 2 we provide the background necessary for the rest of the paper, describing the form of agents in the market. In Section 3 we highlight the limitations of standard package exchange mechanisms for data markets. Sec. 4 then describes the bid structure and winner determination problem in our proposed mechanism, while in Section 5 we detail value-based pricing schemes. In Section 6 details algorithms for winner determination and for determining pricing rules. In Section 7 we consider the effects of replication and undertake a numerical study of our results (Section 8). We discuss related work in Section 9, before concluding.

2 Descriptions of Agents

In our formulation organizations act as data consumers wishing to obtain data from DSPs. Let $N_O = \{1, \ldots, n_O\}$ be the set of organizations and $N_D = \{1, \ldots, n_D\}$ be the set of DSPs. Each organization and DSP is assumed to have quasi-linear preferences.

2.1 Data Service Providers

Each DSP owns, and can provide, data with variable quality. Examples of DSPs from different domains include:

(i) *Datacenters:* In order to transport data to consumers, communication is required. For extreme quantities of data and latency constraints, lossy data compression is required which diminishes the quality of data obtained by consumers. By employing additional resources for communication (e.g., transmission power or bandwidth), a higher data rate is achievable and hence transmission of data with less compression is feasible within the same period of time.

(ii) *IoT-Based Sensor Networks:* Time-series data or images collected from a sensor network may involve a variable number of sensors of different quality. Often, data quality is dependent on how many sensors of each type are employed for data collection. For example, sensors may observe the temperature or humidity in different regions of a city. To improve the quality of the temperature and humidity data, the DSP may exploit a larger number of sensors or increase sampling rates.

(iii) *Astronomical Observatories:* The quality of images collected by a telescope are dependent on their magnification setting and the duration of time the region of space is observed. For example, distances to nearby galaxies are often measured using standard candles, which are a class of stars that dim and brighten periodically due to their chemical composition. The absolute brightness of standard candles (i.e., the expected brightness at a constant distance) are known. As such, the distance to a nearby galaxy can be measured by observing the brightness of standard candles within the galaxy. If the galaxy is viewed for an insufficient amount of time, the periodicity of the standard candles will contain errors. On the other hand, if the standard candles are viewed at an insufficient magnification, the light from the candle will be susceptible to noise from nearby stars, again resulting in errors.

Each DSP j is capable of collecting data $\theta_j \in \Theta_j \subset \mathbb{R}^m$, which is dependent on the available infrastructure. Each element of θ_j corresponds to a data product in the set $M = \{1, \ldots, m\}$, which the DSP j can obtain from its infrastructure. For example, $\theta_{j\ell}$ may correspond to temperature measurements in a particular region l within a city (in the IoT case) or images l of a star (in the observatory example). If the DSP j cannot obtain the data $\theta_{j\ell}$ for some $\ell \in M$, then $\theta_{j\ell} = 0$.

2.2 Organizations

The goal of each organization i is to utilize data θ_i from providers to carry out a task. Each element of $\theta_i \in \mathbb{R}^{|M_i|}$ corresponds to a data product in the set $M_i \subseteq \{1, \ldots, m\}$, which organization i desires. For example, an organization i may wish to make a data-dependent decision, based on data θ_i from a datacenter or collected from an IoT-based sensor network. In this case, the decision problem can often be formalized as the optimization problem

$$x_i^*(\theta_i) \in \arg\min_{x_i \in \mathcal{X}_i} C_i(x;\ \theta_i), \tag{1}$$

where $\theta_i \in \Theta$ is the desired data, $x_i^*(\theta_i)$ is the optimal decision variable, $\mathcal{X}_i$ is a constraint set, and $C_i : \mathcal{X}_i \times \Theta \to \mathbb{R}_+$ is a cost function. The cost function may correspond to a control cost or an empirical risk in machine learning systems.

Alternatively, the organization may wish to carry out a regression or parameter estimation task by computing

$$x_i^*(\theta_i) = \psi_i(\theta_i), \tag{2}$$

where $x_i^*(\theta_i)$ is the parameter estimate and $\psi_i : \Theta_i \to \mathcal{X}_i$ is an estimator. For example, θ_i may correspond to a sequence of images of a galaxy containing a standard candle obtained from a telescope, and $x_i^*(\theta_i)$ an estimate of the distance to the galaxy.

An organization is willing to accept data that satisfies a given quality constraint. In the case of a decision problem, an organization may be willing

to accept a lower quality estimate, $\hat{\theta}_i$, of θ_i, which satisfies

$$C_i(x_i^*(\hat{\theta}_i);\ \theta_i) - C_i(x_i^*(\theta_i);\ \theta_i) \leq \bar{\epsilon}_i, \tag{3}$$

for a given quality constraint $\bar{\epsilon}_i > 0$. In other words, the cost of the decision associated with the solution $x_i^*(\hat{\theta}_i)$ should not significantly exceed the cost associated with the decision made using the optimal solution $x_i^*(\theta_i)$.

In the case of parameter estimation, organization i requires that the estimation error should not exceed

$$\|\psi_i(\hat{\theta}_i) - \psi_i(\theta_i)\| \leq \bar{\epsilon}_i, \tag{4}$$

for a given quality constraint $\bar{\epsilon}_i > 0$. For example, to obtain a reliable estimate of a galaxy distance, $\bar{\epsilon}_i$ may be selected to ensure that the uncertainty does not exceed standards required for publication.

Example 1 **Working Example:** An example of organizations and DSPs arises in condition-based maintenance in smart buildings [5]. In this scenario, a DSP consists of a network of temperature sensors, which are connected via wireless communication links to an access point. The DSP can adapt how many sensors communicate their data to the access point and also the level of compression and therefore accuracy at which information is transmitted (e.g., rounding to the nearest decimal place).

An organization then seeks to construct a classifier to detect anomalous temperatures (e.g., due to heating or ventilation faults). In this case, the quality criterion corresponds to the average loss of the classifier.

2.3 Quality Constraint Proxies

To trade data, it is necessary that DSPs guarantee that the data they provide satisfies the data quality constraints of one or more organizations. As a consequence, the DSPs should have access to the cost functions or estimators of the organizations. In practice, this may be undesirable due to the complexity of computing decision variables or parameter estimates. Alternatively, privacy concerns may prevent organizations from providing DSPs details about their decision or estimation problems.

A solution to this issue is for organizations to provide a proxy quality constraint to DSPs. A proxy for organization i satisfies

$$\tilde{C}_i(\hat{\theta}_i, \theta_i) \leq \bar{\epsilon}_i, \tag{5}$$

where for all $\hat{\theta}_i, \theta_i \in \Theta_i$,

$$\begin{aligned} C_i(x_i^*(\hat{\theta}_i);\ \theta_i) - C_i(x_i^*(\theta_i), \theta_i) &\leq \tilde{C}_i(\hat{\theta}_i, \theta_i), \quad \text{(Decision Problem)} \\ \|\psi_i(\hat{\theta}_i) - \psi_i(\theta_i)\| &\leq \tilde{C}_i(\hat{\theta}_i, \theta_i), \quad \text{(Estimation Problem).} \end{aligned} \tag{6}$$

A typical proxy that we will consider in the remainder of the paper has the form

$$\tilde{C}_i(\hat{\theta}_i, \theta_i) = \sum_{\ell=1}^{m} L_{i\ell} \|\hat{\theta}_{i\ell} - \theta_{i\ell}\|, \tag{7}$$

where $L_{i\ell} \geq 0,\ \ell = 1, \ldots, m$. In an estimation problem, this proxy naturally arises when $\psi_i(\cdot)$ is Lipschitz continuous. In the case of a decision problem, the proxy arises when $C_i(x;\ \theta_i)$ is A-Lipschitz in θ_i for all $x \in \mathcal{X}_i$ and $C_i(x_i^*(\theta_i);\ \theta_i)$ is B-Lipschitz in θ_i. In this case, we have

$$\begin{aligned}
&C_i(x_i^*(\hat{\theta}_i);\ \theta_i) - C_i(x_i^*(\theta_i);\ \theta_i) \\
&= \|C_i(x_i^*(\hat{\theta}_i);\ \theta_i) - C_i(x_i^*(\hat{\theta}_i);\ \hat{\theta}_i) + C_i(x_i^*(\hat{\theta}_i);\ \hat{\theta}_i) - C_i(x_i^*(\theta_i);\ \theta_i)\| \\
&\leq \|C_i(x_i^*(\hat{\theta}_i);\ \theta_i) - C_i(x_i^*(\hat{\theta}_i);\ \hat{\theta}_i)\| + \|C_i(x_i^*(\hat{\theta}_i);\ \hat{\theta}_i) - C_i(x_i^*(\theta_i);\ \theta_i)\| \\
&\leq A\|\theta_i - \hat{\theta}_i\| + B\|\theta_i - \hat{\theta}_i\| \\
&\leq \sum_{\ell=1}^{m} (A+B)\|\theta_{i\ell} - \hat{\theta}_{i\ell}\|.
\end{aligned} \tag{8}$$

We note the optimal value function $C_i(x_i^*(\theta_i);\ \theta_i)$ is Lipschitz continuous for many cost minimization problems. Explicit conditions for Lipschitz continuity of optimal value functions can be found, for example, in [6].

2.4 Platform

The main problem that we consider in the remainder of this paper is how to efficiently match organizations with DSPs such that quality constraints are satisfied, and how to perform pricing in this scenario. We focus on direct revelation mechanisms (where sealed bids are submitted directly to the platform), and where only bids arriving within the last T minute period can be matched. Moreover, unmatched bids are withdrawn; that is, the market is not continuous.

In the following sections, we introduce market mechanisms for this platform. In the next section, we first recall the standard package exchange mechanism in [4] and integrate data quality constraints by introducing additional data products. Our proposed mechanism is introduced in Section 4 (bidding and winner determination) and Section 5 (pricing).

3 A Standard Package Exchange Mechanism

A standard package exchange consists of a set of n traders, denoted by $N = \{1, \ldots, n\} = N_O \cup N_D$, which trade t commodities. Each agent may submit a number of bids, where B_i is the set of bids of agent i. A bid k from agent i is given by the tuple $(b_{ik}, \boldsymbol{q}_{ik})$, where:

(i) $\mathbf{q}_{ik} \in \{-1, 0, 1\}^t$ is a vector encoding the quantity of each product required. We have that $q_{ik\ell} \in \{-1, 0\}$ if $i \in N_O$, $k \in \{1, \ldots, |B_i|\}$ and $q_{ik\ell} \in \{0, 1\}$ if $i \in N_D$, $k \in \{1, \ldots, |B_i|\}$.

(ii) $b_{ik} \in \mathbb{R}$ is the amount bid ($b_{ik} \geq 0$ if $i \in N_O,\ k \in \{1, \ldots, |B_i|\}$ and $b_{ik} \leq 0$ if $i \in N_D,\ k \in \{1, \ldots, |B_i|\}$).

To allow for data products of varying qualities, the t commodities correspond to m data products each at r different quality levels; i.e., $t = mr$. For a given bid $k \in B_i$, each data product can only be offered at a single quality level. In other words,

$$\sum_{\ell \in \mathcal{S}_p} |q_{ik\ell}| \in \{0, 1\}, \tag{9}$$

where $\mathcal{S}_p$ is the set of commodities corresponding to the data product $p \in \{1, \ldots, m\}$ at different quality levels.

Example 2 **Working Example (Cont.):** We return to our scenario from Example 1 consisting of DSPs which obtain data from a network of temperature sensors. In this case, the discrete quality levels may correspond to the number of sensors that are utilized to collect measurements, or the compression level.

The impact of discretization of the quality levels can be clearly seen when each quality level corresponds to a different subset of active sensors. In this case, the removal of data from a sensor may lead to a significant performance reduction. In particular, anomaly detection accuracy will in general be lower.

Let $w_{ik} \in \{0, 1\}$ indicate whether agent i's k-th bid is winning. The winner determination problem can then be formulated as

$$\tilde{w}^* \in \arg \max_{w_{ik},\ \substack{i \in \{1,\ldots,n\},\\ k \in B_i}} \sum_{i \in \{1,\ldots,n\}} \sum_{k \in B_i} b_{ik} w_{ik} \tag{10}$$

subject to the constraints

$$\begin{aligned} &\sum_{i \in \{1,\ldots,n\}} \sum_{k \in B_i} q_{ik\ell} w_{ik} \geq 0,\ \ell \in \{1, \ldots, t\} \\ &\sum_{k \in B_i} w_{ik} \in \{0, 1\},\ i \in \{1, \ldots, n\}, \\ &w_{ik} \in \{0, 1\},\ k \in B_i,\ i \in \{1, \ldots, n\}, \end{aligned} \tag{11}$$

where the first constraint guarantees that the quantity of data products at each quality level does not exceed the quantity sold. The second constraint ensures that a buyer or seller does not have more than one bid that is winning. Note that each bid may only include one quality level for each data product due to the constraint in (9).

Irrespective of the pricing rule, there are several drawbacks to the standard package exchange mechanism in the data market context. First, the number of bids submitted by each organization and DSP is dependent on the number of quality levels, r. For a large number of quality levels, the number of bids will be very large. As such, the number of combinations that must be explored in the winner determination grows exponentially[1] in the number of DSPs as $(1 + N_O r)^{N_D}$.

One way to overcome this growth in complexity is to coarsely discretize the quality levels, which could cause a buyer to purchase data at a much higher level of quality (and thus price) than they require. Furthermore, in the case that the number of quality levels is low, inefficiencies are introduced into the market. In the standard package exchange mechanism, the data quality bid by agents must be fixed. In particular, each agent must select a data product to be bought or sold. In the absence of any prior information about supply or demand, a reasonable assumption is that organizations and DSPs select data qualities for their bids uniformly an interval $[l, u]$, where l and u correspond to the minimum and maximum feasible data qualities, respectively. When the actual demand from buyers is concentrated in an interval $[l_d, u_d] \subset [l, u]$, buyers may only be able to obtain data at much higher qualities (and hence higher cost) that they desire. For example, if an organization i requires data products at a quality ϵ_i and DSPs only offer data products at a quality $\epsilon < \epsilon_i$. If $\epsilon \ll \epsilon_i$, the payment of organization i will be significantly higher than it would be if DSPs could offer the data products at the quality ϵ_i.

Example 3 **Working Example (Cont.):** Returning to our working example of a network of temperature sensors, these issues can be clearly illustrated. Suppose that the different qualities of data products offered by a DSP arise from data collected from different subsets of the temperature sensors. In this case, the number of different quality levels, r, is given by

$$r = \sum_{i=1}^{N_S} \binom{N_S}{i}, \tag{12}$$

where N_S is the number of sensors. Observe that, in addition to the impact of the number of DSPs, the number of different quality data products in this scenario may be very large.

If an organization might be satisfied with low accuracy temperature data from *all* sensors (e.g., with a limited number of decimal places), but DSPs can only offer data gathered from a subset of sensors. While the DSP can offer such low accuracy data, it nevertheless does not meet the organization's requirements, and the DSP and organization would not be able to trade, resulting in lower efficiency in the market.

These drawbacks of the standard package exchange mechanism motivate the design of a new market mechanism which allows for more flexible choices

[1] To see this, note that each DSP can only be matched with at most one organization. As such, the matchings can be represented in the form $((o_1, q_1), \ldots, (o_{N_D}, q_{N_D}))$, where o_i denotes the organization matched with DSP i ($o_i = 0$ if not matched) and q_i denotes the quality level. As such, there are $(1 + N_O r)$ ways of matching each DSP, and the claim therefore follows.

of the data qualities associated with each data product. In the next section, we propose a new market mechanism, which addresses both of these issues.

4 Proposed Market Mechanism

A key challenge for the market platform is to match organizations and DSPs such that the quality constraints in (5) are satisfied, as well as assigning prices to each of the agents. In our setup, we assume that each DSP can be matched with at most one organization, while each organization can be matched with multiple DSPs[2]. That is, organizational access to a DSP is exclusive. This is often justified in the real world where organizations want exclusive access to data for competitive advantage. However, we relax this assumption in Sec. 7.1 by allowing "shadow DSPs", which correspond to replicas of the DSPs.

There are three components of the mechanism, which we detail in the remainder of this section: bid structure; the winner determination problem; and the pricing rule.

4.1 Bids

Each agent may submit a number of bids, where B_i is the set of bids of agent i. We follow the convention in [4], where products or money that are given are associated with a positive quantity and products or money that are taken are associated with a negative quantity[3]. This allows us to treat buyers and sellers in the same manner.

A bid k from agent i (either an organization or DSP) is given by the tuple $(\mathbf{L}_{ik}, \bar{\epsilon}_{ik}, \mathbf{q}_{ik}, b_{ik}(\cdot))$, where:

(i) $\mathbf{L}_{ik} = (L_{ik1}, \ldots, L_{ikm})$ are the proxy parameters in (7) for each data product associated to bid k of agent i. The values in $\mathbf{L}_{ik}$ are zero if the agent is a DSP (i.e, $i \in N_D$, for any $k \in B_i$).

(ii) $\bar{\epsilon}_{ik}$ is the quality constraint in (5) for the data product associated to bid k of agent i. If i is a DSP and k is their bid, i.e., $i \in N_D,\ k \in B_i$, then $\bar{\epsilon}_{ik} = \infty$.

(iii) $\mathbf{q}_{ik} \in \{-1, 0, 1\}^m$ is the quantity of each data product required. We have that $q_{ik\ell} \in \{-1, 0\}$ if $i \in N_O,\ k \in B_i$ and $q_{ik\ell} \in \{0, 1\}$ if $i \in N_D,\ k \in B_i$.

(iv) Let $\epsilon_{ik\ell} = \|\hat{\theta}_{ik\ell} - \theta_{i\ell}\|,\ \ell = 1, \ldots, m$. If $q_{ik\ell} = 0$, then $\epsilon_{ik\ell} = \infty,\ i \in N_D,\ k \in B_i$. If $\epsilon_{ik\ell} = 0$, then $i \in N_O,\ k \in B_i$. Let $\boldsymbol{\epsilon}_{ik} = (\epsilon_{ik1}, \ldots, \epsilon_{ikm})$. Then, $b_{ik} : \mathbb{R}^m_+ \to \mathbb{R}$ is the amount bid ($b_{ik}(\boldsymbol{\epsilon}_{ik}) \geq 0$ if $i \in N_O,\ k \in B_i$ and $b_{ik}(\boldsymbol{\epsilon}_{ik}) \leq 0$ if $i \in N_D,\ k \in B_i$), which depends on the quality of the individual data products.

Example 4 **Working Example (Cont.):** In our working example, consider a DSP i with access to a network of temperature sensors. In bid $k \in B_i$, DSP i offers data products in $\ell \in M_{ik}$. Each data product corresponds to a distnct sensor. In this case,

[2]This is also a common assumption in the package exchange literature [4], arguably because the focus there is on physical assets which cannot be distributed amongst multiple organizations.

[3]In other words, sellers (DSPs) make negative bids while buyers (organizations) make positive bids.

$\mathbf{q}_{ik\ell} = 1,\ \ell \in M_{ik}$ and $q_{ik\ell} = 0,\ \ell \notin M_{ik}$. The data $\theta_{i\ell}$ then corresponds to the temperature measurement collected by sensor ℓ. The compressed (lower accuracy) data sent by sensor ℓ to the access point of DSP i corresponds to $\hat{\theta}_{ik\ell}$.

Compared with low quality data, high quality data where $\epsilon_{ik\ell}$ is small, is both more difficult to produce by DSPs and not less valuable (than low quality data) to organizations. As a consequence, the following assumption on the valuation function $b_{ik}(\cdot)$ is natural.

Assumption 1 *$|b_{ik}(\boldsymbol{\epsilon}_{ik})|$ is a non-increasing function in $\epsilon_{ik\ell},\ l \in M$ for $i \in N_O \cup N_D,\ k \in B_i$. Moreover, if $q_{ik\ell} = 0$, then $b_{ik}(\boldsymbol{\epsilon}_{ik})$ does not depend on $\epsilon_{ik\ell}$.*

4.2 Winner Determination Problem

A common objective of the winner determination problem in a market mechanism is to match agents so as to maximize the value of the trades, known as the surplus. This is a means of ensuring that agents with high bids are prioritized. In standard package exchanges (see, e.g., [4]), the amount bid for a given item is fixed; that is, there is no notion of the quality of an item. In our case, due to the dependence of $b_{ik}(\boldsymbol{\epsilon}_{ik})$ on $\boldsymbol{\epsilon}_{ik}$, it is necessary to also select winners based on the quality of the data products that are bought or sold by an organization or DSP.

In the case that a bid from a DSP can be matched with at most one organization, a natural generalization of the standard objective in the winner determination problem is given by

$$\tilde{f}(\tilde{\mathbf{w}}, \boldsymbol{\epsilon}) = \sum_{i \in N_O \cup N_D} \sum_{k \in B_i} b_{ik}(\boldsymbol{\epsilon}_{ik}) w_{ik}, \tag{13}$$

where $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_{ik})_{i \in N_O \cup N_d,\ k \in B_i}$, $w_{ik} \in \{0, 1\}$ and $\tilde{\mathbf{w}} = (w_{ik})_{i \in N_O \cup N_d,\ k \in B_i}$. However, with this formulation it is difficult to enforce the constraint that bid k of an organization i can only be matched with bid k' of a DSP i' if

$$\epsilon_{i'k'\ell} \leq \epsilon_{ik\ell},\ \ell \in M. \tag{14}$$

To resolve this issue, we instead write the objective of the winner determination problem as

$$\begin{aligned} f(\mathbf{w}, \boldsymbol{\epsilon}) = & \sum_{i \in N_O} \sum_{i' \in N_D} \sum_{k \in B_i} \sum_{k' \in B_{i'}} b_{i'k'}(\boldsymbol{\epsilon}_{i'k'}) w_{iki'k'} \\ & + \sum_{i \in N_O} \sum_{k \in B_i} \sum_{i' \in N_D} \sum_{k' \in B_{i'}} \frac{b_{ik}(\boldsymbol{\epsilon}_{ik}) w_{iki'k'}}{\sum_{i' \in N_D} \sum_{k' \in B_{i'}} w_{iki'k'}}, \end{aligned} \tag{15}$$

where $\mathbf{w} = (w_{iki'k'})$, $w_{iki'k'} \in \{0, 1\}$ and $w_{iki'k'} = 1$ if and only if bid k of organization i is matched to bid k' of DSP i'. The quotient in the second

summation arises due to the fact that an organization may be matched with more than one DSP. Under the constraint that only a single bid from an agent is winning, w_{ik} in (13) can be computed from $w_{iki'k'}$ via

$$\begin{aligned} w_{ik} &= \sum_{i' \in N_D} \sum_{k' \in B_{i'}} \frac{w_{iki'k'}}{\sum_{i' \in N_D} \sum_{k' \in B_{i'}} w_{iki'k'}}, \ i \in N_O, \ k \in B_i, \\ w_{i'k'} &= \sum_{i \in N_O} \sum_{k \in B_i} w_{iki'k'}, \ i' \in N_D, \ k' \in B_{i'}. \end{aligned} \tag{16}$$

In this case, to ensure that $i \in N_O, \ k \in B_i, \ i' \in N_D, \ k' \in B_{i'}$ can only be matched if (14) holds, we introduce the constraint

$$\begin{aligned} &w_{iki'k'}(\epsilon_{ik\ell} - \epsilon_{i'k'\ell}) \geq 0, \\ &\qquad i, i' \in N_O \cup N_D, \ k \in B_i, \ k' \in B_{i'}, \ell \in M. \end{aligned} \tag{17}$$

To ensure that at most one bid from each agent is winning, we also introduce the constraints

$$\begin{aligned} \sum_{k \in B_i} \mathbf{1} \left\{ \sum_{i' \in N_D} \sum_{k' \in B_{i'}} w_{iki'k'} > 0 \right\} \leq 1, \ i \in N_O, \\ \sum_{k' \in B_{i'}} \sum_{i \in N_O} \sum_{k \in B_i} w_{iki'k'} \leq 1, \ i' \in N_D. \end{aligned} \tag{18}$$

In other words, there is at most one bid k' from each DSP $i' \neq i$ matched to organization i, and at most one bid k from any organization i matched to DSP i'.

We also require that any match with a bid of an organization i satisfies the quality constraint $\bar{\epsilon}_{ik}$ for all data products $\ell \in M$. This constraint is enforced via

$$\begin{aligned} &w_{iki'k'} \sum_{\ell=1}^{m} |q_{ik\ell}| L_{ik\ell} \epsilon_{ik\ell} \leq \bar{\epsilon}_{ik}, \\ &\qquad i, i' \in N_O \cup N_D, \ k \in B_i, \ k' \in B_{i'}, \ell \in M. \end{aligned} \tag{19}$$

In summary, the winner determination problem is given by

$$\begin{aligned} \mathbf{w}^* \in \arg\max f(\mathbf{w}, \boldsymbol{\epsilon}) := &\sum_{i \in N_O} \sum_{i' \in N_D} \sum_{k \in B_i} \sum_{k' \in B_{i'}} b_{i'k'}(\boldsymbol{\epsilon}_{i'k'}) w_{iki'k'} \\ &+ \sum_{i \in N_O} \sum_{k \in B_i} \sum_{i' \in N_D} \sum_{k' \in B_{i'}} \frac{b_{ik}(\boldsymbol{\epsilon}_{ik}) w_{iki'k'}}{\sum_{i' \in N_D} \sum_{k' \in B_{i'}} w_{iki'k'}}, \end{aligned} \tag{20}$$

subject to the following constraints:

$$
\begin{aligned}
& w_{iki'k'} \in \{0,1\},\ i,i' \in N_O \cup N_D,\ k \in B_i,\ k' \in B_{i'}, \\
& w_{iki'k'} = 0,\ i \in N_O,\ i' \in N_O, k \in B_i,\ k' \in B_{i'}, \\
& w_{iki'k'} = 0,\ i \in N_D,\ i' \in N_D, \\
& \sum_{i \in N_O} \sum_{k \in B_i} w_{iki'k'} \leq 1,\ i' \in N_D,\ k' \in B_{i'}, \\
& \sum_{k \in B_i} \mathbf{1}\left\{ \sum_{i' \in N_D} \sum_{k' \in B_{i'}} w_{iki'k'} > 0 \right\} \leq 1,\ i \in N_O, \\
& \sum_{k' \in B_{i'}} \sum_{i \in N_O} \sum_{k \in B_i} w_{iki'k'} \leq 1,\ i' \in N_D, \\
& w_{iki'k'} \sum_{\ell=1}^{m} |q_{ik\ell}| L_{ik\ell} \epsilon_{ik\ell} \leq \bar{\epsilon}_{ik}, \\
& \qquad i,i' \in N_O \cup N_D,\ k \in B_i,\ k' \in B_{i'}, \ell \in M, \\
& w_{iki'k'}(\epsilon_{ik\ell} - \epsilon_{i'k'\ell}) \geq 0, \\
& \qquad i,i' \in N_O \cup N_D,\ k \in B_i,\ k' \in B_{i'}, \ell \in M, \\
& \epsilon_{ik\ell} \in [0,\infty),\ q_{ikl} \neq 0,\ i \in N_O \cup N_D,\ k \in B_i, \ell \in M, \\
& \epsilon_{ik\ell} = 0,\ q_{ik\ell} = 0,\ i \in N_O,\ k \in B_i,\ \ell \in M, \\
& \epsilon_{ik\ell} = \infty,\ q_{ik\ell} = 0,\ i \in N_D,\ k \in B_i,\ \ell \in M.
\end{aligned} \tag{21}
$$

The surplus in (15) involves $\sum_{i \in N_O} \sum_{i' \in N_D} |B_i||B_{i'}|$ binary variables $w_{iki'k'}$ and $m \sum_{i \in N_O \cup N_D} |B_i|$ continuous variables. As a consequence, it is necessary to develop an efficient means of solving the winner determination problem described by (20). We address this problem in Sec. 6.1 via a reformulation as a bi-level optimization problem.

5 Pricing

Agents trade the quantities and qualities of data products specified in the winning bids. That is, recalling (16), bid k of agent i of quantity $\mathbf{q}_{ik}$ is traded if $w_{ik}^{S*} = 1$, where $\mathbf{w}^{S*}$ is the optimal matching in (20) for agents in $S = N_O \cup N_D$. We now describe how winning bids are priced.

Define the surplus arising from the winner determination problem by

$$
v(S) = \sum_{i \in S} \sum_{k \in B_i} b_{ik}(\boldsymbol{\epsilon}_{ik}^{S*}) w_{ik}^{S*}. \tag{22}
$$

The amount an agent pays depends on its bid and on the rule used to divide the surplus. Suppose that the quantity of the surplus allocated to agent i is the value $\psi_i(S, v(S))$. In this case, agents pay the amount they bid for any

winning bids minus their share of the surplus. More precisely, the payment by agent i is given by

$$\pi_i(S, v(S)) = \sum_{k \in B_i} w_{ik}^{S*} b_{ik}(\boldsymbol{\epsilon}_{ik}^{S*}) - \psi_i(S, v(S)). \tag{23}$$

Unlike competitive prices and core payments, payments based on values — as per (23) — always exist and are unique. As observed in [4], a key advantage of payments based on values is therefore that they can be used in settings where competitive prices and core payments cannot.

5.1 Standard Pricing Rules

A number of choices for the values $\psi_i,\ i \in S = N_O \cup N_D$ have been proposed for auctions and package exchanges. A standard choice for the values is based on the Shapley value [4], which is defined by

$$\psi_i^{\text{Shapley}}(S, v) = \sum_{T \subseteq S \setminus \{i\}} \frac{s!(s-t-1)!}{s!} \left(v(T \cup \{i\}) - v(T)\right), \tag{24}$$

where $s = |S|$, $t = |T|$, and $v(T)$ is the surplus with agents in $T \subseteq S$ participating; i.e.,

$$v(T) = \sum_{i \in T} \sum_{k \in B_i} b_{ik}(\boldsymbol{\epsilon}_{ik}^{T*}) w_{ik}^{T*}. \tag{25}$$

Another common choice for the values is the Vickrey value, given by

$$\psi_i^{\text{Vickrey}}(S, v) = v(S) - v(S \setminus \{i\}). \tag{26}$$

In [4], Lindsay proposed the balanced winner contribution (BWC) value. Let N_W denote the agents $i \in S$ with a winning bid (i.e., $\exists k \in B_i$ such that $w_{ik}^{S*} = 1$), and $N_L = S \setminus N_W$. The BWC value is then defined by

$$\psi_i^{\text{BWC}}(S, v) = \sum_{T \subseteq S \setminus \{i\}} \frac{t!(s-t-1)!}{s!} \left(v(N_L \cup T \cup \{i\}) - v(N_L \cup T)\right). \tag{27}$$

The following properties of the Shapley, Vickrey and BWC pricing rules are known [4], and relevant for picking an appropriate pricing rule. For definitions of budget balance, individual rationality and incentive compatibility see, for example, [7].

Proposition 1 *The BWC, Shapley and VCG pricing rules always exist.*

Proposition 2 *The BWC and Shapley rules are budget balanced. The Vickey rule is not budget balanced.*

Proposition 3 *The BWC, Shapley and VCG rules are ex post individually rational.*

Proposition 4 *The BWC and Vickrey rules satisfies no payments for losing bidders. This does not hold for the Shapley rule.*

Proposition 5 *The VCG rule is incentive compatible. This does not hold for the Shapley and BWC rules.*

Proposition 6 *The BWC rule satisfies the balanced winner contribution property; namely,*

$$\begin{aligned}&\psi_i^{\mathrm{BWC}}(N_W, v) - \psi_i^{\mathrm{BWC}}(N_W \setminus \{j\}, v)\\ &= \psi_j^{\mathrm{BWC}}(N_W, v) - \psi_j^{\mathrm{BWC}}(N_W \setminus \{i\}, v),\end{aligned} \tag{28}$$

for all $i \in N_W$ *and* $j \in N_W$.

5.2 Modified Balanced Winner Contribution Pricing

The BWC value, introduced in [4], can be viewed as a variation on the Shapley value, where only winning bids have a non-zero value. A natural modification of the BWC value can be obtained by replacing the set of losing bids N_L with a subset $N_E \subset N_L$. The effect of this modification is to allow *some* losing bids to be assigned a non-zero value, leading to the modified BWC (mBWC) rule. A useful application of this observation is in the case of thin markets, which we study further in Sec. 7.2.

More precisely, let $N_E \subset N_L$ and $\overline{N}_E = N_L \setminus N_E$. We define the mBWC rule for N_E as

$$\begin{aligned}&\psi_i^{\mathrm{mBWC},N_E}(S, v)\\ &= \sum_{T \subset S \setminus \{i\}} \frac{t!(s-t-1)!}{s!}(\overline{N}_E \cup T \cup \{i\}) - v(\overline{N}_E \cup T)).\end{aligned} \tag{29}$$

Observe that if $N_E = N_L$, the mBWC rule corresponds to the Shapley rule. Similarly, by setting $N_E = \emptyset$, the mBWC rule corresponds to the BWC rule.

As far as we are aware, the mBWC rule has not been previously studied in the literature. The following properties hold and can be established via standard arguments.

Proposition 7 *The mBWC pricing rule always exists, is budget balanced, and is ex post individually rational.*

6 Algorithms

6.1 Winner Determination Problem

As in other exchanges, a key challenge for the mechanism detailed in Sec. 4 is to solve the winner determination problem (given in (20)), requiring optimization of the data quality $\boldsymbol{\epsilon}_{ik}$ for each winning bid, in addition to determining whether or not an agent's bid is winning.

To make the winner determination problem amenable to lower complexity algorithms, we establish the following proposition.

Proposition 8 *Let $i \in N_O,\ k \in B_i$ be a winning bid for data products in $M_i \subseteq M$. Suppose i is matched with bid $k_{M_j} \in B_{i_{M_j}}$ of $i_{M_j} \in N_D$ for data products $M_j \subseteq M$ for $j \in J$, where $M_j \cap M_{j'} = \emptyset,\ j \neq j',\ j, j' \in J$ and $\cup_{j=1}^{J} M_j = M_i$. Then, the set of data qualities $\boldsymbol{\epsilon}_{ik}, \boldsymbol{\epsilon}_{i_{M_j} k_{M_j}},\ j \in J$ maximizing the surplus contains a solution satisfying*

$$\epsilon^*_{ik\ell} = \epsilon^*_{i_{M_j} k_{M_j} \ell},\ \ell \in M_j, \tag{30}$$

where $\boldsymbol{\epsilon}^* \in S_{ik}$ with

$$S_{ik} = \arg \max_{\substack{\boldsymbol{\mu} \in \mathbb{R}^m_+: \\ \sum_{\ell \in M_i} L_{ik} \mu_{ik\ell} \leq \bar{\epsilon}_{ik}}} b_{ik}(\boldsymbol{\mu}) + \sum_{j=1}^{J} b_{i_{M_j} k_{M_j}}(\boldsymbol{\mu}). \tag{31}$$

Proof In order for $\boldsymbol{\epsilon}^*_{ik}, \boldsymbol{\epsilon}_{i_{M_j} k_{M_j}},\ j \in J$ to be feasible, they must satisfy

$$\epsilon_{i_{M_j} k_{M_j} \ell} \leq \epsilon_{ik\ell},\ \ell \in M_j. \tag{32}$$

Under Assumption 1, any increase in $\epsilon_{i_{M_j} k_{M_j} \ell},\ \ell \in M_j$ will not lead to a lower value of $b(\boldsymbol{\epsilon}_{i_{M_j} k_{M_j}})$ as $i_{M_j} \in N_D$. On the other hand, any decrease in $\epsilon_{ik\ell}$ will not lead to a lower value of $b(\boldsymbol{\epsilon}_{ik})$ as $i \in N_O$. As a consequence, the set of $\boldsymbol{\epsilon}_{ik}, \boldsymbol{\epsilon}_{i_{M_j} k_{M_j}}$ maximizing the surplus contains a solution satisfying $\boldsymbol{\epsilon}^*_{ik\ell} = \boldsymbol{\epsilon}^*_{i_{M_j} k_{M_j} \ell},\ \ell \in M_j$.

As only at most one bid of each agent is winning and each DSP can be matched to at most one organization, it follows that the contribution of bid k of organization i and bid k_{M_j} of DSP i_{M_j} for $j \in J$ to the total surplus is given by $b_{ik}(\boldsymbol{\epsilon}_{ik}) + \sum_{j=1}^{J} b_{i_{M_j} k_{M_j}}(\boldsymbol{\epsilon}_{i_{M_j} k_{M_j}})$. As such, under the assumption that bid k of organization i and bid k_{M_j} of DSP i_{M_j} for $j \in J$ are winning, the surplus is maximized by selecting data qualities via (31). □

A key observation from Prop. 8 is that there exists a solution to the winner determination problem in (20) where the data quality of a data $\ell \in M$ bought by an organization is the same as the data quality sold by the matched DSP. As a consequence of the proposition, the number of data quality variables $\epsilon_{ik\ell},\ i \in N_O \cup N_D,\ k \in B_i,\ \ell \in M$ to be optimized is reduced. The case of an organization $1 \in N_O$ buying a single data product is matched with DSP $2 \in N_D$ is illustrated in Fig. 1.

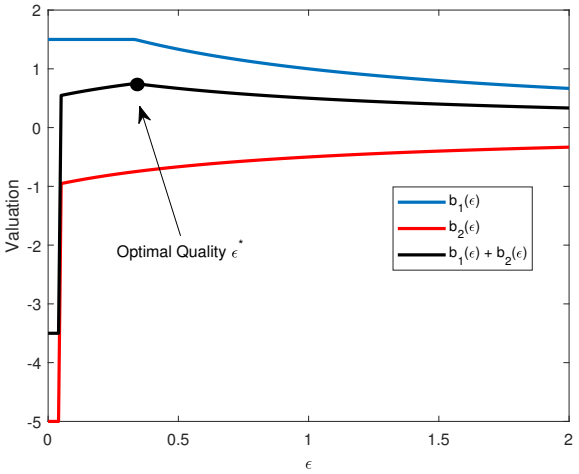

Fig. 1 Illustration of the maximization problem in (8). The blue curve is the function $b_1(\epsilon) = \min\{2/(1+\epsilon), 1.5\}$ for the organization and the red curve is the function $b_2(\epsilon) = -1/(1+\epsilon) \cdot \mathbf{1}\{\epsilon \geq 0.05\} - 5 \cdot \mathbf{1}\{\epsilon < 0.05\}$ for the DSP. The total surplus is shown in the black curve with the optimal quality ϵ^* illustrated by the dot.

A further implication of Prop. 8 is that the winner determination problem (given in (20)) can be reformulated as a bi-level optimization problem. Let $\boldsymbol{\epsilon}^*_{ik}(\mathbf{w}),\ i \in N_O \cup N_D,\ k \in B_i$ be the solution to (31). The optimal matching $\mathbf{w}^*$ is then obtained by solving

$$\begin{aligned}\mathbf{w}^* \in \arg\max_{\mathbf{w}} & \sum_{i\in N_O}\sum_{i'\in N_D}\sum_{k\in B_i}\sum_{k'\in B_{i'}} b_{i'k'}(\boldsymbol{\epsilon}^*_{i'k'}(\mathbf{w}))w_{iki'k'} \\ & + \sum_{i\in N_O}\sum_{k\in B_i}\sum_{i'\in N_D}\sum_{k'\in B_{i'}} \frac{b_{ik}(\boldsymbol{\epsilon}^*_{ik}(\mathbf{w}))w_{iki'k'}}{\sum_{i'\in N_D}\sum_{k'\in B_{i'}} w_{iki'k'}} \end{aligned} \tag{33}$$

subject to the following constraints:

$$\begin{aligned} & w_{iki'k'} \in \{0,1\},\ i,i' \in N_O \cup N_D,\ k \in B_i,\ k' \in B_{i'}, \\ & w_{iki'k'} = 0,\ i \in N_O,\ i' \in N_O, k \in B_i,\ k' \in B_{i'}, \\ & w_{iki'k'} = 0,\ i \in N_D,\ i' \in N_D, \\ & \sum_{k\in B_i} \mathbf{1}\left\{\sum_{i'\in N_D}\sum_{k'\in B_{i'}} w_{iki'k'} > 0\right\} \leq 1,\ i \in N_O, \\ & \sum_{k'\in B_{i'}}\sum_{i\in N_O}\sum_{k\in B_i} w_{iki'k'} \leq 1,\ i' \in N_D. \end{aligned} \tag{34}$$

There remains the issue of finding an efficient solution of the bi-level optimization problem in (33). In the absence of parametric bids, a solution can be

obtained via global search methods (e.g., simulated annealing or genetic algorithms). In particular, the global search algorithm selects a candidate matching $\mathbf{w}$ and evaluates the surplus via (31). A generic algorithm of this form is detailed in Alg. 1.

Algorithm 1 Winner Determination

Require: $i \in N_O \cup N_d,\ k \in B_i$

function WINNERDETERMINATION($\mathbf{L}_{ik}, \bar{\epsilon}_{ik}, b_{ik}(\cdot), \mathbf{q}_{ik}$)
  $c \leftarrow 0$, $s_{best} \leftarrow 0$, $\boldsymbol{\epsilon}_{\text{best}} \leftarrow \mathbf{0}$, $\mathbf{w}_{\text{best}} \leftarrow \mathbf{0}$
  while not converged **do**
    Sample $\mathbf{w}_c$ via the global search algorithm.
    for all $i \in N_O,\ i' \in N_D,\ k \in B_i,\ k' \in B_{i'}$ **do**
      if $w_{c,iki'k'} = 1$ **then**
        Obtain $\boldsymbol{\epsilon}_{c,ik}$ and $\boldsymbol{\epsilon}_{c,i'k'}$ by solving (31).
      else
        $\boldsymbol{\epsilon}_{c,ik} \leftarrow \boldsymbol{\epsilon}_{c,i'k'} \leftarrow 0$
      end if
    end for
    if $f(\boldsymbol{\epsilon}_c, \mathbf{w}_c) > s_{\text{best}}$ and the constraints in (34) are satisfied **then**
      $s_{\text{best}} \leftarrow f(\boldsymbol{\epsilon}_c, \mathbf{w}_c)$
      $(\boldsymbol{\epsilon}_{\text{best}}, \mathbf{w}_{\text{Best}}) \leftarrow (\boldsymbol{\epsilon}_c, \mathbf{w}_c)$
    end if
    $c \leftarrow c + 1$
  end while
  return $(\boldsymbol{\epsilon}_{\text{best}}, \mathbf{w}_{\text{best}})$
end function

6.2 mBWC Pricing Rule Approximation

Evaluation of the mBWC pricing rule of (29) requires the enumeration of a large number of subsets of $N_O \cup N_i$ and the corresponding bids. This problem also arises in the computation of the Shapley rule for which sampling-based approximation techniques have been developed [8]. A similar approach has been applied to evaluation of the BWC pricing rule [4]. In this section, we adapt these approximation techniques to the mBWC pricing rule.

Let N_L be the set of losing bids, N_W be the set of winning bids, and N_E the set of preferred bids detailed in the definition of the mBWC pricing rule in (29). To evaluate the mBWC surplus allocation for bid $k \in B_i$ of agent $i \in N_O \cup N_D$, for a set $T \subseteq S$, let $\pi(T)$ be the set of $|T|!$ permutations. Let $O_{N_E \cup N_W}$ be an arbitrary ordering of the elements in $N_W \cup N_E$ and $\text{Pre}^i(O_{N_E \cup N_W}) = \{O_{N_E \cup N_W}(1), \ldots, O_{N_E \cup N_W}(k-1)\}$ when $i = O_{N_E \cup N_W}(k)$. The mBWC price approximation is then detailed in Alg. 2.

It is known that a similar sampling based approximation of the Shapley value is unbiased [8]. For Alg. 2, the following result holds, which can be established using the same argument as in [8].

Proposition 9 *Let N_E be the set of preferred bids. If the winner determination problem is solved optimally, the estimator $\hat{\psi}_i$ in Alg. 2 is an unbiased estimator of $\psi_i^{\mathrm{mBWC},N_E}(S,v)$.*

Algorithm 2 mBWC Price Approximation

Require: $i \in N_O \cup N_D$, $k \in B_i$, $c_{\max}$, N_L, s_{best}, N_E

function mBWC($\mathbf{L}_{ik}, \bar{\epsilon}_{ik}, b_{ik}(\cdot), \mathbf{q}_{ik}$)
  $\hat{\psi}_i \leftarrow 0,\ i \in N_D \cup N_O$
  for all $c = 1, \ldots, c_{\max}$ **do**
    Uniformly sample $\tilde{O} \in \pi(N_W \cup N_E)$ with probability $\frac{1}{(|N_W|+|N_E|)!}$
    Set O_L to be an arbitrary ordering of $N_L \setminus N_E$
    $O \leftarrow (O_L, \tilde{O})$
    Calculate $\mathrm{Pre}^i(O)$
    for all agents $i \in N_O \cup N_d$ **do**
      if $v(\mathrm{Pre}^i(O)) > s_{\text{best}}$ **then**
        $v(\mathrm{Pre}^i(O)) \leftarrow s_{\text{best}}$
      end if
      if $v(\mathrm{Pre}^i(O) \cup \{i\}) > s_{\text{best}}$ **then**
        $v(\mathrm{Pre}^i(O) \cup \{i\}) \leftarrow s_{\text{best}}$
      end if
      $x(O)_i \leftarrow v(\mathrm{Pre}^i(O) \cup \{i\}) - v(\mathrm{Pre}^i(O))$
      if $x(O)_i < 0$ **then**
        $x(O)_i = 0$
      end if
      $\hat{\psi}_i \leftarrow \hat{\psi}_i + x(O)_i$
      $\hat{\psi}_i \leftarrow \frac{\hat{\psi}_i}{c_{\max}}$
    end for
  end for
  for all agents $i \in N_O \cup N_d$ **do**
    $\pi_i \leftarrow s_{\text{best}} - \hat{\psi}_i$
  end for
  return $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$
end function

We note that Alg. 1 may not yield an optimal solution for the winner determination problem. As such, individual rationality and budget balance for the Vickrey, BWC, Shapley, and mBWC rules are not guaranteed. This issue also arises in the evaluation of the BWC rule and has been addressed in [4]. To

ensure individual rationality, Lines 16-17 guarantee that the allocated surplus is non-negative. On the other hand, budget balance is promoted via Lines 9-14.

7 Impact of Replication

A distinct feature of data markets as opposed to markets for other kinds of goods is that data products can be replicated at low cost. Recent work has explored the impact of replication in the context of submodular auction coalitional games [9, 10] and for data markets for the purpose of machine learning tasks [11] with a single buyer. In the remainder of this section, we study the impact of replication in the context of the market mechanism with data quality constraints developed in Sec. 4.

7.1 Exclusivity

For organizations in competition, access to data is a valuable means of obtaining a competitive advantage. For this reason, it can be desirable for organizations to request that the DSPs only sell their data products to at most one organization. As a consequence, DSPs should not replicate their bids or form *shadow identities*, where a DSP $i \in N_D$ sells its data under another identity $j \in N_D$.

If all shadow DSPs truthfully reveal the DSP they are associated with, then the market can easily ensure that data products are sold to only a single organization. On the other hand, if shadow DSPs are not truthful about which DSP they are associated with, it is desirable that the pricing rule is designed to ensure that this behavior is disincentivized.

For our purposes, the notion of data exclusivity is defined as follows.

Definition 1 A DSP $i \in N_D$ is said to honor a strict exclusivity agreement if it only makes a single bid to sell a specific combination of data products. That is, it does not make any other bids (either in B_i or via a shadow identity $j \in N_D$) to sell the same combination of data products.

The following proposition shows that DSPs have an incentive not to have replicated data (via shadow identities) under the Vickrey rule.

Proposition 10 *Suppose that bid k of DSP i is winning with zero replications; i.e., there are no further shadow identities. Further, suppose that in the case DSP i has two shadow identities (i.e., one replication), denoted by i_1, i_2, and only one of the two identities is winning. Under the Vickrey rule, the surplus allocated to DSP i is maximized with zero replications.*

Proof In the case of two shadow identities, the surplus of DSP i is given by $\sum_{j=1}^{2}(v(S) - v(S \setminus \{i_j\}))$. As only one of the two identities, say i_1, is winning, when

identity i_1 is excluded from the set of bidders, it will be replaced by identity i_2, yielding an allocated surplus for i_1 of zero. On the other hand, identity i_2 is not winning, and hence its allocated surplus is also zero. As a consequence, the total surplus allocated to DSP i is zero. □

Proposition 10 shows that the payoff for DSPs is maximized with n replications only when all of the corresponding shadow identities are winning. Indeed, whenever at least one shadow identity is losing, the surplus of at least one of the winning identities will be zero. This provides evidence that the Vickrey rule, despite the lack of budget balance, may be desirable in scenarios where strict exclusivity agreements are in place.

A similar disincentive for replication is also present in the case of the BWC pricing rule.

Proposition 11 *Suppose that i' is a shadow identity of i and that $i \in N_W$. Further, suppose that $i \in N_W$, $i' \in N_L$, where i is matched with $j \in N_W$. Moreover, j is the only feasible match for i. Then, under the BWC pricing rule, agent i has no incentive to replicate its data to form the shadow identity i'.*

Proof If $i' \in T$ but $i, j \notin T$, then

$$\begin{aligned} &v(N_L \cup T \cup \{i\}) = v(N_L \cup T) \\ &= v(N_L \cup T \cup \{i\} \setminus \{i'\}) = v(N_L \cup T \setminus \{i'\}). \end{aligned} \tag{35}$$

On the other hand, if $i', j \in T$ but $i \notin T$, then

$$\begin{aligned} &v(N_L \cup T \cup \{i\}) = v(N_L \cup T) \\ &= v(N_L \cup T \cup \{i\} \setminus \{i'\}) \geq v(N_L \cup T \setminus \{i'\}). \end{aligned} \tag{36}$$

Finally, if $j \in T$, but $i, i' \notin T$, then

$$\begin{aligned} &v(N_L \cup T \cup \{i\}) = v(N_L \cup T) \\ &= v(N_L \cup T \cup \{i\} \setminus \{i'\}) \geq v(N_L \cup T \setminus \{i'\}). \end{aligned} \tag{37}$$

It then follows that, under the BWC pricing rule, the surplus allocated to identity i' is zero and the surplus allocated to i is less than the surplus allocated to i when the shadow identity i' is not in S. □

7.2 Thin Markets

In some scenarios, it may be desirable for DSPs to replicate their data and corresponding bids. For example, when multiple organizations may efficiently utilize the data of a single DSP and no exclusivity agreements are in place. This scenario arises in thin markets, where there are an insufficient number of DSPs to serve all organizations.

The conclusions from Sec. 7.1 suggest that for markets where there are more DSP bids than organizations the Vickrey and BWC rules penalize replication. As a consequence, when there is uncertainty about the thinness of the market,

agents have incentives to avoid replication. Note that as bids are sealed, such uncertainty is likely to be present.

It can therefore be desirable to modify the pricing rules to signal the thinness of the market and incentivize DSPs to replicate their data. One such modification is to indicate to DSP bidders with a data product m that bids on this data product will not be included in N_L if they are losing. More precisely, let N_E be the set of replicated DSP bids that the market mechanism believes should be present to reduce the thinness of the market.

To incentivize replication of DSP bids in N_E, the market mechanism can utilize the mBWC pricing rule with preferred bids N_E. As a consequence, DSPs with bids in N_E will obtain a non-zero allocated surplus even if their bids in N_E are losing. The effect of the mBWC pricing rule with replication is illustrated in Sec. 8.3.

8 Numerical Examples

8.1 Comparison with the Standard Package Exchange

Consider two organizations $N_O = \{1, 2\}$ and two DSPs $N_D = \{3, 4\}$ offering a single common data product. In the context of our working example (see Example 1), each organization may be responsible for building management and each DSP is a provider with sensors in or nearby the buildings. Suppose the bids have parameters:

$$
\begin{aligned}
&L_1 = 0.01,\ L_2 = 0.01,\ L_3 = 0,\ L_4 = 0\\
&\bar{\epsilon}_1 = 1,\ \bar{\epsilon}_2 = 1,\ \bar{\epsilon}_3 = \infty,\ \bar{\epsilon}_4 = \infty\\
&b_1(\epsilon_1) = \frac{2}{1+\epsilon_1},\ b_2(\epsilon_2) = \frac{1.5}{1+\epsilon_2},\\
&b_3(\epsilon_3) = -\frac{0.1}{1+\epsilon_3}\mathbf{1}\{\epsilon_3 \geq 0.01\} - 1000 \cdot \mathbf{1}\{\epsilon_3 < 0.01\},\\
&b_4(\epsilon_4) = -\frac{1}{1+\epsilon_4}\mathbf{1}\{\epsilon_4 \geq 0.01\} - 1000 \cdot \mathbf{1}\{\epsilon_4 < 0.01\}\\
&q_1 = -1,\ q_2 = -1,\ q_3 = 1,\ q_4 = 1.
\end{aligned}
\tag{38}
$$

For the DSPs, the bid functions $b_3(\cdot), b_4(\cdot)$ in our working example (Example 1 correspond to the cost of communicating the data from sensors to the DSPs access point. In the context of our working example, higher costs are incurred as the quality increases since data must be transmitted at a higher rate or for a longer period of time. As a consequence, more energy is consumed by the sensors during communication [5].

Suppose that the platform allows for bids with quality levels in the interval $[0, 1]$. In the standard package exchange mechanism, detailed in Sec. 3, agents can only submit bids for the single data product as a discrete set of quality levels in $\{\frac{k}{r-1} : k = 0, \ldots, r-1\}$, where $r \in \mathbb{N}$ is the number of quality levels each agent can bid on.

With our mechanism in this scenario, all agents are winning with a total surplus of 2.38. On the other hand, there is a reduction in the surplus in the standard package exchange mechanism due to the limited number of quality levels that can be bid on[4]. This is illustrated in Fig. 2, which shows that at a large value of r is required to approach the surplus achieved by the proposed mechanism. Aside from increasing the communication requirements of the agents, the number of combinations of bids grows on the order of $O((1 + N_O r)^{N_D})$ leading to an increased search complexity.

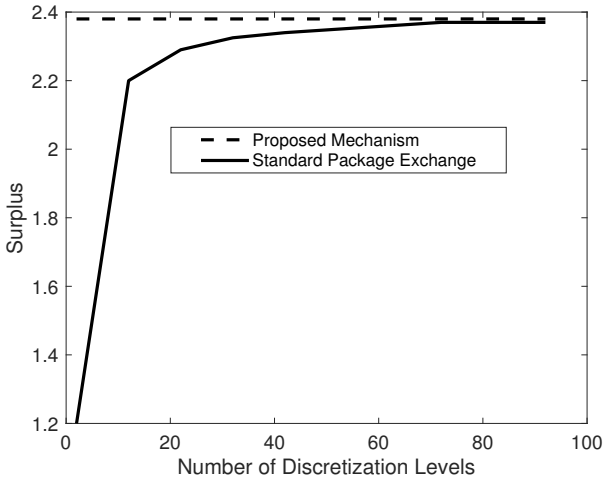

Fig. 2 Plot of surplus for varying numbers of quality discretization levels, r, in the standard package exchange.

8.2 Replication and Exclusivity

To illustrate the effect of DSPs replicating their data under the Vickrey and BWC pricing rules, consider two organizations $N_O = \{1, 2\}$ and two DSPs $N_D = \{3, 4\}$, offering a single common data product. Suppose the bids have parameters:

$$
\begin{aligned}
&L_1 = 0.01,\ L_2 = 0.01,\ L_3 = 0,\ L_4 = 0 \\
&\bar{\epsilon}_1 = 1,\ \bar{\epsilon}_2 = 1,\ \bar{\epsilon}_3 = \infty,\ \bar{\epsilon}_4 = \infty \\
&b_1(\epsilon_1) = \frac{2}{1+\epsilon_1},\ b_2(\epsilon_2) = \frac{1.5}{1+\epsilon_2}, \\
&b_3(\epsilon_3) = -\frac{0.1}{1+\epsilon_3}\mathbf{1}\{\epsilon_3 \geq 0.01\} - 1000 \cdot \mathbf{1}\{\epsilon_3 < 0.01\},
\end{aligned}
$$

[4]The solution to the winner determination problem for the standard package exchange mechanism in Sec. 3 was obtained via numerical linear integer programming methods in the PuLP python package [12].

$$b_4(\epsilon_4) = -\frac{1}{1+\epsilon_4}\mathbf{1}\{\epsilon_4 \geq 0.01\} - 1000 \cdot \mathbf{1}\{\epsilon_4 < 0.01\}$$
$$q_1 = -1,\ q_2 = -1,\ q_3 = 1,\ q_4 = 1. \tag{39}$$

With zero replications, all agents are winning with a total surplus of 2.38.

Suppose that DSP 4 performs data replication and introduces n shadow providers, such that $N_D = \{3, 4, 5, \ldots, 4+n\}$. The bid of the shadow provider $j = 5, \ldots, 4+n$ is given by

$$L_j = 0,\ \bar{\epsilon}_j = \infty,$$
$$b_j(\epsilon_j) = -1\frac{1}{1+\epsilon_j}\mathbf{1}\{\epsilon_j \geq 0.1\} - 1000 \cdot \mathbf{1}\{\epsilon_j < 0.1\}$$
$$q_j = 1. \tag{40}$$

The impact of the additional shadow providers with Vickrey pricing is shown in Table 1. Consistent with Prop. 10, we observe that the total surplus and total payoff of DSP 4 decreases with $n > 0$.

Table 1 also shows the impact of additional shadow providers with BWC pricing. As expected from Prop. 11, the allocated surplus decreases as the number of replications increases. However unlike the Vickrey rule, the allocated surplus remains non-zero. Nevertheless, the BWC pricing rule is budget balanced, which is not the case for the Vickrey rule.

Table 1 Impact of replications on DSP 4 with Vickrey and BWC pricing in scenario (39).

Replications n	0	1	2
Surplus Vickrey	0.495	0	0
Surplus BWC	0.27	0.12	0.1

8.3 Incentivizing Replication in Thin Markets

To illustrate how the mBWC pricing rule can incentivize DSP data replication in thin markets, consider the case where $N_O = \{1, 2\}$ and $N_D = \{3, 7\}$, with

$$L_1 = 0.01,\ L_2 = 0.01,\ L_3 = 0$$
$$\bar{\epsilon}_1 = 1,\ \bar{\epsilon}_2 = 1,\ \bar{\epsilon}_3 = \infty$$
$$b_1(\epsilon_1) = \frac{2}{1+\epsilon_1},\ b_2(\epsilon_2) = \frac{1.5}{1+\epsilon_2},$$
$$b_3(\epsilon_3) = -\frac{0.1}{1+\epsilon_3}\mathbf{1}\{\epsilon_3 \geq 0.01\} - 1000 \cdot \mathbf{1}\{\epsilon_3 < 0.01\},$$
$$q_1 = -1,\ q_2 = -1,\ q_3 = 1, \tag{41}$$

and where DSP 7 cannot be matched with either organization in N_O. In the absence of exclusivity constraints, we consider the scenario where it is desirable for DSP 3 to replicate its data *twice*.

This can be incentivized using the mBWC rule as follows. We introduce the shadow DSPs $4, 5, 6$, which replicate the bid of DSP 3. In the mBWC rule, we set N_E as the two winning bids and one of the losing bids; e.g., $N_E = \{3, 4\}$. Table 2 shows the impact of replications for the scenario described by (41) under the mBWC rule with preferred bids N_E. As expected, when one replication is present, the total surplus and the allocated surplus for DSP 3 increases as the shadow identity is matched with an organization. For two replications, which has been incentivized by the pricing rule, there is a reduction in the total surplus allocated to the DSP. Nevertheless, the reduction is much less than the corresponding reduction arising from the BWC rule. For three replications, the DSP surplus remains similar to the case of two replications. In other words, there is no incentive for the DSP to replicate their bids more than the number proposed by the platform.

Table 2 Impact of replications on DSP 3 surplus for the scenario in Eqn. (41) with mBWC and BWC pricing.

Replications n	0	1	2	3
Total Surplus	1.9	3.3	3.3	3.3
mBWC Surplus	1.5	1.69	1.28	1.28
BWC Surplus	1.5	1.69	0.54	0.48

9 Discussion and Related Work

Early work on data markets was motivated by participatory sensing [13]. There has recently been a renewed interest in the design of data markets [9–11, 14–18]. This interest is largely motivated by machine learning applications [19], where data is bought for the purpose of training models for classification or regression.

A key assumption in existing work is the presence of a single buyer. As a consequence, the value of data for the buyer can be directly evaluated via the accuracy of the trained machine learning model. Moreover, sellers are homogeneous in the sense that they provide the same type of data. In contrast, we have considered the scenario where multiple buyers (i.e., organizations) have potentially different data requirements. As such, a variant of a combinatorial auction or a package exchange mechanism is required.

Combinatorial auctions have been extensively studied, in part motivated by the problem of spectrum auctions [20]. A generalization of combinatorial auctions allowing for multiple sellers is known as a package exchange [3], which have been applied in environmental credit trading [2]. In [21], a pricing scheme based on the Vickrey rule was introduced, which allowed for budget balance. In [4], the BWC rule was introduced, which modifies the Shapley rule to ensure that only winners are allocated a surplus. An alternative approach has also been developed by [21] in the form of a modified VCG pricing rule, which finds the nearest prices that ensure budget balance.

In contrast with standard package exchanges, our mechanism allows for agents to reveal bids with flexible data quality requirements. This yields a significantly higher surplus, which results from a larger number of matches between organizations and DSPs.

Unlike other goods (e.g., spectrum or environmental credits), data can be easily replicated. In data markets with a single buyer and multiple sellers, the work in [9–11] has studied how pricing rules can disincentivize data replication. In particular, it has been observed that the Vickrey rule disincentivizes data replication. In this paper, we have extended this observation to our mechanism which supports multiple buyers. In addition, we have shown that the BWC rule also provides disincentives for data replication.

On the other hand, for thin markets where DSPs are uncertain about demand, it can be desirable for the market to incentivize data replication. We are not aware of any work on package exchanges addressing this problem. As such, we have introduced the mBWC rule, which provides incentives for DSPs to replicate their data.

The challenges in providing incentives for data sharing, discovery and integration [22, 23]. Data valuation in the context of machine learning has been investigated in [1, 24–27]. Surveys of data and information markets are available in [28, 29]. The impact of externalities on data markets have been investigated in [30, 31]. The impact of data leakage on market efficiency has recently been investigated in [32]. Bilateral data exchange has been considered in [33]. Dynamic arrival of buyers and sellers has been considered in [34]. Decentralized data markets have been studied in [35].

Applications of data markets include electricity retail [36] and energy forecasting [37, 38], wind power forecasting [39]. Data licenses have been considered in [40] and business models for data markets have been surveyed in [17].

10 Conclusions

Building on package exchanges, we have proposed a market mechanism for trading data in the presence of multiple buyers and sellers. In our scenario, each buyer and seller may have heterogeneous data demands and quality constraints, which can only be incorporated in standard package exchange mechanisms with additional communication and computational complexity. A second contribution of our work is the introduction of the mBWC pricing rule, which provides incentives for data replication in thin markets and complements standard Vickrey and BWC rules which are suitable in the presence of data exclusivity constraints. Given the complexity of winner determination and pricing in data markets, we believe the primary avenue of future work involves investigating additional algorithms and heuristics to speed up the computation process.