Received: 22 March 2021 | Accepted: 18 January 2022

DOI: 10.1111/poms.13676

POMS

ORIGINAL ARTICLE

An Approximation Scheme for Data Monetization

Sameer Mehta[1] | **Milind Dawande**[2] | **Ganesh Janakiraman**[2] | **Vijay Mookerjee**[2]

[1] Rotterdam School of Management, Erasmus University, Rotterdam, Zuid-Holland, The Netherlands

[2] Naveen Jindal School of Management, The University of Texas at Dallas, Richardson, Texas, USA

Correspondence
Sameer Mehta, Rotterdam School of Management, Erasmus University, Rotterdam, Zuid-Holland, The Netherlands.
Email: mehta@rsm.nl

Handling editor: Dr. Kalyan Singhal

Abstract

The unprecedented rate at which data are being generated has led to the growth of data markets where valuable data sets are bought and sold. A salient feature of this market is that a data-buyer (agent) is endowed with multidimensional private information, namely, her "ideal" record that she values the most and how her valuation for a given record changes as its distance from her ideal record changes. Consequently, the revenue-maximization problem faced by a data-seller (principal), who serves multiple buyers, is a multidimensional mechanism-design problem, which is well recognized as being difficult to solve. Our main result in this paper is an *approximation scheme* that guarantees a revenue within as close a positive amount from the optimal revenue as desired. The scheme generates a posted-price menu consisting of a set of item–price pairs—each entry in the menu consists of an item, that is, a set of records from the data set, and the price corresponding to that item. As a trade-off, the length of the menu resulting from the scheme increases as the desired guarantee gets closer to zero. For convenience in practice, data-sellers may want the ability to limit the length of the menu used by the scheme. To facilitate this, we extend our analysis to obtain a general approximation guarantee corresponding to a menu of any given length. We also demonstrate how the seller can exploit buyers' preferences to generate intuitive and useful rules of thumb for an effective practical implementation of the scheme.

KEYWORDS
approximation scheme, data monetization, multidimensional mechanism design

1 | INTRODUCTION

Modern businesses rely heavily on the use of data to produce and sell goods and services. Firms use data for a variety of purposes, for example, to market products, build accurate models, improve decision making, and target customers. Fueled primarily by the intelligence that data-driven analytics provides in making critical business decisions, the business of monetizing data has grown significantly over the past decade, and has spurred the formation of data markets where firms can buy and sell valuable data. With firms increasingly engaging in the trade of data, the importance of effective pricing mechanisms for data monetization has increased significantly.

Real-world data-sellers such as BookYourData,[1] MegaLeads,[2] DirectMail,[3] and TelephoneLists[4] offer customized data sets to help buyers target specific individuals/businesses of their interest. Such data sets help marketers in narrowing down individuals for targeted advertising, in communicating relevant product offerings, in deciding their bids for online advertising, and so forth. Below, we elaborate on how buyers can purchase specific data sets from BookYourData to execute their advertising campaigns.

BookYourData offers a wide variety of data sets for marketers based on job levels, job titles, job functions, and industries. Figure 1 illustrates some of the data sets with contact information on professionals in the healthcare industry. For example, the *Cardiology* data set contains 13,568 records of U.S.-based cardiologists, and the entire data set is offered at a price of $1633. Each record typically contains detailed information, including name, physical address, phone number, email address, website. In addition to these "ready-made" lists, marketers can also build and purchase customized lists by filtering the data sets for their specific needs.

While data pricing firms such as the ones mentioned above offer similar products, they use a wide variety of pricing policies, including all-or-nothing pricing, price schedules based on the quantity of data (i.e., the number of records)

Accepted by Dr. Kalyan Singhal.

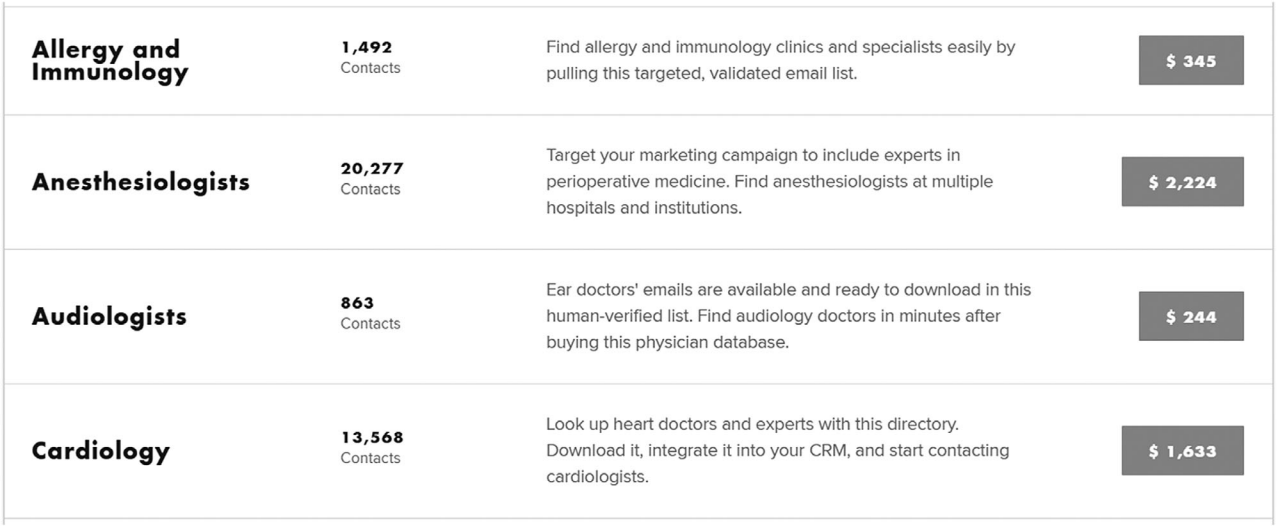

FIGURE 1 Examples of data sets offered for marketing purposes at BookYourData.
Source: https://www.bookyourdata.com/ready-made-lists/healthcare-professionals

purchased and the extent of filtering used, and differential pricing based on characteristics of data such as geographical location; see Mehta et al. (2021). Thus, one can reasonably conclude that while the problem of optimally pricing data sets is important for practitioners, it is challenging and not well understood.

In a recent paper, Mehta et al. (2021) develop a utility framework (which we summarize in Section 3) for the trade of data between buyers and a seller, and the corresponding pricing of the data by the seller. For data sets that can be arranged in a tabular (row–column) format, the data-seller's revenue-maximization problem is formulated as a multidimensional mechanism-design problem—it is well known, however, that such problems are difficult to solve optimally. In view of this difficulty, the authors (i) focus on a special case wherein the data set exhibits a simple "uniform" structure (broadly, the records in the data set are uniformly distributed over the space of records), and obtain an optimal mechanism under that case; (ii) show that this optimal mechanism can be implemented as a *price–quantity schedule* (i.e., a schedule in which the price for any set of records in the data set depends only on the number of records in that set and not on the identity of the records); and (iii) examine the performance of the optimal price–quantity schedule for the general case where the data set does not exhibit that special structure. To put it succinctly, the authors use the optimal price–quantity schedule as a heuristic pricing strategy and examine its performance.

In this paper, we analyze the general multidimensional mechanism-design problem formulated in Mehta et al. (2021), namely, the data-seller's revenue maximization problem. Our analysis offers two main results for this problem, which we now briefly describe. Our first result is an *approximation scheme*, which guarantees a revenue that is arbitrarily close to the optimal revenue (Theorem 1). Specifically, for an arbitrary constant $\eta > 0$, we obtain a pricing mechanism whose revenue to the data-seller differs from the optimal revenue by at most an additive constant that is proportional to $\sqrt{\eta}$. The scheme generates a posted-price menu consisting of a set of item–price pairs. Each entry in the menu consists of an item, that is, a set of records from the data set, and the price corresponding to that item. The seller can publicly post this menu and each buyer can choose the entry (from the menu) of her choice. Note that the posted-price menu generated by our approximation scheme is, in general, not a price–quantity schedule.

As one would expect, the length of the menu (i.e., the number of item–price pairs in it) resulting from the approximation scheme increases as the desired guarantee η gets closer to 0. Thus, it is possible that, for very small values of η, the length of the menu is prohibitively long for convenient use in practice. From a practice point of view, a natural design question for the data-seller then arises: What guarantee do *limited-size* menus provide? In other words, for any desired length L, can we obtain a menu with at most L item–price pairs that provides a revenue guarantee relative to the optimal? Our second result answers this question in the affirmative by obtaining a general approximation guarantee for a menu of length L (Theorem 2). Finally, we also demonstrate how the seller can exploit buyers' preferences to generate intuitive and useful "rules of thumb" for an effective practical implementation of the scheme.

2 | RELATED LITERATURE

Our work is related to two streams of literature: (i) pricing of data and information and (ii) multidimensional mechanism design. There is an extensive literature along both the streams. Therefore, we only review the literature that is closely related to our work.

2.1 | Pricing of data and information

The paper that is closest to ours is Mehta et al. (2021). As mentioned earlier, that paper develops a utility framework, which incorporates multidimensional preferences of the buyer over the data set. This results in a multidimensional mechanism-design problem, which is difficult to optimize. In view of this difficulty, the authors analyze a special case of the problem where the data set exhibits a simple structure and obtain an optimal solution for that special case. In contrast, we investigate the original multidimensional problem formulated in that paper and obtain an approximation scheme that guarantees a revenue that is arbitrarily close to the optimal revenue. Another key technical difference is that Mehta et al. (2021) impose several restrictive assumptions on the utility function of the buyers. In contrast, our analysis does not need any of those assumptions; instead, we only assume that the utility function is Lipschitz continuous in its argument.

There is a rich stream of literature on pricing of information goods that addresses a diverse set of topics including horizontal product differentiation (Choudhary, 2010; Dewan et al., 2003), vertical product differentiation (Bhargava & Choudhary, 2008; Chellappa & Mehra, 2018), payment mechanisms (Chen & Huang, 2016; Masuda & Whang, 2006; Sundararajan, 2004), and bundling (Bakos & Brynjolfsson, 1999; Geng et al., 2005; Wu et al., 2008), among others. In particular, our paper is related to the research stream on payment mechanisms and bundling. The approximation scheme that we develop in Section 4 yields a posted-price menu that consists of item–price pairs, where each item in the menu is generated by bundling an appropriate set of records.

Xue et al. (2016) develop a pricing strategy for the setting where a seller offers a variety of products to customers, who can construct personalized bundles and request quotes for these bundles from the seller. This is in contrast with our setting of data monetization, wherein a seller cannot afford to review personalized bundles from buyers and therefore opts for posted-price mechanisms. Jiang et al. (2011) analyze a multistage dynamic bundle-pricing model in an online setting where customers sequentially add products to their shopping carts and buy multiple products in one transaction. Here, the authors assume that the prices for individual products are already known and the goal is to determine optimal bundle prices considering customers' preferences. In the context of data monetization, it is practically infeasible for sellers to determine and post prices for individual records. Hitt and Chen (2005) analyze a pricing scheme for customized bundling, where consumers can choose up to a fixed number of goods from a larger set of goods for a fixed price, and compare the scheme with other traditional bundle-pricing schemes. In a subsequent study, Wu et al. (2008) take a numerical approach to analyze the problem of customized bundling. In these models of customized bundling, the willingness-to-pay (WTP) of buyers depends only on the *number* of goods they select and not on which goods they select. Our model analyzes a more general setting where the WTP of data-buyers depends on the *set of records* they purchase. Another point of departure from the traditional bundling literature on digital goods is that we consider a setting where consumers have multidimensional private information—this results in a multidimensional mechanism-design problem for the seller. While the solution to such problems is, in general, difficult to characterize, we exploit structural properties of our model and obtain an approximation scheme that is easy to implement.

Li et al. (2014) develop a framework for procuring private data from individuals. In their framework, data-buyers query the database to receive noisy data and data owners are compensated for their privacy loss. Bhargava et al. (2020) consider auction formats for selling sales leads data where buyers have a shared and an exclusive valuation for accessing the data set, and develop heuristic mechanisms for the setting. Agarwal et al. (2019) propose a mathematical model of designing a data marketplace that allows for buying and selling of training data for machine learning tasks.

There is a growing literature on markets of information (see Bergemann & Bonatti, 2019, for an excellent survey). Research in this vein investigates the monetization of information by considering issues such as the sale of supplemental information (Babaioff et al., 2012; Bergemann et al., 2018), competition (Bimpikis et al., 2019; Ma, 2019), and privacy (Eliaz et al., 2019; Ghosh & Roth, 2015).

2.2 | Multidimensional mechanism design

The main problem that we investigate in this paper is a multidimensional mechanism-design problem. The seminal work of Mussa and Rosen (1978) and Myerson (1981) considered information asymmetry with a one-dimensional parameter and the solution procedure they developed for identifying optimal mechanisms has since become standard in the literature. However, characterizing an optimal mechanism for the setting where agents are endowed with multidimensional private information has, in general, proven to be difficult (Belloni et al., 2010; Rochet & Choné, 1998). Another challenge posed by multidimensional setting is that optimal mechanisms are not likely to have attractive practical implementations, as eliciting the multiple types of the agents itself may be impractical. We refer the reader to Rochet and Stole (2003) for a detailed survey of the economics of multidimensional mechanism design. In view of these challenges, there is a growing stream of literature on the development and analysis of approximate mechanisms for multidimensional problems; see, for example, Chawla et al. (2007, 2010). In the development of our approximate scheme, an intermediate step involves discretizing the space of private information and formulating the mechanism-design problem as a linear program. Examples of studies in which a similar approach has been used include Cai et al. (2011) and Lavi and Swamy (2011).

TABLE 1 An illustrative data set consisting of information on users' mobile phone usage

	nonfilterable columns		*filterable* columns			
ID	**Device ID (hashed 8 bit)**	**User email**	**Zip**	**Operating system**	**Default browser**	**Avg Daily Usage (Hours)**
1	x20xkdg3	vifx9@gmail.com	94016	Android	Chrome	4.0
2	nls2354k	moyk23@gmail.com	94015	iOS	Safari	2.0
3	vkxwq23j	bronk1@yahoo.com	94111	iOS	Chrome	2.5
4	m12ls22l	esput21.vk@hotmail.com	94120	iOS	Safari	2.5
5	jsk221mb	th0rp22@yahoo.com	94120	Android	Firefox	1.5
6	zljwls12	e2zbay15@hotmail.com	94131	Android	Chrome	1.0
7	m23jnzp8	mangot.jk@gmail.com	94016	iOS	Chrome	0.5
8	naljh23j	noydi.tk67@gmail.com	94111	iOS	Firefox	2.5
9	op23bct7	rjk8989@yahoo.com	94144	Android	Chrome	3.5
10	qwn440n1	ppl23.mn2@gmail.com	94144	Android	Chrome	3.5

Our approximation scheme is obtained using a decomposition scheme—namely, bounding the performance offered by a "coarser" decomposition of the type space relative to that offered by a "finer" decomposition—that is inspired by a technique developed in Madarász and Prat (2017). In that paper, a profit-maximizing principal does not know the true distribution of an agent's type, which precludes her from designing a "truly" optimal contract. Instead, the principal knows this distribution "approximately" and therefore can only design near-optimal contracts. In contrast, in our data-selling context, the principal (data-seller) knows the true distribution of the agent's (buyer's) type and is able to correctly formulate the multidimensional mechanism-design problem. However, it is difficult to obtain an optimal solution to this problem in full generality, motivating us to seek an approximate solution. Our specific data-selling context imparts a lot of structure to the elements used in Madarász and Prat (2017)—this, in turn, helps us establish sharper results.

3 | MODEL PRELIMINARIES

Our model and notation are the same as in Mehta et al. (2021). We now briefly summarize the model preliminaries.

The focal data set is denoted by $\mathcal{D}$ and is represented in a tabular format consisting of rows and columns. Table 1 represents an illustrative data set that consists of information on the mobile phone usage of various users. The rows in the data set consist of information on entities/individuals, and are referred to as *records*. The columns in the data set represent the attributes of these entities/individuals and are categorized into two sets: (i) filterable columns—the buyer can filter the records of her choice by appropriately selecting the attributes in these columns (e.g., Zip, Operating system, Default browser, Avg daily usage in Table 1) and (ii) nonfilterable columns—these columns contain the contact information that the buyer can use to target the selected entities (e.g., Device ID, User email in Table 1). Henceforth, the terms "columns" and "record" will refer, respectively, to the filterable columns of the data set and the filterable part of a record. Note that two records can have identical filterable columns, in which case, we say that the two records are identical (e.g., Records 9 and 10 in the data set in Table 1 are identical—"94144–Android–Chrome–3.5"). The data set $\mathcal{D}$ consists of N real-valued filterable columns, with no missing entry; thus, each record in the data set is a vector in $\mathbb{R}^N$. Let $\chi \subseteq \mathbb{R}^N$ denote the set of all possible values a record in the data set can take. We assume that the record space χ is discrete; specifically, $\chi = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\}$ for some $M < \infty$ and $\boldsymbol{x}_i \in \mathbb{R}^N, i \in \{1, 2, \ldots, M\}$. Let g_i denote the number of records of value $\boldsymbol{x}_i, i \in \{1, 2, \ldots, M\}$, in the data set. For convenience, we normalize the number of records in the data set to 1; thus, $\sum_{i=1}^{M} g_i = 1$. In the illustrative data set in Table 1, we have $M = 9$ (since there are nine distinct records) and the proportion of records with value "94144–Android–Chrome–3.5" is $2/10$; that is, $g_9 = 2/10$. The set of records that are feasible for the buyer to purchase can be written as:

$$\mathcal{Y} = \left\{ (y_1, y_2, \ldots, y_M) : 0 \leq y_i \leq g_i \; \forall i \in \{1, 2, \ldots, M\} \right\}, \tag{1}$$

where y_i denotes the mass of record $\boldsymbol{x}_i, i \in \{1, 2, \ldots, M\}$, purchased by the buyer. A metric $\rho : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}_+$ measures the distance between any two records; thus, the distance between two records, say $\boldsymbol{x}$ and $\boldsymbol{y}$, is $\rho(\boldsymbol{x}, \boldsymbol{y})$.

As in Mehta et al. (2021), we assume that the purchase decisions of one buyer does not affect those of the other buyers—this allows us to focus our analysis on a single buyer. The ideal record of the buyer—one that she values the most—is denoted by $\bar{\boldsymbol{x}} \in \chi$. This N-dimensional ideal record is private information of the buyer and the seller only knows the distribution of the ideal record. Let f_i denote the probability that the ideal record of the buyer is $\boldsymbol{x}_i, i \in \{1, 2, \ldots, M\}$. We allow for the possibility that the buyer's ideal record may not

be present in the data set. The utility of any record $\boldsymbol{x}$ in the data set to the buyer decreases as its distance $\rho(\bar{\boldsymbol{x}}, \boldsymbol{x})$ from her ideal record $\bar{\boldsymbol{x}}$ increases. The rate of this decrease is characterized by a scalar parameter $t \in [0, \tau]$ and is referred to as the *decay type* of the buyer. This decay type is also private to the buyer and the seller only has distributional knowledge. Let $h(\cdot)$ denote the probability density function of the decay type of the buyer. To summarize, the buyer is endowed with $(N+1)$-dimensional private information—her ideal record $\bar{\boldsymbol{x}}$ (N-dimensional) and her decay type t (one-dimensional). The principal can estimate the distribution of the ideal record and that of the decay type in a variety of ways using historical transaction data. For instance, Paarsch (1992) and Donald and Paarsch (1993) use maximum likelihood estimation, Laffont et al. (1995) use the method of moments, Rezende (2008) employs a linear regression model (least square estimation), and Li and Zheng (2009) use semiparametric Bayesian methods.

The utility to a buyer of decay type t from purchasing a record located at a distance $d \geq 0$ (as measured by the metric ρ) from her ideal record is denoted by $v(d, t)$. Thus, for a buyer of type $(\bar{\boldsymbol{x}}, t)$, the utility obtained from purchasing a record $\boldsymbol{x}$ is given by $v(\rho(\bar{\boldsymbol{x}}, \boldsymbol{x}), t)$. We impose the following assumption on the utility function $v(d, t)$ of the buyers: The function $v(d, t)$ is *Lipschitz continuous* in both of its arguments. That is, there exists finite (Lipschitz) constants κ_d and κ_t, such that

$$|v(d_1, t) - v(d_2, t)| \leq \kappa_d |(d_1 - d_2)| \quad \forall d_1, d_2 \geq 0, \ \forall t \in [0, \tau], \tag{2}$$

$$|v(d, t_1) - v(d, t_2)| \leq \kappa_t |(t_1 - t_2)| \quad \forall d \geq 0, \ \forall t_1, t_2 \in [0, \tau]. \tag{3}$$

It is important to note here that we do not impose any other assumption on the utility function $v(d, t)$. This is in contrast to Mehta et al. (2021), where the authors impose several restrictive assumptions on the utility function (see, e.g., Assumptions P1–P5 and A1–A2 in that paper).

The Lipschitz continuity assumption helps prove Lemma 1 that establishes the following: For any two buyers with types $(\bar{\boldsymbol{x}}, t)$ and $(\bar{\boldsymbol{y}}, s)$, if the distance between their ideal records $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ and the distance between their decay types t and s are both sufficiently small, then the difference in the utilities that these two buyers obtain from consuming any set of records is also sufficiently small. This result enables us to construct a menu for the data-seller that achieves the following property: Buyers whose types are "close" to each other make their respective choices from the menu such that their payments are also "close" to each other; that is, the menu incentivizes neighboring buyers to select items that yield roughly the same revenue to the seller. As we will show in Section 4, such a menu guarantees near-optimal revenue to the seller.

Collectively, the elements of our model such as (a) ideal record, (b) a metric to measure the distance between records in the data set, and (c) decay type, effectively capture the nuances of the data-selling context. Further, our utility function $v(d, t)$ is quite general with mild technical assumptions on its structure. Thus, our modeling setup is quite general and therefore has the potential for wide applicability.

We envision that the buyer's purpose in purchasing the data is to execute advertising campaigns targeted at the individuals or businesses identified in the records of the data. Let $c \geq 0$ denote the targeting cost per record that the buyer incurs upon purchasing the data. The utility to a buyer of type $(\bar{\boldsymbol{x}}, t)$ from purchasing a set of records $S \subseteq \mathcal{D}$ can now be written as:

$$V(S; \bar{\boldsymbol{x}}, t) = \sum_{\boldsymbol{x}_i \in S} \left(v(\rho(\bar{\boldsymbol{x}}, \boldsymbol{x}_i), t) - c \right)^+ \cdot g_i. \tag{4}$$

Let V_{MAX} denote the maximum utility (over all buyer types) that can be obtained from consuming the entire data set. That is,

$$V_{\text{MAX}} = \sup_{(\bar{\boldsymbol{x}}, t) \in \mathcal{X} \times [0, \tau]} V(\mathcal{D}; \bar{\boldsymbol{x}}, t). \tag{5}$$

We are now ready to formulate our main problem.

3.1 | Problem formulation

The goal of the monopolist data-seller is to obtain a mechanism that maximizes his expected revenue. Using the Revelation Principle (Myerson, 1981), we restrict our attention, without loss of generality, to the class of direct mechanisms that are incentive compatible and individually rational for the buyer. Consistent with the notation of Mehta et al. (2021), a direct mechanism $\boldsymbol{\mu}$ is characterized by a pair of functions $(\mathcal{M}^{\mu}, \mathcal{P}^{\mu})$, where $\mathcal{M}^{\mu} : \mathcal{X} \times [0, \tau] \to \mathcal{D}$ is an allocation function and $\mathcal{P}^{\mu} : \mathcal{X} \times [0, \tau] \to \mathbb{R}$ is a payment function. The mechanism-design problem for the seller can be formulated as:

$$\max_{\mu} \ \mathbb{E}_{(\bar{\boldsymbol{x}}, t)} \left[\mathcal{P}^{\mu}(\bar{\boldsymbol{x}}, t) \right] \tag{P$^{\text{MD}}$}$$

subject to:

$$V(\mathcal{M}^{\mu}(\bar{\boldsymbol{x}}, t); \bar{\boldsymbol{x}}, t) - \mathcal{P}^{\mu}(\bar{\boldsymbol{x}}, t) \geq V(\mathcal{M}^{\mu}(\bar{\boldsymbol{y}}, s); \bar{\boldsymbol{x}}, t) - \mathcal{P}^{\mu}(\bar{\boldsymbol{y}}, s) \ \forall (\bar{\boldsymbol{x}}, t), (\bar{\boldsymbol{y}}, s) \in \mathcal{X} \times [0, \tau], \tag{IC-MD}$$

$$V(\mathcal{M}^{\mu}(\bar{\boldsymbol{x}}, t); \bar{\boldsymbol{x}}, t) - \mathcal{P}^{\mu}(\bar{\boldsymbol{x}}, t) \geq 0 \quad \forall (\bar{\boldsymbol{x}}, t) \in \mathcal{X} \times [0, \tau]. \tag{IR-MD}$$

In the above formulation, (IC-MD) and (IR-MD) are the incentive-compatibility and individual-rationality constraints, respectively. Note that P$^{\text{MD}}$ is an $(N+1)$-dimensional mechanism-design problem; it is well known that the analysis of such a problem in full generality is intractable (see,

e.g., Daskalakis et al., 2014). As mentioned earlier, Mehta et al. (2021) solve problem P^{MD} only under a special case—namely, when the records in the data set are "uniformly" distributed over the record space. In this special case, the $(N+1)$-dimensional mechanism-design problem essentially simplifies to a one-dimensional problem, and can therefore be solved by the classical Myersonian approach.

We now investigate problem P^{MD}.

4 | AN APPROXIMATION SCHEME FOR PROBLEM P^{MD}

In this section, we develop an approximation scheme for problem P^{MD}. Specifically, for an arbitrary constant $\eta > 0$, we will obtain a mechanism (a posted-price menu) whose revenue to the data-seller differs from the optimal revenue by at most an additive constant that is proportional to $\sqrt{\eta}$. Thus, as $\eta \to 0$, the difference between the revenue obtained from the mechanism and the optimal revenue approaches 0. As a trade-off, the length of the menu increases as η decreases. To obtain this approximation scheme, we will use, in an intermediate step, a result from Madarász and Prat (2017). As we will see below, while the context of Madarász and Prat (2017) is entirely different from the data-selling context, we are able to adapt a key idea introduced in that paper—namely, bounding the performance offered by a "coarser" decomposition of the type space relative to that offered by a "finer" decomposition.

The following intermediate result will be helpful in our subsequent analysis:

Lemma 1. *Consider any allocation of records $S \in \mathcal{D}$ and buyer types $(\bar{x}, t)$ and $(\bar{y}, s)$. For any $\eta > 0$, if $\rho(\bar{x}, \bar{y}) < \eta$ and $|t - s| < \eta$, then $|V(S; \bar{x}, t) - V(S; \bar{y}, s)| \leq \eta(\kappa_t + \kappa_d)$.*

The proof of Lemma 1 is provided in the Appendix. Essentially, Lemma 1 implies that, for any two buyers with types $(\bar{x}, t)$ and $(\bar{y}, s)$, if the distance between their ideal records $\bar{x}$ and $\bar{y}$ and the distance between their decay types t and s are both sufficiently small, then the difference in the utilities that these two buyers obtain from consuming *any* set of records S is also sufficiently small.

Before describing our approximation scheme in detail, we succinctly summarize the basic ideas. Broadly, the algorithm generates a posted-price menu consisting of a set of item–price pairs. Here, an item consists of a set of records from the data set and a corresponding mass for each of these records. The menu is carefully constructed to achieve the following property: Buyers whose types are "close" to each other (i.e., their ideal records are located "close" to each other in the record space and their decay types are also "close" to each other in the decay-type space) make their respective choices from the menu such that their payments are also "close" to each other. In other words, the menu generated by our algorithm incentivizes neighboring buyers to select items that yield roughly the same revenue to the seller.

The main steps of the algorithm are as follows: Given an arbitrary constant $\eta > 0$, we partition (Section 4.1) the Cartesian product of the record type space and the decay-type space (i.e., $\chi \times [0, \tau]$) into "cells" as follows: Partition the record space into sets such that the maximum distance between any two ideal records within a set is at most η. Similarly, partition the decay-type space into intervals such that the maximum distance between any two decay types within an interval is also at most η. The Cartesian product of the partitioned sets (of the record space) and the partitioned intervals (of the decay-type space) generates a corresponding partition of the space $\chi \times [0, \tau]$ into cells. Next, from each cell of the partition of $\chi \times [0, \tau]$, we arbitrarily select a "representative type," and assign an appropriate probability mass to each of those representative types. Then, we formulate an intermediate (discrete) mechanism-design problem in which we assume that the buying population consists only of representative-type buyers (Section 4.2). This problem can be formulated as a linear program and its solution (which is a menu) can be efficiently obtained. Finally, we appropriately modify the menu obtained from the solution of the linear program via a discounting technique (Section 4.3). This discounted menu, when presented to the entire buying population, achieves the desired approximation guarantee (Theorem 1).

4.1 | Partitioning of the type space and the associated representative-type set

- **Partitioning of the type space**: For any given $\eta > 0$, let $\pi^{REC}(\chi; \eta)$ denote a *partition* of the record space χ (i.e., a decomposition of the elements of the set χ into disjoint nonempty subsets) such that the maximum distance (measured using the metric ρ) between any two ideal records that belong to the same subset is at most η. Similarly, let $\pi^{DEC}([0, \tau]; \eta)$ denote a partition of the decay-type space, $[0, \tau]$, into a finite number of intervals such that the maximum Euclidean distance between any two decay types that belong to the same interval is also at most η. Let $\pi(\chi \times [0, \tau]; \eta) := \pi^{REC}(\chi; \eta) \times \pi^{DEC}([0, \tau]; \eta)$ denote the partition of the space $\chi \times [0, \tau]$. We refer to the elements of the partition $\pi(\chi \times [0, \tau]; \eta)$ as "cells" of that partition. For ease of exposition, we henceforth denote the partitions $\pi(\chi \times [0, \tau]; \eta)$, $\pi^{REC}(\chi; \eta)$, and $\pi^{DEC}([0, \tau]; \eta)$ simply as π, π^{REC}, and π^{DEC}, respectively. Let C_π^{REC} and C_π^{DEC} denote the cardinality of the partitions π^{REC} and π^{DEC}, respectively. The number of cells in the partition π is then $C_\pi^{REC} \cdot C_\pi^{DEC}$.
- **Representative-type set**: For a given partition π, we arbitrarily select a type, $(\bar{x}_i, t_j)$, from the cell corresponding to the ith set of partition π^{REC} and the jth interval of $\pi^{DEC}, i = 1, 2, \ldots, C_\pi^{REC}$, and $j = 1, 2, \ldots, C_\pi^{DEC}$, and refer to the type $(\bar{x}_i, t_j)$ as the *representative type* of that cell of the partition π. We refer to the set $\text{REP}(\pi) = \{(\bar{x}_1, t_1), \ldots, (\bar{x}_i, t_j), \ldots, (\bar{x}_{C_\pi^{REP}}, t_{C_\pi^{DEC}})\}$ as the *representative-type set* of partition π. Next, we associate a probability

weight for each representative type in the set $\text{REP}(\pi)$ as follows: For each representative-type $(\bar{\boldsymbol{x}}_i, t_j) \in \text{REP}(\pi)$, we compute (i) the sum of all the probability masses of the ideal records contained in the ith set of partition π^{REC} and denote this weight by ϕ_i^{REC} and (ii) the integral of the probability density function $h(\cdot)$ over all the decay types contained in jth interval of partition π^{DEC} and denote this weight by ϕ_j^{DEC}. The probability weight associated with the representative-type $(\bar{\boldsymbol{x}}_i, t_j)$ is then simply $\phi_i^{\text{REC}} \cdot \phi_j^{\text{DEC}}$.

The next step in our algorithm is to formulate an intermediate mechanism-design problem and obtain an optimal solution for that problem.

4.2 | Optimal solution of an intermediate problem via LP

Consider any fixed partition π and an associated representative-type set $\text{REP}(\pi)$ for that partition. Suppose that the data-seller wants to design a mechanism to sell his data set $\mathcal{D}$ for the setup in which the buyer's private type can take values only from the representative-type set. That is, the private type of the buyer is $(\bar{\boldsymbol{x}}_i, t_j) \in \text{REP}(\pi)$ with probability $\phi_i^{\text{REC}} \cdot \phi_j^{\text{DEC}}$. For this problem, a direct mechanism consists of:

- an allocation rule $\boldsymbol{\mathcal{M}} : \text{REP}(\pi) \to \mathcal{Y}$ that maps the representative types to the set of feasible records that the representative-type buyer can purchase, and
- A payment rule $\mathcal{P} : \text{REP}(\pi) \to \mathbb{R}$ that specifies the amount that the representative-type buyer pays to the seller for purchasing the set of records specified by $\boldsymbol{\mathcal{M}}$.

Corresponding to the representative-type $(\bar{\boldsymbol{x}}_i, t_j) \in \text{REP}$, the allocation $\boldsymbol{\mathcal{M}}(\bar{\boldsymbol{x}}_i, t_j)$ is a vector $(\mathcal{M}_{ij1}, \mathcal{M}_{ij2}, \ldots, \mathcal{M}_{ijM}) \in \mathcal{Y}$, where the component $\mathcal{M}_{ijk}$ is the mass of the record $\boldsymbol{x}_k \in \boldsymbol{\chi}$. The associated payment for that representative type is $\mathcal{P}(\bar{\boldsymbol{x}}_i, t_j)$. Further, for the representative-type buyer $(\bar{\boldsymbol{x}}_i, t_j)$, we define $v_{ik,j} := v(\rho(\bar{\boldsymbol{x}}_i, \boldsymbol{x}_k), t_j)$ as the utility that she obtains from purchasing a record located at $\boldsymbol{x}_k \in \boldsymbol{\chi}$. For this buyer, the cumulative utility from consuming the set of feasible records $(\mathcal{M}_{ij1}, \mathcal{M}_{ij2}, \ldots, \mathcal{M}_{ijM}) \in \mathcal{Y}$ is

$$\sum_{k=1}^{M} (v_{ik,j} - c)^+ \mathcal{M}_{ijk}. \tag{6}$$

The revenue-maximization problem for the data-seller can now be formulated as the following linear program:

$$\max_{\boldsymbol{\mathcal{M}}, \mathcal{P}} \sum_{i=1}^{C_\pi^{\text{REC}}} \sum_{j=1}^{C_\pi^{\text{DEC}}} \phi_i^{\text{REC}} \cdot \phi_j^{\text{DEC}} \cdot \mathcal{P}_{ij} \tag{P$^{\text{REP}}$}$$

subject to:

$$\sum_{k=1}^{M} (v_{ik,j} - c)^+ \mathcal{M}_{ijk} - \mathcal{P}_{ij} \geq \sum_{k=1}^{M} (v_{ik,j} - c)^+ \mathcal{M}_{rsk} - \mathcal{P}_{rs} \quad \forall i, j, r, s, \tag{IC-REP}$$

$$\sum_{k=1}^{M} (v_{ik,j} - c)^+ \mathcal{M}_{ijk} - \mathcal{P}_{ij} \geq 0 \quad \forall i, j, \tag{IR-REP}$$

$$0 \leq \mathcal{M}_{ijk} \leq g_k \quad \forall i, j, k, \tag{7}$$

$$\mathcal{P}_{ij} \geq 0 \quad \forall i, j. \tag{8}$$

The objective function in the above formulation is the expected revenue to the data-seller from the representative-type buyers. (IC-REP) and (IR-REP) are the incentive compatibility and the individual rationality constraints, respectively, for the representative-type buyers. The constraints in (7) represent feasible allocations to the representative-type buyers. Finally, (8) states that the representative-type buyer pays a nonnegative amount to the seller. Let mechanism $(\boldsymbol{\mathcal{M}}^{\text{REP}}, \mathcal{P}^{\text{REP}})$ denote an optimal solution to problem P$^{\text{REP}}$.

The data-seller can implement the mechanism $(\boldsymbol{\mathcal{M}}^{\text{REP}}, \mathcal{P}^{\text{REP}})$ by offering a *menu* to the representative-type buyers. Simply put, a menu is a list consisting of items and their corresponding prices. For the mechanism $(\boldsymbol{\mathcal{M}}^{\text{REP}}, \mathcal{P}^{\text{REP}})$, each allocation vector $\boldsymbol{\mathcal{M}}^{\text{REP}}(\bar{\boldsymbol{x}}_i, t_j)$, which we denote by $\boldsymbol{\mathcal{M}}_{ij}^{\text{REP}}$, constitutes an item in the menu. The corresponding price for that item is $\mathcal{P}^{\text{REP}}(\bar{\boldsymbol{x}}_i, t_j)$, which we denote by $\mathcal{P}_{ij}^{\text{REP}}$. Let ψ^{REP} denote the menu constructed from the mechanism $(\boldsymbol{\mathcal{M}}^{\text{REP}}, \mathcal{P}^{\text{REP}})$. Thus,

$$\psi^{\text{REP}} = \left\{ \left(\boldsymbol{\mathcal{M}}_{ij}^{\text{REP}}, \mathcal{P}_{ij}^{\text{REP}} \right); i = 1, 2, \ldots, C_\pi^{\text{REC}} \text{ and } j = 1, 2, \ldots, C_\pi^{\text{DEC}} \right\}. \tag{9}$$

Note that the menu ψ^{REP} above is obtained from the solution to an intermediate mechanism-design problem P$^{\text{REP}}$ in which *only* the representative-type buyers are considered. Next, we will appropriately modify ψ^{REP} and consider the revenue to the data-seller when the modified menu is offered to *all* buyer types (and not just to the representative types). When the buyer is presented with a menu, she selects an item from the menu and pays the corresponding price to the seller.

4.3 | The discounted-menu approximation scheme

For $\alpha \geq 0$, consider the discounted menu

$$\psi^{\text{REP-D}} = \left\{ \left(\boldsymbol{\mathcal{M}}_{ij}^{\text{REP}}, \mathcal{P}_{ij}^{\text{REP-D}} \right); i = 1, 2, \ldots, C_\pi^{\text{REC}} \text{ and } j = 1, 2, \ldots, C_\pi^{\text{DEC}} \right\}, \tag{10}$$

where $\mathcal{P}_{ij}^{\text{REP-D}} = (1-\alpha)\mathcal{P}_{ij}^{\text{REP}}$ $\forall i,j$. Thus, the discounted menu $\psi^{\text{REP-D}}$ is obtained from the menu ψ^{REP} by retaining the allocations (i.e., the items) but discounting the corresponding prices by a factor α. Let $\text{REV}(\chi \times [0,\tau]; \psi^{\text{REP-D}})$ denote the revenue to the seller from offering the discounted menu $\psi^{\text{REP-D}}$ to the entire buying population. The following lemma is an application of a more general result in Madarász and Prat (2017)—namely, Theorem 1 in that paper—and helps us assess the revenue guarantee offered by $\psi^{\text{REP-D}}$.

Lemma 2. (*Application of Madarász & Prat, 2017*): *For any $\eta > 0$, setting $\alpha = \sqrt{(2\eta(\kappa_d + \kappa_t))/V_{\text{MAX}}}$ yields the following performance guarantee for the menu $\psi^{\text{REP-D}}$:*

$$\text{REV}(\chi \times [0,\tau]; \psi^{\text{REP-D}}) \geq \text{REV}(\chi \times [0,\tau]; \psi^{\text{OPT-MD}}) - 4\sqrt{\frac{2\eta(\kappa_d + \kappa_t)}{V_{\text{MAX}}}}. \tag{11}$$

We now briefly contrast our context and goal with those in Madarász and Prat (2017). The basic principal–agent setting analyzed in Madarász and Prat (2017) is as follows. A profit-maximizing principal does not know the true distribution of an agent's type, which precludes her from designing a "truly" optimal contract. Instead, the principal knows this distribution "approximately" and therefore can only design near-optimal contracts. For this setting, Madarász and Prat (2017) develop a technique to bound the principal's loss by using an approximate mechanism relative to an optimal mechanism. In contrast, in our data-selling context, the principal (data-seller) knows the true distribution of the agent's (buyer's) type and is able to correctly formulate the mechanism-design problem P^{MD}. However, it is difficult to obtain an optimal solution to P^{MD} in full generality, motivating us to seek an approximate solution.

Our specific data-selling context imparts a lot of structure to the elements used in Madarász and Prat (2017)—this helps us establish sharper results. For instance, unlike in that paper, we are able to formulate the intermediate mechanism-design problem for the representative types (Section 4.2) as a linear program and efficiently solve it. As another example, in our case, the distance between two records is measured using an arbitrary metric ρ, whereas, in Madarász and Prat (2017), the distance between two types is measured using the Euclidean metric. The use of an arbitrary metric is important in the data-selling context, since, typically, the Euclidean distance metric is not an appropriate one. Indeed, as we will see in Remark 3 (Section 6), the structure of the distance metric ρ can be exploited to generate "intelligent" partitions of the space $(\chi \times [0,\tau])$, which in turn can yield solutions with higher revenue for the data-seller.

Lemma 2 yields a lower bound on the revenue accrued to the data-seller by using the discounted menu $\psi^{\text{REP-D}}$. Note that for a buyer of a given type, it is possible that while facing the menu $\psi^{\text{REP-D}}$, she makes a decision that is different from that of her representative type. However, the price discounts and the Lipschitz condition on the utility of the buyer guarantee that the impact of such deviations on the revenue of the seller is not too large. Offering a discounted menu to the buyers introduces the following trade-off for the seller: On the one hand, the seller loses revenue by offering deep discounts (high value of α) to the buyers. On the other hand, deep discounts help relax the buyer's incentive compatibility constraints and make the menu attractive to her. This trade-off is optimized by setting the discount factor $\alpha = \sqrt{(2\eta(\kappa_d + \kappa_t))/V_{\text{MAX}}}$.

Note that $\lim_{\eta \to 0} \sqrt{(2\eta(\kappa_d + \kappa_t))/V_{\text{MAX}}} = 0$. Thus, as η approaches 0, the revenue to the data-seller from offering the menu $\psi^{\text{REP-D}}$ approaches the optimal revenue. Therefore, the discounted menu algorithm indeed yields an approximation scheme. We summarize this scheme below.

For any given $\eta > 0$, the following algorithm lays out the steps to generate a discounted menu that yields near-optimal revenue to the data-seller.

ALGORITHM 1 An approximation scheme

Input: The probability mass function $f(\cdot)$ of the ideal record of the buyer and the probability density function $h(\cdot)$ of decay type of the buyer.

1: Create a partition $\pi^{\text{REC}}(\chi;\eta)$ of the record space χ and create a partition $\pi^{\text{DEC}}([0,\tau];\eta)$ of the decay-type space $[0,\tau]$ (Section 4.1).

2: Obtain the partition $\pi(\chi \times [0,\tau];\eta) := \pi^{\text{DEC}}(\chi;\eta) \times \pi^{\text{REC}}([0,\tau];\eta)$ of the type-space $\chi \times [0,\tau]$.

3: Create a representative-type set $\text{REP}(\pi)$ by arbitrarily selecting one representative type from each cell of the partition $\pi(\chi \times [0,\tau];\eta)$.

4: Compute the probability weight of each representative type in $\text{REP}(\pi)$ using the functions $f(\cdot)$ and $h(\cdot)$.

5: Obtain the menu ψ^{REP} by solving the linear program P^{REP} (Section 4.2).

6: Obtain the discounted menu $\psi^{\text{REP-D}}$ by using the discount factor $\alpha = \sqrt{\frac{2\eta(\kappa_d+\kappa_t)}{V_{\text{MAX}}}}$ (Section 4.3).

Theorem 1. *Algorithm 1 is an approximation scheme for P^{MD}: For any fixed $\eta > 0$, the algorithm delivers a menu $\psi^{\text{REP-D}}$ whose revenue to the data-seller is guaranteed to be within an additive constant $4\sqrt{(2\eta(\kappa_d + \kappa_t))/V_{\text{MAX}}}$ from the optimal revenue.*

Remark 1. **(Computational Complexity of Algorithm 1)**: Recall that the decay-type space is continuous and is specified by its distribution function on the support $[0,\tau]$. Any discussion of the computational complexity of Algorithm 1 would naturally involve the input size of the algorithm, which in turn would require a common agreement on how the distribution function of the decay-type space is specified. For this reason, we avoid a discussion on the complexity of Algorithm 1. If the decay-type space is discrete then problem P^{MD} can be formulated as a linear program. In this case, the input to problem P^{MD} consists of the probability mass functions of the record space and the decay-type space. The size of the linear program is polynomial in the size of this input.

TABLE 2 An Illustrative (a) data set and (b) its numerical representation

(a)					(b)		
Record ID	**Name**	**Email**	**Income**	**Age**	**Record ID**	**Income**	**Age**
1	J. Warner	jw32@hotmail.com	<20k	18–25	1	1	1
2	S. Burns	sburn2@gmail.com	<20k	18–25	2	1	1
3	V. Smith	smitv@gmail.com	<20k	25–40	3	1	2
4	T. Harris	2thar43@hotmail.com	20k–50k	18–25	4	2	1
5	L. Green	lg.2345@yahoo.com	20k–50k	25–40	5	2	2
6	S. Lyon	slyo67@gmail.com	20k–50k	40–60	6	2	3
7	K. Cummins	kat923@hotmail.com	20k–50k	40–60	7	2	3
8	E. Starc	ezstarc81@gmail.com	50k–150k	18–25	8	3	1
9	T. Wade	tom2w@yahoo.com	50k–150k	25–40	9	3	2
10	A. Bairstow	andy.b89@gmail.com	50k–150k	25–40	10	3	2
11	T. Ali	tekk6ali@yahoo.com	50k–150k	40–60	11	3	3
12	N. Anderson	fi2ernat@gmail.com	50k–150k	40–60	12	3	3
13	D. Bracey	dbrac34@gmail.com	50k–150k	>60	13	3	4
14	B. Broad	bybroa187@gmail.com	50k–150k	>60	14	3	4
15	H. Butler	butler.de1@hotmail.com	>150k	25–40	15	4	2
16	T. Cross	tom87cr@hotmail.com	>150k	40–60	16	4	3
17	S. Curran	silv272@yahoo.com	>150k	40–60	17	4	3
18	F. Dunkley	dunkyo9@hotmail.com	>150k	>60	18	4	4
19	L. Barber	larbar832@gmail.com	>150k	>60	19	4	4
20	F. Bernard	bern981@gmail.com	>150k	>60	20	4	4

We end this section by commenting on the special case of a homogeneous decay type; that is, the decay type is the same for all the buyers. In this case, the seller knowing the distribution of the decay type is equivalent to the seller knowing the common decay type of the buyers. That is, if the common decay type of the buyers is, say, $t_0 \in [0, \tau]$, then the buyer's distributional knowledge is $\text{Prob}(t = t_0) = 1$. For this special case, the only simplification that we have in our model is that the problem (P^{MD}) becomes an N-dimensional (instead of $(N + 1)$-dimensional) mechanism-design problem as the seller knows the decay type of the buyers and only the N-dimensional ideal records of the buyers are private to them. However, the procedure to obtain an approximation scheme for the problem (P^{MD}) remains the same. In summary, for the special case of buyers with a homogeneous decay type, our model and results remain effectively unchanged.

Next, we illustrate the approximation scheme on an example data set.

5 | AN ILLUSTRATIVE EXAMPLE

We begin by describing our data set. In our example, the values of the parameters of the scheme have been made to ensure that all the steps of the scheme can be conveniently illustrated and the length of the menu offered by the data-seller is not excessive.

5.1 | Description of the data set

Table 2a shows an illustrative data set that consists of 20 records and four columns. The first two columns—*Name* and *Email*—are non-filterable columns, whereas the last two columns—*Income* and *Age*—are filterable columns (i.e., the buyer can filter the records of her choice using these columns). Thus, in our analysis, we have $N = 2$. The choice of two filterable columns in the records of the data set is deliberate, so that some of the steps of the approximation scheme can be visualized using simple figures. Table 2b shows a numerical representation of the data set obtained by mapping the categorical values in each of two filterable columns to real numbers (see Table 3 for the mapping).

Since the *Income* and *Age* fields can each take four different values, the record space χ consists of 16 elements

TABLE 3 Mapping used to obtain the numerical data set from the actual data set

Income	**Mapped value**	**Age**	**Mapped value**
<20k	1	18–25	1
20k–50k	2	25–40	2
50k–150k	3	40–60	3
>150k	4	>60	4

TABLE 4 Elements of the record space, distribution of records in the data set, and the distribution of the ideal records

Record values (income-age)	No. of records in the data set	Normalized no. of records	Distribution of the ideal records
1-1	2	0.1	0.01
1-2	1	0.05	0.01
1-3	0	0	0.02
1-4	0	0	0.02
2-1	1	0.05	0.05
2-2	1	0.05	0.06
2-3	2	0.1	0.06
2-4	0	0	0.07
3-1	1	0.05	0.06
3-2	2	0.1	0.08
3-3	2	0.1	0.1
3-4	2	0.1	0.1
4-1	0	0	0.06
4-2	1	0.05	0.08
4-3	2	0.1	0.1
4-4	3	0.15	0.12

(record types); that is, $M = 16$. These elements are shown in the first column of Table 4. Corresponding to these elements, the number of records (resp., normalized number of records) in the data set are shown in the second (resp., third) column of that table. Finally, the fourth column of the table shows the distribution of the ideal records. We assume that the decay type t is uniformly distributed over the support $[0, 1]$; that is, $t \sim U(0, 1)$. The distance between any two records—say, $\boldsymbol{x} = (x_1, x_2)$ and $\boldsymbol{y} = (y_1, y_2)$—in the data set is measured using the weighted Manhattan metric, $\rho^{\text{W-MAN}}$, defined as follows:

$$\rho^{\text{W-MAN}}(\boldsymbol{x}, \boldsymbol{y}) = 0.2 \cdot |(x_1 - y_1)| + 0.1 \cdot |(x_2 - y_2)|, \quad (12)$$

For instance, the distance between Record 4 (2-1) and Record 15 (4-2) is $0.2 \cdot |2 - 4| + 0.1 \cdot |1 - 2| = 0.5$.

The utility function of the buyers is defined as $v(d, t) = t \cdot (5 - d)$ and the targeting cost, c, is set to 0.1. Then, for the utility function $v(d, t)$, the Lipschitz constants, κ_d and κ_t, are 1 and 5, respectively, and the maximum utility, V_{MAX} (defined in Equation (5)), over all buyers that can be obtained from consuming the entire data set is 4.65. Figure 2 shows a visual representation of the record space and the decay-type space.

We are now ready to illustrate the approximation scheme described in Section 4. We explain in detail the steps for constructing the menu for $\eta = 0.25$.

5.2 | Approximation scheme for $\eta = 0.25$

- **Partitioning of the type space**: The first step of the scheme is to partition the record space χ and the decay-type space $[0, 1]$ into cells, such that the maximum distance (measured using the metric $\rho^{\text{W-MAN}}$) between any two ideal records in the same cell is less than η $(= 0.25)$ as well as the distance between any two decay types in the same cell is less than η. We create this partition as follows (see Figure 3 for a visual representation of the partition):

$$\pi^{\text{REC}}(\chi; 0.25) = \{(1-1, 2-1), (1-2, 2-2), (1-3, 2-3), (1-4, 2-4), (3-1, 4-1), (3-2, 4-2), (3-3, 4-3), (3-4, 4-4)\}.$$

$$\pi^{\text{DEC}}([0, 1]; 0.25) = \{[0, 0.25), [0.25, 0.5), [0.5, 0.75), [0.75, 1]\}. \quad (13)$$

For this partition, we have $C_\pi^{\text{REC}} = 8$ (the number of partitions in the record space) and $C_\pi^{\text{DEC}} = 4$ (the number of partitions in the decay-type space).

- **Representative-type set**: For the above partition $\pi (= \pi^{\text{REC}} \times \pi^{\text{DEC}})$, the representative-type set, $\text{REP}(\pi)$, is constructed by arbitrarily selecting an ideal record (from each cell in the partition π^{REC}) and a decay type (from each cell in the partition π^{DEC}) as follows (see Figure 4 to visualize the representative ideal records and the representative

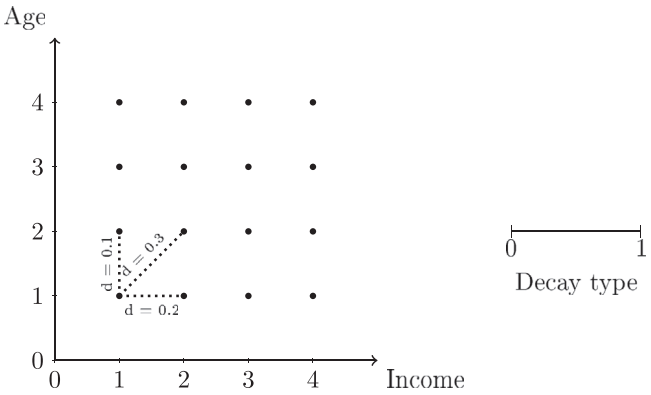

FIGURE 2 Visual representation of the record space and the decay-type space

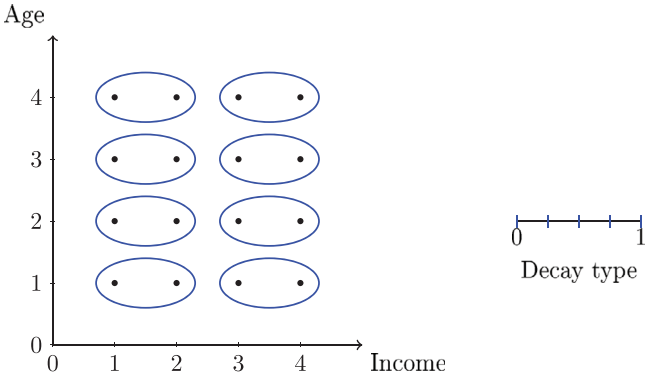

FIGURE 3 A partition of the record space and the decay-type space for $\eta = 0.25$ [Color figure can be viewed at wileyonlinelibrary.com]

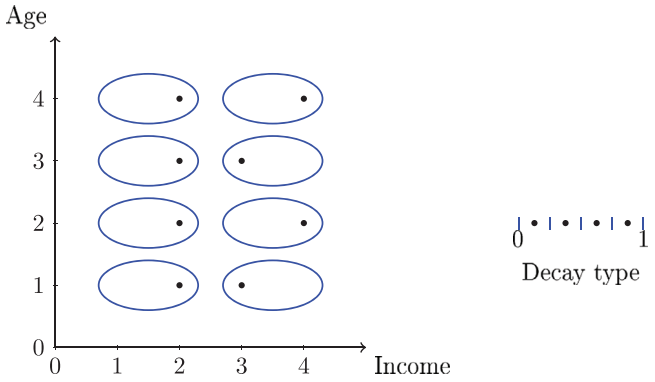

FIGURE 4 Representative ideal records and representative decay types [Color figure can be viewed at wileyonlinelibrary.com]

decay types):

$$\begin{aligned}\text{REP}(\pi) = \{&(2-1,0.2),(2-1,0.4),(2-1,0.6),(2-1,0.8),\\ &(2-2,0.2),(2-2,0.4),(2-2,0.6),(2-2,0.8),\\ &(2-3,0.2),(2-3,0.4),(2-3,0.6),(2-3,0.8),\\ &(2-4,0.2),(2-4,0.4),(2-4,0.6),(2-4,0.8),\\ &(3-1,0.2),(3-1,0.4),(3-1,0.6),(3-1,0.8),\\ &(4-2,0.2),(4-2,0.4),(4-2,0.6),(4-2,0.8),\\ &(3-3,0.2),(3-3,0.4),(3-3,0.6),(3-3,0.8),\\ &(4-4,0.2),(4-4,0.4),(4-4,0.6),(4-4,0.8)\}.\end{aligned} \tag{14}$$

The probability mass associated with each representative type is shown in Table 5.

- **Solution to the intermediate linear program**: We formulate the mechanism-design problem for the representative types as a linear program (described in Section 4.2). The solution to this linear program yields a menu, ψ^{REP}, that consists of item–price pairs, where an item is a bundle of records from the data set.
- **Obtaining the discounted menu**: Finally, we obtain the discounted menu, $\psi^{\text{REP-D}}$, by retaining the items from the menu ψ^{REP}, but discounting the corresponding prices by α, that is, $\mathcal{P}^{\text{REP-D}} = (1-\alpha)\mathcal{P}^{\text{REP}}$. In this case, the optimal value of α is $\sqrt{(2\eta(\kappa_d+\kappa_t))/V_{\text{MAX}}} = 0.8$. Table 6 shows discounted menu for the data-seller. The discounted menu yields an expected revenue of 1.34 to the seller and the additive approximation guarantee, $4\sqrt{(2\eta(\kappa_d+\kappa_t))/V_{\text{MAX}}}$, is 3.2. In other words, the revenue of 1.34 achieved by this menu (which corresponds to $\eta = 0.25$) is within 3.2 of the optimal revenue. Thus, the optimal revenue is at most 4.54.

TABLE 5 Probability mass associated with each representative type

Representative ideal record	Representative decay type	Probability mass ($\phi_i^{\text{REC}} \cdot \phi_i^{\text{DEC}}$)
2-1	0.2	$(0.01+0.05)\cdot 0.25 = 0.015$
2-1	0.4	$(0.01+0.05)\cdot 0.25 = 0.015$
2-1	0.6	$(0.01+0.05)\cdot 0.25 = 0.015$
2-1	0.8	$(0.01+0.05)\cdot 0.25 = 0.015$
2-2	0.2	$(0.06+0.01)\cdot 0.25 = 0.0175$
2-2	0.4	$(0.06+0.01)\cdot 0.25 = 0.0175$
2-2	0.6	$(0.06+0.01)\cdot 0.25 = 0.0175$
2-2	0.8	$(0.06+0.01)\cdot 0.25 = 0.0175$
2-3	0.2	$(0.02+0.06)\cdot 0.25 = 0.02$
2-3	0.4	$(0.02+0.06)\cdot 0.25 = 0.02$
2-3	0.6	$(0.02+0.06)\cdot 0.25 = 0.02$
2-3	0.8	$(0.02+0.06)\cdot 0.25 = 0.02$
2-4	0.2	$(0.02+0.07)\cdot 0.25 = 0.015$
2-4	0.4	$(0.02+0.07)\cdot 0.25 = 0.015$
2-4	0.6	$(0.02+0.07)\cdot 0.25 = 0.015$
2-4	0.8	$(0.02+0.07)\cdot 0.25 = 0.015$
3-1	0.2	$(0.06+0.06)\cdot 0.25 = 0.03$
3-1	0.4	$(0.06+0.06)\cdot 0.25 = 0.03$
3-1	0.6	$(0.06+0.06)\cdot 0.25 = 0.03$
3-1	0.8	$(0.06+0.06)\cdot 0.25 = 0.03$
4-2	0.2	$(0.08+0.08)\cdot 0.25 = 0.04$
4-2	0.4	$(0.08+0.08)\cdot 0.25 = 0.04$
4-2	0.6	$(0.08+0.08)\cdot 0.25 = 0.04$
4-2	0.8	$(0.08+0.08)\cdot 0.25 = 0.04$
3-3	0.2	$(0.1+0.1)\cdot 0.25 = 0.05$
3-3	0.4	$(0.1+0.1)\cdot 0.25 = 0.05$
3-3	0.6	$(0.1+0.1)\cdot 0.25 = 0.05$
3-3	0.8	$(0.1+0.1)\cdot 0.25 = 0.05$
4-4	0.2	$(0.1+0.12)\cdot 0.25 = 0.0550$
4-4	0.4	$(0.1+0.12)\cdot 0.25 = 0.0550$
4-4	0.6	$(0.1+0.12)\cdot 0.25 = 0.0550$
4-4	0.8	$(0.1+0.12)\cdot 0.25 = 0.0550$

TABLE 6 Menu from the approximation scheme corresponding to $\eta = 0.25$

Discounted menu $\psi^{\text{REP-D}}$	
Item (record IDs) $\mathcal{M}^{\text{REP}}$	Price $\mathcal{P}^{\text{REP-D}}$
{1, 2, 4, 5}	0.06
{3, 9, 13, 15}	0.07
{6, 7, 11, 12, 17}	0.07
{6, 7, 13, 14}	0.07
{3, 8, 9, 10, 15, 16}	0.07
{1, 2, …, 20 (entire data set)	2.5

This concludes our illustrative example for $\eta = 0.25$.

Theorem 1 shows that our approximation scheme (namely, Algorithm 1) for Problem P^{MD} is capable of guaranteeing a revenue that is within as close a positive amount from the optimal revenue as desired. However, it is important to note that, the smaller the value of the input parameter η (which indicates the desired closeness to the optimal revenue), the finer is the partition of the Cartesian product of the record space and the decay-type space that is needed by the scheme. Consequently, the number of representative types created by the scheme increases, which may yield a longer discounted menu $\psi^{\text{REP-D}}$ (i.e., a menu with a higher number of item–price pairs). Thus, a very small value of η can make the discounted menu prohibitively long for use in practice. A natural design question for the data-seller then arises: What guarantee do *limited-size* menus provide? In other words, for any desired length L, can we obtain a menu with at most L item–price pairs that provides a revenue guarantee relative to the optimal? We discuss this question in the next section.

6 | MENUS OF LIMITED SIZE

To obtain a performance guarantee for menus of bounded length, we examine the results from the previous section in greater detail. The following result generalizes Lemma 1 (Section 4):

Lemma 3. *Consider any set of records $\mathcal{S} \in \mathcal{D}$ and buyer types $(\bar{\boldsymbol{x}}, t)$ and $(\bar{\boldsymbol{y}}, s)$. For any $\eta_d, \eta_t > 0$, if $\rho(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) < \eta_d$ and $|t - s| < \eta_t$, then $|V(\mathcal{S}; \bar{\boldsymbol{x}}, t) - V(\mathcal{S}; \bar{\boldsymbol{y}}, s)| \leq \eta_t \kappa_t + \eta_d \kappa_d$.*

The proof of Lemma 3 is a straightforward generalization of that of Lemma 1 and is therefore avoided for brevity. Similarly, the result in Lemma 2 can be generalized as follows:

Lemma 4. *For any $\eta_d, \eta_t > 0$, setting $\alpha = \sqrt{(2(\eta_d \kappa_d + \eta_t \kappa_t))/V_{\text{MAX}}}$ yields*

$$\text{REV}(\boldsymbol{\chi} \times [0, \tau]; \psi^{\text{REP-D}}) \geq \text{REV}(\boldsymbol{\chi} \times [0, \tau]; \psi^{\text{OPT-MD}}) - 4\sqrt{\frac{2(\eta_d \kappa_d + \eta_t \kappa_t)}{V_{\text{MAX}}}}. \tag{15}$$

Next, recall that the length of the discounted menu $\psi^{\text{REP-D}}$ (obtained via Algorithm 1) depends on the number of representative types, which in turn depends on the number of cells created by the corresponding partition of the Cartesian product of the record space and the decay-type space. Below, we describe a simple partition that generates at most L cells. The following notation will be convenient for our discussion: For a given N-dimensional vector $\boldsymbol{x}$, let $\boldsymbol{x}[i]$ denote the ith component of that vector, $i \in \{1, 2, \ldots, N\}$.

- Compute the minimum and maximum values that the records in the record space take along each dimension (i.e., each column of the data set) as follows:

$$\omega_i^L := \min\{\boldsymbol{x}_1[i], \boldsymbol{x}_2[i], \ldots, \boldsymbol{x}_M[i]\}, \quad \forall i \in \{1, 2, \ldots, N\},$$
$$\omega_i^H := \max\{\boldsymbol{x}_1[i], \boldsymbol{x}_2[i], \ldots, \boldsymbol{x}_M[i]\}, \quad \forall i \in \{1, 2, \ldots, N\}. \tag{16}$$

  Thus, $[\omega_i^L, \omega_i^H]$ is the range of values that the records in the record space can take along the ith dimension, $i = 1, 2, \ldots, N$. Let $\boldsymbol{\Gamma} := \bigtimes_{i=1}^{N} [\omega_i^L, \omega_i^H]$ denote the Cartesian product of these intervals.
- For $i \in \{1, 2, \ldots, N\}$, partition the interval $[\omega_i^L, \omega_i^H]$ into $\lfloor L^{\frac{1}{N+1}} \rfloor$ equal intervals. Such a partitioning generates $(\lfloor L^{\frac{1}{N+1}} \rfloor)^N$ regions in the $\boldsymbol{\Gamma}$ space. Let η_d^L denote the maximum distance between any two records that belong to the same region. Since all the regions are identical N-dimensional hyperrectangles, η_d^L can be computed simply by evaluating the length of the longest diagonal of any region. To this end, consider the region that has $(\omega_1^L, \omega_2^L, \ldots, \omega_N^L)$ as one of its vertices. The endpoints of the longest diagonal in this region are the vertices $(\omega_1^L, \omega_2^L, \ldots, \omega_N^L)$ and $\left(\frac{\omega_1^H - \omega_1^L}{\lfloor L^{\frac{1}{N+1}} \rfloor}, \frac{\omega_2^H - \omega_2^L}{\lfloor L^{\frac{1}{N+1}} \rfloor}, \ldots, \frac{\omega_N^H - \omega_N^L}{\lfloor L^{\frac{1}{N+1}} \rfloor}\right)$. Therefore,

$$\eta_d^L = \rho\left((\omega_1^L, \omega_2^L, \ldots, \omega_N^L), \left(\frac{\omega_1^H - \omega_1^L}{\lfloor L^{\frac{1}{N+1}} \rfloor}, \frac{\omega_2^H - \omega_2^L}{\lfloor L^{\frac{1}{N+1}} \rfloor}, \ldots, \frac{\omega_N^H - \omega_N^L}{\lfloor L^{\frac{1}{N+1}} \rfloor}\right)\right). \tag{17}$$

- Partition the support $[0, \tau]$ of the decay-type space also into $\lfloor L^{\frac{1}{N+1}} \rfloor$ equal intervals. Thus, the maximum distance, η_t^L, between any two decay types that belong to the same interval is simply

$$\eta_t^L = \frac{\tau - 0}{\lfloor L^{\frac{1}{N+1}} \rfloor} = \frac{\tau}{\lfloor L^{\frac{1}{N+1}} \rfloor}. \tag{18}$$

- The above partitioning of the record space and the decay-type space generates a total of $(\lfloor L^{\frac{1}{N+1}} \rfloor)^{N+1} \leq L$ cells of the space $\boldsymbol{\chi} \times [0, \tau]$.

Using the above partition of $\boldsymbol{\chi} \times [0, \tau]$, let $\psi_L^{\text{REP-D}}$ denote the discounted menu obtained via Algorithm 1. Note that this menu has a length of at most L (i.e., it has at most L item–price pairs). Then, an application of Lemma 2 yields the following:

Theorem 2. *The menu $\psi_L^{\text{REP-D}}$ offers the following performance guarantee:*

$$\text{REV}(\boldsymbol{\chi} \times [0,\tau]; \psi_L^{\text{REP-D}}) \geq \text{REV}(\boldsymbol{\chi} \times [0,\tau]; \psi^{\text{OPT}}) - 4\sqrt{\frac{2\left(\kappa_d \eta_d^L + \kappa_t \eta_t^L\right)}{V_{\text{MAX}}}}, \quad (19)$$

where η_d^L and η_t^L are as defined in (17) and (18), respectively.

Theorem 2 gives us an absolute performance guarantee for the case when the data-seller desires to use a menu of length at most L. As expected, the larger the upper bound L on the length of the menu, the better is the performance guarantee.

In the remarks below, we elaborate on some important aspects of our framework and analysis.

Remark 2. **(Optimizing the Partition)**: Theorem 2 states the performance guarantee that is offered by the discounted menu $\psi_L^{\text{REP-D}}$ obtained via partitioning each dimension of the space $\boldsymbol{\Gamma} \times [0,\tau]$ into the same number, namely, $\lfloor L^{\frac{1}{N+1}} \rfloor$, of intervals. One can generalize by partitioning the ith dimension of the $\boldsymbol{\Gamma}$ space into $\ell_i \geq 1$ equal intervals, $i \in \{1,2,\ldots,N\}$, and the type space into $\ell_{N+1} \geq 1$ equal intervals. The optimal values of $\ell_i, i = 1,2\ldots N+1$, can then be determined by solving the following nonlinear integer program:

$$\min_{\ell_1,\ell_2,\ldots,\ell_{N+1}} \kappa_d \cdot \rho\left((\omega_1^L, \omega_2^L, \ldots, \omega_N^L), \times \left(\frac{\omega_1^H - \omega_1^L}{\ell_1}, \frac{\omega_2^H - \omega_2^L}{\ell_2}, \ldots, \frac{\omega_N^H - \omega_N^L}{\ell_N}\right)\right) + \kappa_t \cdot \frac{\tau}{\ell_{N+1}} \quad \text{P}^{\text{INT}}$$

$$\text{s.t.} \quad \prod_{i=1}^{N+1} \ell_i \leq L, \ \ell_i \in \mathbb{Z}_{++} \ \forall i \in \{1,2,\ldots,N+1\}. \quad (20)$$

We note that the naive solution $\ell_1 = \ell_2 = \cdots = \ell_{N+1} = \lfloor L^{\frac{1}{N+1}} \rfloor$ is a feasible solution to P^{INT}. Thus, an optimal solution to problem P^{INT} results in a partition of the type space that generates a menu with a (weakly) tighter revenue guarantee than that obtained via the naive solution.

More generally, note that partitioning each dimension of the space $\boldsymbol{\Gamma} \times [0,\tau]$ into intervals generates a partition of that space into hyperrectangles. Algorithm 1 does not require the cells of the partition to be hyperrectangles. Future work can focus on more sophisticated partitioning methods that result in better revenue guarantees.

Remark 3. **(Exploiting the Distance Metric)**: The structure of the distance metric ρ can also be exploited by the data-seller to generate intuitive and useful rules of thumb aimed at obtaining a good partition of the type space. We provide a simple illustrative example.

Suppose the data set $\mathcal{D}$ consists of two columns (i.e., $N = 2$), *Income* and *Age*. For simplicity, suppose that the lowest value of the *Income* column and the *Age* column, across all the records in the data set, is 0. Thus, $\omega_1^L = \omega_2^L = 0$. Further, suppose that the highest value in the *Income* column and the *Age* column is normalized to 1. Thus, $\omega_1^H = \omega_2^H = 1$. Suppose that the buyer is more sensitive to changes in the values in the *Income* field relative to changes in the values in the *Age* field; for instance, the buyer is a firm selling a high-value product (targeted at high-income individuals) that appeals to customers of all ages. Such preferences of the buyer can be suitably modeled by using the weighted Manhattan metric, $\rho^{\text{W-MAN}}$, defined for the two columns, as follows:

$$\rho^{\text{W-MAN}}(\boldsymbol{x},\boldsymbol{y}) = \alpha_{\text{INCOME}}|(x_1 - y_1)| + \alpha_{\text{AGE}}|(x_2 - y_2)|, \quad (21)$$

where $\alpha_{\text{INCOME}} > \alpha_{\text{AGE}} > 0$. For the metric $\rho^{\text{W-MAN}}$, the optimal partitioning problem P^{INT} becomes:

$$\min_{\ell_1,\ell_2,\ell_3} \kappa_d \cdot \left(\alpha_{\text{INCOME}}\left(\frac{1}{\ell_1}\right) + \alpha_{\text{AGE}}\left(\frac{1}{\ell_2}\right)\right) + \kappa_t \cdot \frac{\tau}{\ell_3} \quad (22)$$

$$\text{s.t.} \quad \ell_1\ell_2\ell_3 \leq L, \quad \ell_1,\ell_2,\ell_3 \in \mathbb{Z}_{++}. \quad (23)$$

Since $\alpha_{\text{INCOME}} > \alpha_{\text{AGE}}$, it is straightforward to verify that the optimal solution $(\ell_1^*, \ell_2^*, \ell_3^*)$ to the above optimization problem satisfies $\ell_1^* \geq \ell_2^*$. This implies that while designing a menu of length at most L, the data-seller should prefer using a finer partition of the *Income* dimension as compared to that of the *Age* dimension. In other words, the dimensions along which changes are more important for the buyer should be partitioned into finer intervals. Such an insight becomes even more useful for data sets in which records have a large number of columns but the buyer is interested only in a few of these.

Remark 4. **(Private Distance Metric)**: Thus far, we have assumed in our analysis that the distance metric, ρ, is public knowledge. Even if this metric is private information, we can obtain an approximation scheme for the corresponding optimal mechanism-design problem. The development is similar to our analysis above. We briefly outline the key steps in formulating the optimal mechanism-design problem with a private distance metric and obtaining an approximation scheme for that problem.

Let $\Omega = \{\rho_1, \rho_2, \ldots, \rho_K\}$ denote a finite set of (completely defined) distance metrics. Along with an ideal record $\bar{\boldsymbol{x}} \in \boldsymbol{\chi}$ and a decay-type $t \in [0,\tau]$, the buyer is endowed with a metric $\rho \in \Omega$. Let θ_i denote the probability that the buyer uses the distance metric $\rho_i, i \in \{1,2,\ldots,K\}$ and $\sum_{i=1}^K \theta_i = 1$. The buyer has private information in three aspects: (i) the ideal record $\bar{\boldsymbol{x}} \in \boldsymbol{\chi} \subseteq \mathbb{R}^N$; (ii) the decay type, $t \in [0,\tau]$;

and (iii) the distance metric $\rho \in \Omega$, and is characterized by the $(N+2)$-dimensional tuple $(\bar{\boldsymbol{x}}, t, \rho)$. The seller only has distributional knowledge about the private information of the buyer. We assume that the ideal record, the decay type, and the distance metric used by the buyer, are independently distributed. The utility to the buyer of type $(\bar{\boldsymbol{x}}, t, \rho)$ from purchasing a set of records $S \subseteq \mathcal{D}$ is given by:

$$V(S; \bar{\boldsymbol{x}}, t, \rho) = \sum_{x_i \in S} \left(v(\rho(\bar{\boldsymbol{x}}, \boldsymbol{x}), t) - c\right)^+ g_i. \tag{24}$$

A direct mechanism $\boldsymbol{\xi}$ is characterized by a pair of functions $(\mathcal{M}^{\boldsymbol{\xi}}, \mathcal{P}^{\boldsymbol{\xi}})$, where $\mathcal{M}^{\boldsymbol{\xi}} : \boldsymbol{\chi} \times [0, \tau] \times \Omega \to \mathcal{D}$ is an allocation function and $\mathcal{P}^{\boldsymbol{\xi}} : \boldsymbol{\chi} \times [0, \tau] \times \Omega \to \mathbb{R}$ is a payment function. Thus, in a direct mechanism $\boldsymbol{\xi}$, if the buyer reveals her type as $(\bar{\boldsymbol{x}}, t, \rho)$, then she receives the set of records $\mathcal{M}^{\boldsymbol{\xi}}(\bar{\boldsymbol{x}}, t, \rho) \subseteq \mathcal{D}$ from the data set and pays $\mathcal{P}^{\boldsymbol{\xi}}(\bar{\boldsymbol{x}}, t, \rho)$ to the seller. The mechanism-design problem for the seller can then be written as:

$$\max_{\boldsymbol{\xi}} \ \mathbb{E}_{(\bar{\boldsymbol{x}}, t, \rho)}\left[\mathcal{P}^{\boldsymbol{\xi}}(\bar{\boldsymbol{x}}, t, \rho)\right] \tag{P$^{\text{METRIC}}$}$$

subject to:

$$V(\mathcal{M}^{\boldsymbol{\xi}}(\bar{\boldsymbol{x}}, t, \rho); \bar{\boldsymbol{x}}, t, \rho) - \mathcal{P}^{\boldsymbol{\xi}}(\bar{\boldsymbol{x}}, t, \rho) \geq V(\mathcal{M}^{\boldsymbol{\xi}}(\bar{\boldsymbol{y}}, s, \rho'); \bar{\boldsymbol{x}}, t, \rho) - \mathcal{P}^{\boldsymbol{\xi}}(\bar{\boldsymbol{y}}, s, \rho')$$
$$\forall\, (\bar{\boldsymbol{x}}, t, \rho), (\bar{\boldsymbol{y}}, s, \rho') \in \boldsymbol{\chi} \times [0, \tau] \times \Omega, \tag{IC-METRIC}$$

$$V(\mathcal{M}^{\boldsymbol{\xi}}(\bar{\boldsymbol{x}}, t, \rho); \bar{\boldsymbol{x}}, t, \rho) - \mathcal{P}^{\boldsymbol{\xi}}(\bar{\boldsymbol{x}}, t, \rho) \geq 0 \ \forall (\bar{\boldsymbol{x}}, t, \rho) \in \boldsymbol{\chi} \times [0, \tau] \times \Omega, \tag{IR-METRIC}$$

where, (IC-METRIC) and (IR-METRIC) are the incentive-compatibility and the individual-rationality constraints.

Given an arbitrary constant $\eta > 0$, we generate a partition of the record space, the decay-type space, and the set of distance metrics (i.e., $\boldsymbol{\chi} \times [0, \tau] \times \Omega$), as follows: Partition the record space $\boldsymbol{\chi}$ into sets, say $\tilde{\pi}^{\text{REC}}(\boldsymbol{\chi}; \eta)$, such that, for *each* distance metric that belongs to Ω, the maximum distance between any two ideal records (measured using that metric) within a set is at most η. Similarly, partition the decay-type space $[0, \tau]$ into intervals, say $\tilde{\pi}^{\text{DEC}}([0, \tau]; \eta)$, such that the maximum Euclidean distance between any two decay types within an interval is also at most η. Then, the desired partition of the space $\boldsymbol{\chi} \times [0, \tau] \times \Omega$ is simply $\tilde{\pi}^{\text{REC}}(\boldsymbol{\chi}; \eta) \times \tilde{\pi}^{\text{DEC}}([0, \tau]; \eta) \times \Omega$. As we have done above, from each cell of this partition, we arbitrarily select a representative type and assign an appropriate probability mass to that type. Next, solve an intermediate mechanism-design problem in which we assume that the buying population consists only of the representative-type buyers: This problem can be formulated as a linear program and its solution (which is a menu) can be efficiently obtained. Finally, we appropriately modify the menu obtained from the solution of the linear program using the same discounting technique as the one above. This discounted menu, when presented to the entire buying population, yields a revenue to the data-seller that differs from the optimal revenue by at most an additive constant that is proportional to $\sqrt{\eta}$.

Finally, the following remark clarifies a few points related to our single-buyer analysis.

Remark 5. **(Single-Buyer Analysis)**: Our communication with several real-world data-sellers revealed that sole (i.e., unshared) access to the data set is not an important business need for their customers. Our model allows for the case where multiple buyers are interested in the same set of records in the data set. Moreover, any subset of records in the data set can be sold to multiple buyers, since the seller can create multiple copies at no significant cost. Recall our assumption in Section 3 that a buyer's purchasing decisions do not affect the utilities of the other buyers. Under this assumption, the analysis for the case of multiple buyers reduces to the case of a single buyer. Put differently, our analysis for a single buyer trivially extends to the case of multiple buyers. This assumption is satisfied under several reasonable data-selling contexts, for example, when the data set is for general-purpose use such as email-marketing or telemarketing, or when the buyers' operations are in disjoint markets; see Mehta et al. (2021).

7 | CONCLUDING REMARKS

Motivated by the strong growth in the buying and selling of data, this paper studies a revenue-maximization problem faced by a data-seller, and develops an approximation scheme for that problem. Our setting considers data-buyers who possess private information in two aspects: (i) their respective ideal records, which they value the most; and (ii) the rate at which their valuation for the records in the data set decays as they move away from their respective ideal records. The output of the approximation scheme is a menu of item–price pairs: Each entry in the menu consists of an item, that is, a set of records from the data set, and the price corresponding to this item. The seller can publicly post this menu and each buyer can choose the entry (from the menu) of her choice. The main property of the scheme is that it guarantees a revenue within as close a positive amount from the optimal revenue as desired. As a trade-off, the length of the menu (i.e., the number of item–price pairs in it) resulting from the scheme increases as the desired distance from the optimal revenue decreases. We also investigate the performance guarantees offered by menus of limited size.

The design and analysis of data-pricing solutions is in its nascent stage, and will continue to provide a fertile avenue for future research. To illustrate, we discuss a challenging extension to the framework we analyzed in this paper. In our analysis, the utility to a buyer from purchasing a record depends only on her decay type and the deviation (captured

via a distance metric) in the characteristics of that record from those of her ideal record. This functional form of buyers' utilities is appropriate in contexts where the records in the data set are "horizontally" differentiated. For example, consider a data set that has information on physicians in a metropolis and the ideal records of different buyers are located in different counties of that metropolis. Here, although the ideal records of the buyers are different, the utility that a buyer obtains from purchasing a record depends only on the distance of that record from her ideal record and not on the identity of the ideal record. It would be interesting to analyze the vertically differentiated setting where the utility function of the buyer also depends on the identity of the ideal record. For instance, consider two types of data-buyers—a luxury-good firm and an essential-good firm—who are interested in purchasing a consumer data set for marketing purpose. The ideal records for both the firms would naturally be different, and the luxury-good firm may have a high WTP for the records in the data set as compared to the essential-good firm. In such a scenario, the utility to a buyer of type $(\bar{\boldsymbol{x}}, t)$ from purchasing a record $\boldsymbol{x}$ in the data set will, in general, depend on the identity of the ideal record $\bar{\boldsymbol{x}}$, and could be written as:

$$v(\rho(\bar{\boldsymbol{x}}, \boldsymbol{x}), t, \bar{\boldsymbol{x}}). \tag{25}$$

For such a utility function, the analysis quickly becomes intractable. A special case to which our analysis in this paper extends is when the decay type of the buyers is homogeneous (say, t_0). For a buyer with ideal record $\bar{\boldsymbol{x}}$, let q denote the utility to that buyer for her ideal record $\bar{\boldsymbol{x}}$. We assume that q is private to the buyer and the seller only has a distributional knowledge of this utility for the ideal record. We also assume that the utility to a buyer from the ideal record, q, is independent of her ideal record, $\bar{x}$. Thus, the role of the WTP q for the ideal record here is akin to the role of the decay type t in our base model. To summarize, a buyer is endowed with two types of private information—her ideal record $\bar{\boldsymbol{x}}$ (N-dimensional) and the utility q (one-dimensional) that she obtains from her ideal record. Consequently, the utility from a record $\boldsymbol{x}$ to a buyer whose ideal record is $\bar{\boldsymbol{x}}$ can now be written as $v(\rho(\bar{\boldsymbol{x}}, \boldsymbol{x}), q)$. Observe that the structure of the utility function here is similar to that of the utility function in our base model except that the second argument of the utility function in the former model denotes the WTP of the buyer instead of her decay type. Throughout our analysis of the base model, the only assumption that we impose on the utility function is that it is Lipschitz continuous in its arguments. Therefore, it is straightforward to verify that if the utility function $v(\rho(\bar{\boldsymbol{x}}, \boldsymbol{x}), q)$ is Lipschitz continuous, then the approximation scheme developed in Section 4 as well as the results in Section 6 continue to hold for this special case of the vertically differentiated setting.

ORCID

Sameer Mehta https://orcid.org/0000-0002-0410-3248
Milind Dawande https://orcid.org/0000-0001-6956-0856
Vijay Mookerjee https://orcid.org/0000-0001-5583-3585

ENDNOTES

[1] https://www.bookyourdata.com/
[2] https://megaleads.com/
[3] https://www.directmail.com/
[4] https://www.telephonelists.biz/

REFERENCES

Agarwal, A., Dahleh, M., & Sarkar, T. (2019). A marketplace for data: An algorithmic solution. In *Proceedings of the* 2019 *ACM Conference on Economics and Computation* (pp. 701–726). ACM.

Babaioff, M., Kleinberg, R., & Paes Leme, R. (2012). Optimal mechanisms for selling information. In *Proceedings of the* 13*th ACM Conference on Electronic Commerce* (pp. 92–109). ACM.

Bakos, Y., & Brynjolfsson, E. (1999). Bundling information goods: Pricing, profits, and efficiency. *Management Science*, *45*(12), 1613–1630.

Belloni, A., Lopomo, G., & Wang, S. (2010). Multidimensional mechanism design: Finite-dimensional approximations and efficient computation. *Operations Research*, *58*(4-part-2), 1079–1089.

Bergemann, D., & Bonatti, A. (2019). Markets for information: An introduction. *Annual Review of Economics*, *11*, 85–107.

Bergemann, D., Bonatti, A., & Smolin, A. (2018). The design and price of information. *American Economic Review*, *108*(1), 1–48.

Bhargava, H. K., & Choudhary, V. (2008). Research note—When is versioning optimal for information goods? *Management Science*, *54*(5), 1029–1035.

Bhargava, H. K., Csapó, G., & Müller, R. (2020). On optimal auctions for mixing exclusive and shared matching in platforms. *Management Science*, *66*(6), 2653–2676.

Bimpikis, K., Crapis, D., & Tahbaz-Salehi, A. (2019). Information sale and competition. *Management Science*, *65*(6), 2646–2664.

Cai, Y., Daskalakis, C., & Weinberg, S. M. (2011). On optimal multidimensional mechanism design. *ACM SIGecom Exchanges*, *10*(2), 29–33.

Chawla, S., Hartline, J. D., & Kleinberg, R. (2007). Algorithmic pricing via virtual valuations. In *Proceedings of the* 8*th ACM Conference on Electronic Commerce* (pp. 243–251). ACM.

Chawla, S., Hartline, J. D., Malec, D. L., & Sivan, B. (2010). Multi-parameter mechanism design and sequential posted pricing. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing* (pp. 311–320). ACM.

Chellappa, R. K., & Mehra, A. (2018). Cost drivers of versioning: Pricing and product line strategies for information goods. *Management Science*, *64*(5), 2164–2180.

Chen, Y.-J., & Huang, K.-W. (2016). Pricing data services: Pricing by minutes, by gigs, or by megabytes per second? *Information Systems Research*, *27*(3), 596–617.

Choudhary, V. (2010). Use of pricing schemes for differentiating information goods. *Information Systems Research*, *21*(1), 78–92.

Daskalakis, C., Deckelbaum, A., & Tzamos, C. (2014). The complexity of optimal mechanism design. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 1302–1318). SIAM.

Dewan, R., Jing, B., & Seidmann, A. (2003). Product customization and price competition on the internet. *Management Science*, *49*(8), 1055–1070.

Donald, S. G., & Paarsch, H. J. (1993). Piecewise pseudo-maximum likelihood estimation in empirical models of auctions. *International Economic Review*, *34*, 121–148.

Eliaz, K., Eilat, R., & Mu, X. (2019). *Optimal privacy-constrained mechanisms*. CEPR Discussion Paper No. DP13536.

Geng, X., Stinchcombe, M. B., & Whinston, A. B. (2005). Bundling information goods of decreasing value. *Management Science*, *51*(4), 662–667.

Ghosh, A., & Roth, A. (2015). Selling privacy at auction. *Games and Economic Behavior*, *91*, 334–346.

Hitt, L. M., & Chen, P.-y. (2005). Bundling with customer self-selection: A simple approach to bundling low-marginal-cost goods. *Management Science*, *51*(10), 1481–1493.

Jiang, Y., Shang, J., Kemerer, C. F., & Liu, Y. (2011). Optimizing e-tailer profits and customer savings: Pricing multistage customized online bundles. *Marketing Science*, *30*(4), 737–752.

Laffont, J.-J., Ossard, H., & Vuong, Q. (1995). Econometrics of first-price auctions. *Econometrica: Journal of the Econometric Society*, *63*, 953–980.

Lavi, R., & Swamy, C. (2011). Truthful and near-optimal mechanism design via linear programming. *Journal of the ACM* (*JACM*), *58*(6), 25.

Li, C., Li, D. Y., Miklau, G., & Suciu, D. (2014). A theory of pricing private data. *ACM Transactions on Database Systems* (*TODS*), *39*(4), 1–28.

Li, T., & Zheng, X. (2009). Entry and competition effects in first-price auctions: Theory and evidence from procurement auctions. *The Review of Economic Studies*, *76*(4), 1397–1429.

Ma, Y. (2019). Monopoly and competition in the markets for information. *SSRN*: https://ssrn.com/abstract=3505101.

Madarász, K., & Prat, A. (2017). Sellers with misspecified models. *The Review of Economic Studies*, *84*(2), 790–815.

Masuda, Y., & Whang, S. (2006). On the optimality of fixed-up-to tariff for telecommunications service. *Information Systems Research*, *17*(3), 247–253.

Mehta, S., Dawande, M., Janakiraman, G., & Mookerjee, V. (2021). How to sell a dataset? Pricing policies for data monetization. *Information Systems Research*, *32*(4), 1281–1297.

Mussa, M., & Rosen, S. (1978). Monopoly and product quality. *Journal of Economic Theory*, *18*(2), 301–317.

Myerson, R. B. (1981). Optimal auction design. *Mathematics of Operations Research*, *6*(1), 58–73.

Paarsch, H. J. (1992). Deciding between the common and private value paradigms in empirical models of auctions. *Journal of Econometrics*, *51*(1-2), 191–215.

Rezende, L. (2008). Econometrics of auctions by least squares. *Journal of Applied Econometrics*, *23*(7), 925–948.

Rochet, J.-C., & Choné, P. (1998). Ironing, sweeping, and multidimensional screening. *Econometrica*, *66*, 783–826.

Rochet, J.-C., & Stole, L. A. (2003). The economics of multidimensional screening. Dewatripont, M., Hansen, L. P., & Turnovsky, S. J. (Eds.), *Advances in Economics and Econometrics, Theory and Applications, Eighth World Congress*, Cambridge University Press, 2003. https://doi.org/10.1017/CBO9780511610240

Sundararajan, A. (2004). Nonlinear pricing of information goods. *Management Science*, *50*(12), 1660–1673.

Wu, S., Hitt, L. M., Chen, P., & Anandalingam, G. (2008). Customized bundle pricing for information goods: A nonlinear mixed-integer programming approach. *Management Science*, *54*(3), 608–622.

Xue, Z., Wang, Z., & Ettl, M. (2016). Pricing personalized bundles: A new approach and an empirical study. *Manufacturing & Service Operations Management*, *18*(1), 51–68.

How to cite this article: Mehta, S., Dawande, M., Janakiraman, G., & Mookerjee, V. (2022). An Approximation Scheme for Data Monetization. *Production and Operations Management*, 1–17. https://doi.org/10.1111/poms.13676

APPENDIX A: PROOF OF LEMMA 1

We want to show that for any allocation of records $S \in \mathcal{D}$ and buyer types $(\bar{x}, t)$ and $(\bar{y}, s)$, if $\rho(\bar{x}, \bar{y}) < \eta$ and $|t - s| < \eta$ for any $\eta > 0$, then $|V(S; \bar{x}, t) - V(S; \bar{y}, s)| \leq \eta(\kappa_t + \kappa_d)$.

Proof. Let

$$z^* := \arg\max_{z \in \chi} |(v(\rho(\bar{x}, z), t) - v(\rho(\bar{y}, z), s))|. \tag{A1}$$

Note that since χ is a finite set, z^* exists. Recall that an allocation of records $S \in \mathcal{D}$ can be written as $(u_1, u_2, \ldots, u_M)$, where u_i denotes the mass of record x_i and $u_i \leq g_i$, $i \in \{1, 2, \ldots, M\}$. Thus, for a given allocation $S \in \mathcal{D}$ and buyer types $(\bar{x}, t)$ and $(\bar{y}, s)$, we have:

$$\begin{aligned}
&|V(S; \bar{x}, t) - V(S; \bar{y}, s)| \\
&= \left| \sum_{i=1}^{M} (v(\rho(\bar{x}, x_i), t) - c)^+ \cdot u_i - \sum_{i=1}^{M} (v(\rho(\bar{y}, x_i), s) - c)^+ \cdot u_i \right| \\
&= \left| \sum_{i=1}^{M} \left((v(\rho(\bar{x}, x_i), t) - c)^+ - (v(\rho(\bar{y}, x_i), s) - c)^+ \right) \cdot u_i \right| \\
&\leq \sum_{i=1}^{M} \left| \left((v(\rho(\bar{x}, x_i), t) - c)^+ - (v(\rho(\bar{y}, x_i), s) - c)^+ \right) \cdot u_i \right|.
\end{aligned} \tag{A2}$$

□

Claim A1. For a given allocation S, and buyer types $(\bar{x}, t)$ and $(\bar{y}, s)$, we have

$$\begin{aligned}
&\sum_{i=1}^{M} \left| \left((v(\rho(\bar{x}, x_i), t) - c)^+ - (v(\rho(\bar{y}, x_i), s) - c)^+ \right) \cdot u_i \right| \\
&\leq \sum_{i=1}^{M} \left| \left(v(\rho(\bar{x}, x_i), t) - v(\rho(\bar{y}, x_i), s) \right) \cdot u_i \right|.
\end{aligned} \tag{A3}$$

Proof. For any $i \in \{1, 2, \ldots, M\}$, note that

- If $v(\rho(\bar{x}, x_i), t) - c \geq 0$ and $v(\rho(\bar{y}, x_i), s) - c \geq 0$, then

$$\begin{aligned}
&\left| (v(\rho(\bar{x}, x_i), t) - c)^+ - (v(\rho(\bar{y}, x_i), s) - c)^+ \right| \\
&= |v(\rho(\bar{x}, x_i), t) - v(\rho(\bar{y}, x_i), s)|.
\end{aligned} \tag{A4}$$

- If $v(\rho(\bar{x}, x_i), t) - c \leq 0$ and $v(\rho(\bar{y}, x_i), s) - c \leq 0$, then

$$\begin{aligned}
&\left| (v(\rho(\bar{x}, x_i), t) - c)^+ - (v(\rho(\bar{y}, x_i), s) - c)^+ \right| \\
&= 0 \leq |v(\rho(\bar{x}, x_i), t) - v(\rho(\bar{y}, x_i), s)|.
\end{aligned} \tag{A5}$$

- If $v(\rho(\bar{x}, x_i), t) \geq c \geq v(\rho(\bar{y}, x_i), s)$, then

$$\begin{aligned}
&\left| (v(\rho(\bar{x}, x_i), t) - c)^+ - (v(\rho(\bar{y}, x_i), s) - c)^+ \right| \\
&= |v(\rho(\bar{x}, x_i), t) - c| \leq |v(\rho(\bar{x}, x_i), t) - v(\rho(\bar{y}, x_i), s)|.
\end{aligned} \tag{A6}$$

- Finally, if $v(\rho(\bar{y}, x_i), s) \geq c \geq v(\rho(\bar{x}, x_i), t)$, then

$$\begin{aligned}
&\left| (v(\rho(\bar{x}, x_i), t) - c)^+ - (v(\rho(\bar{y}, x_i), s) - c)^+ \right| \\
&= |v(\rho(\bar{y}, x_i), s) - c| \leq |v(\rho(\bar{x}, x_i), t) - v(\rho(\bar{y}, x_i), s)|.
\end{aligned} \tag{A7}$$

Thus, the claim holds. □

TABLE B1 Correspondence between the elements of our model and the model in Madarász and Prat (2017)

Model element	Our paper	Madarász and Prat (2017)
True type space	$\chi \times [0,\tau]$	T_v
Set of representative types	$\text{REP}(\pi)$	S_u
Probability distribution of representative types	$\{\phi_i^{\text{REC}} \cdot \phi_j^{\text{REC}}\}_{i,j}$	$g(\cdot)$
Set of feasible alternatives	$\mathcal{Y}$	Y
Buyer's utility function	$v(\rho(\bar{\boldsymbol{x}}, \boldsymbol{x}), t)$	$v(t, y)$
Lipschitz constant for the buyer's utility function	$\kappa_d + \kappa_t$ (from Lemma 1)	K
Approximation index	η	ϵ
Discount factor	α	τ
Menu of item–price pairs	$(\mathcal{M}, \mathcal{P})$	$M = \{y_k, p_k\}_k$
Optimal menu for representative types	ψ^{REP}	$\hat{M}$
Discounted menu	$\psi^{\text{REP-D}}$	$\tilde{M}$
Optimal menu	$\psi^{\text{OPT-MD}}$	M^*
Maximum utility that any buyer can get from consuming all the alternatives	V_{MAX}	1 (normalized)

Therefore,

$$
\begin{aligned}
&|V(\mathcal{S}; \bar{\boldsymbol{x}}, t) - V(\mathcal{S}; \bar{\boldsymbol{y}}, s)| \\
&\leq \sum_{i=1}^{M} |(v(\rho(\bar{\boldsymbol{x}}, \boldsymbol{x}_i), t) - v(\rho(\bar{\boldsymbol{y}}, \boldsymbol{x}_i), s)) \cdot u_i| (\text{using Claim 1}) \\
&\leq \sum_{i=1}^{M} |(v(\rho(\bar{\boldsymbol{x}}, z^*), t) - v(\rho(\bar{\boldsymbol{y}}, z^*), s)) \cdot u_i| (\text{from (A1)}) \\
&\leq |(v(\rho(\bar{\boldsymbol{x}}, z^*), t) - v(\rho(\bar{\boldsymbol{y}}, z^*), s))| \sum_{i=1}^{M} u_i \\
&\leq |v(\rho(\bar{\boldsymbol{x}}, z^*), t) - v(\rho(\bar{\boldsymbol{y}}, z^*), s)| (\text{since } u_i \leq g_i \leq 1 \ \forall i, \text{and} \sum_{i=1}^{M} g_i = 1) \\
&= |v(\rho(\bar{\boldsymbol{x}}, z^*), t) - v(\rho(\bar{\boldsymbol{x}}, z^*), s) + v(\rho(\bar{\boldsymbol{x}}, z^*), s) - v(\rho(\bar{\boldsymbol{y}}, z^*), s)| \\
&\leq |v(\rho(\bar{\boldsymbol{x}}, z^*), t) - v(\rho(\bar{\boldsymbol{x}}, z^*), s)| + |v(\rho(\bar{\boldsymbol{x}}, z^*), s) - v(\rho(\bar{\boldsymbol{y}}, z^*), s)| \\
&\leq \kappa_t |t - s| + \kappa_d |\rho(\bar{\boldsymbol{x}}, z^*) - \rho(\bar{\boldsymbol{y}}, z^*)| \\
&\leq \kappa_t \eta + \kappa_d \rho(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \\
&\leq \kappa_t \eta + \kappa_d \eta.
\end{aligned}
\tag{A8}
$$

This completes the proof of Lemma 1. □

APPENDIX B: CONNECTIONS WITH MADARÁSZ AND PRAT (2017): LEMMA 2 AND THEOREM 1

Proof of Lemma 2 *and Theorem* 1. Lemma 2 is an immediate consequence of Theorem 1 in Madarász and Prat (2017). To see this, in Table B1, we establish a correspondence between the elements of our model and the model in Madarász and Prat (2017).

Using the correspondence described in Table B1 and Theorem 1 in Madarász and Prat (2017), the result follows. Note that Algorithm 1 simply summarizes the steps needed to apply Lemma 2, and Theorem 1 is nothing but a restatement of that lemma in an easy-to-understand language. □