How to Sell a Dataset? Pricing Policies for Data Monetization

Sameer Mehta  
*The University of Illinois at Urbana-Champaign; sameer@illinois.edu*

Milind Dawande, Ganesh Janakiraman, Vijay Mookerjee  
*The University of Texas at Dallas; {milind, ganesh, vijaym}@utdallas.edu*

Forthcoming in *Information Systems Research*

The wide variety of pricing policies used in practice by data-sellers suggests that there are significant challenges in pricing datasets. In this paper, we develop a utility framework that is appropriate for data-buyers and the corresponding pricing of the data by the data-seller. A buyer interested in purchasing a dataset has private valuations in two aspects – her ideal record that she values the most, and the rate at which her valuation for the records in the dataset decays as they differ from her ideal record. The seller allows individual buyers to filter the dataset and select the records that are of interest to them. The multi-dimensional private information of the buyers coupled with the endogenous selection of records makes the seller’s problem of optimally pricing the dataset a challenging one. We formulate a tractable model and successfully exploit its special structure to obtain optimal and near-optimal data-selling mechanisms. Specifically, we provide insights into the conditions under which a commonly-used mechanism – namely, a price-quantity schedule – is optimal for the data-seller. When the conditions leading to the optimality of a price-quantity schedule do not hold, we show that the optimal price-quantity schedule offers an attractive worst-case guarantee relative to an optimal mechanism. Further, we numerically solve for the optimal mechanism and show that the actual performance of two simple and well-known price-quantity schedules – namely, two-part tariff and two-block tariff – is near-optimal. We also quantify the value to the seller from allowing buyers to filter the dataset.

*Key words*: Data Monetization, Multi-Dimensional Mechanism Design, Price-Quantity Schedules

1. Introduction

The use of targeted marketing is a prevalent and growing activity across industries and businesses. There are many ways in which firms execute targeted marketing campaigns – for instance, by outsourcing such campaigns to marketing firms/social-media platforms (Lee et al. 2018), or by acquiring data on their potential customers from data-aggregators and executing the campaigns in-house (ANA 2018). The latter approach, wherein data is transferred from the seller to the buyer, is widely used and is the context in which our work is situated. Major players (known as data brokers) in this data-selling market include Acxiom, Nielsen, Oracle, Teradata, Experian, among others.
In this paper, we develop a utility framework that is appropriate for a data-buyer and the corresponding pricing of the data by the data-seller. To motivate the need for these contributions, we begin by examining several real-world data-selling firms and the pricing policies they use.

- **BookYourData (BYD) — https://www.bookyourdata.com/**

  **Context:** BYD offers ready-made lists of contacts of business individuals across different industries, job titles, job functions, and job levels. Examples include a list of healthcare professionals, a list of CEOs and CFOs, a list of chiefs and VPs of IT firms, a list of computer equipment manufacturers, etc. A record in a list consists of contact information such as name, email address, job function, job department, country, etc. These niche datasets are typically used by marketers who are interested in targeting a specific set of business individuals.

  **Pricing:** BYD provides a variety of filtering options to buyers so that they can select a subset of records of their choice from any given list. For instance, from the list of healthcare professionals, a buyer can use location filters such as country, state, zipcode, specialty filters such as dentists, chiropractors, or website-domain filters such as .com, .gov, .org, to target the healthcare professionals of interest. BYD uses a non-linear price-quantity schedule to price the set of records selected by a buyer (see Figure (1a)). That is, the price for any set of records depends only on the number of records in that set and not on the identity of the records; specifically, the price is an increasing and concave function of the number of records.

- **SalesLead (SL) — http://sales-lead.org/**

  **Context:** SL maintains a variety of datasets of American businesses in the form of profession-based lists and state/province-based lists. For example, the Accountant Sales Leads dataset contains records of US-based accountants, whereas the Alabama Sales Leads dataset contains records of different businesses (accountants, real-estate agents, etc.) based in Alabama. Each record in a dataset consists of contact information such as mailing address, geo-location, email address, phone number, website, Google pagerank, etc. These datasets are available for purchase in the form of buy-all-or-buy-nothing packages.

  **Pricing:** SL uses a simple flat-fee pricing policy. Unlike BYD, the buyers here do not have an option of filtering records of their choice within a dataset; i.e., they can only purchase the entire dataset. Further, although the datasets differ significantly in the number of records they contain, SL often charges the same price for each of them. For example, each of the four datasets shown in Figure (1b) is priced at $49.

- **DirectMail (DM) — https://www.directmail.com/**

  **Context:** Targeted mailing lists are one of the key products that DM offers its clients as part of a marketing solution. The mailing lists include business lists as well as consumer lists such as new
movers and new homeowners. The consumer lists also have information on the lifestyle and interests of the corresponding individuals.

**Pricing:** DM uses a price schedule that is based on the number of records selected and the set of filters used to obtain those records. To illustrate this pricing policy, let us consider the following example: Suppose a buyer is interested in purchasing a new homeowners list. She uses the filters *age*, *gender*, and *homeowners* to select a total of 7,000 records. Then, the price per record is defined by the total quantity (7000) and the three specific filters used by the buyer to arrive at her chosen set of records. The pricing policy is shown in Figure (1c). For our example, the base price per record is equal to $0.045 (calculated from the base schedule) and the filter-based price per record is equal to $0.0035 + $0 + $0.004 = $0.0075 (calculated from the filter-based schedule). Thus, the total price per record is equal to $0.045 + $0.0075 = $0.0525; the buyer pays a total amount of $367.5 for the 7000 records she selected.

- **TelephoneLists (TL) –** [https://www.telephonelists.biz/](https://www.telephonelists.biz/)

**Context:** TL is a telemarketing firm that specializes in offering phone lists as datasets. The dataset consists of information on consumers (contact details, demographics, etc.) as well as businesses (number of employees, sales volume, etc.) in the US and Canada. A key feature of their dataset is the *do-not-call* flag for each record, which is critical information for telemarketers to avoid calling flagged individuals.

**Pricing:** For one-time buyers, TL offers its data based on the desired zipcodes and/or states/provinces. Thus, the extent of filtering here is limited—buyers can select the zipcodes of their choice but cannot filter the records further within those zipcodes. Interestingly, the pricing is based not on the number of records purchased but on the number of zipcodes and/or states/provinces selected by the buyer (see Figure (1d)).

Common across these examples is the structure of the dataset that the firms offer for sale—data is represented in rows and columns; each row is a record, i.e., information about an entity, and, for a given row, each column represents an attribute of that record. The pricing policies in these examples are reasonably simple—BYD offers a price-quantity schedule, SL uses flat-fee pricing, DM employs a price schedule that is based on the quantity as well as the filters used to select the data, and TL charges based on the number of zipcodes/states/provinces. While business needs dictate that pricing should be simple and easily understood, it is not clear if the use of a simple pricing policy results in the seller sacrificing a significant amount of revenue. A natural way to examine this is by obtaining an optimal pricing structure for such datasets, a question that we address.

It is important to clarify that the context we envision in this paper is the selling of publicly available information that the data seller has worked to compile and process. In contrast, there is also
Pricing Policies for Data Monetization

Figure 1 Different pricing policies used by data-sellers.

A stream of literature on market models that expressly consider privacy of data and compensating individuals for the sharing of information; see e.g., Garfinkel et al. (2006), Li et al. (2014), and Cai et al. (2019).

The selling of data is more nuanced than that of information goods like telephone minutes and bandwidth, in the sense that, for a buyer, it is not just the amount of data (i.e., the number of records) that matters but also the “type” of the data. For instance, presented with the same dataset, a buyer interested in targeting those in need for apartment rentals will likely be interested in records that are quite different from those that appeal to a buyer who wants to reach out to healthcare professionals. Thus, the ability to filter the data has an important implication on buyers’ decisions – it allows heterogeneous buyers to endogenously choose the records that interest them.

We will incorporate this notion of filtering in developing an appropriate utility framework for buyers. Also, note that since the amount of data is not the sole criterion for a buyer, it is not immediately clear whether a data-seller should use a pricing policy based only on the amount.

We envision the dataset for sale to consist of many records, one corresponding to each row of the dataset. The attributes of the records form the columns of the dataset. Thus, if there are \( N \) columns
Pricing Policies for Data Monetization

in the dataset, then each record can be viewed as an $N$-dimensional vector. Each buyer has an “ideal record” (which may or may not necessarily be part of the dataset) that she values the most. For instance, in a dataset of healthcare providers, an ideal record for a medical-equipment marketer in Dallas, Texas, might be that of a Cardiologist who has more than 20 years of experience, accepts Medicaid insurance, has a Yelp-rating of at least 4 stars, and is located in Dallas. We employ the notion of a distance metric to measure how close a specific record in the dataset is to a buyer’s ideal record. The ideal record as well as the decay in utility with distance can both differ across buyers. Thus, the buyers are heterogeneous in two aspects – one, their respective ideal records, which are $N$-dimensional vectors, and two, the rate at which their utility for records in the dataset decreases as they move away from their ideal records. Thus, each buyer is endowed with $(N + 1)$-dimensional private information. Each buyer uses her private information and the seller’s pricing policy to purchase the records of her choice. Anticipating the buyers’ decisions, the seller’s goal is to design a pricing policy to maximize his expected revenue.

With this brief overview, we now summarize our main contributions.

1.1. Our Contributions

As is clear from the above discussion, one approach for the buying and selling of data can be based on the formalization of several context-specific characteristics – e.g., the notion of an ideal record for a buyer (which is her private information) and the corresponding heterogeneity across buyers, the decay in a buyer’s utility for a record as a function of its distance from her ideal record (also private to the buyer) and the corresponding heterogeneity across buyers, and the filtering of data to enable the endogenous selection of records. Thus, the formulation of a tractable model to obtain an optimal data-selling mechanism is itself a non-trivial task. Although multi-dimensional mechanism-design problems are well-known to be intractable in general, we successfully exploit the special structure of our model to examine it both analytically and numerically. Here, it is important to clarify that there can be other potential approaches for the buying and selling of data; e.g., negotiations or long-term contracts. Our approach is well-suited due its simplicity, tractability, and scalability.

We show that when the dataset exhibits a special structure, a price-quantity schedule is an optimal data-selling mechanism (Theorem 2 and Corollary 2). Under a price-quantity schedule, the price for a given number of records does not depend on the identity of the records chosen by the buyer. However, as discussed above, the records of interest differ across buyers since each buyer naturally prefers to buy records that are closest to her ideal record. Put differently, for a given cardinality of records, say $q$, different buyers would prefer different sets of $q$ records. Thus, one would expect an optimal pricing policy to specify a price for each potential set of records that a buyer can select. Clearly, it would be impractical to even specify such a set-based pricing schedule
due to its exponential size, let alone compute the optimal pricing policy. Thus, the significance of our result lies in the implication that in the search for an optimal pricing mechanism, the seller can restrict his attention to schedules in which price depends only on the number of records chosen by the buyer and not on the identity of those records.

A natural question then arises: Even when the assumptions that result in the optimality of price-quantity schedules do not hold, can an optimal price-quantity be provably near-optimal? We answer this question in the affirmative by establishing a worst-case (theoretical) bound on the seller’s revenue from an optimal price-quantity schedule with respect to that from an optimal mechanism (Theorem 3). Further, we numerically examine two popular classes of price-quantity schedules: (i) two-part tariffs (specified by a fixed-fee and a price-per-record), and (ii) two-block tariffs (a piece-wise linear function with two different slopes). We obtain the optimal schedule within each of these classes and demonstrate that it achieves an attractive performance relative to the optimal mechanism (Section 6.4 and Section E in the appendix).

We also examine the value that accrues to the seller from allowing the buyers to filter data (Section E.3 in the appendix). For a buyer, the ability to filter the data allows her to endogenously select any subset of records that is of interest to her. In the absence of filtering, the buyer faces a take-it-or-leave-it offer from the seller — either purchase all the records in the dataset at the stated price or buy nothing. Since this mechanism is a feasible solution to the seller’s mechanism-design problem (in which filtering is allowed), it is clear that, under the setting of our analysis, the seller can only benefit by offering the filtering option. This “value of filtering” depends on the fraction of records in the dataset that yield positive utility to the buyer: The lower this fraction, the more important it is for the buyer to use filtering to identify the desired set of records and, thus, higher is the value of filtering.

**Organization of the paper:** We review the literature related to our work in Section 2. Section 3 presents the preliminaries of our model and Section 4 formulates the multi-dimensional mechanism-design problem. In Section 5, we formulate a relaxation of this problem and obtain an optimal mechanism for the relaxed problem. Using the solution of the relaxed problem, we show that, under certain assumptions, a price-quantity schedule is an optimal solution to the multi-dimensional mechanism-design problem. Section 6 analyzes the scenario where these assumptions are not met. Here, we first show that an optimal mechanism may not necessarily belong to the class of price-quantity schedules and then obtain a worst-case performance guarantee offered by an optimal price-quantity schedule. We also develop an approach to numerically evaluate the optimal mechanism and use this as a benchmark to examine the performance of two popular price-quantity schedules as well as to assess the value of filtering to the data seller. Section 7 concludes the paper.
2. Related Literature

Our work is related to the following streams of literature: (i) pricing of information goods, (ii) commodification of data and information, and (iii) multi-dimensional mechanism design. We now review each of these streams.

2.1. Pricing of Information Goods

The classical papers of Mussa and Rosen (1978) and Maskin and Riley (1984) focus on the non-linear pricing of goods. The former work considers the pricing of a single unit of a good and assumes that a buyer’s utility is contingent on her type and the quality of the good purchased. The latter work develops a more-general utility framework, where a buyer’s utility is influenced by her type and the quality as well as the quantity of the good purchased. Spence (1980) develops a utility framework for selling bundles of goods to consumers, and solves the pricing problem for a given set of bundles. In general, the buying and selling of data offers a richer environment than this classical setting in the sense that, even for a dataset of a given quality (e.g., a fixed set of attributes for each record in the dataset), the buyers’ utilities are not only affected by their type and the quantity of records purchased (i.e., how much data?), but also by the specific subset of records purchased (i.e., which records in the data?). Further, in the classical setting, only the type of the buyer is private knowledge, whereas in the data-selling context, the buyer is endowed with multi-dimensional private information – namely, the vector of attributes of her “ideal” record (i.e., one that she values the most) and the rate at which her valuation for a record decays as its “distance” from her ideal record increases. We contribute to this stream by developing an appropriate and tractable utility framework that incorporates these context-specific features.

Two related sub-streams are those that investigate the notion of bundling for pricing information goods (Bakos and Brynjolfsson 1999, Geng et al. 2005, Wu et al. 2008, Wu et al. 2018) and payment mechanisms (Sundararajan 2004, Choudhary 2010, Chen and Huang 2016). In the data-selling context, when the seller allows buyers to filter a dataset, they are able to “bundle” the records of their choice, i.e., select a customized subset of records from that dataset. This is akin to the seller offering all possible bundles of records. Our analysis allows the seller to offer the filtering option to buyers; we also quantify the value of this option to the seller. Since the price for a set of records may depend on the identity of the records, the most general pricing schedule for data is one that lists a separate price for each subset of records. Our analysis begins with this general form and shows how, under certain conditions, the simpler quantity-based pricing mechanism (i.e., price-quantity schedule) is optimal.
2.2. Commoditytion of Data and Information

Agarwal et al. (2019) study the challenges associated with creating a two-sided market for buying and selling data for machine-learning tasks. Bhargava et al. (2020) derive heuristic mechanisms for selling goods (such as sales leads data) for a setting where buyers can either have shared or exclusive access to the dataset; the latter option is more expensive. As will be discussed in Section 3, exclusivity does not play a role in our setting. The fact that the buyers are not concerned about exclusive access to the dataset allows us to focus our analysis on the trade between the seller and one buyer. Besides this contextual difference, our paper also differs from Bhargava et al. (2020) in technical aspects. For instance, our model captures the preferences of the buyers across all the columns in the dataset (by representing a record with \(N\) columns as an \(N\)-dimensional vector) as well as the preferences over different records in the dataset (through the decay parameter \(t\)). On the flip side, Bhargava et al. (2020) model the preferences of a buyer over a single item, which can be viewed either as a single record or the entire dataset. Thus, if one views the entire dataset as a single item, then the model in Bhargava et al. (2020) assumes that the buyer can buy either the entire dataset or nothing at all.

Kushal et al. (2012) analyze a simplistic model for pricing data in which a buyer’s utility does not depend on the identity of the records bought but only on the quantity. Assuming that there is no heterogeneity among data-buyers (i.e., they all have the same willingness-to-pay), the authors analyze two specific models for pricing data, namely unit pricing and step pricing. Muschalle et al. (2012) survey pricing approaches adopted by established data vendors from different market environments. Bergemann et al. (2018) analyze a setting where buyers, endowed with initial private information, seek supplemental data from sellers who provide statistical experiments as information products – i.e., signals that reveal information about the payoff-relevant state – and derive an optimal menu of statistical experiments for the seller. Bimpikis et al. (2019) study the problem of selling information (such as demand forecasts) to competing firms and show that the seller’s strategy is primarily driven by the nature and intensity of competition among the buyers.

Garfinkel et al. (2006) consider a market where the private information of individuals in the form of numeric data is transacted. The authors develop a security mechanism that safeguards private information of the individuals and a market model that compensates individuals for the use of their private information. Li et al. (2014) develop a theoretical framework for selling private data where buyers can purchase noisy queries and the data owners are compensated for the privacy loss that they incur for each query. Cai et al. (2019) propose a framework for trading web browsing histories of individuals by considering their diverse privacy preferences as well as the utility of the end consumers.
2.3. Multi-dimensional Mechanism Design

The seminal work of Myerson (1981) characterizes an optimal mechanism for selling a single unit to buyers who are endowed with single-dimensional private information. While there is no general framework thus far to obtain a similar result for multi-dimensional mechanism design problems, we exploit structural properties of our setting to obtain an optimal mechanism for our multi-dimensional mechanism-design problem, under reasonable conditions. When these conditions do not hold, we obtain a feasible mechanism that offers an attractive performance guarantee. Thus, our work contributes to the literature on approximate mechanisms for multi-dimensional problems; see, e.g., Chawla et al. (2007) and Chawla et al. (2010). Further, to compute an optimal mechanism, we discretize the space of private information and formulate the multi-dimensional 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).

3. Model Preliminaries

We begin by describing the key elements of our model.

- **Dataset:** The focal dataset, denoted by $D$, is assumed to be represented in a tabular form comprised of rows and columns; an illustrative example is shown in Table 1. Each row, which we refer to as a record, consists of information on an entity such as an individual or a firm. The columns of the dataset represent the attributes associated with each record. The columns can be categorized into two sets: (i) filterable columns, consisting of the columns that the buyers use to filter the data (e.g., State, Provider_Type, Has_Ambulance, and Yelp_Rating in Table 1), and (ii) non-filterable columns, typically consisting of columns that provide contact information of the entities (e.g., Name and Phone_No in Table 1). Thus, each record consists of a filterable part and a non-filterable part. Henceforth, we will use the term “columns” to refer to the filterable columns of the dataset. For instance, in the illustrative dataset shown in Table 1, columns 4 to 7 are filterable. Similarly, we will use the term “record” to refer to its filterable part. Note that this nomenclature implies that two records of the dataset can be identical (e.g., records 9 and 10 in Table 1). The dataset in Table 1 consists of 10 records and 4 columns.

We assume that the dataset $D$ consists of $N$ real-valued columns$^1$ and that there is no missing entry. Thus, each record can be viewed as a vector in $\mathbb{R}^N$. Let $\chi_i$ denote the possible range of values of column $i$, $i = 1, 2, \ldots, N$, and let $\chi = \times_{i=1}^{N} \chi_i$ denote the Cartesian product

$^1$The values of categorical attributes such as State, Provider_Type, Is_Ambulance can be easily mapped to real numbers.
<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Phone_No</th>
<th>State</th>
<th>Provider_Type</th>
<th>Has_Ambulance</th>
<th>Yelp_Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Castellanos Fam. Prac.</td>
<td>6269158992</td>
<td>CA</td>
<td>Family Practice</td>
<td>No</td>
<td>4.5</td>
</tr>
<tr>
<td>2</td>
<td>Charles A Leroy</td>
<td>9095921461</td>
<td>CA</td>
<td>Health Clubs</td>
<td>No</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>Edward P Miranda</td>
<td>4153799815</td>
<td>CA</td>
<td>Plastic Surgery</td>
<td>No</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>Fremont Urgent Care</td>
<td>5107961050</td>
<td>CA</td>
<td>Physicians</td>
<td>Yes</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>AFC Exp. Urgent Care</td>
<td>4127815300</td>
<td>PA</td>
<td>Physicians</td>
<td>Yes</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>Ethan Milgrom</td>
<td>2149822264</td>
<td>TX</td>
<td>Physicians</td>
<td>No</td>
<td>4.5</td>
</tr>
<tr>
<td>7</td>
<td>Chantilly Fam. Med.</td>
<td>5713161557</td>
<td>VA</td>
<td>Family Practice</td>
<td>No</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>Regan Fam. Care</td>
<td>6169587326</td>
<td>CA</td>
<td>Family Practice</td>
<td>No</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>Focus MD - Richmond</td>
<td>8048590748</td>
<td>VA</td>
<td>Physicians</td>
<td>No</td>
<td>3.5</td>
</tr>
<tr>
<td>10</td>
<td>Gavin Kole</td>
<td>8146807413</td>
<td>VA</td>
<td>Physicians</td>
<td>No</td>
<td>3.5</td>
</tr>
</tbody>
</table>

Table 1: An illustrative dataset consisting of information on healthcare providers.

of these ranges. Thus, $\chi$ denotes the record-space; i.e., the set of all possible values a record in the dataset can take. Let $g(\cdot)$ denote the density function of records in the dataset $D$. Thus, the number of records in the infinitesimal hypercube $[x,x + dx]$ is equal to $g(x)dx$. A metric $\rho: \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}_+$ measures the distance between any two records; thus, the distance between two records, say $x$ and $y$, is $\rho(x,y)$.

- **Data-Seller**: The monopolistic data-seller’s goal is to design a pricing policy that maximizes his expected revenue. The seller allows each buyer to filter the data through the filterable columns of the dataset and select a set of records to purchase. For each record selected by a buyer, the seller provides the entire record (i.e., the filterable part and the non-filterable part of the record). Note that the same record(s) can be sold to multiple buyers.

- **Data-Buyers**: Buyers are interested in purchasing records of their interest from the dataset. Examples include a furniture retailer interested in purchasing a list of new movers in her area of service and a pharmaceutical marketer interested in targeting physicians in Texas who have over 20 years of experience. We assume that the purchase of a dataset by one buyer does not affect the utilities, and consequently the decisions, of the other buyers. This assumption is satisfied under several reasonable data-selling contexts. For instance, when the dataset offered by the seller is for general-purpose use such as telemarketing or mass-advertising campaigns, or when the business interests of the buyers are non-overlapping, or when buyers operate in different geographies or serve different demographics. Our communication with each of the four data-sellers discussed in Section 1 confirmed that an exclusive access to the records in the dataset is not a concern for their clients. Thus, the purchase decision of one buyer does not affect the utilities and, consequently, the decisions of the other buyers. This assumption allows us to focus our analysis on a single buyer.

Recall from Section 1 the notion of a buyer’s ideal record, namely one she values the most; this record does not necessarily have to be part of the focal dataset. We denote the buyer’s ideal
record by $\bar{x} \in \chi \subseteq \mathbb{R}^N$. The $N$-dimensional vector $\bar{x}$ is private information of the buyer; that is, only the buyer knows the location of $\bar{x}$ in the record-space $\chi$. The seller only has a distributional knowledge of $\bar{x}$; let $F(\cdot)$ and $f(\cdot)$ denote, respectively, the cumulative distribution function and the probability density function of $\bar{x}$. We will interchangeably refer to the buyer’s ideal record as her location type. We assume that the utility of any record $x \in \chi$ to the buyer decreases as its distance $\rho(\bar{x}, x)$ from her ideal record $\bar{x}$ increases. Further, we assume that a parameter $t \in [0, \tau]$, which is (also) private to the buyer, characterizes the rate of this decrease. We refer to $t$ as the buyer’s decay type, of which the seller only has distributional knowledge; let $H(\cdot)$ and $h(\cdot)$ denote, respectively, the c.d.f. and the p.d.f. of the decay type on the support $[0, \tau]$. We assume that the distribution of the decay type is regular; that is, $\frac{h(t)}{1-H(t)}$ is non-decreasing in $t$. In summary, the buyer has private information in two aspects—namely, the location type $\bar{x} \in \chi \subseteq \mathbb{R}^N$ and the decay type, $t \in [0, \tau]$—and is characterized by the $(N + 1)$-dimensional tuple $(\bar{x}, t)$.

Let us denote the utility to a buyer of decay type $t$ from purchasing a record located at a distance $d \geq 0$ from her ideal record by $v(d, t)$. Thus, for a buyer of type $(\bar{x}, t)$, the utility obtained from purchasing a record $\bar{x} \in \chi$ is given by $v(\rho(\bar{x}, \bar{x}), t)$.

As is routine in the mechanism-design literature, we assume certain reasonable properties on a buyer’s utility function $v(d, t), d \geq 0, t \in [0, \tau]$, that help keep the analysis tractable:

P1. Non-negative ($v(d, t) \geq 0$) and $C^2$ (twice continuously differentiable).

P2. Decreasing with distance: $\frac{\partial v(d, t)}{\partial d} \leq 0$ for all $t$.

P3. A higher value of decay type receives a higher utility: $\frac{\partial v(d, t)}{\partial t} \geq 0, v(d, 0) = 0$ for all $d$.

P4. Concave in the decay type (Maskin and Riley 1984): $\frac{\partial^2 v(d, t)}{\partial t^2} \leq 0$ for all $d$.

P5. Non-increasing absolute risk aversion (Maskin and Riley 1984): $\frac{\partial}{\partial t} \left( -\frac{v(d, t)}{v(d, t)} \right) \leq 0$.

Finally, we assume that for every record a buyer purchases, she incurs a (publicly-known) targeting cost$^2$ of $c \geq 0$ per record. The cumulative utility to a buyer of type $(\bar{x}, t)$ from purchasing a set of records$^3$ $S \subseteq D$ can then be written as:

$$V(S; \bar{x}, t) = \int_{S} (v(\rho(\bar{x}, x), t) - c)^+ g(x) dx.$$  \hspace{1cm} (1)

For tractability, we make the following assumptions on a buyer’s cumulative utility $V(S; \bar{x}, t)$:

$^2$ Our underlying assumption is that the buyer’s intended use of the purchased records is to execute targeted marketing campaigns to the individuals/organizations identified in those records. For instance, such a campaign might involve reaching out to these individuals via fliers, phone calls, or e-mails. The parameter $c$ is intended to capture this targeting cost per record. We note that the choice $c = 0$ is allowed in our analysis.

$^3$ We assume that the set of records, $S$, that a buyer can purchase, is Lebesgue-measurable. From a practical viewpoint, this is an innocuous assumption which we impose to make the marginal utility function $v(\cdot, \cdot)$ Lebesgue-integrable.
A1. For any Lebesgue-measurable set $S \subseteq D$, the function $V(S; \bar{x}, t)$ is differentiable and absolutely continuous in $t$.

A2. There exists a positive and finite constant $\Lambda$, such that:

$$|V(D; \bar{x}, t) - V(D; \bar{x}, t')| \leq \Lambda |(t - t')| \quad \forall t, t' \in [0, \tau], \bar{x} \in \chi.$$  

Table 2 below summarizes our main notation.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>The focal dataset that the seller wishes to sell to the buyer.</td>
</tr>
<tr>
<td>$N$</td>
<td>Number of columns in the dataset $D$.</td>
</tr>
<tr>
<td>$\chi \in \mathbb{R}^N$</td>
<td>The record space.</td>
</tr>
<tr>
<td>$\bar{x} \in \chi$</td>
<td>Ideal record of the buyer (private information).</td>
</tr>
<tr>
<td>$t \in [0, \tau]$</td>
<td>Decay rate parameter (private information).</td>
</tr>
<tr>
<td>$\rho$</td>
<td>The distance metric used to measure the distance between any two records in the dataset.</td>
</tr>
<tr>
<td>$v(d, t)$</td>
<td>The utility to the buyer of decay type $t$ from a record that is at distance $d$ from her ideal record.</td>
</tr>
<tr>
<td>$V(S; \bar{x}, t)$</td>
<td>The net utility to the buyer of type $(\bar{x}, t)$ from consuming a set $S \subseteq D$ of records.</td>
</tr>
</tbody>
</table>

We are now ready to formulate the mechanism-design problem for the seller.

4. Problem Formulation

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; that is, mechanisms in which (i) it is optimal for the buyer to reveal her location and decay type truthfully to the seller, and (ii) the buyer obtains a non-negative payoff upon participating.

A direct mechanism $\mu$ is characterized by a pair of functions $(M^\mu, P^\mu)$, where $M^\mu: \chi \times [0, \tau] \rightarrow D$ is an allocation function and $P^\mu: \chi \times [0, \tau] \rightarrow \mathbb{R}$ is a payment function. Thus, in a direct mechanism $\mu$, if the buyer reveals her type as $(\bar{y}, s)$, then she receives the set of records $M^\mu(\bar{y}, s) \subseteq D$ from the dataset, where $M^\mu(\bar{y}, s)$ is assumed to be a Lebesgue-measurable set, and pays an amount $P^\mu(\bar{y}, s)$ to the seller (corresponding to the set of records $M^\mu(\bar{y}, s)$). For the buyer of type $(\bar{x}, t)$, the utility she obtains from purchasing the set of records by reporting her type as $(\bar{y}, s)$ is:

$$V(M^\mu(\bar{y}, s); \bar{x}, t) = \int_{M^\mu(\bar{y}, s)} (v(\rho(\bar{x}, x), t) - o^+) g(x) dx.$$  \hspace{1cm} (2)

As is common in the mechanism-design literature, we assume that the buyer has a quasi-linear utility function. Thus, the net utility to the buyer of type $(\bar{x}, t)$ when she reports her type as $(\bar{y}, s)$
is given by $V(\mathcal{M}^{\mu}(\tilde{y}, s); \tilde{x}, t) - P^{\mu}(\tilde{y}, s)$. The incentive-compatibility constraints can now be stated as follows:

$$V(\mathcal{M}^{\mu}(\tilde{x}, t); \tilde{x}, t) - P^{\mu}(\tilde{x}, t) \geq V(\mathcal{M}^{\mu}(\tilde{y}, s); \tilde{x}, t) - P^{\mu}(\tilde{y}, s) \forall (\tilde{x}, t), (\tilde{y}, s) \in \chi \times [0, \tau]. \quad (\text{IC-MD})$$

The (IC-MD) constraints (where MD signifies multi-dimensional) state that, under the mechanism $\mu$, it is optimal for the buyer to reveal her type truthfully. The individual rationality constraints can be stated as follows:

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

The (IR-MD) constraints impose that the buyer receives a non-negative payoff in the Bayesian Nash Equilibrium upon truthfully revealing her type. The mechanism-design problem for the seller can now be formulated as:

$$\max_{\mu} \mathbb{E}_{(x,t)} [P^{\mu}(\tilde{x}, t)] \quad (P^{\text{MD}})$$

s.t. (IC-MD), (IR-MD).

Notice 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). In view of this difficulty, our structural analysis will proceed as follows:

- We will start by defining a relaxation of problem $(P^{\text{MD}})$. This relaxation is obtained by assuming that the location type, $\tilde{x}$, of the buyer is public knowledge and only the decay type, $t$, is private to the buyer. This assumption results in a single-dimensional mechanism-design problem, of which we will obtain an optimal solution (Theorem 1, Section 5).
- Then, in Section 5.1, we will consider the special case in which the records in the dataset are “uniformly” distributed over the record space. More precisely, for an arbitrary buyer with location type $\tilde{x}$ (i.e., her ideal record is $\tilde{x}$), the mass of data that is within a distance $R (\geq 0)$ from $\tilde{x}$ is independent of $\tilde{x}$, for all values of $R$. For this special case, we establish a useful result: the characterization of the optimal mechanism is independent of the location type and depends only on the decay type of the buyer (Theorem 2). This then enables the important conclusion that an optimal mechanism for problem $P^{\text{MD}}$ can be implemented as a price-quantity schedule (Corollary 2).
- Next, in Section 6, we study the case when the uniformity assumption mentioned above does not hold. In Section 6.1, we observe that, in this case, an optimal solution to problem $(P^{\text{MD}})$ may not necessarily belong to the class of price-quantity schedules. Thus, a natural question arises: Does an optimal price-quantity schedule provide a near-optimal solution to problem
In Section 6.2, we answer this in the affirmative by establishing Theorem 3, which proves an attractive performance guarantee offered by an optimal price-quantity schedule. This performance guarantee is a proper generalization of Theorem 2, in the sense that, when the uniformity assumption is satisfied, Theorem 3 reduces to Theorem 2.

- While Theorem 3 establishes a theoretical guarantee on the performance of the optimal price-quantity schedule, we show that its actual performance relative to an optimal mechanism (evaluated numerically) is even better! To this end, we analyze a discrete setting in which the private information of the buyer is defined by discrete distributions. The advantage of discretization is that the optimal mechanism-design problem can now be formulated as a linear program; we do this in Section 6.3. The solution to this linear program allows us to numerically evaluate the performance of pricing policies of our interest; i.e., price-quantity schedules. We analyze two commonly-used price-quantity schedules, viz., two-part tariffs and two-block tariffs, and demonstrate their attractive performance (Section E of the appendix).

With this overview, we now proceed with our analysis.

5. A Relaxation of Problem \( P_{\text{MD}} \)

Consider the problem obtained by assuming that the location type, \( \bar{x} \), of the buyer is public knowledge (i.e., it is known to the data-seller) and only her decay type, \( t \), is private to her. Thus, the seller’s mechanism-designing problem is now single-dimensional. A direct mechanism \( \sigma(\bar{x}) \) for the relaxed problem\(^4\) consists of the following:

- An (Lebesgue-integrable) allocation function, \( M^\sigma : [0, \tau] \rightarrow \mathcal{D} \), that maps the decay type revealed by the buyer to a set of records in the dataset \( \mathcal{D} \). Since the location type \( \bar{x} \) is publicly-known here (and hence is a parameter), we will use the notation \( M^\sigma(\cdot; \bar{x}) \) to denote the allocation function.

- A payment function, \( P^\sigma : [0, \tau] \rightarrow \mathbb{R} \), that specifies the payment to be made by the buyer corresponding to the set of records allocated to him. Consistent with the notation for the allocation function, we will denote the payment function by \( P^\sigma(\cdot; \bar{x}) \).

For the buyer with location type \( \bar{x} \), the incentive compatibility and the individual rationality constraints under the mechanism \( \sigma \) are:

\[
V(M^\sigma(t; \bar{x}); \bar{x}, t) - P^\sigma(t; \bar{x}) \geq V(M^\sigma(s; \bar{x}); \bar{x}, t) - P^\sigma(s; \bar{x}), \quad \forall s, t \in [0, \tau], \quad \text{(IC-sd}(\bar{x})
\]

\[
V(M^\sigma(t; \bar{x}); \bar{x}, t) - P^\sigma(t; \bar{x}) \geq 0 \quad \forall t \in [0, \tau], \quad \text{(IR-sd}(\bar{x})
\]

\(^4\)To avoid cumbersome notation, we will henceforth avoid stating the dependence of the mechanism on \( \bar{x} \) explicitly and simply use \( \sigma \) instead of \( \sigma(\bar{x}) \).
and the optimal mechanism-design problem for the seller is:

\[
\max_{\sigma} \mathbb{E}_t [P^{\sigma}(t; \bar{x})] \quad (P^{\sigma}(\bar{x}))
\]

\[
\text{s.t. } (\text{ic-sd}(\bar{x})), (\text{ir-sd}(\bar{x})).
\]

Unlike a standard single-dimensional mechanism-design problem in which the seller allocates a single object, our allocation function \( M^\sigma \) needs to allocate a set of records from the dataset to the buyer. Therefore, to apply the steps of the Myersonian approach (Myerson 1981) for solving problem \( P^{\text{sd}}(\bar{x}) \) (sd short for single-dimensional), we need several additional results; see Claims 1 – 5 in Appendix A, which presents the complete derivation of the optimal solution to problem \( P^{\text{sd}}(\bar{x}) \).

Here, we outline the key steps involved in deriving the optimal solution.

- **(Payment in terms of allocation)** Using the constraints (ic-sd(\( \bar{x} \))) and (ir-sd(\( \bar{x} \))) along with the Envelope Theorem (Milgrom and Segal 2002, Corollary 1), we obtain the following relationship between the payment and allocation functions of a mechanism:

\[
P^{\sigma}(t; \bar{x}) = V(M^{\sigma}(t; \bar{x}); \bar{x}, t) - \int_0^t V_i(M^{\sigma}(s; \bar{x}); \bar{x}, s) ds,
\]

where

\[
V_i(S; \bar{x}, t) := \frac{\partial V(S; \bar{x}, t)}{\partial t}.
\]

- **(Problem Simplification)** The relationship between the allocation and the payment function in (3) enables us to formulate the (point-wise) revenue-maximization problem \( P^{\text{sd}}(\bar{x}) \) of the data-seller solely in terms of the allocation function:

\[
\max_{M^{\sigma}(t; \bar{x}) \subseteq \{x \in \prod_i \rho_i(\bar{x}, x, t), c \geq 0\}} \int \left( v(\rho(\bar{x}, x), t) - c - \frac{\partial v(\rho(\bar{x}, x), t)}{\partial t} \left( \frac{1 - H(t)}{h(t)} \right) \right) g(x) dx
\]

\[
\text{s.t. } (\text{ic-sd}(\bar{x})), (\text{ir-sd}(\bar{x})).
\]

- **(Optimizing over sets)** Notice that the above formulation is a set-optimization problem: For every decay type \( t \) that the buyer reveals, an optimal mechanism \( \sigma \) allocates a set of records, \( M^{\sigma}(t; \bar{x}) \), to the buyer in a way that maximizes the data-seller’s expected revenue. Let

\[
w(d, t) := v(d, t) - c - \frac{\partial v(d, t)}{\partial t} \left( \frac{1 - H(t)}{h(t)} \right).
\]

The function \( w(\cdot, \cdot) \) can be viewed as the buyer’s “virtual value” function, a well-known concept in the mechanism-design literature. The integrand of the objective function can then be written as \( w(\rho(\bar{x}, x), t) g(x) \) and the revenue-maximization problem of the data-seller becomes

\[
\max_{M^{\sigma}(t; \bar{x}) \subseteq \{x \in \prod_i \rho_i(\bar{x}, x, t), c \geq 0\}} \int w(\rho(\bar{x}, x), t) g(x) dx
\]

\[
\text{s.t. } (\text{ic-sd}(\bar{x})), (\text{ir-sd}(\bar{x})).
\]
Define $r : [0, \tau] \to \mathbb{R}_+$ as:

$$r(t) = \begin{cases} \max \{d : w(d, t) \geq 0\} & \text{if } \exists d \text{ s.t. } w(d, t) \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The following result states the solution to optimization problem above.

**Theorem 1.** For a given location type $\bar{x}$, the mechanism $\text{opt-sd}$ defined by:

$$M^{\text{opt-sd}}(t; \bar{x}) = \{x \in D : \rho(\bar{x}, x) \leq r(t)\},$$

$$P^{\text{opt-sd}}(t; \bar{x}) = V(M^{\text{opt-sd}}(t; \bar{x}); \bar{x}, t) - \int_0^t V(M^{\text{opt-sd}}(s; \bar{x}); \bar{x}, s) ds,$$

is an optimal solution to problem $P^{\text{sd}}(\bar{x})$.

In words, Theorem 1 states that for a buyer of type $(\bar{x}, t)$, it is optimal to allocate her all the records in the dataset that are within a distance $r(t)$ from her ideal record $\bar{x}$ and charge a price that is equal to the benefit that the buyer enjoys from those records less the information rent$^5$.

Recall that the density of the records in the dataset is given by $g(x)$. Thus, if the data-buyer purchases a set of records, say $S$, from the dataset, then the *quantity of records* purchased by that buyer is simply $\int_S g(x) dx$. Under the opt-sd mechanism defined above, the quantity of records, which we denote by $Q^{\text{opt-sd}}(t; \bar{x})$, purchased by a buyer of type $(\bar{x}, t)$ is given by $\int_{M^{\text{opt-sd}}(t; \bar{x})} g(x) dx$.

This correspondence between the allocation set $M^{\text{opt-sd}}$ and the quantity of records $Q^{\text{opt-sd}}$ enables us to characterize the opt-sd mechanism by the pair of functions $(Q^{\text{opt-sd}}, P^{\text{opt-sd}})$.

Note that buyers with different decay types could purchase the same quantity of data; that is, for a given quantity $q$ and location type $\bar{x}$, there may exist two (or more) decay types, say $t_1, t_2; t_1 \neq t_2$, such that $Q^{\text{opt-sd}}(t_1; \bar{x}) = Q^{\text{opt-sd}}(t_2; \bar{x})$. Let $\Gamma^{\text{sd}}(q; \bar{x}) := \{t : Q^{\text{opt-sd}}(t; \bar{x}) = q\}$ denote the set of decay types of the buyers with location type $\bar{x}$ who consume a quantity $q$ under the opt-sd mechanism. If $\Gamma^{\text{sd}}(q; \bar{x}) \neq \emptyset$, then let $t^{\text{sd}}(q; \bar{x}) := \sup \{t : t \in \Gamma^{\text{sd}}(q; \bar{x})\}$. To generate a price-quantity schedule, let us set the price for any buyer with location type $\bar{x}$ and decay type in $\Gamma^{\text{sd}}(q; \bar{x})$ (thus, the buyer buys quantity $q$ under the opt-sd mechanism) to the price for the buyer with location type $\bar{x}$ and decay type $t^{\text{sd}}(q; \bar{x})$. Then, it follows immediately that, presented with the price-quantity schedule defined in (7) below, each buyer will self-select the quantity that she is allocated under the opt-sd mechanism.

$^5$In the mechanism-design literature, the surplus accrued to an agent by the virtue of the fact that the agent is the only one in the system who knows his/her private valuation (of the object under consideration) is referred to as information rent (see e.g., Krishna 2009).
Corollary 1. For a given location type \( \bar{x} \), the \( \text{opt-sd} \) mechanism can be implemented as a price-quantity schedule:

\[
P_{\text{sd-pq}}(q; \bar{x}) = \begin{cases} 
P_{\text{opt-sd}}(t^{\text{sd}}(q); \bar{x}) & \text{if } \Gamma^{\text{sd}}(q; \bar{x}) \neq \emptyset \\
\infty & \text{otherwise}.
\end{cases}
\]

(7)

Note that under the \( \text{opt-sd} \) mechanism, the set of records allocated to the buyer, \( M^{\text{opt-sd}}(t; \bar{x}) \), and consequently the quantity of records, \( Q^{\text{opt-sd}}(t; \bar{x}) \), depends on both the location type \( \bar{x} \) and the decay type \( t \) of the buyer. Accordingly, Corollary 1 specifies the price-quantity schedule \( P_{\text{sd-pq}} \) for a given location type \( \bar{x} \).

We now examine the analysis above under a special case where the dataset \( D \) is “uniform” (defined precisely below). In this case, we show that (i) the price-quantity schedule becomes independent of the location type \( \bar{x} \), and (ii) the relaxation \( P^{\text{sd}}(\bar{x}) \) can be used to derive an optimal solution for our original problem, namely \( P^{\text{md}} \). Together, these two results enable us to obtain an optimal solution to \( P^{\text{md}} \) that can be conveniently implemented as a price-quantity schedule (which is independent of the location type).

5.1. Analysis of Uniform Datasets

Let \( \bar{R}(\tau) := \max\{d : v(d, \tau) - c \geq 0\} \). Recall that \( \frac{\partial v(d, \tau)}{\partial d} \leq 0 \) (Property P2) and \( \frac{\partial v(d, \tau)}{\partial \tau} \geq 0 \) (Property P3). Thus, for any buyer of type \( t \in [0, \tau] \), the maximum distance from her ideal record that she traverses to select the records of her choice is \( \bar{R}(\tau) \).

Definition (Uniform Dataset): A dataset \( D \) is said to be uniform if the density of records, \( g(x) \), is such that for any two ideal records \( \bar{x}, \bar{y} \in \chi \), the following property holds: For all \( R, 0 \leq R \leq \bar{R}(\tau) \), we have

\[
\int_{\{x \in D : \rho(\bar{x}, x) \leq R\}} g(x) \, dx = \int_{\{x \in D : \rho(\bar{y}, x) \leq R\}} g(x) \, dx.
\]

(8)

In words, in a uniform dataset, for any two buyers with types \((\bar{x}, t)\) and \((\bar{y}, t)\), and for any value of \( R \leq \bar{R}(\tau) \), the quantity of data that lies within a distance of \( R \) from their respective ideal records is equal. We will refer to a dataset that is not uniform as a non-uniform dataset.

Example 1: Table 3a shows a dataset that consists of 6 records and 2 columns. Table 3b shows a numerical representation of the dataset obtained by mapping the categorical values in each of the three columns to real numbers. There are only two potential buyers interested in purchasing this dataset: A and B. The ideal record for buyer A is “Female–High”, i.e., record 3 in the dataset. Similarly, the ideal record for buyer B is “Male–Unemployed”, i.e., record 4 in the dataset. The distance between any two records in the dataset is measured using the Manhattan metric (rectilinear distance or the \( l^1 \) norm). For example, the distance between record 2 (0–2) and record 4 (1–0) is
Pricing Policies for Data Monetization

The utility function $v(\cdot, \cdot)$ is defined as $v(d, t) = t \cdot (4 - d)$ and the targeting cost, $c$, is set to 0.5. Observe that $v(\cdot, \cdot)$ satisfies properties P1–P5 stated in Section 3. Let $\tau = 2$; thus, the decay type of both the buyers can take values in $[0, 2]$. Note that (i) $\bar{R}(\tau) = 3.75 < 4$, and (ii) the distance between any two records in our dataset is an integer. Thus, to show that the uniformity condition (equality (8)) is satisfied for all distances $R \leq 3.75$, it suffices to check that equal number of records lie within integer value distances of 0, 1, 2, and 3 from the respective ideal records of A and B. Table 4 verifies this condition.

<table>
<thead>
<tr>
<th>Record ID</th>
<th>Gender</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Female</td>
<td>Unemployed</td>
</tr>
<tr>
<td>2</td>
<td>Female</td>
<td>Middle</td>
</tr>
<tr>
<td>3</td>
<td>Female</td>
<td>High</td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
<td>Unemployed</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>Low</td>
</tr>
<tr>
<td>6</td>
<td>Male</td>
<td>High</td>
</tr>
</tbody>
</table>

(a) A uniform dataset

<table>
<thead>
<tr>
<th>Record ID</th>
<th>Gender</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

(b) Numerical representation of the dataset

Table 3 An Illustrative example of a uniform dataset.

<table>
<thead>
<tr>
<th>Distance (R)</th>
<th>Number of records that lie within a distance (R) from the ideal record of A</th>
<th>Number of records that lie within a distance (R) from the ideal record of B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1 (Record 3)</td>
<td>1 (Record 4)</td>
</tr>
<tr>
<td>1</td>
<td>3 (Records 2, 3, 6)</td>
<td>3 (Records 1, 4, 5)</td>
</tr>
<tr>
<td>2</td>
<td>3 (Records 2, 3, 6)</td>
<td>3 (Records 1, 4, 5)</td>
</tr>
<tr>
<td>3</td>
<td>5 (Records 1, 2, 3, 5, 6)</td>
<td>5 (Record 1, 2, 4, 5, 6)</td>
</tr>
</tbody>
</table>

Table 4 Verification of uniformity condition.

Proposition 1 below shows that the OPT-SD mechanism exhibits a special structure when $\mathcal{D}$ is a uniform dataset.

**Proposition 1.** If $\mathcal{D}$ is a uniform dataset, then the quantity of data purchased by a data-buyer of type $(\bar{x}, t)$, and the corresponding price that she pays, under the OPT-SD mechanism is independent of her location type $\bar{x}$. Mathematically, if $\mathcal{D}$ is a uniform dataset, then for all $t \in [0, \tau]$ and $\bar{x}, \bar{y} \in \chi$,

$$Q^{\text{OPT-SD}}(t; \bar{x}) = Q^{\text{OPT-SD}}(t; \bar{y}),$$

and

$$\mathcal{P}^{\text{OPT-SD}}(t; \bar{x}) = \mathcal{P}^{\text{OPT-SD}}(t; \bar{y}).$$

This simplification in the structure of the OPT-SD mechanism helps us obtain an optimal solution for the multi-dimensional mechanism-design problem $\mathcal{P}^{\text{SD}}$. The following result states the solution:
Theorem 2. If $\mathcal{D}$ is a uniform dataset, then the mechanism $\text{OPT-MD}$ defined by:

$$
\mathcal{M}^{\text{OPT-MD}}(\bar{x}, t) = \{x \in \mathcal{D} : \rho(\bar{x}, x) \leq r(t)\}, \quad \text{and}
$$

$$
P^{\text{OPT-MD}}(\bar{x}, t) = V(\mathcal{M}^{\text{OPT-MD}}(\bar{x}, t); \bar{x}, t) - \int_0^t V_t(\mathcal{M}^{\text{OPT-MD}}(\bar{x}, s); \bar{x}, s) ds
$$

is an optimal solution to problem $P^{\text{MD}}$. Moreover, for all $t \in [0, T]$ and $\bar{x}, \bar{y} \in \chi$,

$$
Q^{\text{OPT-MD}}(\bar{x}, t) = Q^{\text{OPT-MD}}(\bar{y}, t), \quad \text{and}
$$

$$
P^{\text{OPT-MD}}(\bar{x}, t) = P^{\text{OPT-MD}}(\bar{y}, t),
$$

where $Q^{\text{OPT-MD}}(\bar{x}, t) = \int_{\mathcal{M}^{\text{OPT-MD}}(\bar{x}, t)} g(x) dx$.

Theorem 2 states when $\mathcal{D}$ is a uniform dataset, a buyer of decay type $t$ purchases the set of records that are within a distance $r(t)$ from her ideal record. Furthermore, the quantity of records purchased by the buyer and the corresponding price are both independent of her location type. Thus, $\forall x \in \chi$, we have

$$
Q^{\text{OPT-MD}}(t) := Q^{\text{OPT-MD}}(\bar{x}, t), \quad \text{and}
$$

$$
P^{\text{OPT-MD}}(t) := P^{\text{OPT-MD}}(\bar{x}, t).
$$

Let $\Gamma^{\text{MD}}(q) := \{t : Q^{\text{OPT-MD}}(t) = q\}$ denote the set of decay types of the buyers who consume a quantity $q$ under the $\text{OPT-MD}$ mechanism. If $\Gamma^{\text{MD}}(q) \neq \emptyset$, then let $t^{\text{MD}}(q) := \sup\{t : t \in \Gamma^{\text{MD}}(q)\}$.

Corollary 2. If $\mathcal{D}$ is a uniform dataset, then the optimal mechanism $\text{OPT-MD}$ can be implemented as the following price-quantity schedule

$$
P^{\text{MD-PQ}}(q) = \begin{cases} 
P^{\text{OPT-MD}}(t^{\text{MD}}(q)) & \text{if } \Gamma^{\text{MD}}(q) \neq \emptyset \\ \emptyset & \text{otherwise.} \end{cases}
$$

We will now turn our attention to the analysis of non-uniform datasets; that is, datasets for which condition (8) is not satisfied. The optimality of price-quantity schedules for uniform datasets raises several natural questions: First, are price-quantity schedules optimal for non-uniform datasets as well? Second, if not, then what performance guarantee does an optimal price-quantity schedule offer for such datasets? We examine these questions in the next section.

6. Analysis for Non-Uniform Datasets

We start this section by demonstrating that, for a non-uniform dataset, no optimal mechanism to problem $P^{\text{MD}}$ may belong to the class of price-quantity schedules; that is, each price-quantity schedule may be sub-optimal. Nevertheless, Theorem 3 establishes the usefulness of price-quantity schedules for non-uniform datasets by establishing an attractive performance guarantee offered by an optimal price-quantity schedule. This result is a natural generalization of Theorem 2 in the sense that, for a uniform dataset, Theorem 3 reduces to Theorem 2.
6.1. Sub-optimality of price-quantity schedules for non-uniform datasets

For a non-uniform dataset $D$, the following simple but illustrative example demonstrates that an optimal price-quantity schedule is sub-optimal for problem $P^{MD}$.

Example 2: As shown in Figure 2, the records in the dataset are located at four positions A, B, C, and D, on the real number line. Recall from Section 3 that the number of records is normalized to 1. The mass of records at location A is 0.4 and at each of locations B, C, and D, is 0.2. The distance between any two records is the Euclidean distance between the positions of those records on the number line. Thus, $\rho(A, C) = 2, \rho(A, D) = 3, \rho(B, C) = 1, \rho(A, A) = 0$, etc. The distributional information that the data-seller possesses about the location type (i.e., the ideal record) of the data-buyer is as follows: The ideal record of the data-buyer is located at position A with probability 0.5 and at position C with probability 0.5. The decay type of the data-buyer is publicly known and is deterministically equal to 1. The utility function $v(\cdot, \cdot)$ is defined as $v(d, t) = t \cdot (2 - d)^+ = (2 - d)^+$. Observe that $v(\cdot, \cdot)$ satisfies properties P1–P5 stated in Section 3. Thus, a buyer whose ideal record is located at A receives a utility of 2 (per unit record) from purchasing records located at A, 1 from purchasing records located at B, and 0 from records located at C and D. This fall in the utility that the data-buyer derives from purchasing records that are located farther from her ideal record is depicted by the red-solid line in Figure 2. Similarly, a buyer whose ideal record is located at C receives a utility of 2 from purchasing records located at C, 1 from purchasing records located at B and D each, and 0 from records located at A (depicted by the blue-dotted line in Figure 2). The targeting cost $c$ is equal to 0.

Consider a buyer whose ideal record is located at A. This buyer’s valuation for the records located at positions C and D is 0. Thus, the total mass of records for which this buyer has a positive valuation is $0.4 \times 0.2 + 0.2 \times 1 = 1$. Similarly, consider a buyer whose ideal record is located at C. This buyer’s valuation

![Figure 2: An illustrative example](https://ssrn.com/abstract=3333296)
Pricing Policies for Data Monetization

for records located at position A is 0. The total mass of records for which this buyer has a positive valuation is 0.2 (mass of the records located at C) + 0.2 (mass of the records located at B) + 0.2 (mass of the records located at D) = 0.6. Thus, corresponding to a total quantity of 0.6 records, this buyer would be willing to pay $0.2 \times 2 + 0.2 \times 1 + 0.2 \times 1 = 0.8$. In summary, both the buyers are willing to purchase the same quantity of data (albeit, different sets of records) but have different willingness to pay for that same quantity. Therefore, a price-quantity schedule (which specifies one price for a given quantity of data) cannot fully exploit the heterogeneity in the valuations of the buyers for the same quantity of data.

Consider, instead, the following pricing policy: The data-seller offers two bundles of records – (i) records at positions A and B are bundled together and offered for a price of 1, and (ii) records at positions B, C, and D are bundled together and offered for a price of 0.8. Under this pricing policy, buyers whose ideal record is at position A purchase the first bundle for a price of 1 and the buyers whose ideal record is at position C purchase the second bundle for a price of 0.8. This pricing policy yields a strictly higher revenue than any price-quantity schedule.

6.2. Performance of Price-Quantity Schedules for Non-Uniform Datasets

While the above example shows that an optimal price-quantity schedule is not, in general, optimal for $P_{md}$, can it be provably near-optimal? We answer this question in the affirmative by establishing a worst-case bound on the performance of an optimal price-quantity schedule. Before proceeding, we state a few simplifying assumptions and introduce additional notation needed for our analysis.

Assumption 1: For every data-buyer of type $$(\bar{x}, t)$$, there exists a record $x$ such that $v(\rho(\bar{x}, x), t) - \epsilon \leq 0$.

This assumption implies that not all the records in the dataset are useful (i.e., generate a positive utility) to the buyer. For instance, for any given buyer, a typical real-world dataset will contain records that are not of any interest to that buyer. We make this technical assumption solely to simplify the boundary value of the virtual value function, $w(\cdot, \cdot)$, under the $\text{OPT-SD}$ mechanism; more precisely, under this assumption, $w(r(t), t) = 0 \forall t$.

Recall from Section 5 that $\Gamma_{s\bar{d}}(q; \bar{x})$ denotes the set of decay types of the buyers with location type $\bar{x}$ who consume a quantity $q$ under the $\text{OPT-SD}$ mechanism. Further, recall that if $\Gamma_{s\bar{d}}(q; \bar{x}) \neq \emptyset$, then $t^{s\bar{d}}(q; \bar{x}) = \sup \{t : t \in \Gamma_{s\bar{d}}(q; \bar{x}) \}$. Let

$$\epsilon_t := \sup_{q, \bar{x}, \bar{y}} |t^{s\bar{d}}(q; \bar{x}) - t^{s\bar{d}}(q; \bar{y})|.$$  \hspace{1cm} (14)

Let $\delta(q; \bar{x})$ denote the minimum distance that a buyer, whose ideal record is located at $\bar{x}$, traverses to accumulate a quantity $q$ of records in the dataset. Formally,

$$\delta(q; \bar{x}) = \min d, \text{ subject to } \int_{\{x \in D : \rho(\bar{x}, x) \leq d\}} g(x) \, dx = q.$$
For $\delta(q;\bar{x})$, we assume the following:

**Assumption 2:** There exist a finite constant $\epsilon_\delta$ such that

$$\max_{0 \leq q \leq 1} \left| \delta(q;\bar{x}) - \delta(q;\bar{y}) \right| \leq \epsilon_\delta \forall \bar{x}, \bar{y} \in \chi.$$  \hfill (15)

In words, this assumption imposes that for any two buyers who differ in their location type, the maximum difference in the distances these buyers traverse from their respective ideal records to consume the same quantity of records is bounded. For instance, a well-designed dataset of healthcare professionals in a metropolitan area will typically include records that are evenly spread out over that area and is therefore likely to have a small value of $\epsilon_\delta$.

The constants $\epsilon_\delta$ and $\epsilon_t$ capture the extent of non-uniformity in the dataset. Note here that, for uniform datasets, we have $\epsilon_\delta = \epsilon_t = 0$.

For a given distance $d$ and decay type $t$, let

$$L(d,t) := \frac{\partial v(d,t)}{\partial t} \cdot \frac{1 - H(t)}{h(t)}.$$  \hfill (16)

We assume the following for the function $L(d,t)$:

**Assumption 3:** There exist finite constants $M_\delta, M_t \geq 0$ such that

$$\left| \frac{\partial L(d,t)}{\partial d} \right| \leq M_\delta \forall t, \text{ and}$$

$$\left| \frac{\partial L(d,t)}{\partial t} \right| \leq M_t \forall d.$$  \hfill (17)

(18)

The function $L(d,t)$ can be explained as follows. Recall from (4) that the virtual value, $v(d,t)$, of a buyer of decay type $t$ for a record(s) located at a distance $d$ from her ideal record, is the valuation, $v(d,t) - c$ minus the term $\frac{\partial v(d,t)}{\partial t} \cdot \frac{1 - H(t)}{h(t)}$. This latter term is denoted by $L(d,t)$ in (16). The expectation of $L(d,t)$ with respect to the decay type $t$ is the information rent to the buyer under the opt-sd mechanism. Assumption 3 states that the rate of change of the function $L(d,t)$ with respect to its arguments is bounded. Clearly, this assumption holds for “well-behaved” functions whose values change smoothly.

Let us now formally define the optimal revenue that the data-seller can obtain from making a take-it-or-leave-it offer for purchasing the entire dataset. Using the notation defined in Section 4, the benefit from consuming the entire dataset $\mathcal{D}$ for a buyer of type $(\bar{x},t)$ is $V(\mathcal{D};\bar{x},t)$. Let $F^\mathcal{D}$ denote the distribution of $V(\mathcal{D};\bar{x},t)$ and let $p$ denote the price at which the seller decides to sell the dataset $\mathcal{D}$. Let $\text{REV}^{\text{opt-full}}$ denote the optimal revenue to the data-seller obtained by selling the entire dataset. That is,

$$\text{REV}^{\text{opt-full}} = \max_p p(1 - F^\mathcal{D}(p)).$$  \hfill (19)
Let \( \text{REV}^{\text{opt-pq}} \) denote the revenue of the data-seller from using an optimal price-quantity schedule. Since a take-it-or-leave-it pricing policy belongs to the class of price-quantity schedules, it follows that \( \text{REV}^{\text{opt-pq}} \geq \text{REV}^{\text{opt-full}} \). Finally, let \( \text{REV}^{\text{opt}} \) denote the revenue of the seller under an optimal mechanism (i.e., an optimal solution to our original mechanism-design problem \( P^{\text{md}} \)). Recall that \( \epsilon_t \) and \( \epsilon_\delta \) are as defined in (14) and (15), respectively.

**Theorem 3.** For any dataset \( D \), the worst-case performance guarantee offered by an optimal price-quantity schedule, relative to an optimal mechanism, is:

\[
\frac{\text{REV}^{\text{opt}}}{\text{REV}^{\text{opt-pq}}} \leq 1 + \frac{\epsilon_\delta M_\delta + \epsilon_t M_t}{\text{REV}^{\text{opt-full}}}.
\]

(20)

Theorem 3 provides us a way to assess the performance of an optimal price-quantity schedule (relative to an optimal mechanism, over all mechanisms) based on the extent of non-uniformity in the dataset. To see this, note that, when \( D \) is a uniform dataset, we have \( \epsilon_\delta = \epsilon_t = 0 \), and consequently \( \text{REV}^{\text{opt}} = \text{REV}^{\text{opt-pq}} \), which is the result we had established earlier in Theorem 2 (Corollary 2). The second term in the performance guarantee depends on the extent of non-uniformity in the dataset, as measured by the constants \( \epsilon_\delta \) and \( \epsilon_t \).

We now present a brief overview of the non-trivial steps involved in the proof of Theorem 3. Recall from Corollary 1 (Section 5) that the \( \text{OPT-SD} \) mechanism can be implemented as a price-quantity schedule, \( \mathcal{P}^{\text{sd-pq}}(q; \bar{x}) \), that is contingent on the location type of the buyer. Put differently, when the location type, \( \bar{x} \), of a buyer is known, offering her a type-specific price-quantity schedule, \( \mathcal{P}^{\text{sd-pq}}(q; \bar{x}) \), is an optimal mechanism. Using the price-quantity schedules for each \( \bar{x} \), we construct a price-quantity schedule, \( \mathcal{P}^{\text{bound}}(q) \), in which, for any quantity \( q \), the rate of change of price with respect to \( q \) is the lowest such rate, over all location types \( \bar{x} \), of the schedules \( \mathcal{P}^{\text{sd-pq}}(q; \bar{x}) \). The schedule \( \mathcal{P}^{\text{bound}}(q) \) possesses two interesting properties: (i) it does not depend on the location type of the buyers, and (ii) the quantity of records that a buyer of type \( (\bar{x}, t) \) purchases under the schedule \( \mathcal{P}^{\text{bound}}(q) \) is at least as much as that under the schedule \( \mathcal{P}^{\text{sd-pq}}(q; \bar{x}) \). These two properties, along with Assumptions 1, 2, and 3, help us obtain a bound on the ratio of the revenue \( \text{REV}^{\text{opt}} \) from the optimal mechanism to that of the revenue \( \text{REV}^{\text{bound}} \) from the schedule \( \mathcal{P}^{\text{bound}}(q) \). However, computing the revenue \( \text{REV}^{\text{bound}} \) from the schedule \( \mathcal{P}^{\text{bound}}(q) \) is analytically as well as computationally difficult. On the other hand, computing the revenue \( \text{REV}^{\text{opt-full}} \) of the optimal take-it-or-leave-it policy is relatively easy. Therefore, using the fact that the revenue \( \text{REV}^{\text{opt-pq}} \) from the optimal price-quantity schedule is greater than or equal to \( \max\{\text{REV}^{\text{bound}}, \text{REV}^{\text{opt-full}}\} \), we obtain the performance guarantee in Theorem 3.

Next, using an example, we illustrate the strength of the worst-case bound in Theorem 3.

**Example 3:** The records in the dataset are located on the circumference of a unit circle, symmetrically across the diameter \( AB \); see Figure 3. The density (mass) of records at an angle \( \theta \) from
OB, $\theta \in [-\pi, \pi]$, is $\frac{1}{2\pi} + \alpha - \frac{2\alpha}{\pi}$, where $\alpha \in [0, \frac{1}{2\pi}]$ is a parameter. The parameter $\alpha$ captures the extent of non-uniformity of the distribution of records in the dataset: for $\alpha = 0$, the records are uniformly distributed over the circle (i.e., equation (8) is satisfied); as $\alpha$ increases, a larger mass of records is concentrated at point B relative to that at point A. The distance between any two records is the length of the shorter of the two arcs along the circle that joins their locations. The utility function $v(\cdot, \cdot)$ is defined as $v(d, t) := t \cdot (\pi - d)$. Observe that this utility function satisfies properties P1-P5 stated in Section 3. The value of the per-record targeting cost, $c$, takes values from the set $\{0.05, 0.1, 0.2\}$. The location type of the buyers is uniformly distributed over the circumference of the circle and the decay type of the buyers is uniformly distributed in $[0, 1]$. For this setting, we evaluate the bound in Theorem 3 for different values of $\alpha$ in its range $[0, \frac{1}{2\pi}]$; see Figure 4.

For $\alpha = 0$, the records in the dataset are uniformly distributed and the value of the bound is 1; that is, an optimal price-quantity schedule is an optimal mechanism. As the value of $\alpha$ increases, the distribution of the records in the dataset becomes increasingly non-uniform. It is worth noting that even for the highest possible value of $\alpha$, the value of the worst-case bound is below 2. We also note that for a fixed value of $\alpha$, the bound increases with an increase in the targeting cost, $c$. Specifically, as $c$ increases, while the values $\epsilon_b, M_b, e_t$, and $M_t$, remain unaffected, the revenue $\text{Rev}_{\text{opt-full}}$ from selling the entire dataset decreases (since the willingness-to-pay of the buyer decreases). Thus, the worst-case bound in Theorem 3 increases.

While we now have a theoretical (worst-case) assessment of the performance of an optimal price-quantity schedule, a better evaluation of the actual performance of price-quantity schedules can be obtained via a numerical study. To this end, we develop an approach to compute an optimal mechanism for non-uniform datasets and then use this as a benchmark to examine two popular price-quantity schedules, namely two-part tariffs and two-block tariffs.

6.3. Optimal mechanism for non-uniform datasets

We start with the key elements of our setting.
Pricing Policies for Data Monetization

\[ r_{\text{min}} \leq r \leq r_{\text{max}}. \]

**Parameter \( \alpha \):** Extent of non-uniformity in the dataset

### Figure 4

The value of the worst-case performance guarantee (RHS of (20) in Theorem 3) with respect to the extent of non-uniformity in the dataset.

- **Distribution of records:** We assume that the record space \( \chi \) (see Section 3) is discretized; specifically, \( \chi = \{x_1, x_2, \ldots, x_M\} \) for some \( M < \infty \). Each record in the dataset \( D \) takes a value from \( \chi \). As before, we refer to the value \( x_i \) (say) of a record as its "location". Further, as we had done before for convenience, we normalize the number of records in \( D \) to 1. Let \( g_i \) denote the mass of records at location \( x_i \), and \( \sum_{i=1}^{M} g_i = 1 \). Using this notation, the set of records that are feasible for a buyer to purchase can be written as:

\[ \mathcal{Y} = \left\{ (y_1, y_2, \ldots, y_M) : y_i \leq g_i \quad \forall i \in \{1, 2, \ldots, M\} \right\}, \]

where \( y_i \) denotes the mass of record \( x_i \), \( i \in \{1, 2, \ldots, M\} \), purchased by the buyer.

- **Private information:** The location type of the data-buyer (i.e., the data-buyer’s ideal record) \( \bar{x} \in \{x_1, x_2, \ldots, x_M\} \). The seller has distributional knowledge of the buyer’s location type: Let \( f \) denote the probability mass function of the buyer’s location type over \( \{x_1, x_2, \ldots, x_M\} \); \( f_i \) denotes the probability that the location type of the buyer is \( x_i \), \( i \in \{1, 2, \ldots, M\} \). Similar to the record space, we assume that the space of the decay types \( [0, \tau] \) of the buyer is also discretized and takes \( T(< \infty) \) distinct values \( \{t_1, t_2, \ldots, t_T\} \). The seller has distributional knowledge of the buyer’s decay type: Let \( h \) denote the probability mass function of the decay type over the set \( \{t_1, t_2, \ldots, t_T\} \); \( h_j \) denotes the probability that the decay type of the buyer is \( t_j \), \( j \in \{1, 2, \ldots, T\} \).

The cumulative utility to the buyer of type \((x_i, t_j)\), \( i \in \{1, 2, \ldots, M\} \) and \( j \in \{1, 2, \ldots, T\} \), from consuming the set of records \( \mathcal{Y} \) is

\[
\sum_{k=1}^{M} (v(\rho(x_i, x_k), t_j) - c)^+ y_k. \tag{21}
\]
For easy exposition, we define \( v_{ik,j} := v(\rho(x_i, x_k), t_j) \). A direct mechanism for our discretized record and type space consists of:

- An allocation rule \( \mathcal{M} : \{x_1, x_2, \ldots, x_M\} \times \{t_1, t_2, \ldots, t_T\} \rightarrow \mathcal{Y} \) that maps the type of the data-buyer to the set of feasible records that the buyer can purchase, and
- A payment rule \( \mathcal{P} : \{x_1, x_2, \ldots, x_M\} \times \{t_1, t_2, \ldots, t_T\} \rightarrow \mathbb{R} \) that specifies the amount that the buyer pays to the seller for purchasing the set of records specified by \( \mathcal{M} \).

The allocation rule \( \mathcal{M}(x_i, t_j) \) is a vector \( \mathcal{M}_{ij} := (\mathcal{M}_{ij1}, \mathcal{M}_{ij2}, \ldots, \mathcal{M}_{ijM}) \), where the component \( \mathcal{M}_{ijk} \) is the mass of the record \( x_k \) allocated to the buyer of type \( (x_i, t_j) \). The payment \( \mathcal{P}(x_i, t_j) \) is a scalar that can be succinctly denoted by \( \mathcal{P}_{ij} \). The revenue-maximization problem for the data-seller can now be formulated as the following linear program:

\[
\max_{\mathcal{M}_{ij}, \mathcal{P}_{ij}} \sum_{i=1}^{M} \sum_{j=1}^{T} \sum_{t=1}^{T} f_i h_j \mathcal{P}_{ij} \tag{P_{MID-LP}}
\]

subject to:

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

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

\[
0 \leq \mathcal{M}_{ijk} \leq g_k \quad \forall i, j, k \quad \tag{22}
\]

\[
\mathcal{P}_{ij} \geq 0 \quad \forall i, j. \quad \tag{23}
\]

In the above formulation, the objective function represents the expected revenue to the data-seller. (IC-LP) and (IR-LP) are the incentive compatibility and the individual rationality constraints, respectively, for the data-buyer. The constraints in (22) represent feasible allocations to the buyer. Finally, (23) states that the buyer pays a non-negative amount to the seller. The linear program \( \text{(P}_{\text{MID-LP}} \) consists of \( O(M^2T) \) decision variables and \( O(M^2T^2) \) constraints.

Next, to highlight the effectiveness of price-quantity schedules even for non-uniform datasets, we examine the performance of two widely-used price-quantity schedules: namely, Two-Part Tariffs and Two-Block Tariffs. We explain these two schemes next.

6.4. Popular Price-Quantity Schedules: Two-Part Tariffs and Two-Block Tariffs

A two-part tariff is characterized by two parameters, \((p_f, p_u)\), where \(p_f\) denotes the fixed price and \(p_u\) denotes the per-record price. Precisely, in our context, the price that the data-buyer pays for purchasing a quantity \( q \) of records, denoted by \( P^{tp}(q) \), is:

\[
P^{tp}(q) = \begin{cases} 
0 & \text{if } q = 0, \\
p_f + p_u \cdot q & \text{if } q > 0.
\end{cases}
\]
A two-block tariff is characterized by three parameters, $(\lambda, p_\lambda, p)$. If $\lambda = 0$, then the two-block tariff reduces to a two-part tariff with fixed price $p_\lambda$ and per-record price $p$. If $\lambda > 0$, then the buyer pays a per-record price of $\frac{p_\lambda}{\lambda}$ for purchasing the first $\lambda$ quantity of records and a per-record price of $p$ for purchasing the records in excess of $\lambda$. Thus, the price that a buyer pays for purchasing a quantity of $q$ records, denoted by $P^{tn}(q)$, is:

\[
P^{tn}(q) = \begin{cases} 
0 & \text{if } q = 0, \\
 p_\lambda + p \cdot q & \text{if } \lambda = 0, \\
 \frac{p_\lambda}{\lambda} \cdot q & \text{if } \lambda > 0 \text{ and } q \leq \lambda, \\
 p_\lambda + p \cdot (q - \lambda) & \text{if } \lambda > 0 \text{ and } q \geq \lambda.
\end{cases}
\]

Below, we summarize the results of our numerical evaluation of (i) the performance of the optimal two-part tariff and the optimal two-block tariff developed above, relative to the optimal mechanism, and (ii) the value to the data-seller for providing the option of filtering the dataset to the buyers. The details of this analysis are in Appendix E.

Over a comprehensive test bed of instances, we show that on average, the revenue of both the two-part tariff and the two-block tariff is within 10% of the optimal revenue. We observe that the two-block tariff outperforms the two-part tariff since the former is a proper generalization of the latter. To assess the attractiveness of the two-block tariff over the two-part tariff, we consider three different classes of values of the targeting cost: low, intermediate, and high. The main takeaway of this analysis is that the attractiveness of an optimal two-block tariff over an optimal two-part tariff is accentuated for intermediate values of the targeting cost (as opposed to extreme values). For such values, there is more heterogeneity across buyers in the fraction of records that offers them a positive utility. An optimal two-block tariff better discriminates the heterogeneity in buyers’ willingness-to-pay by charging a steep price-per-record for low-quantity buyers and offering a discount for high-quantity buyers.

The provision of the filtering option (by the data-seller) enables buyers to endogenously select subsets of records that are of interest to them. In the absence of the filtering option, the buyers face a take-it-or-leave-it offer from the seller; that is, they either have to purchase the entire dataset at the stated price or buy nothing. We observe that as the targeting cost increases, the value to the seller from offering the filtering option increases. Note that with an increase in the targeting cost, the utility of each record to the buyer decreases, and, therefore, so does the fraction of records in the dataset that yield positive utility to the buyer. This, coupled with the high heterogeneity in buyers’ valuations for the entire dataset, makes a take-it-or-leave-it offer less attractive to the buyer and thereby increases the seller’s value from offering the filtering option.
Remark 1. (Multiple Ideal Records): Our analysis thus far assumed that the data-buyers are endowed with a single ideal record. It is plausible that, in general, buyers may possess multiple ideal records. In this remark, we explain how our analysis can be extended for the case of multiple ideal records.

Recall from Section 3 that the metric $\rho : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}^+$ is used to evaluate the utility of a record in the dataset to a buyer (by measuring the distance between that record and the ideal record of the buyer). To account for the setting when buyers possess multiple ideal records, we generalize the notion of a metric as follows. For a non-empty set $I \subseteq \chi$ in the record space and a record $x$ in the dataset $D$, define the function $\Omega$ as:

$$\Omega(I, x) := \inf \{\rho(x_i, x) | x_i \in I\}.$$  \hspace{1cm} (24)

This generalization equips us to model the setting where buyers possess multiple ideal records. We now describe the key changes in our model primitives.

Let $I_K = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K\} \subseteq \chi, K \geq 1$, denote the set of ideal records of a buyer and let $t$ denote her decay type. Thus, the buyer has private information in two aspects – namely, the set of ideal records $I_K \subseteq \chi$ and the decay type, $t \in [0, \tau]$ – and is characterized by the tuple $(I_K, t)$. For a buyer of type $(I_K, t)$, the utility obtained from purchasing a record $x \in D$ is now given by $v(\Omega(I_K, x), t)$, and the utility from consuming a set of records $S$ is given by:

$$V(S; I_K, t) = \int_S (v(\Omega(I_K, x), t) - c)^+ g(x)dx.$$ \hspace{1cm} (25)

We note a few important points here:

- Recall that the utility to a buyer of decay type $t$ from purchasing a record located at a distance $d$ from her ideal record was given by $v(d, t)$. In the setting where buyers possess multiple ideal records, the utility function is still given by $v(d, t)$. The only difference in this new setting is that the distance $d$ is measured by the function $\Omega$, whereas in the earlier setting where the buyers possessed a single ideal record, the distance was measured by the metric $\rho$. Note that the only way in which the metric $\rho$ enters our analysis thus far is through the utility function $v(d, t)$, where $d$ is the distance (as measured by the metric $\rho$) between the (single) ideal record of the buyer and a given record.

- The assumptions P1-P5 are related to the structural properties of the function $v(d, t)$ and do not depend on the metric $\rho$ that is used to measure $d$. Therefore, these assumptions remain unchanged if the distance $d$ is measured using the function $\Omega$.

- Assumptions A1 and A2 can be restated as follows:
With these changes in the model primitives, all our results thus far continue to hold for multiple ideal records (with appropriate modifications in the notation).

7. Concluding Remarks
The business of monetizing data has grown significantly over the past decade and has created unprecedented opportunities for firms to market their products, advance their predictive abilities, and target customers with surgical precision. The growth in this business is primarily fuelled by the intelligence that data-driven analytics provides in making critical business decisions. Thus, the economics of data monetization has become a subject of significant practical and theoretical importance. In this study, we addressed two fundamental aspects related to data monetization: (i) the development of a utility framework that is appropriate, parsimonious, and tractable, and (ii) the analysis of optimal and near-optimal pricing mechanisms.

Our study focused on a monopolistic data-seller who is interested in monetizing a dataset. The ongoing explosive growth in the supply and demand of data has led to the emergence of data-selling platforms – firms such as Data & Sons, Dawex, and Lotame cater to both sides of the market: sellers list their datasets on the platform and buyers purchase data from one or more sellers. Here, the pricing of the data can be done by the seller or by the platform on behalf of the seller. Other firms such as BIG and Narrative enable auction-based purchase and sale of data – buyers browse datasets from different sellers and bid their valuations, and sellers manage their transactions by selecting buyers of their choice. Such two-sided settings become richer due to several interesting constraints: On the supply-side, some sellers may want to sell their data only as a monolithic unit (i.e., avoid sale of proper subsets of records) and some may want to restrict sale to a limited number of buyers. On the demand-side, buyers may want exclusive access to data or may have budget and/or minimum-volume constraints. Further, from a mechanism-design perspective, the analysis of two-sided data-selling platforms presents interesting theoretical challenges. For instance, different from the setting we analyzed in this study, one may face a context wherein buyers’ valuation of a dataset depends on the number of other buyers who have also purchased that dataset. In this case, the private information of a buyer is a function that is endogenous to the number of buyers who have
shared access to the dataset. We believe that our work can serve as a foundation for future work on optimal and/or approximate mechanisms for such settings and, more generally, on the design of efficient data-selling platforms.

References


Online Appendix to "How to Sell a Dataset? Pricing Policies for Data Monetization"

Appendix A  Proof of Theorem 1

Recall that in problem $P^\infty(\bar{x})$, the location type $\bar{x}$ of the buyer is publicly-known. Let

$$U^\sigma(s; \bar{x}, t) = V(M^\sigma(s; \bar{x}); \bar{x}, t) - \mathcal{P}^\sigma(s; \bar{x})$$

denote the net utility to the buyer of decay type $t$ when she reveals her type as $s$, and let $U^\sigma(t; \bar{x}, t)$ denote the net utility to that buyer under an incentive-compatible mechanism $\sigma$. The incentive compatibility constraints, (IC-SD($\bar{x}$)), imply that:

$$U^\sigma(\bar{x}, t) = \max_{s \in [0, \tau]} V(M^\sigma(s; \bar{x}); \bar{x}, t) - \mathcal{P}^\sigma(s; \bar{x}).$$

From Assumption A1 (Section 3), we know that for any Lebesgue-measurable set $S \subseteq \mathcal{D}$, the function $V(S; \bar{x}, t) - \mathcal{P}^\sigma(s; \bar{x})$ is absolutely continuous and differentiable in $t$. Let

$$V_i(S; \bar{x}, t) := \frac{\partial V(S; \bar{x}, t)}{\partial \bar{t}}.$$

Next, note that for a fixed set $S \subseteq \mathcal{D}$ and any $t', t \in [0, \tau]$, we have

$$|V(S; \bar{x}, t) - V(S; \bar{x}, t')| = \left|\int_S \left( (v(\rho(\bar{x}, x), t) - c^+) - (v(\rho(\bar{x}, x), t') - c^+) \right) g(x) dx \right|$$

$$\leq \left|\int_{\mathcal{D}} \left( (v(\rho(\bar{x}, x), t) - c^+) - (v(\rho(\bar{x}, x), t') - c^+) \right) g(x) dx \right| \quad \text{(since } \frac{\partial v(d, \bar{t})}{\partial \bar{t}} \geq 0 \text{ (Property P3))}$$

$$= |V(D; \bar{x}, t) - V(D; \bar{x}, t')|$$

$$\leq \Lambda |t - t'| \quad \text{(Assumption A2; Section 3)}.$$

Consequently, $\left|\sup_{S} V_i(S; \bar{x}, t)\right| \leq \Lambda$. This implies that the function $\sup_{S} |V_i(S; \bar{x}, t)|$ is integrable (with respect to $t$) over $[0, \tau]$. Thus, using the envelope theorem (Milgrom and Segal 2002, Corollary 1), we know that the function $U^\sigma(\bar{x}, t)$ is absolutely continuous in $t$ and therefore can be written as:

$$U^\sigma(\bar{x}, t) = U^\sigma(\bar{x}, 0) + \int_0^t V_i(M^\sigma(s; \bar{x}); \bar{x}, s) ds. \quad (26)$$

Using (26), the payment function can be written as:

$$\mathcal{P}^\sigma(t; \bar{x}) = \mathcal{P}^\sigma(0; \bar{x}) - V(M^\sigma(0; \bar{x}); \bar{x}, 0) + V(M^\sigma(t; \bar{x}); \bar{x}, t) - \int_0^t V_i(M^\sigma(s; \bar{x}); \bar{x}, s) ds.$$
Since \( v(d, 0) = 0 \) from Property (P3), we have \( V(M^\sigma(0; \bar{x}); \bar{x}, 0) = 0 \) from (1). Thus, constraint \((\text{IR-sd}(\bar{x}))\) for a buyer of decay type \( t = 0 \) yields \( P^\sigma(0; \bar{x}) = 0 \). The payment function then simplifies to:

\[
P^\sigma(t; \bar{x}) = V(M^\sigma(t; \bar{x}); \bar{x}, t) - \int_0^t V_i(M^\sigma(s; \bar{x}); \bar{x}, s)ds.
\]  

(27)

This relationship between the allocation and payment functions helps us to formulate the optimization problem \( P_m^\sigma(\bar{x}) \) solely in terms of the allocation function.

The expected revenue of the data-seller is:

\[
\mathbb{E}_t [P^\sigma(t; \bar{x})] = \int_0^\tau V(M^\sigma(s; \bar{x}); \bar{x}, s)h(s)ds - \int_0^\tau \int_0^s V_i(M^\sigma(s; \bar{x}); \bar{x}, s)h(t)dsdt
\]

\[
= \int_0^\tau V(M^\sigma(s; \bar{x}); \bar{x}, s)h(s)ds - \int_0^\tau \int_0^s V_i(M^\sigma(s; \bar{x}); \bar{x}, s)h(t)dtds
\]

\[
= \int_0^\tau V(M^\sigma(s; \bar{x}); \bar{x}, s)h(s)ds - \int_0^\tau V_i(M^\sigma(s; \bar{x}); \bar{x}, s)(1 - H(s))ds
\]

\[
= \int_0^\tau \left( V(M^\sigma(t; \bar{x}); \bar{x}, t) - V_i(M^\sigma(t; \bar{x}); \bar{x}, t) \left( \frac{1 - H(t)}{h(t)} \right) \right) h(t)dt.
\]

Therefore, the revenue-maximization problem \( P_m^\sigma(\bar{x}) \) of the data-seller can be written as:

\[
\max_{M^\sigma(t; \bar{x})} \int_0^\tau \left( V(M^\sigma(t; \bar{x}); \bar{x}, t) - V_i(M^\sigma(t; \bar{x}); \bar{x}, t) \left( \frac{1 - H(t)}{h(t)} \right) \right) h(t)dt
\]

s.t. \((1C-\text{sd}(\bar{x})), (\text{IR-sd}(\bar{x}))\).

Consider the following relaxation of problem \( P_m^\sigma(\bar{x}) \) obtained by neglecting the \((1C-\text{sd}(\bar{x}))\) and \((\text{IR-sd}(\bar{x}))\) constraints:

\[
\max_{M^\sigma(t; \bar{x})} \int_0^\tau \left( V(M^\sigma(t; \bar{x}); \bar{x}, t) - V_i(M^\sigma(t; \bar{x}); \bar{x}, t) \left( \frac{1 - H(t)}{h(t)} \right) \right) h(t)dt. \quad (P_{\text{sd-relax}}(\bar{x}))
\]

The above problem can be solved through point-wise maximization; that is, we fix a \( t \) and obtain the optimal allocation rule \( M^\sigma(t; \bar{x}) \) for that \( t \) by solving:

\[
\max_{M^\sigma(t; \bar{x})} V(M^\sigma(t; \bar{x}); \bar{x}, t) - V_i(M^\sigma(t; \bar{x}); \bar{x}, t) \left( \frac{1 - H(t)}{h(t)} \right).
\]

Using the definition of \( V(M^\sigma(t; \bar{x}); \bar{x}, t) \) from (1), we get:

\[
\max_{M^\sigma(t; \bar{x})} \int_{M^\sigma(t; \bar{x})} \left( (v(\rho(\bar{x}, x), t) - c)^+ - \frac{\partial}{\partial t} (v(\rho(\bar{x}, x), t) - c)^+ \right) \left( \frac{1 - H(t)}{h(t)} \right) g(x)dx.
\]
Let \( D^*_\text{pos}(\bar{x}, t) := \{ \mathbf{x} \in D : v(\rho(\bar{x}, \mathbf{x}), t) - c \geq 0 \} \). That is, \( D^*_\text{pos}(\bar{x}, t) \) denotes the set of all records in the dataset for which the buyer of type \((\bar{x}, t)\) has a non-negative valuation. Note that, \( \forall \mathbf{x} \in D \setminus D^*_\text{pos}(\bar{x}, t) \),

\[
(v(\rho(\bar{x}, \mathbf{x}), t) - c)^+ = 0,
\]

\[
\frac{\partial (v(\rho(\bar{x}, \mathbf{x}), t) - c)^+}{\partial t} = 0.
\]

Therefore,

\[
\int_{D^*_\text{pos}(\bar{x}, t)} \left( (v(\rho(\bar{x}, \mathbf{x}), t) - c)^+ - \frac{\partial (v(\rho(\bar{x}, \mathbf{x}), t) - c)^+}{\partial t} \left( \frac{1 - H(t)}{h(t)} \right) \right) g(\mathbf{x}) d\mathbf{x} = 0.
\]

Since \( M^*(t; \bar{x}) = (M^*(t; \bar{x}) \cap D^*_\text{pos}(\bar{x}, t)) \cup (M^*(t; \bar{x}) \cap D \setminus D^*_\text{pos}(\bar{x}, t)) \) and the value of the objective function on the set \( D \setminus D^*_\text{pos}(\bar{x}, t) \) is equal to 0, the problem \( P^{\text{SD-RELAX-Succ}[\bar{x}]} \) can be written as:

\[
\max_{M^*(t; \bar{x}) \subseteq D \setminus D^*_\text{pos}(\bar{x}, t)} \int_{M^*(t; \bar{x})} \left( (v(\rho(\bar{x}, \mathbf{x}), t) - c)^+ - \frac{\partial (v(\rho(\bar{x}, \mathbf{x}), t) - c)^+}{\partial t} \left( \frac{1 - H(t)}{h(t)} \right) \right) g(\mathbf{x}) d\mathbf{x},
\]

which can be further simplified to

\[
\max_{M^*(t; \bar{x}) \subseteq D \setminus D^*_\text{pos}(\bar{x}, t)} \int_{M^*(t; \bar{x})} v(\rho(\bar{x}, \mathbf{x}), t) - c - \frac{\partial v(\rho(\bar{x}, \mathbf{x}), t)}{\partial t} \left( \frac{1 - H(t)}{h(t)} \right) g(\mathbf{x}) d\mathbf{x}.
\]

Define \( w : \mathbb{R}_+ \times [0, T] \to \mathbb{R} \) as \( w(d, t) := v(d, t) - c - \frac{\partial v(d, t)}{\partial t} \left( \frac{1 - H(t)}{h(t)} \right) \). The integrand in the objective function can then be written as \( w(\rho(\bar{x}, \mathbf{x}), t)g(\mathbf{x}) \) and problem \( P^{\text{SD-RELAX-Succ}[\bar{x}]} \) of the data-seller can be succinctly stated as:

\[
\max_{M^*(t; \bar{x}) \subseteq D \setminus D^*_\text{pos}(\bar{x}, t)} \int_{M^*(t; \bar{x})} w(\rho(\bar{x}, \mathbf{x}), t)g(\mathbf{x}) d\mathbf{x}. 
\]

**Claim 1:** The optimal solution to problem \( P^{\text{SD-RELAX-Succ}[\bar{x}]} \) is given by:

\[
M^*(t; \bar{x}) = \{ \mathbf{x} \in D : w(\rho(\bar{x}, \mathbf{x}), t) \geq 0 \}\, . \tag{28}
\]

**Proof:** Note that if \( w(\rho(\bar{x}, \mathbf{x}), t) \geq 0 \) for some \( \mathbf{x} \), then \( v(\rho(\bar{x}, \mathbf{x}), t) - c \geq 0 \implies \mathbf{x} \in D^*_\text{pos}(\bar{x}, t) \). Thus, \( M^*(t; \bar{x}) \subseteq D^*_\text{pos}(\bar{x}, t) \) is a feasible solution to \( P^{\text{SD-RELAX-Succ}[\bar{x}]} \). Consider any arbitrary set \( M^*(t; \bar{x}) \subseteq D \). Then,

\[
\int_{M^*(t; \bar{x})} w(\rho(\bar{x}, \mathbf{x}), t)g(\mathbf{x}) d\mathbf{x}
\]

\[
= \int_{M^*(t; \bar{x}) \cap M^*(t; \bar{x})} w(\rho(\bar{x}, \mathbf{x}), t)g(\mathbf{x}) d\mathbf{x} + \int_{M^*(t; \bar{x}) \cap (D \setminus M^*(t; \bar{x}))} w(\rho(\bar{x}, \mathbf{x}), t)g(\mathbf{x}) d\mathbf{x}
\]

\[
\leq \int_{M^*(t; \bar{x}) \cap (D \setminus M^*(t; \bar{x}))} w(\rho(\bar{x}, \mathbf{x}), t)g(\mathbf{x}) d\mathbf{x} \text{ (since } w(\cdot, t) < 0 \text{ on } D \setminus M^*(t; \bar{x}))
\]

\[
\leq \int_{M^*(t; \bar{x})} w(\rho(\bar{x}, \mathbf{x}), t)g(\mathbf{x}) d\mathbf{x} \text{ (as } w(\cdot, t) \geq 0 \text{ on } M^*(t; \bar{x})).
\]
Therefore, $\mathcal{M}^*(t; \bar{x})$ is an optimal solution to the relaxed problem.

We now show the allocation set $\mathcal{M}^*$ in (28) satisfies a monotonicity property.

**Claim 2:** $\mathcal{M}^*(\hat{t}; \bar{x}) \subseteq \mathcal{M}^*(t; \bar{x}) \ \forall \hat{t} \leq t$, where $\hat{t}, t \in [0, \tau]$.

**Proof:** Observe that

$$w(d, t) = \frac{\partial w(d, t)}{\partial t} = \frac{\partial v(d, t)}{\partial t} - \frac{\partial v(d, t)}{\partial t} \left(1 - \frac{H(t)}{h(t)}\right) - \frac{\partial^2 v(d, t)}{\partial t^2} \left(1 - \frac{H(t)}{h(t)}\right)$$

$$= \frac{\partial^2 v(d, t)}{\partial t^2} \left(1 - \frac{H(t)}{h(t)}\right).$$

Since $t - \frac{1}{2} \frac{H(t)}{h(t)}$ is increasing in $t$ (as we have assumed that the distribution $H(\cdot)$ of the decay type is regular), and $\frac{\partial^2 v(d, t)}{\partial t^2} \leq 0$, we get $\frac{\partial w(d, t)}{\partial t} \geq 0 \ \forall d$.

Consider any $t, \hat{t} \in [0, \tau]$ such that $\hat{t} \leq t$. The claim trivially holds if $\mathcal{M}^*(\hat{t}; \bar{x}) = \emptyset$. Consider the case when $\mathcal{M}^*(\hat{t}; \bar{x}) \neq \emptyset$. Then, for any $y \in \mathcal{M}^*(\hat{t}; \bar{x})$, we have $w(\rho(\bar{x}, y), \hat{t}) \geq 0$. Since $\frac{\partial w(d, t)}{\partial t} \geq 0 \ \forall d$, we get $w(\rho(\bar{x}, y), t) \geq 0$. Therefore $y \in \mathcal{M}^*(t; \bar{x})$. Since $y$ is arbitrary, we get the desired result.

Next, we present a result that will help us write the allocation set $\mathcal{M}^*$ in a compact manner.

**Claim 3:** For any $t \in [0, \tau]$,

$$w(d, t) \geq 0 \implies \frac{\partial w(d, t)}{\partial d} \leq 0.$$

**Proof:** Note that for any $t$,

$$w(d, t) \geq 0 \implies v(d, t) - c \frac{\partial v(d, t)}{\partial t} \left(1 - \frac{H(t)}{h(t)}\right) \geq 0$$

$$\implies v(d, t) \geq \frac{\partial v(d, t)}{\partial t} \left(1 - \frac{H(t)}{h(t)}\right)$$

$$\implies 1 - \frac{\partial v(d, t)}{\partial t} \left(1 - \frac{H(t)}{h(t)}\right) \frac{1}{v(d, t)} \geq 0. \quad (29)$$

Further, from Property P5, we have,

$$\frac{\partial}{\partial t} \left(-\frac{\partial v(d, t)}{\partial d} / v(d, t)\right) \leq 0 \implies \frac{\partial^2 v(d, t)}{\partial d \partial t} \geq \frac{\partial^2 v(d, t)}{\partial d \partial t} \frac{\partial v(d, t)}{\partial d} \frac{\partial v(d, t)}{\partial t}. \quad (30)$$

Therefore,

$$\frac{\partial w(d, t)}{\partial d} = \frac{\partial v(d, t)}{\partial d} - \frac{\partial^2 v(d, t)}{\partial d \partial t} \left(1 - \frac{H(t)}{h(t)}\right)$$

$$\leq \frac{\partial v(d, t)}{\partial d} - \frac{\partial^2 v(d, t)}{\partial d \partial t} \left(1 - \frac{H(t)}{h(t)}\right) \geq \frac{\partial v(d, t)}{\partial d} \left(1 - \frac{\partial v(d, t)}{\partial t} \left(1 - \frac{H(t)}{h(t)}\right) \frac{1}{v(d, t)} \right) \quad \text{(using 30)}$$

$$\leq 0. \quad \text{(using 29)}$$

Thus, the result follows.
We now present an equivalent representation of the allocation set \( \mathcal{M}^* \). Define \( r : [0, \tau] \to \mathbb{R}_+ \) as:

\[
r(t) = \begin{cases} 
\sup \{d : w(d, t) \geq 0\} & \text{if } \exists d \text{ s.t. } w(d, t) \geq 0 \\
0 & \text{otherwise.}
\end{cases}
\]

and let

\[
\mathcal{R}(t; \bar{x}) = \{x \in \mathcal{D} : \rho(\bar{x}, x) \leq r(t)\}.
\]

**Claim 4:** \( \mathcal{M}^*(t; \bar{x}) = \mathcal{R}(t; \bar{x}) \forall \bar{x}, t. \)

**Proof:** Fix \( \bar{x} \) and \( t \). Consider any \( x \in \mathcal{R}(t; \bar{x}) \). Thus, \( \rho(\bar{x}, x) \leq r(t) \) which implies \( w(\rho(\bar{x}, x), t) \geq 0 \). Consequently, \( x \in \mathcal{M}^*(t; \bar{x}) \). Therefore, \( \mathcal{R}(t; \bar{x}) \subseteq \mathcal{M}^*(t; \bar{x}) \forall \bar{x}, t. \)

Next, consider any \( x \in \mathcal{M}^*(t; \bar{x}) \). Thus, \( w(\rho(\bar{x}, x), t) \geq 0 \). Since \( \frac{\bar{w}(d, t)}{\bar{c}} \leq 0 \) (from **Claim 3**), we get \( \rho(\bar{x}, x) \leq r(t) \). Consequently, \( x \in \mathcal{R}(t; \bar{x}) \). Therefore, \( \mathcal{M}^*(t; \bar{x}) \subseteq \mathcal{R}(t; \bar{x}) \forall \bar{x}, t. \)

Thus, **Claim 4** holds.

Using **Claim 4**, the allocation set \( \mathcal{M}^* \) in (28) can equivalently be written as:

\[
\mathcal{M}^*(t; \bar{x}) = \{x \in \mathcal{D} : \rho(\bar{x}, x) \leq r(t)\}.
\]

Next, we present sufficient conditions under which a mechanism \( \sigma \) is incentive compatible and individually rational for the data-buyers.

**Claim 5:** A mechanism \( \sigma \) satisfies the incentive compatibility (IC-SD(\( \bar{x} \))) and individual rationality (IR-SD(\( \bar{x} \))) constraints if

1. \( \mathcal{M}^*(\hat{t}; \bar{x}) \subseteq \mathcal{M}^*(t; \bar{x}) \forall \hat{t} \leq t. \)
2. \( \mathcal{P}^*(t; \bar{x}) = V(\mathcal{M}^*(t; \bar{x}); \bar{x}, \hat{t}) - \int_0^\infty V_i(\mathcal{M}^*(s; \bar{x}); \bar{x}, s) ds. \)

**Proof:** Since \( \frac{\bar{w}(d, t)}{\bar{c}} \geq 0 \) from Property (P3), we have:

\[
\hat{t} \leq t \implies (v(\rho(\bar{x}, x), t) - c)^+ - (v(\rho(\bar{x}, x), \hat{t}) - c)^+ \geq 0.
\]

Thus, for \( \mathcal{M}^*(\hat{t}; \bar{x}) \subseteq \mathcal{M}^*(t; \bar{x}) \) and \( \hat{t} \leq t \), we have:

\[
\int_{\mathcal{M}^*(t; \bar{x})} \left( (v(\rho(\bar{x}, x), t) - c)^+ - (v(\rho(\bar{x}, x), \hat{t}) - c)^+ \right) g(x) dx \\
\geq \int_{\mathcal{M}^*(\hat{t}; \bar{x})} \left( (v(\rho(\bar{x}, x), t) - c)^+ - (v(\rho(\bar{x}, x), \hat{t}) - c)^+ \right) g(x) dx.
\]

\[
\implies \int_{\mathcal{M}^*(t; \bar{x})} (v(\rho(\bar{x}, x), t) - c)^+ g(x) dx - \int_{\mathcal{M}^*(\hat{t}; \bar{x})} (v(\rho(\bar{x}, x), t) - c)^+ g(x) dx \\
\geq \int_{\mathcal{M}^*(t; \bar{x})} (v(\rho(\bar{x}, x), \hat{t}) - c)^+ g(x) dx - \int_{\mathcal{M}^*(\hat{t}; \bar{x})} (v(\rho(\bar{x}, x), \hat{t}) - c)^+ g(x) dx.
\]

\[
\implies V(\mathcal{M}^*(t; \bar{x}); \bar{x}, t) - V(\mathcal{M}^*(\hat{t}; \bar{x}); \bar{x}, \hat{t}) \geq V(\mathcal{M}^*(t; \bar{x}); \bar{x}, \hat{t}) - V(\mathcal{M}^*(\hat{t}; \bar{x}); \bar{x}, \hat{t}).
\]

\[
\implies V(\mathcal{M}^*(t; \bar{x}); \bar{x}, t) - V(\mathcal{M}^*(\hat{t}; \bar{x}); \bar{x}, \hat{t}) \geq V(\mathcal{M}^*(\hat{t}; \bar{x}); \bar{x}, \hat{t}) - V(\mathcal{M}^*(\hat{t}; \bar{x}); \bar{x}, \hat{t}).
\]
Dividing both sides by $t - \hat{t}$ and taking the limit as $\hat{t} \to t$, we get:

$$V_t(\mathcal{M}^\sigma(t; \bar{x}); \bar{x}, t) \geq V_t(\mathcal{M}^\sigma(\hat{t}; \bar{x}); \bar{x}, \hat{t})$$

(32)

To show that a mechanism $\sigma$ is incentive compatible, we need to show that

$$U^\sigma(t; \bar{x}, t) - U^\sigma(\hat{t}; \bar{x}, t) \geq 0 \quad \forall \hat{t}, t \in [0, \tau].$$

Consider any $\hat{t}, t \in [0, \tau]$. We have

$$U^\sigma(t; \bar{x}, t) - U^\sigma(\hat{t}; \bar{x}, t) = V(\mathcal{M}^\sigma(t; \bar{x}); \bar{x}, t) - P^\sigma(t; \bar{x}) - \left(V(\mathcal{M}^\sigma(\hat{t}; \bar{x}); \bar{x}, \hat{t}) - P^\sigma(\hat{t}; \bar{x})\right)$$

$$= \int_0^t V_t(\mathcal{M}^\sigma(s; \bar{x}); \bar{x}, s) ds - \left(V(\mathcal{M}^\sigma(\hat{t}; \bar{x}); \bar{x}, t) - V(\mathcal{M}^\sigma(\hat{t}; \bar{x}); \bar{x}, \hat{t})\right) + \int_0^\hat{t} V_t(\mathcal{M}^\sigma(s; \bar{x}); \bar{x}, s) ds$$

$$= \int_0^t V_t(\mathcal{M}^\sigma(s; \bar{x}); \bar{x}, s) ds - \int_0^t V_t(\mathcal{M}^\sigma(\hat{t}; \bar{x}); \bar{x}, s) ds$$

$$= \int_0^t (V_t(\mathcal{M}^\sigma(s; \bar{x}); \bar{x}, s) - V_t(\mathcal{M}^\sigma(\hat{t}; \bar{x}); \bar{x}, s)) ds$$

$$\geq 0 \quad \text{(from (32)).}$$

To show that an incentive-compatible mechanism $\sigma$ is individually rational, we need to show that

$$V(\mathcal{M}^\sigma(t; \bar{x}); \bar{x}, t) - P^\sigma(t; \bar{x}) \geq 0 \quad \forall t \in [0, \tau].$$

For the payment rule $P^\sigma(t; \bar{x}) = V(\mathcal{M}^\sigma(t; \bar{x}); \bar{x}, t) - \int_0^t V_t(\mathcal{M}^\sigma(s; \bar{x}); \bar{x}, s) ds$, we have

$$V(\mathcal{M}^\sigma(t; \bar{x}); \bar{x}, t) - P^\sigma(t; \bar{x}) = \int_0^t V_t(\mathcal{M}^\sigma(s; \bar{x}); \bar{x}, s) ds$$

$$= \int_0^t \left(\int_{\mathcal{M}^\sigma(s; \bar{x})} \frac{\partial v(\bar{x}, x, t)}{\partial t} |_{s \to x} ds \right) dx.$$

From property (P3), we have $\frac{\partial v(d, t)}{\partial t} \geq 0$ and $v(d, 0) = 0$ $\forall d$. Therefore, $V_t(\mathcal{M}^\sigma(s; \bar{x}); \bar{x}, s) \geq 0 \quad \forall s \in [0, t]$. Consequently, $\int_0^t V_t(\mathcal{M}^\sigma(s; \bar{x}); \bar{x}, s) ds \geq 0$. 

$\blacksquare$
We are now ready to state an optimal solution to problem $P^\text{sd}(\bar{x})$. Consider the following mechanism (OPT-SD):

$$M^\text{opt-sd}(t; \bar{x}) = \{ x \in D : \rho(\bar{x}, x) \leq r(t) \},$$

$$P^\text{opt-sd}(t; \bar{x}) = V(M^\text{opt-sd}(t; \bar{x}); \bar{x}, t) - \int_0^t V_i(M^\text{opt-sd}(s; \bar{x}); \bar{x}, s)ds.$$

From Claim 1 and Claim 4, we know that $M^\text{opt-sd}$ is an optimal solution to a relaxation of problem $P^\text{sd}(\bar{x})$. Claim 2 establishes the monotonicity property of $M^\text{opt-sd}$ and Claim 3 enables us to write the allocation rule in a compact manner. Finally, $(M^\text{opt-sd}, P^\text{opt-sd})$ satisfy the sufficient conditions presented in Claim 5 and is therefore a feasible solution for problem $P^\text{sd}(\bar{x})$. Consequently, OPT-SD is an optimal mechanism for problem $P^\text{sd}(\bar{x})$.

**Appendix B  Proof of Proposition 1**

We want to show that if the dataset $D$ is a uniform dataset, then the quantity of data purchased by a data-buyer of type $(\bar{x}, t)$ and the corresponding price that she pays under the OPT-SD mechanism is independent of her location type $\bar{x}$. The definition of a uniform dataset (see (8)) states that for any two ideal records $\bar{x}, \bar{y} \in \mathcal{X}$, the following property holds: For all $R$, $0 \leq R \leq \bar{R}(t)$, we have

$$\int_{\{x \in D : \rho(\bar{x}, x) \leq R\}} g(x)dx = \int_{\{x \in D : \rho(\bar{y}, x) \leq R\}} g(x)dx.$$

Recall that

$$r(t) = \begin{cases} \max\{d : v(d, t) - c - \frac{\hat{c}_v(d, t) - \hat{c}_v(d, t)}{\hat{c}_v(d, t)} \left( \frac{1-H(t)}{h(t)} \right) \geq 0\} & \text{if } \exists d \text{ s.t. } w(d, t) \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

and $\bar{R}(t) = \max\{d : v(d, t) - c \geq 0\}$.

Since $\frac{\hat{c}_v(d, t) - \hat{c}_v(d, t)}{\hat{c}_v(d, t)} \left( \frac{1-H(t)}{h(t)} \right) \geq 0$, we have $r(t) \leq \bar{R}(t) \forall t \in [0, \tau]$. Thus, for a uniform dataset, we have

$$\int_{\{x \in D : \rho(\bar{x}, x) \leq r(t)\}} g(x)dx = \int_{\{x \in D : \rho(\bar{y}, x) \leq r(t)\}} g(x)dx$$

(setting $R = r(t)$)

$$\Rightarrow \int_{M^\text{opt-sd}(t; \bar{x})} g(x)dx = \int_{M^\text{opt-sd}(t; \bar{y})} g(x)dx$$

(definition of $M^\text{opt-sd}$),

$$\Rightarrow Q^\text{opt-sd}(t; \bar{x}) = Q^\text{opt-sd}(t; \bar{y}).$$

Thus, under the OPT-SD mechanism, the quantity of data purchased by the buyer of type $(\bar{x}, t)$ is independent of $\bar{x}$.

Define the function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ as follows:

$$\psi(z) = \int_{\{x \in D : \rho(\bar{x}, x) \leq z\}} g(x)dx.$$
We assume that the distance metric $\rho$ is such that the function $\psi(\cdot)$ is differentiable. Thus, we have
\[
V(M^{\text{opt-md}}(t; \bar{x}); \bar{x}, t) = \int_{M^{\text{opt-md}}(t; \bar{x})} (v(\rho(\bar{x}, \bar{x}), t) - c)^+ g(\bar{x}) d\bar{x}
\]
\[
= \int_{\{\bar{x} \in D : \rho(\bar{x}, \bar{x}) \leq r(t)\}} (v(\rho(\bar{x}, \bar{x}), t) - c)^+ g(\bar{x}) d\bar{x}
\]
\[
= \int_0^t (v(z, t) - c)^+ d\psi(z). \tag{33}
\]
Thus, $V(M^{\text{opt-md}}(t; \bar{x}); \bar{x}, t)$ is independent of $\bar{x}$. Consequently, $V_s(M^{\text{opt-md}}(s; \bar{x}); \bar{x}, s)$ is also independent of $\bar{x}$ for all $s$. Therefore, the payment rule under the opt-md mechanism,
\[
\mathcal{P}^{\text{opt-md}}(t; \bar{x}) = V(M^{\text{opt-md}}(t; \bar{x}); \bar{x}, t) - \int_0^s V_s(M^{\text{opt-md}}(s; \bar{x}); \bar{x}, s) ds
\]
is also independent of $\bar{x}$. The result follows. \qed

**Appendix C  Proof of Theorem 2**

We want to show that if $D$ is a uniform dataset, then the mechanism opt-md defined by:
\[
M^{\text{opt-md}}(\bar{x}, t) = \{x \in D : \rho(\bar{x}, x) \leq r(t)\}
\]
\[
\mathcal{P}^{\text{opt-md}}(\bar{x}, t) = V(M^{\text{opt-md}}(\bar{x}, t); \bar{x}, t) - \int_0^t V_s(M^{\text{opt-md}}(s; \bar{x}); \bar{x}, s) ds
\]
is an optimal solution to problem $P^{\text{md}}$.

**Proof:** Note that the optimization problem $P^{\text{md}}(\bar{x})$, defined at the start of Section 5, is one where the location type of the data-buyer, $\bar{x}$, is public knowledge and only the decay type $t$ is private to the buyer. On the other hand, the optimization problem $P^{\text{md}}$ is one where *both* the location and the decay type of the buyer is private to the buyer.

Let $\text{rev}^{\text{opt-md}}(\bar{x})$ denote the optimal value of problem $P^{\text{md}}(\bar{x})$ and let $\text{rev}^\mu$ denote the objective value of problem $P^{\text{md}}$ under mechanism $\mu$. Clearly,
\[
\text{rev}^\mu \leq \mathbb{E}_{\bar{x}} \left[ \text{rev}^{\text{opt-md}}(\bar{x}) \right].
\]

We will first show that the optimal solution of $P^{\text{md}}(\bar{x})$, namely $(M^{\text{opt-md}}(\cdot; \bar{x}), \mathcal{P}^{\text{opt-md}}(\cdot; \bar{x}))$, is feasible for $P^{\text{md}}$. Since $(M^{\text{opt-md}}(\cdot; \bar{x}), \mathcal{P}^{\text{opt-md}}(\cdot; \bar{x}))$ is an optimal solution to $P^{\text{md}}(\bar{x})$, it satisfies the following (ic-md($\bar{x}$)) and (ir-md($\bar{x}$)) constraints for $\bar{x} \in \mathcal{X}$ and for all $s, t \in [0, t]$:
\[
V(M^{\text{opt-md}}(t; \bar{x}); \bar{x}, t) - \mathcal{P}^{\text{opt-md}}(t; \bar{x}) \geq V(M^{\text{opt-md}}(s; \bar{x}); \bar{x}, t) - \mathcal{P}^{\text{opt-md}}(s; \bar{x}), \tag{IC-md($\bar{x}$)}
\]
\[
V(M^{\text{opt-md}}(t; \bar{x}); \bar{x}, t) - \mathcal{P}^{\text{opt-md}}(t; \bar{x}) \geq 0. \tag{IR-md($\bar{x}$)}
\]
Claim 6: If $\mathcal{D}$ is a uniform dataset, then for all $t \in [0, \tau]$ and $\bar{x}, \bar{y} \in \chi$,

$$V(\mathcal{M}^{\text{opt-td}}(t; \bar{x}); \bar{x}, t) \geq V(\mathcal{M}^{\text{opt-td}}(t; \bar{y}); \bar{x}, t).$$

Proof: Note that

$$V(\mathcal{M}^{\text{opt-td}}(t; \bar{x}); \bar{x}, t) = \int_{\mathcal{M}^{\text{opt-td}}(t; \bar{x})} (v(\rho(x, x), t) - c^+) g(x) dx,$$

$$V(\mathcal{M}^{\text{opt-td}}(t; \bar{y}); \bar{x}, t) = \int_{\mathcal{M}^{\text{opt-td}}(t; \bar{y})} (v(\rho(x, x), t) - c^+) g(x) dx.$$

Define $S(t; \bar{x}, \bar{y}) := \mathcal{M}^{\text{opt-td}}(t; \bar{x}) \cap \mathcal{M}^{\text{opt-td}}(t; \bar{y})$. Then,

$$V(\mathcal{M}^{\text{opt-td}}(t; \bar{x}); \bar{x}, t) = \int_{S(t; \bar{x}, \bar{y})} (v(\rho(x, x), t) - c^+) g(x) dx + \int_{\mathcal{M}^{\text{opt-td}}(t; \bar{x}) \setminus S(t; \bar{x}, \bar{y})} (v(\rho(x, x), t) - c^+) g(x) dx.$$

Similarly,

$$V(\mathcal{M}^{\text{opt-td}}(t; \bar{y}); \bar{x}, t) = \int_{S(t; \bar{x}, \bar{y})} (v(\rho(x, x), t) - c^+) g(x) dx + \int_{\mathcal{M}^{\text{opt-td}}(t; \bar{y}) \setminus S(t; \bar{x}, \bar{y})} (v(\rho(x, x), t) - c^+) g(x) dx.$$

Thus, to prove Claim 6, it suffices to show that

$$\int_{\mathcal{M}^{\text{opt-td}}(t; \bar{x}) \setminus S(t; \bar{x}, \bar{y})} (v(\rho(x, x), t) - c^+) g(x) dx \geq \int_{\mathcal{M}^{\text{opt-td}}(t; \bar{y}) \setminus S(t; \bar{x}, \bar{y})} (v(\rho(x, x), t) - c^+) g(x) dx \quad (34)$$

Recall that

$$\mathcal{M}^{\text{opt-td}}(t; \bar{x}) = \{ x \in \mathcal{D} : \rho(x, x) \leq r(t) \}, \text{ and}$$

$$\mathcal{M}^{\text{opt-td}}(t; \bar{y}) = \{ x \in \mathcal{D} : \rho(y, x) \leq r(t) \}.$$

Thus, for any $x \in \mathcal{M}^{\text{opt-td}}(t; \bar{x}) \setminus S(t; \bar{x}, \bar{y})$, $\rho(x, x) \leq r(t)$, and consequently, $v(\rho(x, x), t) \geq v(r(t), t)$, since $\frac{v(d,t)}{d} \leq 0$ (Property P3). The LHS of (34) can then be lower-bounded as follows:

$$\int_{\mathcal{M}^{\text{opt-td}}(t; \bar{x}) \setminus S(t; \bar{x}, \bar{y})} (v(\rho(x, x), t) - c^+) g(x) dx \geq \int_{\mathcal{M}^{\text{opt-td}}(t; \bar{y}) \setminus S(t; \bar{x}, \bar{y})} (v(r(t), t) - c^+) g(x) dx$$

$$= (v(r(t), t) - c^+) \int_{\mathcal{M}^{\text{opt-td}}(t; \bar{y}) \setminus S(t; \bar{x}, \bar{y})} g(x) dx$$
Similarly, for any $x \in \mathcal{M}^{opt-sd}(t;\bar{y}) \setminus \mathcal{S}(t;\bar{x},\bar{y})$, $\rho(\bar{x}, x) \geq r(t)$, and consequently, $v(\rho(\bar{x}, x), t) \leq v(r(t), t)$. Therefore, the RHS of (34) can then be upper-bounded as follows:

$$\int_{\mathcal{M}^{opt-sd}(t;\bar{y}) \setminus \mathcal{S}(t;\bar{x},\bar{y})} (v(\rho(\bar{x}, x), t) - c)^+ g(x) dx \leq \int_{\mathcal{M}^{opt-sd}(t;\bar{y}) \setminus \mathcal{S}(t;\bar{x},\bar{y})} (v(r(t), t) - c)^+ g(x) dx = (v(r(t), t) - c)^+ \int_{\mathcal{M}^{opt-sd}(t;\bar{y}) \setminus \mathcal{S}(t;\bar{x},\bar{y})} g(x) dx.$$

Since, $D$ is a uniform-dataset, Proposition 1 yields

$$\int_{\mathcal{M}^{opt-sd}(t;\bar{x}) \setminus \mathcal{S}(t;\bar{x},\bar{y})} g(x) dx = \int_{\mathcal{M}^{opt-sd}(t;\bar{y}) \setminus \mathcal{S}(t;\bar{x},\bar{y})} g(x) dx. \quad \implies \quad \int_{\mathcal{M}^{opt-sd}(t;\bar{x}) \setminus \mathcal{S}(t;\bar{x},\bar{y})} g(x) dx = \int_{\mathcal{M}^{opt-sd}(t;\bar{y}) \setminus \mathcal{S}(t;\bar{x},\bar{y})} g(x) dx.$$

Thus,

$$\int_{\mathcal{M}^{opt-sd}(t;\bar{x}) \setminus \mathcal{S}(t;\bar{x},\bar{y})} (v(\rho(\bar{x}, x), t) - c)^+ g(x) dx \geq \int_{\mathcal{M}^{opt-sd}(t;\bar{y}) \setminus \mathcal{S}(t;\bar{x},\bar{y})} (v(\rho(\bar{x}, x), t) - c)^+ g(x) dx.$$

This completes the proof of Claim 6. \hfill \blacksquare

Recall from Proposition 1 that if $D$ is a uniform dataset, then for a data-buyer of decay type $t$, the price $\mathcal{P}^{opt-sd}(t;\bar{x})$ she pays for purchasing her optimal set of records is independent of her location type $\bar{x}$. That is,

$$\mathcal{P}^{opt-sd}(t;\bar{x}) = \mathcal{P}^{opt-sd}(t;\bar{y}) \forall \bar{y} \in \chi. \quad (35)$$

Using Claim 6 and (35), the (IC-sd($\bar{x}$)) and (IR-sd($\bar{x}$)) constraints, for all $\bar{x}, \bar{y} \in \chi$, then imply:

$$V(\mathcal{M}^{opt-sd}(t;\bar{x});\bar{x}, t) - \mathcal{P}^{opt-sd}(t;\bar{x}) \geq V(\mathcal{M}^{opt-sd}(s;\bar{y});\bar{x}, t) - \mathcal{P}^{opt-sd}(s;\bar{y}), \forall s, t \in [0, \tau],$$

$$V(\mathcal{M}^{opt-sd}(t;\bar{x});\bar{x}, t) - \mathcal{P}^{opt-sd}(t;\bar{x}) \geq 0 \forall t \in [0, \tau].$$

Thus, the allocation and payment functions $(\mathcal{M}^{opt-sd}, \mathcal{P}^{opt-sd})$ satisfy the (IC-MD) and (IR-MD) constraints of problem $P^{MD}$.

Consider the following mechanism $\text{OPT-MD}$:

$$\mathcal{M}^{opt-md}(\bar{x}, t) = \mathcal{M}^{opt-sd}(t;\bar{x})$$

$$\mathcal{P}^{opt-md}(\bar{x}, t) = \mathcal{P}^{opt-sd}(t;\bar{x})$$

Clearly, the mechanism $\text{OPT-MD}$ is a feasible solution to problem $P^{MD}$. Moreover,

$$\text{REV}^{opt-md} = \mathbb{E}_\bar{x} [\text{REV}^{opt-sd}(\bar{x})].$$

Thus, $\text{OPT-MD}$ is an optimal solution to $P^{MD}$. \hfill \blacksquare
Appendix D  Proof of Theorem 3

Let \( \text{REV}^{\text{opt-pq}} \) denote the revenue to the data-seller from using an optimal price-quantity schedule. Similarly, let \( \text{REV}^{\text{opt}} \) denote the revenue to the seller under an optimal mechanism. We want to derive an upper bound on the ratio \( \frac{\text{REV}^{\text{opt}}}{\text{REV}^{\text{opt-pq}}} \).

Recall from Theorem 1 that the \( \text{OPT-SD} \) mechanism for a given location type \( \bar{x} \) can be characterized by the pair of functions \( (Q^{\text{opt-sd}}(t; \bar{x}), P^{\text{opt-sd}}(t; \bar{x})) \), where

\[
Q^{\text{opt-sd}}(t; \bar{x}) = \int_{M^{\text{opt-sd}}(t; \bar{x})} g(x) \, dx \\
= \int_{\{x \in D; (x, \bar{x}) \in r(t)\}} g(x) \, dx, \text{ and} \\
P^{\text{opt-sd}}(t; \bar{x}) = V(M^{\text{opt-md}}(\bar{x}, t); \bar{x}, t) - \int_0^t V_i(M^{\text{opt-md}}(\bar{x}, s); \bar{x}, s) \, ds.
\]

From Corollary 1, we know that for a location type \( \bar{x} \), the \( \text{OPT-SD} \) mechanism can be implemented as a price-quantity schedule:

\[
P^{\text{sd-pq}}(q; \bar{x}) = \begin{cases} 
P^{\text{opt-sd}}(t^{sn}(q; \bar{x}); \bar{x}) & \text{if } Q^{\text{opt-sd}}(t^{sn}(q; \bar{x}); \bar{x}) = q \\
\infty & \text{otherwise},
\end{cases}
\]

where \( t^{sn}(q; \bar{x}) = \sup\{t : t \in \Gamma^{sn}(q; \bar{x})\} \). Further, from the definition of \( \delta(q; \bar{x}) \), it follows that \( \delta(q; \bar{x}) = r(t^{sn}(q; \bar{x})) \).

Claim 7: For all \( \bar{x} \),

\[
\frac{dP^{\text{sd-pq}}(q; \bar{x})}{dq} = v(\delta(q; \bar{x}), t^{sn}(q; \bar{x})) - c.
\]

Proof: We have

\[
\frac{dP^{\text{sd-pq}}(q; \bar{x})}{dq} = \frac{dP^{\text{opt-sd}}(t; \bar{x})}{dt} \bigg|_{t = e^{p}(q; \bar{x})} \frac{dt^{sn}(q; \bar{x})}{dq}.
\]

Also,

\[
\frac{dQ^{\text{opt-sd}}(t^{sn}(q; \bar{x}); \bar{x})}{dt} \bigg|_{t = e^{p}(q; \bar{x})} \frac{dt^{sn}(q; \bar{x})}{dq} = 1 \quad \text{(definition of } t^{sn})
\]

\[
\implies \frac{dt^{sn}(q; \bar{x})}{dq} = \frac{1}{\frac{dQ^{\text{opt-sd}}(t; \bar{x})}{dt} \bigg|_{t = e^{p}(q; \bar{x})}}.
\]

Therefore,

\[
\frac{dP^{\text{sd-pq}}(q; \bar{x})}{dq} = \frac{dP^{\text{opt-sd}}(t; \bar{x})}{dt} \bigg|_{t = e^{p}(q; \bar{x})} \frac{dQ^{\text{opt-sd}}(t; \bar{x})}{dt} \bigg|_{t = e^{p}(q; \bar{x})}.
\]

We will now compute \( \frac{dP^{\text{opt-sd}}(t; \bar{x})}{dt} \) and \( \frac{dQ^{\text{opt-sd}}(t; \bar{x})}{dt} \). For any \( r \geq 0 \), let

\[
Q(r; \bar{x}) = \int_{\{x \in D; (x, \bar{x}) \leq r\}} g(x) \, dx.
\]
Thus,

\[ Q_{\text{OPT-SD}}^{t}(t; \bar{x}) = Q(r(t); \bar{x}). \]

\[ \Rightarrow \frac{dQ_{\text{OPT-SD}}^{t}(t; \bar{x})}{dt} = Q'(r(t); \bar{x})r'(t). \]

Note that,

\[ V(M_{\text{OPT-SD}}^{t}(t; \bar{x}), \bar{x}, t) = \int_{\{x \in D : r(\bar{x}, x) \leq r(t)\}} (v(\rho(\bar{x}, x), t) - c) g(x)dx \]

\[ = \int_{0}^{r(t)} (v(z, t) - c)Q'(z; \bar{x})dz, \]

and

\[ \int_{0}^{r(t)} V_{1}(M_{\text{OPT-SD}}^{t}(s; \bar{x}), \bar{x}, s)ds = \int_{0}^{r(t)} \left( \int_{\{x \in D : r(\bar{x}, x) \leq r(s)\}} \frac{\partial v(\rho(\bar{x}, x), t)}{\partial t} \bigg|_{t-s} g(x)dx \right)ds \]

\[ = \int_{0}^{r(t)} \left( \int_{0}^{r(s)} \frac{\partial v(z, t)}{\partial t} \bigg|_{t-s} Q'(z; \bar{x})dz \right)ds \]

\[ = \int_{0}^{r(t)} \left( \int_{0}^{r(s)} v_{2}(z, s)Q'(z; \bar{x})dz \right)ds. \]

Thus,

\[ P_{\text{OPT-SD}}^{t}(t; \bar{x}) = V(M_{\text{OPT-SD}}^{t}(t; \bar{x}), \bar{x}, t) - \int_{0}^{r(t)} V_{1}(M_{\text{OPT-SD}}^{t}(s; \bar{x}, \bar{x}, s))ds \]

\[ = \int_{0}^{r(t)} (v(z, t) - c)Q'(z; \bar{x})dz - \int_{0}^{r(t)} \left( \int_{0}^{r(s)} v_{2}(z, s)Q'(z; \bar{x})dz \right)ds. \]

Next,

\[ \frac{dP_{\text{OPT-SD}}^{t}(t; \bar{x})}{dt} \]

\[ = (v(r(t), t) - c)Q'(r(t); \bar{x})r'(t) + \int_{0}^{r(t)} v_{2}(z, t)Q'(z; \bar{x})dz - \int_{0}^{r(t)} v_{2}(z, t)Q'(z; \bar{x})dz \]

\[ = (v(r(t), t) - c)Q'(r(t); \bar{x})r'(t). \]

Therefore,

\[ \frac{dP_{\text{OPT-SD}}^{t}(t; \bar{x})}{dt} = v(r(t), t) - c. \]

\[ \Rightarrow \frac{dP_{\text{OPT-SD}}^{t}(q; \bar{x})}{dq} = \frac{dP_{\text{OPT-SD}}^{t}(t; \bar{x})}{dt} \big|_{t = t^{*0}(q; \bar{x})} = v(r(t^{*0}(q; \bar{x}), t^{*0}(q; \bar{x})) - c \]

\[ = v(\delta(q; \bar{x}), t^{*0}(q; \bar{x})) - c. \]
This completes the proof of Claim 7.

Next, consider the price-quantity schedule $P^{\text{bound}}(q)$ defined through its derivative as follows:

$$\frac{dP^{\text{bound}}(q)}{dq} = \min_{\bar{x}} \frac{dP^{\text{sd-pq}}(q; \bar{x})}{dq}, \text{ and } P^{\text{bound}}(0) = 0.$$

Let $Q^{\text{bound}}(\bar{x}, t)$ denote the quantity purchased by a buyer of type $(\bar{x}, t)$ under the price-quantity schedule $P^{\text{bound}}(\cdot)$.

Claim 8: $Q^{\text{bound}}(\bar{x}, t) \geq Q^{\text{opt-sd}}(t; \bar{x})$.

Proof: Let $U(q; \bar{x}, t)$ denote the benefit to the buyer of type $(\bar{x}, t)$ when she consumes a quantity $q$ of records.

$$U(q; \bar{x}, t) = V(S; \bar{x}, t) \text{ s.t. } \int_S g(x) dx = q.$$

Let $\pi^{\text{bound}}(q; \bar{x}, t)$ denote the net utility (benefit less the price) to the data buyer of type $(\bar{x}, t)$ when she purchases a quantity $q$ of records under the price-quantity schedule $P^{\text{bound}}(\cdot)$. Similarly, let $\pi^{\text{sd-pq}}(q; \bar{x}, t)$ denote the net utility to the data buyer of type $(\bar{x}, t)$ when she purchases a quantity $q$ of records under the price-quantity schedule $P^{\text{sd-pq}}(\cdot; \bar{x})$. Thus,

$$\pi^{\text{bound}}(q; \bar{x}, t) = U(q; \bar{x}, t) - P^{\text{bound}}(q), \text{ and } \pi^{\text{sd-pq}}(q; \bar{x}, t) = U(q; \bar{x}, t) - P^{\text{sd-pq}}(q; \bar{x}).$$

Furthermore,

$$Q^{\text{bound}}(\bar{x}, t) = \arg \max_q \pi^{\text{bound}}(q; \bar{x}, t), \text{ and } Q^{\text{opt-sd}}(t; \bar{x}) = \arg \max_q \pi^{\text{sd-pq}}(q; \bar{x}, t). \tag{36}$$

Consider any $q \leq Q^{\text{opt-sd}}(t; \bar{x})$. We have

$$\pi^{\text{bound}}(Q^{\text{opt-sd}}(t; \bar{x}); \bar{x}, t) - \pi^{\text{bound}}(q; \bar{x}, t)
= \int_q \left( \frac{d\pi^{\text{bound}}(q; \bar{x}, t)}{dq} \right) dq
= \int_q \left( \frac{dU(q; \bar{x}, t)}{dq} - \frac{dP^{\text{bound}}(q)}{dq} \right) dq
\geq \int_q \left( \frac{dU(q; \bar{x}, t)}{dq} - \frac{dP^{\text{sd-pq}}(q; \bar{x})}{dq} \right) dq.$$
Since \( Q^\text{BOUND}(x; t) \) maximizes \( \pi^\text{BOUND}(q; x; t) \). Since \( \pi^\text{BOUND}(Q^\text{OPT-SD}(t; \bar{x}); \bar{x}, t) - \pi^\text{OPT-SD}(q; \bar{x}, t) \geq 0 \) \( \forall q \leq Q^\text{OPT-SD}(t; \bar{x}) \), it follows that \( Q^\text{BOUND}(\bar{x}, t) \geq Q^\text{OPT-SD}(t; \bar{x}) \).

From Assumption 1, for a buyer of type \( p \), recall \( w(d, t) = v(d, t) - c - \frac{\bar{c}_u(d, t) - 1 - H(t)}{\bar{c}_t} \) and \( r(t) = \sup \{ d : w(d, t) \geq 0 \} \).

Since \( \frac{\bar{c}_u(d, t) - 1 - H(t)}{\bar{c}_t} \geq 0 \) \( \forall d, t \), the above assumption implies that under the OPT-SD mechanism, \( w(r(t), t) = 0 \). Therefore, \( v(r(t), t) - c = \frac{\bar{c}_u(r(t), t) - 1 - H(t)}{\bar{c}_t} \).

From Claim 7 and the result from the above assumption, we have

\[
\frac{dP^{SD-pq}(q; \bar{x})}{dq} = \frac{\partial v(\delta(q; \bar{x}), t^m(q; \bar{x}))}{\partial t} \frac{1 - H(t^m(q; \bar{x}))}{h(t^m(q; \bar{x}))}.
\]

Next,

\[
\text{REV}^\text{BOUND} = E_{(\bar{x}, t)} \left[ \pi^\text{BOUND}(x; t) \left( \int_0^{\pi^\text{BOUND}(q; \bar{x}; t)} \frac{dP^q}{dq} dq \right) \right]
\]

\[
= E_{(\bar{x}, t)} \left[ \pi^\text{BOUND}(x; t) \left( \int_0^{\pi^\text{BOUND}(q; \bar{x}; t)} \min_Y \frac{dP^{SD-pq}(q; \bar{y})}{dq} dq \right) \right] \quad \text{(definition of } P^\text{BOUND})
\]

\[
\geq E_{(\bar{x}, t)} \left[ \pi^\text{OPT-SD}(x; t) \left( \int_0^{\pi^\text{OPT-SD}(q; \bar{x}; t)} \min_Y \frac{dP^{SD-pq}(q; \bar{y})}{dq} dq \right) \right] \quad \text{(from Claim 8)}
\]

\[
= E_{(\bar{x}, t)} \left[ \pi^\text{OPT-SD}(x; t) \left( \int_0^{\pi^\text{OPT-SD}(q; \bar{x}; t)} \min_Y L(\delta(q; \bar{y}), t^m(q; \bar{y})) dq \right) \right] \quad \text{(definition of } L(d, t))
\]

Also,

\[
\text{REV}^\text{OPT} \leq E_{(\bar{x}, t)} \left[ \pi^\text{OPT-SD}(x; t) \left( \int_0^{\pi^\text{OPT-SD}(q; \bar{x}; t)} \frac{dP^{SD-pq}(q; \bar{x})}{dq} dq \right) \right]
\]

\[
= E_{(\bar{x}, t)} \left[ \pi^\text{OPT-SD}(x; t) \left( \int_0^{\pi^\text{OPT-SD}(q; \bar{x}; t)} L(\delta(q; \bar{y}), t^m(q; \bar{y})) dq \right) \right]
\]

\[
= E_{(\bar{x}, t)} \left[ \pi^\text{OPT-SD}(x; t) \left( \int_0^{\pi^\text{OPT-SD}(q; \bar{x}; t)} \min_Y L(\delta(q; \bar{y}), t^m(q; \bar{y})) + \epsilon_d M_\delta + \epsilon_t M_t dq \right) \right].
\]
Therefore,

\[
\frac{\text{REV}^\text{opt}}{\text{REV}^\text{bound}} \leq \frac{\mathbb{E}_{(x,t)} \left[ \int_0^{Q^\text{opt-sd}(t;\bar{x})} \left( \min_y L(q;\bar{y}), t^\text{sd}(q;\bar{y}) \right) dq \right]}{\mathbb{E}_{(x,t)} \left[ \int_0^{Q^\text{opt-sd}(t;\bar{x})} \left( \min_y L(q;\bar{y}), t^\text{sd}(q;\bar{y}) \right) dq \right]}
\]

\[
= 1 + \frac{(\epsilon_\delta M_\delta + \epsilon_t M_t) \mathbb{E}_{(x,t)} [Q^\text{opt-sd}(t;\bar{x})]}{\mathbb{E}_{(x,t)} \left[ \int_0^{Q^\text{opt-sd}(t;\bar{x})} \left( \min_y L(q;\bar{y}), t^\text{sd}(q;\bar{y}) \right) dq \right]}
\]

\[
\leq 1 + \frac{(\epsilon_\delta M_\delta + \epsilon_t M_t)}{\text{REV}^\text{bound}} \quad (\text{since } Q^\text{opt-sd}(t;\bar{x}) \leq 1).
\]

Thus,

\[
\text{REV}^\text{opt} \leq \text{REV}^\text{bound} + (\epsilon_\delta M_\delta + \epsilon_t M_t).
\]

Recall from (19) that the optimal revenue to the data-seller obtained by selling the entire dataset at one price is \(\text{REV}^\text{opt-full}\). Clearly,

\[
\text{REV}^\text{opt-pq} \geq \max \{\text{REV}^\text{opt-full}, \text{REV}^\text{bound}\}.
\]

We have,

\[
\frac{\text{REV}^\text{opt-pq}}{\text{REV}^\text{opt-full}} \leq \frac{\text{REV}^\text{bound}}{\max \{\text{REV}^\text{opt-full}, \text{REV}^\text{bound}\}} + \frac{\epsilon_\delta M_\delta + \epsilon_t M_t}{\max \{\text{REV}^\text{opt-full}, \text{REV}^\text{bound}\}}
\]

\[
\leq 1 + \frac{\epsilon_\delta M_\delta + \epsilon_t M_t}{\text{REV}^\text{opt-full}}.
\]

This completes the proof of Theorem 3. ■

Appendix E  Numerical Analysis

In this section, we discuss the numerical evaluation of (i) the performance of the optimal two-part tariff and the optimal two-block tariff, relative to the optimal mechanism, and (ii) the value to the data-seller for providing the option of filtering the dataset to buyers. We begin by describing our test bed.

E.1  Test bed

Each record in the dataset consists of 1 attribute (column), i.e. \(N = 1\). Thus, together with the private information about the decay type of the buyer, problem \(P^\text{md}\) becomes a 2-dimensional
mechanism-design problem. Since $N = 1$, each record in the dataset is a scalar; we let the record space $\chi \subseteq [0,1]$ and discretize the space equidistantly, so that the location type and each record in the dataset takes a value from the set $\{0,0.1,0.2,\ldots,1\}$. Similarly, the space of decay types is also equidistantly discretized; the decay type of the buyer can take a value from the set $\{0,0.1,0.2\ldots,1\}$.

The location type and the decay type of the buyer are private to her and the seller only has distributional knowledge; the location type follows a (discrete) triangular distribution\(^1\) over the record space and the decay type follows a (discrete) uniform distribution over the space of decay types. The number of records in the dataset are normalized to 1. This induces a distribution of the records in the dataset over the record space; we assume that this distribution of records follows a triangular distribution. Let $m_{\text{rec}}$ and $m_{\text{loc}}$ denote, respectively, the mode of the triangular distribution of the records in the dataset and the location type of the buyer. In our analysis, we vary $m_{\text{rec}}$ and $m_{\text{loc}}$ to take values from the set $\{0.1,0.5,0.9\}$.

The utility function $v(\cdot,\cdot)$ is defined as $v(d,t) = t \cdot (1 - d)$. Observe that $v(\cdot,\cdot)$ satisfies properties P1–P5 stated in Section 3. The targeting cost $c$ takes a value from the set $\{0.1,0.3,0.5,0.7,0.9\}$.

<table>
<thead>
<tr>
<th>Parameter description</th>
<th>Value(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Record space</td>
<td>${0,0.1,0.2,\ldots,1}$</td>
</tr>
<tr>
<td>Decay type space</td>
<td>${0,0.1,0.2,\ldots,1}$</td>
</tr>
<tr>
<td>Mode of the triangular distribution of records ($m_{\text{rec}}$)</td>
<td>${0.1,0.5,0.9}$</td>
</tr>
<tr>
<td>Mode of the triangular distribution of location type ($m_{\text{loc}}$)</td>
<td>${0.1,0.5,0.9}$</td>
</tr>
<tr>
<td>Targeting cost ($c$)</td>
<td>${0.1,0.3,0.5,0.7,0.9}$</td>
</tr>
</tbody>
</table>

**Table 5** Parameter values for numerical analysis.

### E.2 Performance of Two-Part Tariffs and Two-Block Tariffs

For the test bed described above, we optimize the parameters of both these tariffs by performing a line search over the set of feasible values that each parameter can take, and obtain the optimal two-part tariff and the optimal two-block tariff. Then, we compute the revenue from these schedules and benchmark it with the revenue obtained from the optimal mechanism by solving the linear program $P\text{MD-LP}^\text{MD-LP}$. Table 6a and Table 6b show the ratio of the seller's revenue from the optimal two-part tariff to the revenue from the optimal mechanism, for extreme values of the targeting cost $c$. In each table, we compute the performance of the two-part tariff for three different values of the mode $m_{\text{rec}}$ of

---

\(^1\) See [https://en.wikipedia.org/wiki/Triangular_distribution](https://en.wikipedia.org/wiki/Triangular_distribution) for details.
Pricing Policies for Data Monetization

the distribution of the records in the dataset and the mode $m_{loc}$ of the distribution of the location type of the buyer. When the targeting cost is low, a large fraction of records in the dataset yield a positive utility to the buyers. Further, there is low heterogeneity in buyers’ valuations for the entire dataset. Consequently, the seller can extract most of the surplus through a take-it-or-leave-it offer (i.e., charging a fixed price) for the entire dataset. The two-part tariff, being a proper generalization of the fixed-price policy, performs even better. On the other hand, when the targeting cost is high, relatively few records in the dataset yield a positive utility to the buyers. In this case, while a fixed-price policy for the entire dataset is significantly suboptimal, the optimal two-part tariff uses a lower fixed price and appropriately adjusts the per-unit price to again yield near-optimal revenue. Since the class of two-block tariffs is a generalization of the class of two-part tariffs, the performance of an optimal two-block tariff is even better; see Tables 7a and 7b.

<table>
<thead>
<tr>
<th>Mode of the distribution of location type ($m_{loc}$)</th>
<th>0.1</th>
<th>0.5</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode of the distribution of the records ($m_{rec}$)</td>
<td>0.1</td>
<td>91.85%</td>
<td>92.48%</td>
</tr>
<tr>
<td>Mode of the distribution of the records ($m_{rec}$)</td>
<td>0.5</td>
<td>92.74%</td>
<td>93.68%</td>
</tr>
<tr>
<td>Mode of the distribution of the records ($m_{rec}$)</td>
<td>0.9</td>
<td>94.24%</td>
<td>92.48%</td>
</tr>
</tbody>
</table>

(a) Low targeting cost ($c = 0.1$)

<table>
<thead>
<tr>
<th>Mode of the distribution of location type ($m_{loc}$)</th>
<th>0.1</th>
<th>0.5</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode of the distribution of the records ($m_{rec}$)</td>
<td>0.1</td>
<td>95.67%</td>
<td>94.86%</td>
</tr>
<tr>
<td>Mode of the distribution of the records ($m_{rec}$)</td>
<td>0.5</td>
<td>93.60%</td>
<td>94.89%</td>
</tr>
<tr>
<td>Mode of the distribution of the records ($m_{rec}$)</td>
<td>0.9</td>
<td>92.23%</td>
<td>94.86%</td>
</tr>
</tbody>
</table>

(b) High targeting cost ($c = 0.9$)

Table 6 Performance of the optimal two-part tariff for extreme values of the targeting cost.

<table>
<thead>
<tr>
<th>Mode of the distribution of location type ($m_{loc}$)</th>
<th>0.1</th>
<th>0.5</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode of the distribution of the records ($m_{rec}$)</td>
<td>0.1</td>
<td>93.76%</td>
<td>93.60%</td>
</tr>
<tr>
<td>Mode of the distribution of the records ($m_{rec}$)</td>
<td>0.5</td>
<td>94.87%</td>
<td>95.77%</td>
</tr>
<tr>
<td>Mode of the distribution of the records ($m_{rec}$)</td>
<td>0.9</td>
<td>95.14%</td>
<td>93.60%</td>
</tr>
</tbody>
</table>

(a) Low targeting cost ($c = 0.1$)

<table>
<thead>
<tr>
<th>Mode of the distribution of location type ($m_{loc}$)</th>
<th>0.1</th>
<th>0.5</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode of the distribution of the records ($m_{rec}$)</td>
<td>0.1</td>
<td>95.67%</td>
<td>94.95%</td>
</tr>
<tr>
<td>Mode of the distribution of the records ($m_{rec}$)</td>
<td>0.5</td>
<td>93.60%</td>
<td>94.59%</td>
</tr>
<tr>
<td>Mode of the distribution of the records ($m_{rec}$)</td>
<td>0.9</td>
<td>92.30%</td>
<td>94.95%</td>
</tr>
</tbody>
</table>

(b) High targeting cost ($c = 0.9$)

Table 7 Performance of the optimal two-block tariff for extreme values of the targeting cost.

The attractiveness of an optimal two-block tariff over an optimal two-part tariff is accentuated for intermediate values of the targeting cost; see Table 8. Here, there is more heterogeneity across buyers in the fraction of records that offers them a positive utility. An optimal two-block tariff better discriminates the heterogeneity in buyers’ willingness-to-pay by charging a steep price-per-record for low-quantity buyers and offering a discount for high-quantity buyers.
Table 8 Performance of (a) two-part tariff and (b) two-block tariff for intermediate values of the targeting cost.

<table>
<thead>
<tr>
<th>(a) two-part tariff (c = 0.5)</th>
<th>(b) two-block tariff (c = 0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode of the distribution of location type ($m_{loc}$)</td>
<td>Mode of the distribution of location type ($m_{loc}$)</td>
</tr>
<tr>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>84.95%</td>
<td>80.12%</td>
</tr>
<tr>
<td>83.27%</td>
<td>85.76%</td>
</tr>
<tr>
<td>80.12%</td>
<td>82.50%</td>
</tr>
</tbody>
</table>

E.3 Value of Filtering

The provision of the filtering option (by the data-seller) enables buyers to endogenously select subsets of records that are of interest to them. In the absence of the filtering option, the buyers face a take-it-or-leave-it offer from the seller; that is, they either have to purchase the entire dataset at the stated price or buy nothing. For a pricing policy $P$, we define the seller’s value from offering the filtering option, denoted $V^P_{filter}$, as the incremental benefit to the seller from using the policy $P$ instead of an optimal take-it-or-leave-it policy. Precisely,

$$V^P_{filter} = \frac{\text{Revenue from pricing policy } P}{\text{Revenue from the optimal take-it-or-leave-it pricing policy}} - 1.$$ 

Figure 5 demonstrates the increase in the value of filtering to the seller when he uses the optimal mechanism (Figure 5a) and when he uses an optimal two-block tariff (Figure 5b), under different distributions of the location type and the records in the dataset. As the targeting cost increases,
the utility of each record to the buyer decreases and, therefore, so does the fraction of records in the
dataset that yield positive utility to the buyer. This, coupled with the high heterogeneity in buyers’
valuations for the entire dataset, makes a take-it-or-leave-it offer less attractive to the buyer and
thereby increases the seller’s value from offering the filtering option.