Buying Data using Data: a Buyer's Perspective on an Infinite Supply Market

Anonymous Author(s)

ABSTRACT

place holder place holder

ACM Reference Format:

Anonymous Author(s). 2020. Buying Data using Data: a Buyer's Perspective on an Infinite Supply Market. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

place holder for greg.

Our main contributions can be summarized as follows.

- We introduce a buying strategy for players in a data market including a dataset allocation and price prediction (Section 3).
- We provide a novel data market simulation to be publicly available upon acceptance (Section 4.1) based on a model defined Section 2.
- Using the simulation, we demonstrate a proof-of-concept of our approach, showing its superiority over several baseline methodologies (Section 4.3).

In addition, we discuss related work in Section 5 and conclude in Section 6.

2 DATA MARKET

Next, we layout our data market setting partially following [1].

2.1 Data Market Model

A *data market* $\mathcal{M}$ is a composition of three sets of entities, namely *buyers* $\mathcal{B}$, *sellers* $\mathcal{S}$, and *mediator* m. We shall refer to the first two entities also as the *players* $\mathcal{A} = \mathcal{B} \cup \mathcal{S}$ in the market. Players may be individuals or groups (*e.g.,* companies or organizations) interested in trading datasets. We use $\mathcal{D}$ to denote the set of authorized datasets in $\mathcal{M}$. Usually, a dataset D is associated with a domain and several of other properties that represent it including *e.g.*, the features it contains. Each buyer $b \in \mathcal{B}$ is interested in a set of datasets $D_b \subseteq \mathcal{D}$ whereas each seller offers a set of products $D_s \subseteq \mathcal{D}$. In addition, each player has a budget L_a, which is updated according to the trasactions of a player. The mediator is in charge of the transactions between the different players in the market.

Example 1. *RS: complete*

In what follows, each player $a \in \mathcal{A}$ has a different utility from a dataset, according to which they can set prices. A market is a temporal ecosystem. As such, we assume that the market has a finite horizon T and each interaction between the different entities takes place in a discrete timestamp $t < T$. Each timestamp t represents a set of transactions supervised by the mediator. Players price their products according to...

RS: say something about the value of data (find references)

A *transaction* is when a buyer b acquires a dataset D from a seller s. The buyer pays $p(D)$, L_b and L_s are updated, and the buyer is no longer interested in buying D. Note, that the seller is still willing to sell D, as the inventory of a dataset is assumed to be unlimited. Finally, at the end of each timestamp, buyers and sellers update their datasets pricing.

In practice, the player can not simply purchase any product she wishes, even if she has a large enough budget for two primary reasons. First, a the "buying" player and the "selling" player may not be synchronized (*e.g.,* the seller demands a price that the buyer is not willing to pay). Second, when an auction is in order, the player may not win the auction for the product she wishes to purchase. RS: say something about auctions

2.2 Sellers

In this paper, we focus on the buyers side, aiming to maximize their utility. In what follows, we describe next the methodology we use to represnet and simulate the behavior of sellers in the marker.

3 BUYING DATA IN A DATA MARKET

Focusing on buyers, we identify two main factors that affect buying data in the setting defined in Section 2. The first focuses on selecting the set of datasets a buyer is willing to buy at time t. The second is pricing in the form of cost estimation and bids.

3.1 Datasets Allocation Optimization Strategy

Next, we define the *datasets allocation* problem with respect to a single player a before timestamp t in the market horizon T.

Recall that a player a has a $\mathcal{D}$=$\langle D_1, D_2, \ldots, D_n \rangle$ (set of relevant datasets), L (budget), and a valuation value v_i for each relevant dataset D_i. At timestamp 0, the set of datasets $\mathcal{D}$, its corresponding valuations V and an initial L are set. While the valuation values

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, July 2017, Washington, DC, USA

© 2020 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

of datasets stays constant throughout the horizon, the $\mathcal{D}$ and L change with respect to the interaction in the market. We denote the datasets set available at time t and the current budget at time t as $\mathcal{D}^t$ and L^t, respectively. We assume that the only thing that changes the datasets is that a player has purchased a dataset and thus, $\mathcal{D}^t \subseteq \mathcal{D}^{t'}, t' < t$ and $\mathcal{D}^0 = \mathcal{D}$. The budget changes with respect to the revenue from each purchased product and its cost.

At time t the player has to select a subset of datasets $\bar{\mathcal{D}} \subseteq \mathcal{D}^{t-1}$ that maximizes her future revenues (recall Section 2.1). To simplify the notation, we denote that number of datasets available at time t as $m \leq n$. In what follows, let $C = \langle c_1, c_2, \ldots, c_m \rangle, c_i \in \mathbb{R}$ and $W = \langle w_1, w_2, \ldots, w_m \rangle, w_m \in \{0, 1\}$ represent a realization of costs and win indicators of the relevant datasets after timestamp t has completed, respectively.

In practice, a player has to allocate a subset of products at the beginning of time t. Thus, the player does not know the actual cost of datasets when the dataset allocation takes place. Accordingly, the player has to estimate the costs $\hat{C} = \langle \hat{c}_1, \hat{c}_2, \ldots, \hat{c}_m \rangle$ and wining indicators $\hat{W} = \langle \hat{w}_1, \hat{w}_2, \ldots, \hat{w}_m \rangle$. Using these estimations, the *datasets allocation* problem can be formalized as follows:

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \sum_{i=1}^{m} (v_i - \hat{c}_i) \cdot \hat{w}_i \cdot X_i \\ \text{subject to} \quad & \sum_{i=1}^{m} \hat{c}_i \cdot \hat{w}_i \cdot X_i \leq L \\ & X_i \in \{0, 1\} \; i = 1, \ldots, m. \end{aligned} \tag{1}$$

Recalling that $\hat{w}_i \in \{0, 1\}$ the player can actually decrease the size of $\mathcal{D}^{t-1}$ to the set of product she estimates she would win, *i.e.*, $win(\mathcal{D}^{t-1}) = \{p_i \in \mathcal{D}^{t-1} | \hat{w}_i = 1\}$. We denote the size of $win(\mathcal{D}^{t-1})$ as m_{win}

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \sum_{i=1}^{m_{win}} (v_i - \hat{c}_i) \cdot X_i \\ \text{subject to} \quad & \sum_{i=1}^{m_{win}} \hat{c}_i \cdot X_i \leq L \\ & X_i \in \{0, 1\} \; i = 1, \ldots, m_{win}. \end{aligned} \tag{2}$$

Solving Eq. 2 is RS: hard? can we prove it?. In addition, simultaneous auctions combining infinite supply with multiple buyers and sellers is far from easy elaborate. Thus, in this paper we focus on an action-free market as described next.

3.2 An Auction-Free Market

we begin with problem definition assuming auctions and than relax the auction such that each buyer is allocated to a seller and in case there is a match (the price the buyer is willing to pay is higher that the price limit set by the seller), the buyer gets the product.

relax the problem, without w

3.3 Price Prediction and Bidding

estimating costs

biding strategies

4 EVALUATION

bla bla bla

4.1 A Data Market Simulation

4.2 Empirical Settings

4.3 Results

5 RELATED WORK

complete

6 CONCLUSIONS

complete

REFERENCES

[1] Raul Castro Fernandez, Pranav Subramaniam, and Michael J Franklin. 2020. Data Market Platforms: Trading Data Assets to Solve Data Problems [Vision Paper]. (2020).