Data Markets with Dynamic Arrival of Buyers and Sellers

Dmitry Moor
University of Zurich
Zurich, Switzerland
dmoor@ifi.uzh.ch

ABSTRACT

We propose a market design solution for a market for distributed data. The main challenges addressed by our solution are (1) different data providers produce different databases that can be joined to produce answers for users' queries; (2) data providers have high fixed costs for producing their databases; and (3) buyers and sellers can arrive dynamically to the market. Our design relies on using a Markov chain with states corresponding to different numbers of allocated databases. The transition probabilities between different states are governed by the payments suggested by the market platform to the data providers. The main challenge in this setting is to guarantee *dynamic incentive compatibility*, i.e., to ensure that buyers and sellers are not incentivized to arrive late to the market or to misreport their costs or values. To achieve this, we disentangle the payments suggested by the market platform to the sellers from the posted prices exposed to the buyers. We prove that the buyer-optimal payments that are exposed to sellers are non-increasing which prevents late arrivals of sellers. Further, we demonstrate that the posted prices exposed to buyers constitute a martingale process (i.e., late arrivals lead to the same expected price). Finally, we show that our design guarantees zero expected average budget deficit and we perform a number of simulations to validate our model.

CCS CONCEPTS

• **Applied computing** → **Economics**; *Marketing*;

KEYWORDS

Market Design, Data Markets, Dynamic Markets

ACM Reference Format:
Dmitry Moor. 2019. Data Markets with Dynamic Arrival of Buyers and Sellers. In *The 14th Workshop on the Economics of Networks, Systems and Computation (NetEcon'19), June 28, 2019, Phoenix, AZ, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3338506.3340270

1 INTRODUCTION

Many datasets on the Web are unstructured. This means that they can be easily interpreted by humans but not by machines. Imposing some structure on the data by publishing it as a database and linking it to other databases can help machines to make sense of the content of the data. This significantly reduces the effort of humans for the search, analysis and making predictions based on this data by delegating many of these tasks to the machine. This naturally results in great benefits for society (see Bernstein et al. [2016]).

The technology for producing and querying such structured and distributed data already exists and used in numerous areas (e.g., [W3C 2014]). Despite of all its potential benefits, this technology is not highly utilized. One of the reasons for that is the lack of financial incentives of data providers to publish their data in a structured format. This happens because the high fixed costs that the data providers incur for producing their databases, structuring the data and linking it to the datasets of other data providers can never be recouped, [Moor et al. 2019]. As a result, a different system of incentives is required to compensate the data providers. In this paper, we propose such a system by designing a market for distributed data.

1.1 Call for Data Markets

In recent years, there were numerous attempts to design a market for data. Koutris et al. [2015] aim at designing a market for selling different views of a database while satisfying a *no-arbitrage* constraint. However, their approach does not easily extend to domains when users join data produced by multiple data providers.

Moor et al. [2015, 2019] and Agarwal et al. [2019] emphasize the importance of joining data coming from different data providers. They argue that the combinatorial preferences of buyers is a crucial feature for data markets as many databases can complement each other. As a result, the buyer who can access more databases gets a more precise and thus, valuable answer for his query.

However, none of these studies consider the dynamics of the data market. While in many combinatorial markets the dynamics may not play a critical role, data markets are inherently dynamic.[1] This means that both buyers and sellers in these markets arrive regularly and can strategically delay their arrivals if they expect to be better off by doing so. Due to the combinatorial nature of preferences of buyers such delays can have a dramatic effect on the operation of the market. Indeed, the late arrivals of sellers may result in a very low surplus reached by the buyers who arrive earlier and thus, can access only very few databases. In our work, we focus on both of these aspects, i.e., on the complementary nature of the data and on the dynamics of the market.

1.2 Overview of our approach

In this paper, we propose a model for a dynamic data market. We focus on the following challenges: (1) the data providers have high

[1] For example, combinatorial spectrum auctions typically happen once in several years (Cramton [2013]). Within this time frame the technology can change dramatically making it impractical for the bidders to misreport their bids based on the expected outcome of one of the future auctions.

NetEcon'19, June 28, 2019, Phoenix, AZ, USA

ACM ISBN 978-1-4503-6837-7/19/06…$15.00
https://doi.org/10.1145/3338506.3340270

fixed costs for producing their databases; (2) the databases can be complementary for the buyers, i.e., joining two databases generates some additional value for the buyers; (3) buyers and sellers arrive to the market over time and can strategically decide when to arrive.

We adopt a similar approach as proposed by Moor et al. [2019], i.e., we design a market that aims at optimizing the surplus of buyers while guaranteeing that the sellers' costs for producing their databases are compensated. In contrast to [Moor et al. 2019], we design a market that uses posted prices for both sellers and buyers. The rationale for this design decision is twofold. On the one hand, the market with posted prices has a very simple interface for possibly non-sophisticated buyers and sellers. On the other hand, the restriction of using the posted prices results in a much simpler strategic behavior of buyers and sellers. Indeed, in this case, they do not have to compute their optimal bid but can simply respond to the proposed posted prices.[2]

The market platform in our market plays the role of a *regulator*, i.e., it decides on which sellers to allocate, which queries to execute and how much the buyers need to pay to the sellers. Thus, on the one hand, the market platform computes payments for the sellers and allows the sellers to respond to these payments. If the seller's cost is smaller than the proposed payment, then the seller gets allocated, i.e., she creates and delivers her database to the market platform. On the other hand, the market platform computes the posted prices (per query) that are exposed to the buyers. The market platform then receives the buyers' queries, executes them, collects the respective amounts of money from the buyers and transfers them to the sellers.

We demonstrate how to compute the payments suggested to the sellers and the posted prices for the buyers in a way that neither sellers nor buyers have an incentive to strategically delay their arrival or misreport their costs or values. This guarantees *dynamic incentive compatibility*. Furthermore, we argue that while the traditional notion of *budget balancedness* is incompatible with dynamic incentive compatibility, our market design still satisfies *zero expected average budget deficit*. In other words, we show that, as the number of databases grows, the expected budget deficit per seller decreases to zero. Finally, we validate our approach via simulations.

2 FORMAL MODEL

We assume that time is discrete and we consider an infinite horizon problem where $t = 0, 1, 2, \ldots, \infty$ are the consecutive time steps. We let $N \in \mathbb{N}$ be the maximum number of databases that the market platform can allocate.

Sellers. Data providers arrive to the market independently at different time steps. At every time step at most one data provider with the new database can arrive with probability r.[3] Each data provider can produce a single database.

We let $\theta_i = \langle a_i, c_i \rangle$ be the type of the data provider i. Here, $a_i \in \mathbb{N}$ is the arrival time of the data provider, i.e., the time step when the data provider obtains her data; c_i is the fixed cost that the data provider incurs for producing the database out of her data. We assume that all c_i are drawn independently from the cumulative distribution $F(c)$, $f(c)$ is the corresponding density function. We assume that c_i includes mainly the *labor* cost for producing the database, i.e., costs for setting up the database, structuring the data, linking the data against other existing databases, etc.[4] We assume that θ_i is a private knowledge of the data provider and let $\hat{\theta}_i = \langle \hat{a}_i, \hat{c}_i \rangle$ be the reported type of the data provider.

Consider the data provider i who obtains her data at time t, i.e., $a_i = t$. This data provider can decide to structure her data and to produce a database. We assume that the database can be produced immediately after the data provider gets her data. As the data provider is strategic, she can decide to deliver her database to the market at a different *reported arrival time* $\hat{a}_i \neq a_i$ if she expects to be better off by doing so. We impose the following assumption on early arrivals:

Assumption 1 (Sellers' Limited Misreports). *For every data provider i it must hold $\hat{a}_i \geq a_i$.*

This assumption is not too restrictive as data providers cannot produce and deliver their databases before they obtain the actual data (which happens at time t).

Let $X_t \in \mathbb{N}$ denote the number of databases allocated at time t, $X_0 = 0$. Also, let $p(X_t)$ be the payment that the market platform is willing to pay for the new database when $X_t - 1$ databases have already been allocated. Then,

$$X_{t+1} = X_t + \sum_{\hat{\theta}_i : \hat{a}_i = t+1} \mathbb{1}\{\hat{c}_i \leq p(X_t + 1)\}.$$

Informally, this means that a new database is allocated at time $t + 1$ if there is an arrival of a new data provider at time $t + 1$ and the cost of the data provider is not larger than the payment $p(X_t + 1)$ proposed by the market platform.

We assume that data providers have quasi-linear utility functions, i.e., the present value of the utility of the data provider i who obtains her data at time a_i but decides to deliver it at time $\hat{a}_i$ is $u_i(\theta_i, \hat{\theta}_i) = -c_i + \delta^{\hat{a}_i - a_i} p(X_{\hat{a}_i})$; here $p(X_{\hat{a}_i})$ is the payment paid by the market platform to the data provider; $\delta \in (0, 1)$ is the constant discount rate for money. [5]

Buyers. Generally speaking, at every time step multiple buyers with different queries can arrive. Each buyer is willing to pay a certain amount of money for an answer for his query. To keep our model simple, instead of considering the demand of each buyer separately, we consider an aggregate demand of all buyers. In other words, we assume that at every time step there is a single risk-neutral *aggregate* buyer willing to get an answer for his question by submitting a query. In what follows, we will always refer to the aggregate buyer as simply a "buyer".

A buyer who arrives with his question at time t can strategically submit his query late at time $\hat{t} \neq t$ if he expects to be better off by doing so. We assume that the buyer cannot submit his query before he gets his question to ask:

Assumption 2 (Buyers' Limited Misreports). *Buyers cannot arrive earlier, i.e., $\hat{t} \geq t$.*

[2] In what follows we will use the word (posted) *payments* for sellers and *posted prices* for buyers.

[3] In practice, this can be achieved by making time intervals small enough. Considering a continuous time model with a Poisson arrival process is a possible future extension.

[4] We assume zero marginal costs, i.e., the electricity costs, the costs of maintaining the data, etc.

[5] Notice, that the sellers discount only their future payments but not the costs. This follows from the fact that these are the "labor" costs and must be indexed over time with the same rate δ (i.e., c_i is the constant present value of the future labor costs).

In this setting, the instantaneous utility of the buyer who gets his question at time t but submits his query at time $\hat{t}$ is $\mathcal{U}_t(\hat{t}) = \gamma^{\hat{t}-t}\big(V(X_{\hat{t}}) - \tau_{\hat{t}}(X_{\hat{t}})\big)$, where $\gamma \in [0, \delta)$ is the discount factor for the buyer's utility;[6] $\tau_{\hat{t}}(X_{\hat{t}})$ is the posted price faced by the buyer at time $\hat{t}$ if $X_{\hat{t}}$ databases are allocated. Observe that in our setting, the posted prices $\tau_t(X_t)$ depend on the number of allocated databases and thus, constitute a stochastic process (see Section 4 for more details). The expected value $V(.)$ of the buyer for the answer for his query depends on the number of allocated databases X_t. We assume that $V(.)$ is concave and strictly increasing. This reflects the fact that the larger is the number of available databases, the more informative (and thus, valuable) an answer for the buyer's query can be. Furthermore, the marginal value of an additional database becomes smaller as the number of allocated databases grows. Thus, such a shape of $V(.)$ captures the complementarity aspect of the buyers' preferences and the diminishing value of additional databases. Important here is that all databases are assumed to be *homogeneous*, i.e., they have similar values for possibly different groups of individual buyers. This assumption excludes the "junk" data, i.e., the data that has no value for any individual buyer. We elaborate on this value model and show how such an aggregate buyer can be constructed in Appendix B. We also assume that $V(.)$ is known by the market platform.[7]

Remark 1. *In practice, the value of each individual (not aggregate) buyer for his query can depend not only on the number of allocated databases X_t but also on the identities of those databases. While these preferences of individual buyers may be very diverse (and generally unknown), the aggregate preferences are typically much simpler to predict. This idea was discussed by Bakos and Brynjolfsson [1999] who suggested bundling of information goods as a way to obtain consumers' valuations for those goods. With this interpretation, in our model the buyers pay for an access to a bundle of databases. Under mild assumptions one can let the value of the buyers for such an access be concave and strictly increasing in the number of databases in the bundle.*

Market Platform. Our design relies on modeling the dynamics of the market via a Markov chain. The states of this Markov chain correspond to different numbers of allocated databases. The transition probabilities are defined by the arrival rates of the sellers and the payments suggested by the market platform to the sellers. To compute these payments, we adopt a similar approach as proposed by Moor et al. [2019], i.e., we aim at optimizing the total expected future discounted utility of buyers while guaranteeing that the fixed costs of the allocated sellers are compensated. The rationale for such a market design objective comes from the fact that in data markets, the sellers can be "monopolists" for their data. Thus, the market platform should play the role of a *regulator* that prevents the rent extracting behavior of the sellers (see Moor et al. [2019]).

Formally, we can think about our market platform as a Markov chain with $N+1$ states. A state is characterized by the number of databases being allocated at this state. Assume that at time t the market platform is in the state $X_t \in \{0, 1, ..., N\}$. At this state, the market platform announces the payment $p(X_t + 1)$ for the seller arriving next. Data providers observe the proposed payments and decide whether to produce their databases. We set explicitly $p(N+1) = 0$ to indicate that in the terminal state, no further databases can be allocated.

We impose a number of constraints on our market design.

Definition 1 (DIC for Sellers). *The mechanism is **dynamic incentive compatible for sellers** if for any seller i and $\forall\theta_i, \hat{\theta}_i$ that satisfy Assumption 1 we have $u_i(\theta_i, \theta_i) \geq \mathbb{E}_{X_{\hat{a}_i}}[u_i(\theta_i, \hat{\theta}_i)|X_{a_i}]$.*

In words, we say that the mechanism is dynamic incentive compatible for sellers, if neither seller can expect to get a higher utility at any of the future states $X_{\hat{a}_i}$ by misreporting her cost or by delaying her arrival.

Definition 2 (DIC for Buyers). *The mechanism is **dynamic incentive compatible for buyers** if $\exists t^* > 0$ s.t., for any $t \geq t^*$ we have $\mathcal{U}_t(t) \geq \mathbf{E}_{X_{\hat{t}}}[\mathcal{U}_t(\hat{t})|X_t]$ for any $\hat{t}$ that satisfy Assumption 2.*

In words, we say that the mechanism is dynamic incentive compatible for buyers if once the market gets sufficiently large (i.e., many databases are available), the buyers cannot expect to get a higher utility by delaying their arrival. The latter definition rules out some corner cases that can occur when the market just starts operating, i.e., during the interval $[0, t^*]$ when only very few databases are available.

Given these design constraints, we can now formally define the transitions of the Markov chain. Let i, j be the states of the Markov chain and let $P = [P_{ij}]_{(N+1)\times(N+1)}$ be the stochastic transition matrix of this Markov chain with

$$P_{ij} = \begin{cases} rF\big(p(i+1)\big), & \texttt{if } j = i+1 \\ 1 - rF\big(p(i+1)\big), & \texttt{if } j = i \\ 0, & \texttt{otherwise.} \end{cases} \quad (1)$$

Thus, the Markov chain transitions from the state i to the state $i+1$ if there is an arrival of a seller (with probability r) and the cost of the seller is not larger than the payment proposed to this seller (which happens with probability $F\big(p(i+1)\big)$). Let $P^n = P \cdot P^{n-1}$, $n = 2, 3, ...$

A commonly used property of *budget balancedness* (see, e.g., Mas-Colell et al. [1995]) can be informally stated as follows: A mechanism is budget balanced if the total amount of money paid to the sellers net the total amount collected from the buyers is equal to zero. Observe that in our setting, the notion of budget balancedness is not compatible with DIC for Buyers. Indeed, assume that at some time step the posted price for the buyer is $\tau_0 > 0$ and all N databases are already allocated. If at some time step $t^* > 0$ the mechanism is budget balanced, then for any $\epsilon > 0$ and for any $t \geq t^*$ the amount of money that should be collected from buyers is smaller than ϵ. Let us choose $\epsilon < \tau_0$. Then, the posted price at any time $t \geq t^*$ must be smaller than ϵ and consequently, smaller than τ_0. Thus, the buyer who does not discount the future strongly (i.e., $\gamma \approx 1$) would always prefer to wait until t^* to submit his query. This violates the DIC for Buyers.

Thus, instead of focusing on the traditional notion of budget balancedness, we aim at achieving *zero expected average budget*

[6] Buyers are typically not willing to wait for a long time before getting their queries answered. Consequently, γ is normally much smaller than δ. We also assume that γ is a common knowledge.

[7] Similarly to [Moor et al. 2019], the buyers' side of the market is thick and one can easily sample buyers to learn their valuations. In practice, such learning can be performed by the market platform by iteratively updating its belief about $V(.)$ when observing the responses of the buyers for the posted prices. The design of the respective learning procedure, however, is outside the scope of this paper.

deficit, i.e., we show that the *shortfall* per seller decreases as the number of databases increases. We can define this property formally in the following way: Let $\tilde{p}(t)$ be the present value (at time t) of all the past payments that have been already paid to the allocated sellers up to time t. Similarly, let $\tilde{\tau}(t)$ be the present value of all the payments made by the buyers up to time t. Thus, the *budget deficit* at time t can be defined as $BD(t) = \tilde{p}(t) - \tilde{\tau}(t)$.

Definition 3 (Expected Budget Deficit). *The expected budget deficit is*

$$\mathbb{E}[BD] = \lim_{t\to\infty} \mathbb{E}_{\theta_i}[BD(t)],$$

where the expectation is over different types of sellers θ_i.

In words, the expected budget deficit is equal to the expected residual amount of money that even in the limit cannot be collected from the buyers to fully compensate the sellers. Zero average expected budget deficit requires that this loss per-seller becomes negligibly small as the market grows, i.e.,

Definition 4 (Zero Expected Av. Budget Deficit). *The mechanism has **zero average expected budget deficit** if*

$$\lim_{N\to\infty} \frac{\mathbb{E}[BD]}{N} = 0.$$

The mechanism we propose maximizes the expected surplus of buyers and, thus, must be individually rational for buyers in expectation. It is also individually rational for sellers as they can always opt-out if the proposed payment is smaller than their cost.

3 COMPUTING PAYMENTS TO SELLERS

Remember that the market platform maximizes the total expected future discounted surplus of buyers. Let v_k^* be the maximal expected total future discounted surplus of buyers when k databases are already allocated. Consider the Bellman equations for the market platform:

$$v_N^* = \frac{V(N)}{1-\gamma} \tag{2}$$

$$v_{k-1}^* = \max_{p(k)} \Big\{ V(k-1) + \gamma r F\big(p(k)\big)\big(v_k^* - p(k)\big) + \Big(1 - rF\big(p(k)\big)\Big)\gamma v_{k-1}^* \Big\} \tag{3}$$

for $k = 1, ..., N$. Informally, the maximal expected total future discounted surplus of buyers in the state $k-1$ of the Markov chain is equal to the immediate "reward" in this state, i.e., $V(k-1)$, plus the discounted expected future maximal surplus in the next state. The latter depends on whether the Markov chain stays in the state $k-1$ (i.e., if no allocation happens) or if it transitions to the state k.

The first-order conditions imply

$$v_k^* - p^*(k) - v_{k-1}^* = \frac{F\big(p^*(k)\big)}{f\big(p^*(k)\big)}. \tag{4}$$

Now, we can rewrite

$$v_{k-1}^* = V(k-1) + r\gamma \frac{F^2(p^*(k))}{f(p^*(k))} + \gamma v_{k-1}^* \tag{5}$$

Equations (4) and (5) constitute a system of $2N$ non-linear equations with $2N$ unknowns.[8] The solution of these equations gives us the payments for sellers $p^*(k)$ at every state k of the Markov chains (along with the values v_k^*).

Now, we claim that the sellers have no incentive to arrive late. This follows from the fact that in such a setting the payments proposed by the market platform can only decrease with time. This proves dynamic incentive compatibility for sellers. The following theorem states this formally.

Definition 5. *We say that a distribution $f(c)$ is **strongly regular** if $\frac{F(c)}{f(c)}$ is monotone and strictly increasing.*

Theorem 1. *If $f(.)$ is strongly regular, then the mechanism is dynamic incentive compatible for sellers.*

Proof. We first show that if $f(.)$ is strongly regular, then $p(1), p(2), ...$ weakly decreases with time. The Bellman equations can be rewritten as follows:

$$v_{k-1}^* - v_{k-2}^* = \frac{1}{1-\gamma}\Big[V(k-1) - V(k-2) + \gamma r\Big(\frac{F^2\big(p^*(k)\big)}{f\big(p^*(k)\big)} - \frac{F^2\big(p^*(k-1)\big)}{f\big(p^*(k-1)\big)}\Big)\Big]$$

for $k = 1, ..., N$. Using Equation (4) we can rewrite:

$$p^*(k-1) + \frac{F\big(p^*(k-1)\big)}{f\big(p^*(k-1)\big)} + \frac{\gamma r}{1-\gamma}\frac{F^2\big(p^*(k-1)\big)}{f\big(p^*(k-1)\big)} = \frac{1}{1-\gamma}\Big[V(k-1) - V(k-2) + \gamma r \frac{F^2\big(p^*(k)\big)}{f\big(p^*(k)\big)}\Big].$$

By induction, we see that $p(N+1) = 0, p(N) > 0$; the l.h.s. is a strictly increasing function while the r.h.s. gets larger as $k \to 0$ (due to the concavity of $V(.)$ and the induction hypothesis $p^*(k) \geq p^*(k+1)$). Thus, the solution for $p^*(k-1)$ must also get larger as $k \to 0$, i.e., $p^*(k-1) \geq p^*(k)$.

Finally, for any $\theta_i, \hat{\theta}_i$ that satisfy Assumption 1 we have

$$u_i(\theta_i, \hat{\theta}_i) = -c_i + \delta^{\hat{a}_i - a_i} p(X_{\hat{a}_i}) \leq -c_i + \delta^{\hat{a}_i - a_i} p(X_{a_i}) \leq -c_i + p(X_{a_i}) = u_i(\theta_i, \theta_i).$$

Q.E.D. □

4 COMPUTING PRICES FOR BUYERS

Observe that if we set the posted prices for the buyer equal to the payments for sellers, i.e., $\tau_t(X_t) = p(X_t)$, then the mechanism cannot satisfy DIC for Buyers. Indeed, as we have shown in Theorem 1, the payments $p(X_t)$ can only decrease with time. In this case, the posted price would also only decrease. This would incentivize the buyers to arrive late which violates DIC for Buyers. Therefore, we need to disentangle the posted prices exposed to the buyer from the payments paid to the sellers. There are two main requirements to constructing such posted prices:

R1. The posted prices τ_t must guarantee DIC for Buyers;
R2. τ_t and $p(X_t)$ must satisfy zero expected average budget deficit.

In this section, we show how to construct a pricing scheme satisfying these two requirements.

First, the solution of Equations (4) and (5) allows us to compute the transition matrix P as defined in Equation (1). Now, let $\tilde{\pi}(k)$ denote the present value of the *future total payment* to sellers when

[8] We solve it with the Newton method.

the Markov chain is in the state k. Thus, for each k we can compute the expected value of $\tilde{\pi}(k)$ at this state as follows:

$$\mathbb{E}[\tilde{\pi}(k)] = \delta r F\big(p(k+1)\big)\big(p(k+1) + \mathbb{E}[\tilde{\pi}(k+1)]\big) + \delta\Big(1 - rF\big(p(k+1)\big)\Big)\mathbb{E}[\tilde{\pi}(k)], \quad \forall k = 0, 1, ..., N. \tag{6}$$

In words, if the Markov chain is in the state k, then two things can happen. Either an allocation happens, and therefore, the Markov chain transitions to the state $k+1$. In this case, the market platform must make an "immediate" payment $p(k+1)$ and expects to make a future payment of $\mathbb{E}[\tilde{\pi}(k+1)]$. Alternatively, no allocation happens. In this case, the market platform stays in the state k and expects to make a future payment of $\mathbb{E}[\tilde{\pi}(k)]$. Thus, the present value of the expected future total payment to the sellers in state k is equal to the discounted convex combination of the two aforementioned terms. Equations (6) constitute a system of $N+1$ linear equations with $N+1$ unknowns $\mathbb{E}[\tilde{\pi}(k)]$, $k = 0, ..., N$.

Now remember, that P^n_{ij} is the probability that the Markov chain transitions from the state i to the state j within n time intervals. Thus, to satisfy the requirement R1 at time $t = 0$ we must have

$$\tau_0(0) + \delta\big(P_{00}\tau_1(0) + P_{01}\tau_1(1) + ...\big) + \delta^2\big(P^2_{00}\tau_2(0) + P^2_{01}\tau_2(1) + P^2_{02}\tau_2(2) + ...\big) + ... = \mathbb{E}[\tilde{\pi}(0)]. \tag{7}$$

To satisfy the requirement R2 we compute the prices in a way that at any time step t and any allocation of databases X_t at time t, the expected future posted price at any possible future time interval is equal to the current posted price (i.e. to the posted price at time t). Thus, for $t = 0$ we set $P_{00}\tau_1(0) + P_{01}\tau_1(1) + ... = \tau_0(0)$, $(P^2_{00}\tau_2(0) + P^2_{01}\tau_2(1) + P^2_{02}\tau_2(2) + ...) = \tau_0(0)$ etc. for any $\hat{t} > t$. Now, we can simplify the Equation (7): $\tau_0(0) = (1-\delta)\mathbb{E}[\tilde{\pi}(0)]$. Generally, if at time t the Markov chain is in the state k, we set

$$\tau_t(k) = (1-\delta)\Big(\mathbb{E}[\tilde{\pi}(k)] + \tilde{p}(t) - \frac{\tilde{\tau}(t-1)}{\delta}\Big). \tag{8}$$

Finally, the overall mechanism looks as follows:

Dynamic Data Market Mechanism

Payments to Sellers: At time $t = 0$, solve Equations (4) and (5):

$$v^*_{k-1} = V(k-1) + r\gamma\frac{F^2(p^*(k))}{f(p^*(k))} + \gamma v^*_{k-1},$$

$$v^*_k - p^*(k) - v^*_{k-1} = \frac{F\big(p^*(k)\big)}{f\big(p^*(k)\big)}, \qquad k = 1, ..., N.$$

Here, $p^*(k)$ is the payment proposed by the market platform for the k'th database, $k = 1, ..., N$.

Allocation of Sellers: At each time $t > 0$, a seller with cost c_i may arrive and respond to $p^*(X_{t-1}+1)$. If $c_i \le p^*(X_{t-1}+1)$, the seller is allocated, $X_t = X_{t-1} + 1$. Otherwise, the seller is not allocated, $X_t = X_{t-1}$.

Posted Prices for Buyers: At each time t, the market platform computes the posted price for this interval according to Equation (8):

$$\tau_t(X_t) = (1-\delta)\Big(\mathbb{E}[\tilde{\pi}(X_t)] + \tilde{p}(t) - \frac{\tilde{\tau}(t-1)}{\delta}\Big).$$

The dynamic incentive compatibility for buyers follows from the following theorem.

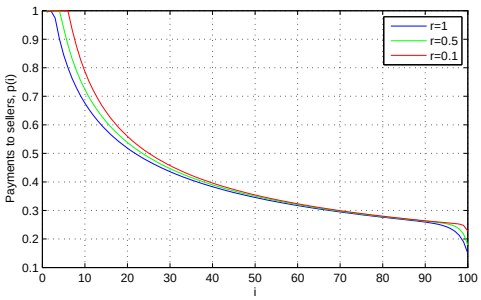

Figure 1: Payment for the i'th allocated database for different arrival rates r of sellers.

Theorem 2. *The process $\tau_t(X_t)$ is a martingale.*

Proof. See Appendix A. □

To complete the proof of DIC for Buyers, observe that $\gamma V(X_{t+1}) \le \gamma V(X_t + 1) = \gamma(V(X_t + 1) - V(X_t)) + \gamma V(X_t)$. From concavity of $V(.)$ it follows that as X_t gets sufficiently large, the difference $(V(X_t + 1) - V(X_t))$ gets small. This fact together with Theorem 2 proves that if the market is large enough, the buyers do not get more value from delaying their arrival. Formally,

$$\mathbf{E}[\boldsymbol{\mathcal{U}}_t(\hat{t} = t+1)|X_t] = \mathbf{E}[\gamma V(X_{t+1})|X_t] - \gamma\mathbf{E}[\tau_{\hat{t}}(X_{\hat{t}})|X_t] \le \underbrace{\gamma\mathbb{E}[V(X_t+1) - V(X_t)]}_{\texttt{Goes to 0 as } X_t \texttt{ grows.}} + \gamma\boldsymbol{\mathcal{U}}_t(t).$$

Finally, the following theorem shows that the proposed mechanism satisfies zero expected average budget deficit.

Theorem 3. *The dynamic data market mechanism has zero expected average budget deficit.*

Proof. See Appendix A. □

5 EXPERIMENTS

To validate our model, we carry out a number of simulations of the proposed market under different simulation scenarios. We assume that for all scenarios, the costs of sellers are drawn from the uniform distribution, $c_i \sim U[0, 1]$. We further assume that the value of the buyer is $V(X_t) = \sqrt{X_t}$. The discount rate is $\delta = 0.9$.

Payments for Sellers. First, we perform a simulation with $N = 100$ databases while varying r and γ. Figure 1 illustrates the payment $p(i)$ for the newly arriving database $i \le N$ when $(i-1)$ databases are already allocated. Here, we vary $r \in \{0.1, 0.5, 1.0\}$ while fixing $\gamma = 0.9$. In line with our results proved in Theorem 1, the payments decrease over time. From this figure, we also see that as the arrival rate r gets smaller, the market platform suggests higher payments to the sellers. This result follows form the fact that as the probability of arrival of a seller decreases, the opportunity cost of the market platform for waiting increases. Indeed, if at time t the seller does not arrive, and $X_t = X_{t-1}$, then the buyer enjoys a smaller value of $V(X_{t-1})$ instead of the value $V(X_{t-1}+1)$ he could have enjoyed if the seller arrived at time t and delivered her database. Thus, the market platform "loses" the possible higher value of the buyer and consequently, has a higher opportunity cost for waiting. Due to the increased opportunity costs, the market platform increases the payments.

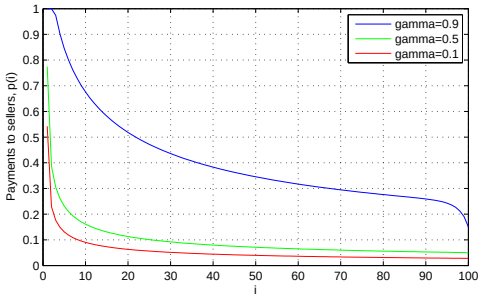

Figure 2: Payment for the i'th allocated database for different discount factors γ of buyers.

Now, let us look into the dependency of the payments to the sellers on the discount factor γ of the buyer. Figure 2 illustrates the payments of the market platform for the i'th database when $(i-1)$ databases are already allocated. Here, we fix $r = 1$ and vary $\gamma \in \{0.1, 0.5, 0.9\}$. The figure demonstrates that the stronger the buyer discounts the future, the smaller the payments proposed to the sellers by the market platform. The explanation of this phenomenon comes from a similar opportunity cost argument: Indeed, stronger discounting of the future value of the buyer decreases the opportunity cost of "losing" the future buyer's surplus. Thus, the opportunity cost of not allocating the seller now gets smaller. Consequently, the payments suggested by the market platform to the sellers must also decrease.

Posted Prices for the Buyer. We illustrate the posted prices exposed to the buyer by generating 10 trajectories corresponding to the process $\tau_t(X_t)$. To achieve this, we sample 10 different arrival scenarios and costs c_i. We set $\gamma = 0.5$, $r = 1$, $N = 100$. We then let the simulated sellers arrive to the market and respond to the suggested payments. At each time step t we compute the number of allocated databases X_t as well as the posted price $\tau_t(X_t)$ according to Equation (8). Figure 3 (top) illustrates the different trajectories corresponding to the martingale process of the posted price $\tau_t(X_t)$ while Figure 3 (bottom) demonstrates the respective trajectories of the process X_t. From comparing the Figure 3 (top) with the Figure 3 (bottom) we see that if an allocation does not happen at time t (i.e., the trajectory of X_t has a plateau), then the posted price $\tau_t(X_t)$ decreases. If an allocation happens at time t, then there is a respective spike in the posted price.

Expected Average Budget Deficit. We illustrate the convergence of the expected average budget deficit, $\frac{\mathbb{E}[BD]}{N}$, to zero as the number of allocated databases N grows. Figure 4 illustrates our findings. Here, we sample 1000 different trajectories corresponding to different arrivals and costs of sellers. We then compute the mean values and the standard errors of the resulting expected average budget deficit. As expected, the result goes in hand with our Theorem 3.

6 CONCLUSIONS

In this paper, we have studied the dynamics of the combinatorial data market. We proposed a mechanism that optimizes the expected future discounted surplus of buyers while compensating the fixed costs of allocated sellers and satisfying the two key properties: dynamic incentive compatibility and zero expected average budget deficit. We further studied the proposed mechanism in a simulation environment. Our results confirm our intuition regarding the

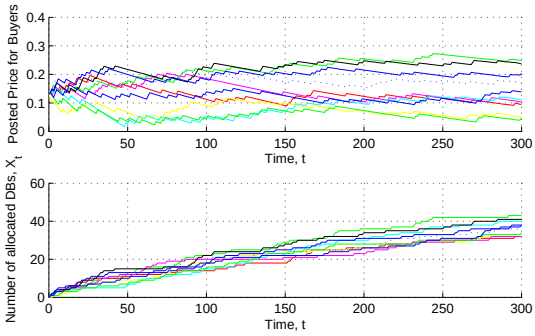

Figure 3: Trajectories of the posted price $\tau_t(X_t)$ (top) and the number of allocated databases X_t (bottom). For every trajectory X_t, the respective trajectory $\tau_t(X_t)$ is depicted with the same color.

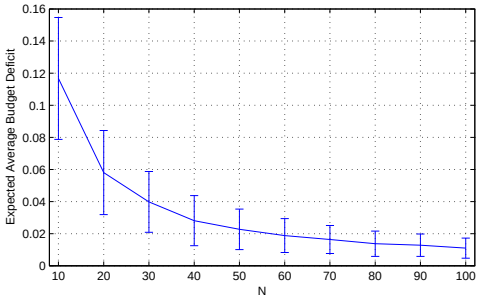

Figure 4: Expected Average Budget Deficit for different numbers of N. Here, $r = 1$, $\gamma = 0.5$.

changes in prices and in the budget deficit when slightly changing the parameters of the mechanism. In future work, we are planning to expand these simulations and to study a number of further economic properties of the proposed mechanism.

REFERENCES

Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. A marketplace for data: An algorithmic solution. Technical report, Massachusetts Institute of Technology, Working Paper, 2019.

Yannis Bakos and Erik Brynjolfsson. Bundling information goods: Pricing, profits, and efficiency. *Management Science*, 45(12):1613–1630, 1999. URL http://EconPapers.repec.org/RePEc:inm:ormnsc:v:45:y:1999:i:12:p:1613-1630.

Dirk Bergemann, Alessandro Bonatti, and Alex Smolin. The design and price of information. *American Economic Review*, 108(1):1–48, January 2018. doi: 10.1257/aer.20161079. URL http://www.aeaweb.org/articles?id=10.1257/aer.20161079.

Abraham Bernstein, James Hendler, and Natalya Noy. A new look at the semantic web. *Commun. ACM*, 59(9):35–37, August 2016. ISSN 0001-0782. doi: 10.1145/2890489. URL http://doi.acm.org/10.1145/2890489.

P. Cramton. Spectrum auction design. *Review of Industrial Organization*, 42(2):030–190, March 2013.

Paraschos Koutris, Prasang Upadhyaya, Magdalena Balazinska, Bill Howe, and Dan Suciu. Query-based data pricing. In *Journal of the ACM (JACM)*, volume 62, October 2015.

Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory*. Oxford University Press, New York, 1995.

Dmitry Moor. Data Markets with Dynamic Arrival of Buyers and Sellers. Working paper. https://www.ifi.uzh.ch/en/ce/people/moor.html, 2019.

Dmitry Moor, Tobias Grubenmann, Sven Seuken, and Abraham Bernstein. A double auction for querying the web of data. In *The Third Conference on Auctions, Market Mechanisms and Their Applications*, 2015.

Dmitry Moor, Sven Seuken, Tobias Grubenmann, and Abraham Bernstein. The design of a combinatorial data market. Technical report, Faculty of Business, Economics and Informatics, University of Zurich, Working Paper, 2019.

W3C. Linked Data. https://www.w3.org/standards/semanticweb/data, 2014.

A PROOFS

Lemma 1. *For any state $k = 0, \ldots, N$ the following inequality holds:*

$$\mathbb{E}[\tilde{\pi}(k)] - \mathbb{E}[\tilde{\pi}(k+1)] \leq p(k+1).$$

Proof. Follows from Equation (6). We can rewrite

$$\Big(1 + \frac{1-\delta}{\delta r F\big(p(k+1)\big)}\Big)\mathbb{E}[\tilde{\pi}(k)] - \mathbb{E}[\tilde{\pi}(k+1)] = p(k+1). \quad (9)$$

Here, $1 + \frac{1-\delta}{\delta r F\left(p(k+1)\right)} \geq 1$. Therefore,

$$\mathbb{E}[\tilde{\pi}(k)] - \mathbb{E}[\tilde{\pi}(k+1)] \leq p(k+1).$$

□

Lemma 2. *For any time $t > 0$ it holds $\tilde{p}(t) - \tilde{\tau}(t) > 0$.*

Proof. We proceed by induction. For $t = 1$, the payment $p(1)$ is maximal and the result holds, i.e., $BD(1) = p(1) - \frac{\tau_0(0)}{\delta} - \tau_1(\ell) > 0$. Consider an arbitrary time $t > 1$ and $X_t = \ell$. We have

$$\begin{aligned}
BD(t) =& \tilde{p}(t) - \tilde{\tau}(t) = \tilde{p}(t) - \frac{\tilde{\tau}(t-1)}{\delta} - \\
& (1-\delta)\Big(\mathbb{E}[\tilde{\pi}(\ell)] + \tilde{p}(t) - \frac{\tilde{\tau}(t-1)}{\delta}\Big) = \\
& \delta\tilde{p}(t) - \tilde{\tau}(t-1) - (1-\delta)\mathbb{E}[\tilde{\pi}(\ell)] = \\
& \delta\Big(\tilde{p}(t) - \frac{\tau(t-1)}{\delta}\Big) - (1-\delta)\mathbb{E}[\tilde{\pi}(\ell)] = \\
& \delta\Big(\frac{\tilde{p}(t-1)}{\delta} + p(X_t) - \frac{\tilde{\tau}(t-1)}{\delta}\Big) - (1-\delta)\mathbb{E}[\tilde{\pi}(\ell)] = \\
& BD(t-1) + \delta p(X_t) - (1-\delta)\mathbb{E}[\tilde{\pi}(\ell)].
\end{aligned}$$

Now, consider two cases: $p(X_t) = p(\ell)$ and $p(X_t) = p(\ell+1)$ (i.e., dependent on whether there is an allocation has happened at time t).

In the former case, using Equation (9) we can rewrite:

$$\begin{aligned}
BD(t) =& BD(t-1) + \Big(\delta + \frac{1-\delta}{rF\big(p(\ell)\big)}\Big)\mathbb{E}[\tilde{\pi}(\ell-1)] - \\
& \delta\mathbb{E}[\tilde{\pi}(\ell)] - (1-\delta)\mathbb{E}[\tilde{\pi}(\ell)] = \\
& BD(t-1) + \Big(\delta + \frac{1-\delta}{rF\big(p(\ell)\big)}\Big)\mathbb{E}[\tilde{\pi}(\ell-1)] - \mathbb{E}[\tilde{\pi}(\ell)] \\
& \geq BD(t-1) \geq 0.
\end{aligned}$$

In the latter case,

$$\begin{aligned}
BD(t) =& BD(t-1) + \Big(\delta + \frac{1-\delta}{rF\big(p(\ell)\big)}\Big)\mathbb{E}[\tilde{\pi}(\ell)] - \\
& \delta\mathbb{E}[\tilde{\pi}(\ell+1)] - (1-\delta)\mathbb{E}[\tilde{\pi}(\ell)] = \\
& BD(t-1) + \Big(\delta + \frac{1-\delta}{rF\big(p(\ell)\big)} - (1-\delta)\Big)\mathbb{E}[\tilde{\pi}(\ell)] - \\
& \delta\mathbb{E}[\tilde{\pi}(\ell+1)] \geq BD(t-1) \geq 0.
\end{aligned}$$

Q.E.D. □

Theorem 3 1. *The mechanism has zero expected average budget deficit.*

Proof. Let us consider the budget deficit at time t, i.e., $BD(t) = \tilde{p}(t) - \tilde{\tau}(t)$. We know that $BD(0) = 0 - \tilde{\tau}(0) = -(1-\delta)\mathbb{E}[\tilde{\pi}(0)]$. The expected budget deficit at time $t = 1$ is

$$\begin{aligned}
\mathbb{E}[BD(1)] =& rF\big(p(1)\big)\Big(p(1) - \tau_1(1)\Big) - \frac{\tau_0(0)}{\delta} - \Big(1 - rF\big(p(1)\big)\Big)\tau_1(0) = \\
& rF\big(p(1)\big)\Big(p(1) - \tau_1(1) + \tau_1(0)\Big) - \frac{\tau_0(0)}{\delta} - \tau_1(0).
\end{aligned}$$

Observe, that $\tau_1(1) = \tau_1(0) + (1-\delta)(p(1) + \mathbb{E}[\tilde{\pi}(1)] - \mathbb{E}[\tilde{\pi}(0)])$. Thus, we can rewrite

$$\begin{aligned}
\mathbb{E}[BD(1)] =& rF\big(p(1)\big)\Big(p(1) - (1-\delta)(p(1) + \mathbb{E}[\tilde{\pi}(1)] - \mathbb{E}[\tilde{\pi}(0)])\Big) - \\
& \frac{\tau_0(0)}{\delta} - \tau_1(0) = \\
& rF\big(p(1)\big)\Big(\mathbb{E}[\tilde{\pi}(0)] - \mathbb{E}[\tilde{\pi}(1)]\Big) + \delta rF\big(p(1)\big)\Big(p(1) + \mathbb{E}[\tilde{\pi}(1)]\Big) - \\
& \delta rF\big(p(1)\big)\mathbb{E}[\tilde{\pi}(0)] - \frac{\tau_0(0)}{\delta} - \tau_1(0) = \\
& rF\big(p(1)\big)\Big(\mathbb{E}[\tilde{\pi}(0)] - \mathbb{E}[\tilde{\pi}(1)]\Big) + \delta rF\big(p(1)\big)\Big(p(1) + \mathbb{E}[\tilde{\pi}(1)]\Big) + \\
& \delta\Big(1 - rF\big(p(1)\big)\Big)\mathbb{E}[\tilde{\pi}(0)] \\
& -\delta\mathbb{E}[\tilde{\pi}(0)] - \frac{\tau_0(0)}{\delta} - (1-\delta)\Big(\mathbb{E}[\tilde{\pi}(0)] - \frac{\tau_0(0)}{\delta}\Big) \\
& = rF\big(p(1)\big)\Big(\mathbb{E}[\tilde{\pi}(0)] - \mathbb{E}[\tilde{\pi}(1)]\Big) + BD(0).
\end{aligned}$$

From Lemma 1 it follows that $\mathbb{E}[BD(1)] \leq BD(0) + rF\big(p(1)\big)p(1)$. Now, consider the expected budget deficit at time $t > 1$:

$$\begin{aligned}
\mathbb{E}[BD(t)] =& \mathbb{E}[BD(t-1)] + \\
& \sum_{\ell} \Pr(state = \ell)\Big[rF\big(p(\ell+1)\big)\big(p(\ell+1) - \tau_t(\ell+1)\big) + \\
& \Big(1 - rF\big(p(\ell+1)\big)\Big)\big(-\tau_t(\ell)\big)\Big].
\end{aligned}$$

Observe, that

$$\tau_t(\ell+1) = \tau_t(\ell) + (1-\delta)\Big(\mathbb{E}[\tilde{\pi}(\ell+1)] - \mathbb{E}[\tilde{\pi}(\ell)] + p(\ell+1)\Big).$$

Thus, we can rewrite

$$\begin{aligned}
& rF\big(p(\ell+1)\big)\Big(p(\ell+1) - \tau_t(\ell+1)\Big) + \\
& \Big(1 - rF\big(p(\ell+1)\big)\Big)(-\tau_t(\ell)) = \\
& rF\big(p(\ell+1)\big)\Big(p(\ell+1) - \tau_t(\ell+1) + \tau_t(\ell)\Big) - \tau_t(\ell) = \\
& rF\big(p(\ell+1)\big)\Big(\delta p(\ell+1) - (1-\delta)(\mathbb{E}[\tilde{\pi}(\ell+1)] - \mathbb{E}[\tilde{\pi}(\ell)])\Big) \\
& \qquad - \tau_t(\ell) =
\end{aligned}$$

$$rF\big(p(\ell+1)\big)\Big(\mathbb{E}[\tilde{\pi}(\ell)] - \mathbb{E}[\tilde{\pi}(\ell+1)]\Big) +$$
$$\delta rF\big(p(\ell+1)\big)\Big(p(\ell+1) + \mathbb{E}[\tilde{\pi}(\ell+1)]\Big) +$$
$$\delta\Big(1 - rF\big(p(\ell+1)\big)\Big)\mathbb{E}[\tilde{\pi}(\ell)] - \delta\mathbb{E}[\tilde{\pi}(\ell)] - \tau_t(\ell) =$$
$$rF\big(p(\ell+1)\big)\Big(\mathbb{E}[\tilde{\pi}(\ell)] - \mathbb{E}[\tilde{\pi}(\ell+1)]\Big) +$$
$$\mathbb{E}[\tilde{\pi}(\ell)](1-\delta) - \tau_t(\ell) \leq$$
$$rF\big(p(\ell+1)\big)\Big(\mathbb{E}[\tilde{\pi}(\ell)] - \mathbb{E}[\tilde{\pi}(\ell+1)]\Big) - \frac{1-\delta}{\delta}BD(t-1) <$$
$$rF\big(p(\ell+1)\big)\Big(\mathbb{E}[\tilde{\pi}(\ell)] - \mathbb{E}[\tilde{\pi}(\ell+1)]\Big).$$

Here, the last inequality follows directly from Lemma 2. Now, we can rewrite

$$\mathbb{E}[BD(t)] < \mathbb{E}[BD(t-1)] +$$
$$\max_{\ell}\Big\{rF\big(p(\ell+1)\big)\Big(\mathbb{E}[\tilde{\pi}(\ell)] - \mathbb{E}[\tilde{\pi}(\ell+1)]\Big)\Big\} \leq$$
$$\mathbb{E}[BD(t-1)] + \max_{\ell}\Big\{rF\big(p(\ell+1)\big)p(\ell+1)\Big\}.$$

Which implies that the expected budget deficit grows slower than linearly. Thus,

$$\lim_{N\to\infty}\frac{\mathbb{E}[BD(t)]}{N} = 0.$$

Q.E.D. □

Theorem 2 1. *The process $\tau_1(X_1), \tau_2(X_2), \ldots$ is a martingale.*

Proof. Let $X_t = s$. We want to show that $\mathbb{E}[\tau_{t+1}(\ell)|s] = \tau_t(s)$. Precisely,

$$\mathbb{E}[\tau_{t+1}(\ell)|s] = rF\big(p(s+1)\big)\tau_{t+1}(s+1) +$$
$$\Big(1 - rF\big(p(s+1)\big)\Big)\tau_{t+1}(s) =$$
$$(1-\delta)\Big[rF\big(p(s+1)\big)\Big(\mathbb{E}[\tilde{\pi}(s+1)] + \frac{\tilde{p}(t)}{\delta} + p(s+1) - \tilde{\tau}(t)\Big)$$
$$+ \Big(1 - rF\big(p(s+1)\big)\Big)\Big(\mathbb{E}[\tilde{\pi}(s)] + \frac{\tilde{p}(t)}{\delta} - \tilde{\tau}(t)\Big)\Big] =$$
$$(1-\delta)\Big[rF\big(p(s+1)\big)\Big(\mathbb{E}[\tilde{\pi}(s+1)] + p(s+1)\Big) +$$
$$\Big(1 - rF\big(p(s+1)\big)\Big)\mathbb{E}[\tilde{\pi}(s)] + \frac{\tilde{p}(t)}{\delta} - \tilde{\tau}(t)\Big] =$$
$$(1-\delta)\Big[\frac{\mathbb{E}[\tilde{\pi}(s)]}{\delta} + \frac{\tilde{p}(t)}{\delta} - \tilde{\tau}(t)\Big] =$$
$$\frac{1-\delta}{\delta}\Big[\mathbb{E}[\tilde{\pi}(s)] + \tilde{p}(t) - \delta\tilde{\tau}(t-1) - \tau_t(s)\Big] =$$
$$\frac{1-\delta}{\delta}\Big[\frac{\tau_t(s)}{1-\delta} - \tau_t(s)\Big] = \tau_t(s).$$

Q.E.D. □

B VALUE MODEL

We assume that buyers can acquire data to make certain predictions about the state of the world. If the prediction of a buyer is good, then he gets a high reward $R_H \in \mathbb{R}^+$. Otherwise, the buyer receives a low reward $R_L \in \mathbb{R}^+$, $R_L < R_H$.[9] For simplicity, we assume that the reward is the same for all buyers. Due to the inherent uncertainty about the world, a buyer without any additional information faces a lottery in which he can obtain the high reward with $\Pr(R_H) = \tilde{P}_0$ or the low reward with $\Pr(R_L) = 1 - \tilde{P}_0$. Let R_N denote the expected reward of buyers when N databases are allocated. Thus, the expected reward of the buyer who does not acquire any data is

$$R_0 = \tilde{P}_0 R_H + (1 - \tilde{P}_0)R_L. \tag{10}$$

We assume that if the buyer can make a better prediction about the state of the world, then he has a higher chance to receive the high reward. Thus, in order to improve his prediction, the buyer can purchase data by submitting a query against one or multiple databases.

We let $\tilde{P}_i$ be the probability that $i = 0, 1, 2, \ldots$ databases joined together allow the buyer to make a good prediction of the state of the world (and consequently, to receive the high reward R_H). Thus, the fraction of buyers who can receive the high reward R_H by querying a single available database is $\tilde{P}_1$. Similarly, the fraction of buyers who are interested in joining exactly two available databases is $\tilde{P}_2$, etc. Implicit here is the *homogeneity* assumption, i.e., the assumption that this probability does not depend on identities of the databases. For example, if there are two databases available for the buyers, then the fraction $\tilde{P}_1$ of all buyers can receive the high reward by querying only against the first database, and the fraction $\tilde{P}_1$ of buyers can get the high reward by querying only against the second database. Naturally, these two subsets can overlap if the data is substitutable for the buyers.[10] However, we restrict our attention to the case when the queries themselves are *fixed*. This means that if the buyer wants to join two databases, he can only decide on the databases to join, but cannot decide to join three databases. Notice also, that typically the number of different databases N is much larger than the number of databases relevant for answering each buyer's query.

In what follows, we show that as the number of available databases N increases, the aggregate value of buyers for a yet another database also increases. Generally speaking, allocating an additional complementary database may result in either concave or convex aggregate value function $V(.)$. However, we demonstrate that for larger numbers of N, the aggregate value function $V(.)$ must be concave. This follows from the fact, that as N grows, the buyers still typically join only a very small number of databases compared to N. Therefore, there is no exponential growth in the values of buyers for answers for their queries *on average*.

Joining up to two databases, $i \leq 2$. First, consider the case when there is only a single database available for buyers (i.e., $N = 1$). In this case, $R_1 = \tilde{P}_1 R_H + (1 - \tilde{P}_1)R_0$. Consequently, the willingness of the buyers to pay for this database is

$$\omega(1) = R_1 - R_0 = \tilde{P}_1(R_H - R_0). \tag{11}$$

Let us now consider the case when buyers can access two databases (i.e., $N = 2$). In this case, the probability to get the high reward is $\Pr(R_H) = (2\tilde{P}_1 - \tilde{P}_1^{\,2}) + \tilde{P}_2$, where the first term corresponds to the fraction of buyers who submit queries against a single database, and the second term corresponds to the fraction of buyers who need to join both databases to make a good prediction. Thus,

[9] This is similar to the model of Bergemann et al. [2018].

[10] This model obviously excludes the case of the "junk" data, i.e., the data that is not valued by any buyer.

$R_2 = (2\tilde{P}_1 - \tilde{P}_1^{\,2} + \tilde{P}_2)R_H + \big(1 - (2\tilde{P}_1 - \tilde{P}_1^{\,2} + \tilde{P}_2)\big)R_0$. Therefore, the willingness of buyers to pay for the second database is

$$\omega(2) = R_2 - R_1 = \big(\tilde{P}_1(1 - \tilde{P}_1) + \tilde{P}_2\big)(R_H - R_0). \tag{12}$$

Observe, that if $\tilde{P}_2 \leq \tilde{P}_1^2$, then $\omega(2) \leq \omega(1)$. Intuitively, this means that if the two databases are not strongly complementary, the value function $V(.)$ is still concave even for small N. However, if the two databases are strong complements (i.e., $\tilde{P}_2 > \tilde{P}_1^2$), then concavity of $V(.)$ can be violated for small N. Nevertheless, as we show next, the value function $V(.)$ is still concave for large N.

Generally, for any $k > 2$ we have

$$\omega(k) = \left[\tilde{P}_1(1 - \tilde{P}_1)^{k-1} + (1 - \tilde{P}_2)^{C^2_{k-1}}\tilde{P}_2 \sum_{j=1}^{k-1}(1 - \tilde{P}_2)^{j-1}\right](R_H - R_0). \tag{13}$$

The intuition behind this formula is as follows. If there are already $k-1$ databases allocated, then there are two kinds of buyers who can benefit from allocating the k'th database. The first kind corresponds to the buyers who are willing to access only the new k'th database and who have not benefited from accessing any of the previously allocated $k-1$ databases. The fraction of such buyers is $\tilde{P}_1(1 - \tilde{P}_1)^{k-1}$. The second kind corresponds to the buyers who want to join two databases and who could not benefit from joining any two of $k-1$ already allocated databases. The fraction of such buyers is $(1 - \tilde{P}_2)^{C^2_{k-1}}\tilde{P}_2 \sum_{j=1}^{k-1}(1 - \tilde{P}_2)^{j-1}$. Naturally, the buyers of both kinds are willing to pay up to $R_H - R_0$ for the new database.

From Equation (13) it follows that the willingness to pay for the k'th database decreases as k increases. Indeed, the chance that none of the available individual databases are helpful for the buyers gets smaller as the number of databases increases. Observe also that there are $C^2_{k-1} = (k-1)(k-2)/2$ possible ways to join two out of $k-1$ databases that are already available. Allocating the k'th database leads to $k-1$ additional ways to join the data (and therefore leads to a higher potential reward). However, the chance that none of the already available C^2_{k-1} combinations is helpful for the buyers gets very small much faster than the linear growth due in the new combinations to join the data.

Joining up to ℓ databases, $i \leq \ell$. In this case, we have

$$\omega(k) = \sum_{i=1}^{\ell} \tilde{P}_i(1 - \tilde{P}_i)^{C^i_{k-1}} \cdot \sum_{j=1}^{C^{i-1}_{k-1}}(1 - \tilde{P}_i)^{j-1}. \tag{14}$$

The analysis performed in the previous paragraph extends to this setting.

C ACKNOWLEDGMENTS

The author would like to thank members of the Computation and Economics Research Group of the University of Zurich for the insightful discussions and the reviewers of NetEcon'19 workshop for their helpful comments. This research is supported by the SNSF (Swiss National Science Foundation) project #153598.