DATA SALES AND DATA DILUTION*

Ernest Liu† Song Ma‡ Laura Veldkamp§
Princeton & NBER Yale & NBER Columbia & NBER

Abstract

We explore indicators of market power in a data market. Markups cannot measure competition, because most data products' marginal cost is zero, making the markup infinite. Yet, data monopolists may not exert monopoly power because they cannot commit to restricting data sales to future customers. This limited commitment and strategic substitutability of data undermine sellers' monopoly power. But data subscriptions restore this monopoly power. Evidence from online data markets supports the model's insight that subscriptions indicate market power. Model and evidence reveal that data subscriptions are better for consumers because they sustain the incentive to invest in high-quality data.

Keywords. Market Power. Data markets. Data economy. Technological Change. Market Structure.

JEL. C6. D4. D5. L1.

*We thank Christopher Tonetti, Itay Goldstein, and Liyan Yang for discussing our paper, and Anton Korinek and participants at the FRA Conference, Econometric Society Meeting, and GSU-RFS FinTech Conference for helpful comments. We thank the Brookings Institution for financial support of this project.

†ernestliu@princeton.edu

‡song.ma@yale.edu

§lv2405@columbia.edu

One of the largest concerns that economists and policymakers have about the new digital economy is the market power of firms that sell data. The fact that data has a large fixed cost component and is free to replicate suggests the emergence of natural monopolies. However, little is known about how this market functions and prices data. We use theory to understand what indicators of market power to look for and collect new empirical evidence on data marketplaces to measure that market power and its welfare consequences.

It is not obvious how to identify market power in a market where every seller has a monopoly over their data set and where the marginal cost of producing additional data copies is zero. Therefore our empirical exploration of data markets needs to be guided by theory. We build a dynamic model of a monopolist data seller with two key features: Information that others know generates less value and sellers cannot commit not to sell more data in the future. These are realistic features of a data market. The first, commonly called the "strategic substitutability" of information (S. J. Grossman and J. E. Stiglitz 1980), arises in many settings where users of data choose quantities and market clearing determines the price. The second assumption, a classic commitment friction, is particularly relevant in data markets where one could easily transform data to make it non-identical, but functionally equivalent.

Our model teaches us that, even a data monopolist may have limited power to extract rent from their customers when the data seller cannot commit to a price schedule. The reason is that a data seller competes with its future self. If a data seller cannot commit not to sell the data to a firm's competitors, the firm's willingness to pay for the data declines. This force keeps data prices low. In this type of environment, we should worry less about the excessive profits of data monopolies and worry more about whether data is underprovided. Since we observe that many data producers sell data subscriptions, rather than data ownership, we add that feature to our model. We find that subscriptions for data allow a firm to re-capture much of its lost revenue from the lack of commitment.

But if data subscriptions allow firms to capture more surplus from consumers, why don't all firms sell data subscriptions? We use the model to identify three features of a data seller firm that make subscriptions less attractive to it: financial frictions, a small market, and high data depreciation.

Our theory thus provides us a way to understand the prevalence and force of monopoly power— it directs us to examine data sales models. Specifically, we should look for the prevalence of data sales versus data subscriptions and patterns in how these trade types are used. Therefore, how data is sold becomes the centerpiece of our empirical analysis.

To measure activity in data markets, we hand-collect a novel data set from Datarade, one of the largest online data marketplaces that connects buyers and sellers of data. The evidence about the geographic, industry, and data type coverage of this market place paints a nuanced picture of the way in which data is traded. Across over 3,600 data products, we find that 46% offer an option to buy the data for a one-time fee. However, over 90% offer a subscription or usage-based payment system. These fractions do not sum to one because many sellers offer multiple purchasing options. This finding suggests that at least half of all data providers have significant abilities to extract rents.

To test the predictions of our model, we need to merge the data marketplace evidence with company-level characteristics of the data sellers and the characteristics of their data products. Some of these data sellers are publicly listed companies, but many are private. We use a variety of data sources—Crunchbase, Pitchbook, Compustat, and CRSP to collect information on these companies background information and financing history. We use Edgar 10-K filings, combined with data product descriptions on Datarade, to fill in the characteristics of the markets in which they sell their data.

The model predicts that data sellers should choose one-time fees if they are financially constrained. If they do not urgently need cash, the subscription model of selling data is typically more profitable because it resolves the commitment problem. However, one-time fees bring in more revenue early in the life of the firm. The data confirms this prediction. We find a significant correlation between the way in which it sells its data and the age of the firm, the number of rounds of VC funding it has received, and the total amount of that funding. The older, better-funded firms are more likely to extract surplus, through the use of data subscriptions.

The model also predicts that when the market of data buyers is small, there is less scope to erode the value of data with future sales. Therefore, data that pertains to a more specialized group of potential buyers could be sold for a one-time fee, with little loss.

The data marketplace evidence also confirms this prediction. We determine the size of the market for data sales by comparing the textual similarity of data descriptions with the universe of firms' 10-K reports and then determining the industries with the greatest similarities. Then, we compute the number of publicly listed firms in those industries to determine the size of the market for the data. We find that this market size positively predicts data subscriptions and is negatively correlated with data sales.

We acknowledge that it is also possible that some settings lend themselves to complementarity in the use of data. In settings like speculative attacks or price-setting, the value of data might rise as others acquire it. In such settings, the data sellers' lack of commitment will be less costly, because data gains in value when more copies are sold. In contrast, dynamic complementarity, where an investor wants to learn now data that others will acquire later, still decreases the value of data over time.

Finally, one might object that data is not a durable good. It does depreciate. Information becomes stale. The rate of depreciation of data depends on the rate at which the environment changes. We explore the role of data depreciation in Section 3.2.

These results inform ongoing debates about data policy. If even a monopolist data seller has little market power in the market for data because it cannot commit not to compete against itself, then we should craft a very different data competition policy than a world where data-owning firms can extract extensive rents. Our results call into question even the idea that market power should be eliminated. Developing useful data and algorithms are like innovation. If we do not provide some monopoly rents, the incentive to provide a high-quality product disappears. It may well make sense to regulate other harms. However, monopoly rents alone do not imply an undesirable outcome for consumers.

Related literature Our work builds on the insights of the literature on the dynamic Coase (1972) conjecture. When selling a durable good, a monopolist who lacks commitment not to lower future prices is forced to compete with its future self. As consumers become very patient, such a firm is unable to obtain any rents, despite its monopoly power (Fudenberg and Tirole (1991), Chapter 10).

What we add to this well-known problem is three-fold. First, we connect data to semi-durable goods. Second, we introduce strategic substitutability between users of data: Data that others have is less valuable. Not only is a firm competing with its own lower prices, as in a standard durable goods problem, it also suffers from its inability to commit not to sell to others. Third, we quantify the strength of this force in data marketplaces.

Externalities also arise in multilateral contracting with a principal's lack of commitment power in Arnott and Stiglitz (1991, 1993) and Segal (1999). In Brunnermeier and Oehmke (2013), Green and Liu (2021), and DeMarzo and He (2021), the externality is debt dilution among creditors.

While the logic of debt dilution and information leakage have similarities, our mechanism has important differences as well. First, information is non-rival and can be replicated at near zero marginal cost. Second, information depreciates as the state of the world changes and old information becomes less relevant. Finally, information can be sold or licensed as a subscription, in a way that debt cannot.

The data economy literature is more similar in topic, but more different in its tools. Acemoglu et al. (2021) and Bergemann and Bonatti (2022) explore whether static data markets are efficient. Ichihashi (2020) show how firms can use consumer data to price discriminate. Jones and Tonetti (2020), Cong, Xie, and Zhang (2021) and Farboodi and Veldkamp (2022) build models of the data economy, but without market power in data markets. Existing work on the digital economy does explore whether data can be a source of market power (Kirpalani and Philippon 2020). Lambrecht and Tucker (2015) take a strategy perspective on whether data has the necessary features to confer market power. However, none of these consider the dynamic commitment problem of a data seller with market power, that we explore and quantify.

1 A Model of a Market for Data Purchases

Our model has two parts. The first part, describing households who purchase goods from producers that can utilize data, is there because we need households with utility functions in order to make welfare statements. This part of the model is constructed to make

the willingness to pay for data decreasing in the number of other agents that buy the data. For all the non-welfare results, it would be sufficient to simply assume this relationship directly. The idea that data is a strategic substitute is an old one. It goes back to the work of S. Grossman and J. Stiglitz (1980). Of course, that paper was written about information used to choose portfolios of risky assets. But the idea of strategic substitutability in information acquisition or data purchases holds much more broadly. Hellwig, Kohls, and Veldkamp (2012) show that information is a strategic substitute in most settings where actions are strategic substitutes. Markets where quantities are chosen and prices clear markets are such a setting. If other agents demand more of a good or sell more of a product, that moves prices adversely and makes other less willing to take the same action. While we take strategic substitutability as a payoff primitive in this model, we sketch an oligopolistic goods market in the appendix to show why this form arises.

The second part of the model describes the problem of the data seller who lacks the power to commit not to engage in future data sales. This is where the model's novel ideas lie. One reason a firm might not be able to commit to restrict its sale of data is that proving the equivalence of two data sets is not easy. The seller could give the data set a different name, create linear combinations of the variables, or even add a small amount of noise to data. Although the information content of the new data set would be nearly equivalent to the original, it might be difficult to enforce a contract prohibiting the sale of identical data.

One might object that most data providers are not true monopolists. In many cases, buyers could obtain substitutable data from another source. However, since we are exploring whether market power might not be as effective as one might think, we start from a setting with an extreme degree of market power and see how much commitment problems remedy that power.

1.1 Model Assumptions

Households and data buyers Time is discrete $t = 0, 1, 2, \ldots, \infty$. There are three types of players: a representative consumers, goods-producing firms, and a monopoly data supplier. The representative consumer has preferences over a measure-one continuum of

goods, indexed by i

$$U = \sum_{t=0}^{\infty} \beta^t u_t, \qquad u_t = \int_0^1 \frac{\sigma}{\sigma - 1} q_{it}^{\frac{\sigma-1}{\sigma}} - p_{it} q_{it} \, \mathrm{d}i, \tag{1.1}$$

where $\sigma > 1$ governs the elasticity of substitution across goods.

There is a measure-2 continuum of goods-producing firms—twice as many firms as goods. Firms choose prices to maximize expected profit. At each date t, two firms are randomly selected to produce each good. This randomness simplifies our exposition by ensuring that firms face uncertainty in whom to compete with in the future. Once matched, two firms produce perfectly substitutable goods and compete as in standard Bertrand price competition.

Goods-producing firms use data to reduce their marginal cost of production. Let n be the measure of firms that have data. A firm without data has a marginal cost of $c = 1$. A firm with data can use the data to optimize its operations and has a marginal cost is $c = 1/z$, where $z > 1$ is the quality of the data.

Data sellers The data supplier is a monopolist who maximizes the expected present discounted value of profits. Data sellers choose data quality z and the number of copies of the data to sell n. At an ex-ante stage, the data producer chooses the data quality z with a one-time, convex fixed cost $F(z)$. We assume that $F(z) = \frac{1}{2}\left(\left(\frac{\sigma-1}{\sigma} z\right)^{\sigma-1} - 1\right)^2 / 2$. The functional form is chosen to simplify expressions. Once produced, the data can be sold to multiple firms with zero marginal cost.

Data sales take place over time. At each date t, the data supplier chooses how many additional copies to sell in that period. Data buyers can generate an infinite stream of profits from the data. But that profit depends on the quality of the data z and on how many firms n have acquired the same data in that period.

Then time moves on to $t + 1$ and the game repeats. Future payoffs are discounted at the rate $\beta \leq 1$. In the limit as $\beta \to 1$, future profits are valued just as highly as current ones. For now, data sellers and buyers share the same discount rate, β, and data does not depreciate. We later relax these assumptions.

1.2 Discussion of Model Assumptions

An important feature of the model is that data purchasers cannot resell data. In reality, most data is sold with a contract that forbids data buyers from selling the purchased data to others. But this stands in contrast to the assumption that data sellers cannot use contracts to commit themselves.

While these assumptions comport with real features of data contracts, they do raise the question of why commitment is one-sided. One reason could be that there is one seller and many buyers. If a buyer violates the contract, the seller has a strong incentive to sue. However, if the seller were to commit to sell few copies and violated that contract, each buyer might find it optimal to wait for other buyers to sue. In other words, contract enforcement is costly. Enforcing contractual restrictions on data sellers could be subject to a collective action problem.

1.3 Equilibrium

Equilibrium Definition: At the start of the game, data sellers choose data quality z to maximize the expected present value of their profits, discounted to time-0. Then, in each period t,

1. the data supplier makes take-it-or-leave-it offers to sell data to a chosen number of goods producers;
2. goods producers decide whether to buy the data or not, taking as given the others' past, current and expected future choices;
3. good producers are randomly matched and choose prices to maximize their one-period profit;
4. households choose their basket of goods to maximize (1.1), taking all prices as given;
5. time moves on to $t+1$.

Differentiating household utility (1.1) with respect to the quantity of each good and setting that to zero yields a first-order condition, which can be re-arranged in the form

of a demand curve, $q_i = p_i^{-\sigma}$. The consumer surplus associated with each variety i is $\frac{\sigma}{\sigma-1} q_i^{\frac{\sigma-1}{\sigma}} - p_i q_i$.

For each product variety, there are three possible market configurations in each period: 1) both producers have data; 2) one producer has data and the other does not, and 3) neither producer has data. Let n denote the measure of firms that have data in that period. The probability of any firm having data is $n/2$ and the probability that two randomly selected firms have data is $n^2/4$ (case 1). Similarly, the probability that a single firm does not have data is $(2-n)/2$. The probability that two randomly matched firms both lack data is $(2-n)^2/4$ (case 3). The fraction of varieties for which one firm has data and the other does not is $n(2-n)/2$ (case 2).

In cases 1) and 3), the two firms producing the variety have symmetric marginal costs. Symmetric firms that engage in price competition make zero profits.

In case 2), one producer has a lower marginal cost than its competitor. In this case, the firm with data maximizes its profit, $q_i(p_i - c)$, by charging price $p_d = \min\{\frac{\sigma}{\sigma-1}/z, 1\}$. That is, there are two possible pricing regimes. In one, the firm charges the unconstrained monopolistic price $\frac{\sigma}{\sigma-1}/z$. In the other regime, i.e., when the unconstrained monopolistic price is above the competitor's marginal cost 1, the firm engages in limit pricing and charges the marginal cost of its competitor. As we show later, the functional form imposed on the data supplier's cost function $F(z)$ for improving data quality ensures that, in equilibrium, $\frac{\sigma}{\sigma-1}/z \leq 1$, thereby ensuring that we are always in the monopoly prcing regime, and $p_d = \frac{\sigma}{\sigma-1}/z$. We use the subscript d here to denote the price and quantity for a firm that has data when its competitor does not. This implies a markup of $\frac{\sigma}{\sigma-1}$. The firm without data sells nothing because its marginal cost of 1 exceeds this price.

Substituting the price p_d into the household demand curve implies that the quantity sold is $q_d = \left(z\frac{\sigma-1}{\sigma}\right)^{\sigma}$. This generates revenue for the firm with data of $p_d q_d = \left(z\frac{\sigma-1}{\sigma}\right)^{\sigma-1}$. The firm's profit is $\frac{1}{\sigma}\left(z\frac{\sigma-1}{\sigma}\right)^{\sigma-1}$ when its competitor does not have data.

The expected value of data, for one period, is the probability that the buyer's competitor will be uninformed, times the profit of having data when a competitor does not. We

call this one-period expected value $\pi(n;z)$:

$$\pi(n;z) = \frac{1}{\sigma}\left(z\frac{\sigma-1}{\sigma}\right)^{\sigma-1}(1-n/2). \tag{1.2}$$

For our subsequent analysis, it is useful to define $x \equiv (z\frac{\sigma-1}{\sigma})^{\sigma-1}$ as a monotonically transformed measure of data quality, and define $a \equiv 1/\sigma$, $b = a/2$, so that the per-period profit of each goods producer with access to data can be written as

$$\pi(n;x) = x(a-bn). \tag{1.3}$$

Substitutability arises here because the goods producing firm makes zero profit in every case, except that case where it has data and its competitor does not. This is what makes the firm's expected value of data decline in the number of other firms that also have data. This is surely extreme. But it is a simple way of capturing an externality that is much more prevalent that this specific model mechanism. Appendix B works out a richer equilibrium model of oligopolistic firms that use data to forecast demand, that justifies this substitutability assumption.

The reason for building out the household part of the model, rather than just assuming a π function, is to be able to derive welfare. We return to the welfare calculation in Section 6. For the rest of the model solution, we simply use the fact that the data buyers' (goods producer's) profit function π, that is increasing in quality z and decreasing in data sold n, i.e. $\pi_z > 0$, $\pi_n < 0$. The assumption that expected value is decreasing in data sold n captures the strategic substitutability of information. The parameter b governs the strength of the externality. If b is large, then data substitutability is strong. If b is close to zero, the strategic effect disappears.

1.4 Commitment Solution

We first explore a solution where a firm can commit to the quality and quantity of data. It can tell customers exactly how many copies of the data will be sold. The data seller will never sell any more copies of the data than the committed number n. This ability to

commit will allow the firm to choose a higher price up front and will maximize the firm's revenue. After presenting this solution, we compare the price and revenue to the solution when the firm cannot commit.

When the data producer can commit to selling quantity n of data with quality x (recall $x \equiv (z\frac{\sigma-1}{\sigma})^{\sigma-1}$) at the start and never sell the data again, the firm chooses x and n to maximize its profit. The buyers' willingness to pay is the present discounted value of their profits, discounted at rate β, which is $\pi(n;x)/(1-\beta)$. The monopolist seller charges a price equal to this willingness to pay. The sellers profit is this price times quantity n, minus the one-time fixed cost F of producing data of quality x:

$$V = \max_{x,n} \quad n \cdot \frac{\pi(n;x)}{1-\beta} - F(x). \tag{1.4}$$

Conditioning on data quality x, the value of the data producer is concave in n: selling to more clients n could bring more profits but could also reduces the willingness to pay by each client. There is a Laffer curve that plots the relationship between quantity and revenue; the optimal choice n^* reflects the point at which the Laffer curve is maximized.

The solution to the problem with commitment is that the firm chooses to sell data quantity and quality

$$n^* = a/2b, \quad x^* = 1 + \frac{a^2}{4b(1-\beta)}.$$

Notice that when data is a stronger strategic substitute (b large), the firm chooses to sell less of it. Given that less data will be sold, the investment in data quality is also lower.

Note that when setting up the problem, equation (1.4) implicitly rules out the possibility that the data seller would commit to a time-varying sequence of new data issuance (subject to the constraint that $\{n_t\}$ is a non-decreasing sequence). This is without loss of generality, as it is never optimal for the seller to choose a time-varying sequence $\{n_t\}$; instead, the seller would prefer making all sales n^* upfront. This is because n^* coincides with the maximizer of $n \cdot \pi(n;x)$, implying that $n_t = n^* \forall t$ is the maximizer to $\max_{\{n_t\}} \sum_{t=0}^{\infty} \beta^t n_t \cdot \pi(n_t;x)$, and the constraint, that $\{n_t\}$ must be non-decreasing, does not bind.

Under the optimal choice, the value obtained by selling data given quality x is

$$xn^*(a - bn^*)/(1 - \beta) = \frac{xa^2}{4b(1 - \beta)}.$$

The value of the data-selling firm, net of the investment cost for data, is

$$V^{commit} = \frac{a^2}{4b\,(1 - \beta)} + \frac{1}{2}\left(\frac{a^2}{4b\,(1 - \beta)}\right)^2.$$

This is the maximum value the data seller can achieve. It will be the benchmark, against which we compare the imperfect commitment solutions. Figure 1 plots the data seller's revenue, as a function of the number of copies of the data it sells. If the firm sells zero data copies, it has zero revenue. But if the firm sold $n = a/b$ copies of the data, it would earn a price of $x\,(a - bn)\,/(1 - \beta) = 0$ per units. This is also zero revenue. The peak revenue is achieved half way in between these two points at $n = a/(2b)$. Since data is sold at date 1 and never again, this is a one-period profit realized at date 1.

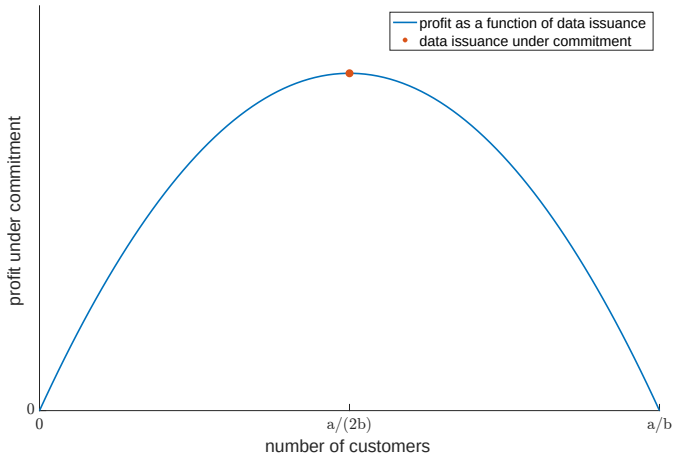


Figure 1: Equilibrium choice of n that maximizes profit under commitment

The formula for the curve is the total profit expression in (1.4), plus the one-time investment cost $F(x)$.

This relationship between firm revenue and quantity is similar to the idea of a Laffer curve in public finance that describes the relationship between government taxation and government revenue.

1.5 Non-commitment Dynamic Game

The problem is that after selling the data to n firms at time t, the data producer has the incentive to sell to more clients at $t+1$. Doing so reduces the profitability of prior clients. Knowing that future copies will be sold makes prior clients are unwilling to pay the non-commitment price.

We now study this more realistic game without commitment. Note that while the firm cannot commit to the number of copies of data it sells in the future, it does commit to the quality of the data as the quality is chosen up-front and does not change over time (an assumption we relax later).

Non-commitment problem At each time t, the data producer solves a recursive problem, with the total number of data copies already sold n and data quality x as state variables. The value function represents the seller's present discounted value of the revenue derived from their data. The seller is a monopolist. So they can make a take-it-or-leave-it offer to the data buyer for the buyer's willingness to pay for the data.

The buyer, however, has a willingness to pay that is based on their rational expectation of the seller's future data sales. The buyer will earn $\pi(n;x)$ profits from their data in the first period. But will earn only $\pi(\tilde{n};x)$ dollars of profit the following period, if an additional $\tilde{n}-n$ customers buy the data. Thus, at date t, the buyer's willingness to pay is $\sum_{\tau=1}^{\infty}\beta^{\tau-t}\mathbb{E}_t\left[\pi(n_\tau;x)\right]$, where n_τ is the total purchases of the data, up to and including date τ. Next, we rewrite this willingness to pay, as a function of the optimal selling strategy of the data seller. We don't yet know that strategy. We will use a placeholder strategy function g and then solve for the optimal selling strategy, as a fixed point of the buyers' and seller's problem.

Let $g\left(n,x\right)$ denote the data seller's optimal choice of new total data sales, given that n firms have already purchased the data. As we show below, the optimal choice depends only on n and is invariant to the data quality x, so we suppress the argument and simply write $g(n)$. The number of new clients being sold to is $g(n)-n$. Let $g^2\left(n\right)\equiv g\left(g\left(n\right)\right)$, $g^0\left(n\right)=n$ and define $g^k\left(n\right)$ to be the operation g performed k times on n. Then g^k represents how many total copies of the data the data seller will choose to sell k periods from

now, if there are already n total buyers today. Note that n is a stock of total past buyers and $g^k(n)$ is a new stock of buyers. If a firm decided to sell no new copies of data, then this would be represented as $g^k(n) = n$.

Substituting $g^{t-1}(n)$ for n_τ in the sum above, the buying firms' total stream of profits from data can be expressed as

$$\bar{\pi}(n;x) = \sum_{t=1}^{\infty} \beta^{t-1} \mathbb{E}\left[\pi\left(g^{t-1}(n);x\right)|n,x\right] \tag{1.5}$$

This profit $\bar{\pi}(n;x)$ incorporates firms' conditional rational expectations that their ability to extract value from this data will decline over time, given the current state variables (n,x). It anticipates the future path of data sales. This present discounted revenue from data is also the buyer's willingness to pay for data.

Since the data seller is a monopolist, the revenue-maximizing choice is to charge each firm their willingness to pay for data. Giving this willingness to pay, the data producer, who has already sold n data copies, chooses to sell $\tilde{n} - n$ additional copies of the data in period t. This choice earns the seller a price of $\bar{\pi}(\tilde{n},x)$ earned for each of the $(\tilde{n}-n)$ additional copies of the data sold that period. The seller's optimal choice should maximize this current revenue, plus the discounted present value of future revenue. This choice problem can be written recursively as,

$$V(n,x) = \max_{\tilde{n}} \left\{(\tilde{n}-n)\,\bar{\pi}(\tilde{n},x) + \beta V(\tilde{n},x)\right\}. \tag{1.6}$$

Definition 1.1. Given data quality x, a *Markov perfect equilibrium (MPE)* is the pair of functions $\{\bar{\pi}(\cdot;x), g(\cdot)\}$ such that:

1. the goods producers' willingness to pay for data $\bar{\pi}(n;x)$ is consistent with their rational expectation of the future sequence of data sales, satisfying (1.5);

2. the policy function for the data supplier $g(n)$ solves the problem solves (1.6).

Optimal data sales, without commitment In principle, the dynamic game involving a non-commitment data supplier and a sequence of data buyers (i.e., the goods producer) is

difficult to analyze, as the MPE involves a fixed point in the two functions $\{\bar{\pi}(\cdot; x), g(\cdot)\}$. However, under our tractable formulation, the recursive problem of the data supplier (1.6) is quadratic in the state variable n, thereby enabling us to solve for the equilibrium in closed-form.

The solution to this model shows how firm commitment problems result in more data sales, lower prices and reduced profits. Importantly, the solution also tells us where to look for evidence of this commitment problem: declining data prices, over time.

Proposition 1. The data issuance policy function $g(n)$ is characterized by an equilibrium scalar δ:

$$g(n) = n + (1-\delta)\left(\frac{a}{b} - n\right) \qquad \text{with} \qquad \delta = \frac{1-\sqrt{1-\beta}}{\beta}. \tag{1.7}$$

Firms' willingness to pay is characterized by an equilibrium scalar $\xi \equiv \sqrt{1-\beta}$

$$\bar{\pi}(n, x) = \xi \pi(n, x) = \frac{(a - bn)}{\sqrt{1-\beta}} x. \tag{1.8}$$

Given data quality x, the data provider's value function is

$$V(n, x) = \frac{b\delta}{2\sqrt{1-\beta}} \left(\frac{a}{b} - n\right)^2 x \tag{1.9}$$

All proofs are in Appendix A.

Interpretation. The Markov perfect equilibrium is captured by the two endogenous variables (δ, ξ), respectively parametrizing the data producer's and data buyers' equilibrium strategy.

Note $\bar{n} \equiv \frac{a}{b} = 2$ is the maximum total sales; data is worthless to goods producers when every potential competitor has access to it. Given existing sales n at the beginning of each period, the total sales at the end of each period $g(n)$ is a weighted average between $\bar{n}$ and n. Intuitively, $1 - \delta$ captures how aggressive the data producer sells to new clients. When $\delta = 1$, $g(n) - n = 0$, meaning the data producer does not sell to new clients. A lower δ translates to more aggressive sales.

On the other hand, ξ scales how firms value data, reflecting their expectation about

future data sales. When firms anticipate no future data sales, $\xi = 1$, and $\bar{\pi}(n) = \frac{a-bn}{1-\beta}$ coincides with the present discounted future flow value under the commitment solution. Absent commitment, firms anticipate future sales and thereby place a proportional discount ξ on the value of data.

Note that the path of total data issuance is $g^t(0) = \left(1-\delta^t\right)\frac{a}{b}$, meaning the path of new sales at each time period is

$$g^t(0) - g^{t-1}(0) = (1-\delta)\frac{a}{b}\delta^{t-1}$$

which decays to zero exponentially at rate $\delta \equiv \frac{1-\sqrt{1-\beta}}{\beta}$.

The path of sales price is $\bar{p}\left(g^t(0)\right) = \frac{\xi\left(a-bg^t(n)\right)}{1-\beta}$, where $\xi \equiv \sqrt{1-\beta}$, which simplifies to

$$\bar{\pi}\left(g^t(0)\right) = \frac{a\delta^t}{\sqrt{1-\beta}}$$

which also decays to zero exponentially at rate δ.

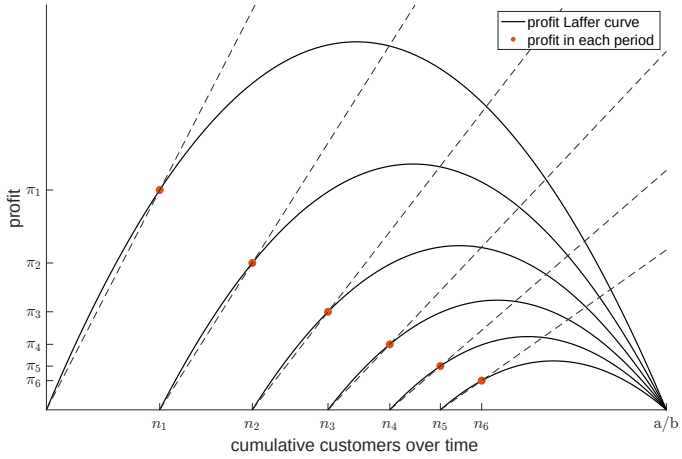


Figure 2: Equilibrium paths of data sales, prices and profit, without commitment

Figure 2 shows the equilibrium path without commitment. Specifically, the outermost curve plots $(n-0)\,\bar{\pi}(n)$, which is the initial Laffer curve. The red dot reflects the equilibrium choice n_1 in the first period and the corresponding profit. The second curve shows $(n-n_1)\,\bar{\pi}(n)$ as a function of n, which is the Laffer curve in the second period, and the red dot on the curve reflects the equilibrium choice n_2 in the second period and the cor-

responding profit. The slopes of the dashed lines reflect the sequence of equilibrium willingness to pay $\bar{\pi}(n_t)$.

Ex-ante stage: Choosing Data Quality We now solve the data supplier's ex-ante problem of choosing data quality, noting that the data acquisition cost can be written as $F(x) = (x-1)^2/2$, where $x := \left(\frac{\sigma-1}{\sigma}z\right)^{\sigma-1}$:

$$\max_x V(0,x) - (x-1)^2/2$$

Proposition 1 implies that

$$x^{non-commit} = V(0) + 1 = 1 + \frac{1}{\sigma}\frac{1}{2\left(1+\sqrt{1-\beta}-\beta\right)}. \tag{1.10}$$

1.6 The discount rate and the value of commitment

An important implication of the model is that the value of commitment relates monotonically with the discount rate β. For a given data quality x, the value that the data supplier can extract from goods producers is $V(0,x)$ when lacking commitment; the value is $\pi(n^*,x)/(1-\beta)$ when the data supplier can commit. Proposition 1 implies that

$$\lim_{\beta\to 0}\frac{V(0,x)}{\pi(n^*,x)/(1-\beta)} = 1, \quad \lim_{\beta\to 1}\frac{V(0,x)}{\pi(n^*,x)/(1-\beta)} = 0.$$

That is, relative to the case with commitment power, the value obtained by the data supplier who cannot commit is vanishing as $\beta \to 1$. This is because as the goods producer anticipate future data sales, their willingness to pay relative to the commitment case declines towards zero (i.e., $\frac{\bar{\pi}(n,x)}{\pi(n,x)} = \xi = \sqrt{1-\beta} \to 0$ as $\beta \to 1$. Intuitively, low discounting implies that existing data buyers expect more negative externalities arising from future data sales. This expectationcauses the data buyer to discount the value of the data by more. The value of data to the goods producers that buy it is lower than in the full-commitment case, because the buyers know that additional future copies will be sold. The non-commitment value of data is the commitment value times $\sqrt{1-\beta}$. In this solution, there are opposing

forces from the buyer's patience and the seller's patience. We will disentangle those two forces in Section 3.1.

The preceeding discussion holds the data quality x constant. It is also useful to consider the ratio of data quality without and with commitment. Let $\chi \equiv x^{non-commit} / x^{commit}$. Combining solutions reveals that $\chi = \frac{2\left(\sqrt{1-\beta}-(1-\beta)\right)}{\beta}$. To decode this expression, consider two limiting cases. One case is where agents are so impatient that they almost ignore the future ($\beta \to 0$). In this impatient case, the data seller has little incentive to invest in data quality even under commitment ($x^{commit} = 1$), and the lack of commitment does not penalizes data quality further $\lim_{\beta \to 0} \chi = 1$. The other limiting case is where all agents are so patient, that they hardly discount the future at all ($\beta \to 1$). In this case, the data producer has a strong incentive to investment in data quality, but the data quality when lacking commitment is vanishing relative to the data quality under commitment: $\lim_{\beta \to 1} \chi = 0$. The reason the data seller invests little relative to the commitment case is precisely that, as the patient data buyers anticipate future sales, ξ, their willingness to pay relative to the commitment case, converges to zero.

2 Data Subscriptions

Many data sellers do not charge a one-time fee. Instead, they charge an ongoing fee per period for continuously updated data. A data buyer who does not pay the subscription fee keeps their old data, but loses access to updates. We show that this pricing scheme allows a firm without the ability to commit to achieve the sequence of flow revenue closer to the full-commitment outcome. When old data is worth little, the firm can achieve full-commitment revenue. In the section that follows, we will explore why a firm might not choose the subscription model, as the basis of a model testing strategy.

Data Subscription Model As before, a new data buyer in each period t has marginal cost $1/z$. What changes is that data can become obsolete. Suppose that goods producers who have previously bought the data, but do not make a new purchase at date t, face a probability $1 - \alpha$ with which the data become obsolete. The value of obsolete data drops

to zero, and the marginal cost of the goods producer goes back to 1. The probabilistic obsolescence retains the tractability of our analysis.

Given data quality choice x, the goods producers' willingness to pay for data $\bar{\pi}(n;x)$ is

$$\bar{\pi}\left(n;x\right)=\sum_{t=1}^{\infty}(\alpha\beta)^{t-1}\pi\left(g^{t-1}\left(n\right);x\right) \tag{2.1}$$

where $g(n)$ captures the seller's data issuance policy function and is the solution to the Bellman equation

$$V\left(n;x\right)=\max_{\widetilde{n}}\left(\widetilde{n}-\alpha n\right)\bar{\pi}\left(\widetilde{n};x\right)+\beta V\left(\widetilde{n};x\right). \tag{2.2}$$

This is a model of subscription because the goods producers need to make repeated payments in order to continue to access data. The expected number of periods for which the data stays valid is $1/(1-\alpha)$. A higher rate of obsolescence implies a greater frequency that data needs to be renewed.

Proposition 2. In the subscription model characterized by equations (2.1) and (2.2), the data issuance policy function $g(n)$ is characterized by an equilibrium scalar δ:

$$g\left(n\right)=\frac{2\left(1-\alpha\beta\right)\left(1-\alpha\beta\delta\right)}{\left(2-\alpha-\alpha\beta\right)-\alpha\beta\left(1-\alpha+\delta\left(1-\alpha\beta\right)\right)}\left(1-\delta\right)+\delta n,\text{ with }\quad\delta=\frac{1-\sqrt{1-\alpha^{2}\beta}}{\alpha\beta} \tag{2.3}$$

Firms' willingess to pay is

$$\bar{\pi}\left(\widetilde{n};x\right)=\frac{x}{\sigma}\left(\frac{\left(2-\alpha-\alpha\beta\right)}{\left(2-\alpha-\alpha\beta\right)-\alpha\beta\left(1-\alpha+\delta\left(1-\alpha\beta\right)\right)}-\frac{1}{1-\alpha\beta\delta}\widetilde{n}/2\right) \tag{2.4}$$

Given data quality x, the data provider's value function is

$$\begin{aligned}V\left(0\right)=&2\frac{x}{\sigma}\left(1-\alpha\beta\delta\right)\left(\frac{\left(2-\alpha-\alpha\beta\right)}{\left(2-\alpha-\alpha\beta\right)-\alpha\beta\left(1-\alpha+\delta\left(1-\alpha\beta\right)\right)}\right)^{2} &(2.5)\\ &\times\rho\left(1-\delta\right)\left[\frac{1-\rho}{1-\beta\delta}+\frac{\delta\rho}{1-\beta\delta^{2}}\right] &(2.6)\end{aligned}$$

where $\rho\equiv\frac{1-\alpha\beta}{2-\alpha-\alpha\beta}$.

A Special Case: Subscription Restores Full Commitment. When the data becomes obsolete after each period ($\alpha = 0$), the commitment solution is restored.

In this limiting case, subscription is a requirement that a data user pay each period, instead of paying a one-time lump sum for all future usage. Hence the problem of the data producer becomes a repeated static decision of how many users there will be in that period. By removing the dynamic element of the decision, the commitment problem disappears.

As before, a data producer chooses an up-front investment in data quality x, at a cost $F(x)$. After choosing data quality once, that quality remains fixed. Then, each period, the data producer decides on how many copies of the data to sell n. However, selling data in this case means selling a right to use the data for one period. The number of users n does not cumulate over time.

The buyer's value of data is now the one-period profit that it generates. So $\pi(x; n)$ is the buyer's willingness to pay for data of quality x sold to n buyers.

Each period t, conditioning on data quality x, the producer solves $\max_n n\pi(x; n)$. The optimal number of data copies sold n is time-invariant. The data quality choice problem can be equivalently formulated as $\max_x \{\max_n np(x; n)\}$, which coincides with the data producer's problem under commitment.

In what follows, we assume $\alpha = 0$ when referring to the subscription model. Future results will explore weaker versions of data subscriptions.

3 Testable Predictions of the Model

While the previous section laid out the solution to the data sales and data subscription models, we ultimately want to ask whether the model's mechanism is consistent with data markets. This section derives predictions to test. These are not meant to capture all the relevant considerations in a data seller's choice of business model. Rather, the predictions below are the considerations that are indicative of our mechanism at work.

3.1 Financial Constraints, Discount Rates and the Choice of Data Sales or Subscription

Clearly, the discount rate or degree of patience plays a crucial role in commitment problems. However, data buyers and sellers may not share the same discount rate. In particular, financial constraints on the part of data sellers may induce them to behave as if they were less patient. When unable to borrow, a firm may value immediate cash flows more highly than later ones. This type of financial constraint might explain why some firms choose the one-time data payment model, instead of a data subscription model. We introduce differential discount rates and explore the optimal data pricing model next.

Next, we consider a models of data sales with and without commitment, and compare them to a market with data subscriptions. Throughout, we maintain the following assumptions: Suppose data producer's discount rate is γ, and buyer's discount rate is β. As before, $\pi(n;x) = x(a - bn)$. there is an ex-ante choice of data quality x with cost $x^2/2$.

Different buyer and seller discount rates, with commitment Consider the case of data sales, where data buyers pay for their data up front for the right to use it in perpetuity. With commitment, data buyers still value the data the same as before. Once they own the data, they can earn $\pi(n;x)$, in perpetuity, which means the data seller can charge $\frac{\pi(n;x)}{1-\beta}$. The data seller's discount rate γ does not matter for this problem because all the revenue is earned in the first period. The cost of data quality is also incurred in the first period. Just as before, the solution is $n^* = a/2b$ copies of the data are sold, at quality level $x^* = \frac{a^2}{4b(1-\beta)}$.

Different buyer and seller discount rates, without commitment Without commitment, the seller earns revenue gradually, by selling more copies of the data over time. In this problem, the seller's discount rate is relevant. As before, we solve by backwards induction. Given data quality x, the data seller's choice for how many data points $\tilde{n}$ to sell in the current period, given that n data points were already sold is

$$V(n) = \max_{\tilde{n}} \left\{ (\tilde{n} - n)\, \bar{\pi}(\tilde{n};x) + \gamma V(\tilde{n}) \right\} \tag{3.1}$$

where the data buyers' willingness to pay $\bar{\pi}(\widetilde{n}; x)$ is (1.5), which takes the same form as before because the buyers' discount rate is still β.

The solution to this problem has the same characterization as before, except that the data provider sells more or less data each period, depending on his discount rate δ. Let $g(n)$ be the $\widetilde{n}$ that solves (3.1), given that n copies of the data were already sold.

Proposition 3. The data issuance policy function $g(n)$ satisfies $g(n) = (1-\delta)\frac{a}{b} + \delta n$ with $\delta = \frac{1-\sqrt{1-\gamma}}{\gamma}$. Firms' willingness to pay is $\bar{\pi}(n; x) = \frac{a-bn}{1-\beta\delta}$. The data provider's value function is

$$V(n; x) = \frac{b\delta}{2(1-\beta\delta)}\left(\frac{a}{b} - n\right)^2 x.$$

The initial value of the data seller, after paying the cost to invest in data quality x but before having sold any data, is $V(0; x)$. The ex-ante data quality choice maximizes the initial value net of the investment cost in data quality: $\max_x V(0; x) - x^2/2$. Applying Proposition 3 and taking the first-order condition of this problem, we find that the firm's optimal choice of data quality is $x = \frac{a^2}{2b\left(1-\beta+\sqrt{1-\gamma}\right)}$. We can then plug that choice back into the data seller's initial value function to find the value of the data to the seller with discount rate γ:

$$V^{sales} = \frac{1}{2}\left(\frac{a^2}{2b\left(1-\beta+\sqrt{1-\gamma}\right)}\right)^2. \tag{3.2}$$

We learn that patient data sellers (γ close to 1) extract more surplus from selling data. However, selling to a patient buyer with β close to 1 generates even more value.

Discount rates and data sales vs. subscription. One of the key determinants of the pricing model a seller selects for their data is how impatient they are. Later, we interpret financial constraints as a source of data sellers' impatience. Since a one-time fee delivers revenue up front, it makes sense that an impatient data seller prefers to sell data, rather than adopt a subscription model. The preference for selling data becomes even stronger if setting up the subscription model requires fixed upfront investments. We tease out

this relationship between patience and subscriptions vs. sales formally, in order to tie the model as closely as possible to the data analysis.

Subscription provides the data producer a payoff stream from the commitment solution n^{commit}, with an upfront cost $\eta \geq 0$. This implies that

$$V^{subscription} = \max_{n,x} x\frac{n\left(a-bn\right)}{1-\gamma} - F\left(x\right) - \eta = \frac{1}{2}\left(\frac{a^2}{4b\left(1-\gamma\right)}\right)^2 - \eta \tag{3.3}$$

and $x^* = \frac{1}{4b(1-\gamma)}$.

The data producer prefers outright sales over subscription if $V^{sales} \geq V^{subscription}$, which is true iff $\left(\frac{1}{1-\beta+\sqrt{1-\gamma}}\right)^2 - \left(\frac{1}{2(1-\gamma)}\right)^2 + \frac{8b^2}{a^4}\eta > 0$. The left-hand side of this inequality is increasing in the discount rate of the buyers (β), decreasing in the discount rate of the seller (γ), and decreasing in the fixed cost η of setting up subscription services.

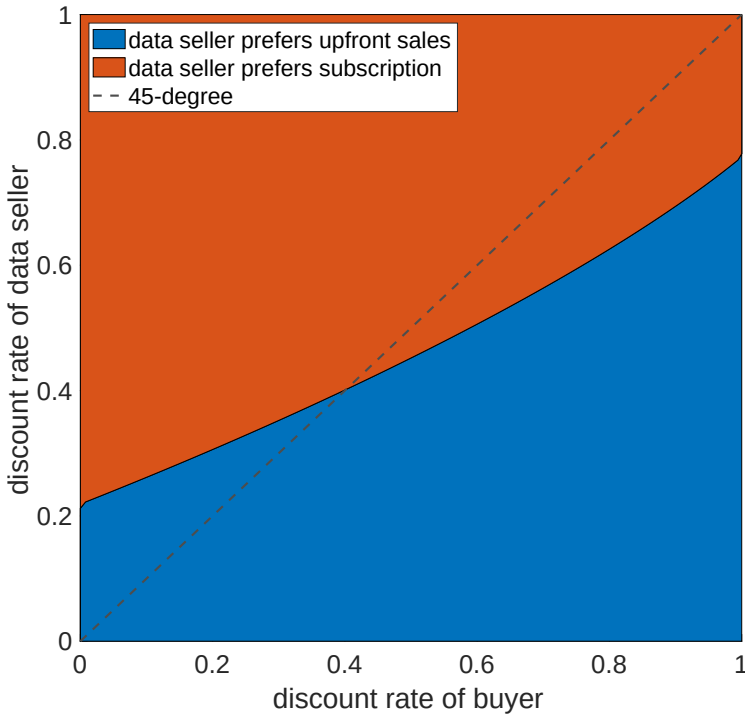


Figure 3: Data contract choices that maximize seller revenue, without commitment. Lower discount rates mean that there is a stronger preference for early cash flows. Depreciation of old data is $\omega \leq 1/z$.

Figure 3 illustrates, for a given cost η, the set of discount rates for buyers and sellers that would result in the choice of a seller to sell the data outright, or to license it with

a subscription fee. More patient sellers wait for the subscription revenue because it is higher.

H1: Financially constrained data sellers are more likely to sell data, rather than license it.

3.2 Data Depreciation and Data Sales

So far, we have characterized data as being like a durable good. At the same time, data loses value over time as well. Next, we consider how data depreciation affects our predictions. Data depreciates for many reasons. First, data may become inaccurate over time due to changes in the environment or context in which the data was collected. For example, data on consumer behavior from five years ago may not be applicable today as consumer preferences may have evolved. Second, data can degrade over time due to corruption of electronic files. Lastly, legal requirements or regulations may require data to be deleted or destroyed after a certain period. Appendix C formalizes data depreciation and formally links it to volatility in the economic environment.

Suppose data quality depreciates exponentially at rate $\lambda < 1$. Given initial quality x, the flow payoff to data buyers is

$$x\left(a - bn_1\right), \lambda x\left(a - bn_2\right), \lambda^2 x\left(a - bn_3\right) \ldots$$

Given the anticipated sequence of future data sales, the willingness to pay by a borrower with discount rate β for the data that depreciates at rate λ is therefore identical to the willingness to pay by a borrower with discount rate $\lambda\beta$ for data that do not depreciate. Because the seller's flow profit in equilibrium also scales proportionally with the quality of data (c.f. equation 3.1), the equilibrium in a model with data depreciation is therefore isomorphic to one without data depreciation, but where the data buyers have discount rate $\beta\lambda$ and the seller has discount rate $\gamma\lambda$, where $\lambda < 1$ captures the effect of data depreciation.

If we shrink each player's rate of time preference by a common factor $\lambda < 1$, the data seller is more likely to favor data sales over subscription. This can be seen from Figure 3:

starting with a given discount rate (β, γ) in the red region and moving towards the origin, the seller switches from preferring subscrition to upfront sales.

H2: Data that depreciates faster is more likely to be sold, instead of licensed.

3.3 Market Size and Data Sales or Subscriptions

The size of a market that a firm might sell to will determine the firm's initial revenue. In this model, we might think of a large market as one where the goods producers have lots of market power, in other words, a lower σ. This market power doesn't matter directly for the firms' choice of sales or subscription. But it might affect the choice, through the financial constraint. If we think that a low firm discount factor γ reflects the fact that a firm is financially constrained and needs cash flow earlier, then initial high profits of the firm should relax this constraint and raise γ. With less need for early cash flows, the firm would be more likely to choose the data subscription model.

H3: Data that is relevant to a large set of firms is more likely to be licensed, instead of sold.

3.4 The Decline in New Customers and New Revenue

In the data subscription model, the prediction is extreme. The results in Section 2 show that the number of data copies n is time-invariant. That means that after the first period, there are no new customers. Obviously, that is unrealistic, in part because our model is one with no customer acquisition costs or time. In reality, it would still take time to acquire the optimal number of customers. But after reaching that market size, data subscriptions should stagnate.

When data firms sell data, there are new customers over time, because of the lack of commitment. The maximum total data that a data seller would ever sell is a/b. Proposition 1 tells us that when the number of data copies reaches $n = a/b$, data becomes worthless to product firms. When $0 < \delta < 1$, δ regulates how quickly the firm converges to this maximum. When data sellers are patient, β is close to one and therefore δ is close

to 1. This makes for slow convergence. The patient data seller sells a little bit of additional data every period. But impatient sellers have β close to zero and δ close to zero. These data sellers sell most of their data quickly and approach a/b data points sold quickly. The value of such data deteriorates rapidly as the market is saturated with this data.

H4: New customers and new revenue should grow more slowly for data firms than for comparable non-data firms.

The decline in price comes from two sources. First, is the classic force from the Coase conjecture that applies more generally to durable goods. When the most eager buyers have already purchased the data, the remaining market has a lower value of data, which makes the optimal price for a seller lower. Second, is the new force, the strategic substitutability of data. Data that others already know typically generates less profit for the user. While one or neither force could be at work for a non-data firm, both forces operate to push prices down over time for a given data set.

Of course, not all data exhibits equal strategic substitutability. Financial data is known to be nearly worthless if it is widely known. Similarly, strategic business data typically offers profits only if that market strategy is not saturated. In contrast, weather data is something everyone can benefit from, to all be better prepared for the day. The difference is not in the type of data, as much as it is the use of that data. If some trader discovered a profitable trading strategy, based on the weather, then this would be less valuable if other acquired weather data and used it for that same purpose. Data has strategic substitutability when the uses of data are to inform actions that themselves exhibit substitutability. The finance data is a strategic substitute because buying risky assets is less profitable when others buy the same assets at the same time. Investors compete with each other. The typical weather forecast is not a strategic substitute because people do not typically compete with each other in the use of an umbrella.

4 Data and Descriptive Statistics

4.1 Data Products and Providers

We obtain data on the market for data from Datarade (https://datarade.ai/), a global data trading platform that helps companies discover, compare, and connect with data providers across the globe. As of Spring 2023, Datarade is one of the largest data markets, hosting more than 3,000 different data products, provided by more than 2,000 data providers, spanning dozens of major categories.

Datarade provides detailed information on each data product. For each of the products hosted on Datarade, the information page reports key statistics that are useful for our empirical study. Figure A.1 in the Appendix provides an example of the information page. First, they report the data provider company, which allows us to later merge with company-level information. Second, a menu of data transaction methods is provided, including one-off purchase, licensing (monthly or yearly), and usage-based pricing. Sometimes, the price level is available, though the coverage is limited.

Datarade also tags each data product with one or more data categories. There are 527 different data categories which are finely defined. In our example, the product is associated with five different tags: Location Data, Foot Traffic Data, Mobile Location Data, Raw Location Data, Mobility Data. The description of the data product is extensive (see Figure A.2 for the continued example). It describes, not only the data contents, but also potential use cases. In our example, the data description lists uses such as consumer insights, market intelligence, advertising, and retail analytics. We use these descriptions to map each product to industries that may use it. Figure 4 presents a word cloud of the descriptions to provide an overview of how data suppliers describe their data.

We link data sellers from Datarade to company-level data sets Crounchbase and PitchBook to extract their characteristics. Out of the 2,458 data sellers in our data, we are able to find 1,701 (69%) of them in these databases, using the fuzzy name-matching algorithm. Through these company-level data sets, we can determine the geographic location, funding history, survival, and executives' background, of each data seller.

Figure 4: Word Cloud of Product Descriptions on Datarade

Notes. This is a word cloud generated by the text of all the data descriptions on Datarade. URL address: https://datarade.ai/data-products/lifesight-foot-traffic-data-global-mobile-location-data-2-lifesight.

For funding history, an important characteristic for our analysis, we observe whether these companies obtained any funding from venture investors (as of March 2023), the total rounds of investments, and the total amount of investment. For the unmerged companies, we impute the funding variables to be 0. The assumption that data sellers that have never received documented venture funding are unfunded, is one that we verified by hand checking and internet searches.

4.2 Measuring Market Size and Depreciation

To how each data product is relevant to usages by industrial sectors, we obtain data from Compustat, CRSP, and Edgar 10-K filings. We use this information to determine which industries each specific data product is relevant for, the size of the market for the data product, and to impute data depreciation.

Identifying Relevant Industries for Each Data Product First, we create a mapping from each product to its relevant industries using data product descriptions from Datarade and business descriptions (item 1) and Management Discussion and Analysis (MD&A, item 7) from 10-K. In specific, for each data product, we calculate its textual similarities with all the 10-K filings in 2020, and find the top 20 filings with the highest similarity scores. The industry span of these 20 firms, in our perspective, captures the industries that may find the data useful. Let $\mathcal{J}_i$ be the set of SIC industries (at the three-digit SIC level) determined to be relevant to data product i. Each data product can be linked to six to fifteen different industries. That is, the size of set $\mathcal{J}$ ranges from six to fifteen.

Measuring Market Size The market size for a data product is the number of publicly-listed firms in relevant industries. Let m_j be the number of firms in industry j. Then, for a given data product i, the market size is

$$MarketSize_i = \sum_{j \in \mathcal{J}_i} m_j \tag{4.1}$$

Measuring Data Depreciation The depreciation of a data product is the weighted average of the depreciation rates for the industries relevant to that data (i.e., $j \in \mathcal{J}_i$). For each industry, the idea of measuring data depreciation is to capture the ability for information from the past to predict future business performance. If past information is more predictive of future business activities in that industry, we consider that the depreciation of data's value is low; while if past information is less predictive of future business activities, we consider data in that industry to depreciate fast.

To implement this idea, we calculate the within-industry average R^2 when using lagged

ROA to predict the same firm's ROA in the next period using the following model for each firm l,

$$ROA_{lt} = \alpha + \beta\ ROA_{l,t-1} + \varepsilon. \tag{4.2}$$

Denote the R^2 from this regression as R_l^2, and the industry-level R_j^2 is calculated using all firms within the same industry j. Then, for each data product i, we can define stability to be the average R^2 over all related industries for i:

$$Stability_i = \frac{1}{MarketSize_i} \sum_{j \in \mathcal{J}_i} R_j^2. \tag{4.3}$$

A higher R^2 in this predictive model for returns suggests greater stability and a *low* data depreciation rate. In contrast, a low R^2 suggests a rapidly changing environment. This translates into a high depreciation rate of data (see Appendix C for a formal mapping between volatility and data depreciation). Therefore, we define $Depreciation_i \equiv 1/Stability_i$. Finally, we group data products into 10 deciles based on the depreciation rates calculated.

4.3 Describing the Data Market

We start by describing the data on the Datarade market – the company locations, types of data they sell, and the pricing model they use. Datarade itself is a Germany-based startup, but nearly half of the data providers headquarter in the US.[1] Other popular headquarter locations for data providers are the UK, India, and Germany.[2]

Each data product is tagged with four categories, on average. Categories include B2B contact data, company data, or business website data. Categories overlap. The category B2B contact data (and similarly, B2B leads data, B2B marketing data) is the most popular on the platform, consisting of 15% of the sample. Other business intelligence data, such as company data or point of interest data, are also popular. Table 1 details the top ten categories of data products traded on Datarade and their market shares.

This type of data is well-suited to examine the commitment problems described by

[1] See https://www.crunchbase.com/organization/datarade.

[2] Table A.1 in the appendix provides more details on the geographic locations of the data providers.

Table 1: Key Data Categories on the Market for Data

Category	Product Count	Percentages
B2B Contact Data	530	14.39%
B2B Leads Data	527	14.31%
B2B Marketing Data	522	14.17%
Company Data	513	13.93%
B2B Email Data	457	12.41%
Firmographic Data	303	8.23%
B2B Decision Maker Data	302	8.20%
Global POI Data	268	7.28%
Point Of Interest POI Data	268	7.28%
Business Website Data	241	6.54%

Notes. This table presents top data categories on the Datarade platform.

our model for three reasons. First, these data categories describe types of data that are durable, not ephemeral insights. That is important because it suggests that the threat of a seller selling this durable data to others is a relevant problem. Second, these data products are suitable for multiple users (not firm-specific). Finally, they fit the model because future data sales would likely decrease the value of data for earlier buyers. For example, if more competitors obtain the contact list of potential customers, then their ability to profit from contacting that same group of customers diminishes.

4.4 Data Sellers' Funding / Financial Frictions

A key determinant of whether data sellers offer data subscriptions, with high market power, or one-time fees, with more competitive pricing, is their impatience. We interpret firms' impatience as deriving from their financial constraints. Therefore, next, we explore the ability of the data sellers on Datarade to access financing.

Table 2 reports the funding status of the Datarade data sellers. 696 companies (28.3%) obtained venture funding, with the average number of funding rounds being 0.96. Conditional on obtaining funding (in the 575 sample), the average number of funding rounds is 3.4. The total funding obtained by a provider is highly skewed, with an average of just over 20 million USD. Our data providers are quite mature—the median age is 12 years

Table 2: Financial Information about Data Sellers: Summary Statistics

Variable	count	mean	std	25%	50%	75%
Obtained VC Funding (0/1)	2458	0.234	0.423	0	0	0
No. of Funding Rouds	2458	0.961	2.065	0	0	1
Total Funding (mil. USD)	2458	$22.8m	$29.3m	0	0	0
Founding Year	1584	2003	26	2001	2011	2015
Age (as of Mar, 2023)	1584	19.918	25.589	8	12	22

Notes. This table presents summary statistics for the financial information linked to the data providers on the Datarade platform. Source: Crunchbase.

old.

5 Testing Model Predictions: Data Subscriptions and Data Sales

Recall that the main prediction of the theory was that data subscriptions were associated with market power in the data marketplace, whereas one-time sales were likely to yield less revenue for firms. This section measures the extent of data purchases versus subscriptions and provides empirical tests of the theory's prediction. These results do not establish any causal relationships. Instead, these are novel empirical facts that inform us about the features of data markets. They also support the prediction that firms choose data sales in some cases and licensing or subscriptions in others.[3]

The main variable in our analysis is the transaction model adopted by each product, which is a measure to capture the trade-offs faced by the provider. We report this information in Table 3. One-off-purchase is the most popular transaction model offered by data products, available for 46% of the products. Yearly licensing is offered for 38% of the products, while monthly licensing is available for 28% of the products. Usage-based pricing is available for 37% of the products. These different models could be simultaneously

[3]In our empirical evidence, many firms offer both data sales and subscriptions. This is why the means of the indicator variables in Table 3 add up to more than one. In our simple model, this is only a knife-edge case. However, we could easily extend the model by adding data buyers that are heterogeneous in their rates of time preference. In such an environment, some firms will choose to offer both sales and subscriptions to segment the market.

Table 3: Summary Statistics of Transactions Models by Products

Pricing Model	count	mean	std	25%	50%	75%
One-Off-Purchase (0/1)	3683	0.464	0.499	0	0	1
Monthly Licensing (0/1)	3683	0.277	0.448	0	0	1
Yearly Licensing (0/1)	3683	0.384	0.487	0	0	1
Usage-Based (0/1)	3683	0.370	0.483	0	0	1

Notes. This table summarizes the transactions models available at the product-level from Datarade.

available for a product. About 60% of the products that offer one-time purchasing also offer yearly licensing.

5.1 Financial Constraints and Data Transaction Model

The first prediction of the model (H1) that we test pertains to the relationship between data sellers' financial conditions and the type of pricing models they adopt. To do this, we estimate

$$TransactionModel_i = \alpha + \beta \cdot Financing_i + \theta_{age} + \theta_{category} + \varepsilon_i. \tag{5.1}$$

The dependent variables are dummy variables indicating whether the data product offers a certain transaction model (one-off-purchase, licensing, or usage-based). The key explanatory variable *Financing* takes multiple forms. In one specification, financing measures the total number of rounds of VC financing; in another, it measures the total amount of VC financing obtained by the data provider. In this model and others in this section, we control for fixed effects of provider company age and its primary data category as tagged on Datarade. Standard errors are clustered at the level of data categories to account for correlations of transaction models among similar data products. Our model predicts a negative β coefficient for the one-off-purchase model, and positive β coefficients for licensing and usage-based models.

Consistent with the model's predictions, Table 4 shows that more financially constrained firms are more likely to choose data sales. In Panel A, we present the results using the logarithm of the number of funding rounds of the provider as the explanatory variable. In columns (1)-(3), we find that the number of funding rounds is associated with

Table 4: Providers' Financial Condition and Transaction Models

Panel A: Total Funding Rounds and Transaction Model

	(1) One-Off-Purchase	(2) Monthly Licensing	(3) Yearly Licensing	(4) Usage-Based
ln(No. of Funding Rounds)	-0.075***	0.092***	0.109***	0.011
	(0.016)	(0.014)	(0.015)	(0.015)
Observations	3,683	3,683	3,683	3,683
R-squared	0.006	0.012	0.014	0.000

Panel B: Total Funding Amounts and Transaction Models

	(1) One-Off-Purchase	(2) Monthly Licensing	(3) Yearly Licensing	(4) Usage-Based
ln(Total Funding Amt)	-0.003*	0.012***	0.012***	0.005***
	(0.001)	(0.001)	(0.001)	(0.001)
Observations	3,683	3,683	3,683	3,683
R-squared	0.001	0.023	0.019	0.003

Notes. This table correlates the type of data transaction models available for each data product with the funding status of the providers. Panel A presents the analysis using the logarithm of total funding rounds, and Panel B presents the analysis using the logarithm of the total funding amount received. All specifications control for provider age and primary data category fixed effects. *$<$ 0.1, **$<$ 0.05, ***$<$ 0.01.

a lower probability of offering one-off-purchase models but a higher probability of offering licensing, both monthly and yearly. We do not find a statistically meaningful relation between the financing variable and the usage-based transaction model.

In terms of economic magnitude, going from a firm with one round of financing to a firm with two rounds, the probability of using the one-off-purchase model goes down by $0.075 \times (\ln 2 - \ln 1) = 5.2$ percentage points (pp), which is an 11.2% from the base rate of 46.4%. Applying a similar calculation to column (2), we find that going from one round of financing to two rounds of financing is associated with a 6.4 pp (23.0% from the base) increase in the probability of using monthly licensing, and 7.5 pp (19.7% from the base) increase of the probability of using yearly licensing.

In Panel B, we show that the results are robust to the use of the logarithm of total funding amounts as the key explanatory variable. The patterns are consistent with Panel

A regarding one-off-purchases and licensing models, while in this case, the total funding amount is also predictive of the adoption of the usage-based model. The economic magnitudes are also sizable. For example, going from no-funding to 1 million funding is associated with a 4.1pp (8.8% from the base) decrease in the probability of using one-off purchases; associated with 16.6 pp (59.9% from the base) increase in the use of monthly licensing, and 16.6pp (43.2% from the base) increase in the use of yearly licensing.

5.2 Data Depreciation

Next, we explore (H2), that data depreciation makes data sales more likely. In our model, data about environments that change quickly (fast-changing finance data vs. slow-moving consumer tastes) is like an environment with a higher discount rate. As Figure 3 shows, data providers may have more commitment power and there is likely less loss from the one-time fee model.

To test this, we again use the transaction model as a proxy for a provider's commitment power. We use the following model:

$$TransactionModel_i = \alpha + \beta \cdot Depreciation_i + \theta_{age} + \theta_{category} + \varepsilon_i. \tag{5.2}$$

In this empirical model, the key explanatory variable is $Depreciation_i$, which is calculated using the steps outlined in Section 4.2—we map each product to its closely related industries using textual similarities between data product descriptions and public firm 10-K filings, and then calculate the industry-level R^2 using lagged ROA to predict concurrent ROA. For interpretation purposes, we cut the products into 10 deciles based on the average R^2 in matched industries, lower R^2 is high depreciation.

Table 5 shows that the information depreciation rate associated with a data product positively correlates with the use of one-off-purchases, and negatively predicts the use of licensing models. The economic magnitude is sizable. For example, the 0.027 in column (1) suggests that raising the depreciation rate by one decile predicts a 2.7 percentage points higher rate of using one-off-purchase. This increase is 5.8% of the base rate reported in Table 3. Thus, data that depreciates is more likely to be sold with a single transaction,

Table 5: Data Depreciation Rate and Transaction Models

	(1) One-Off-Purchase	(2) Monthly Licensing	(3) Yearly Licensing	(4) Usage-Based
Depreciation	0.027***	-0.012***	-0.011***	0.023***
	(0.003)	(0.003)	(0.003)	(0.003)
Observations	3,667	3,667	3,667	3,667
R-squared	0.025	0.006	0.004	0.018

Notes. This table correlates the type of data transaction models available for each data product with the information depreciation rate of connected industries. The equation estimated is (5.1). Depreciation is defined in Section 4.2. All specifications control for provider age and primary data category fixed effects. *$<$ 0.1, **$<$ 0.05, ***$<$ 0.01.

rather than a subscription, which allays fears about data seller market power. The same change is associated with a 1.2 pp decrease (4.3% from the base) in the probability of using monthly licensing, and a 1.1 pp (2.9% from the base) decrease in the probability of using yearly licensing.

5.3 Potential Market Size

Next, we explore H3, that a larger market size makes data subscriptions or licensing more likely. In the model, the market size is the potential buyers that may find the data product useful and thus may make a purchase. This potential market size further lowers a data provider's commitment power and loosens the financial constraint, both of which make a data subscription more profitable than data sales.

$$TransactionModel_i = \alpha + \beta \cdot MarketSize_i + \theta_{age} + \theta_{category} + \varepsilon_i. \tag{5.3}$$

The key explanatory variable in this model is *MarketSize*, which intends to capture the number of potential buyers of a data product. To achieve this goal, we use the total number of public firms in the connected industries (as defined in Section 4.2). A larger number of connected firms means a greater potential market size. We take a logarithm of this counting variable.

Table 6: Potential Market Size and Transaction Models

	(1) One-Off-Purchase	(2) Monthly Licensing	(3) Yearly Licensing	(4) Usage-Based
Market Size	-0.015	0.144***	0.182***	0.010
	(0.018)	(0.016)	(0.017)	(0.017)
Observations	3,683	3,683	3,683	3,683
R-squared	0.000	0.022	0.030	0.000

Notes. This table correlates the type of data transaction models available for each data product with the potential market size of each product. The equation estimated is (5.3). Market size is defined in Section 4.2. All specifications control for provider age and primary data category fixed effects. *$<$ 0.1, **$<$ 0.05, ***$<$ 0.01.

Table 6 shows the relationship between market size and pricing models. We find that market size, though having little correlation with the use of one-off-purchases and usage-based transaction models, strongly correlates with the use of the licensing model in transactions. The coefficient of 0.144 in column (2) means that a 10% increase in the potential market size is associated with a 1.4 percentage point increase in the likelihood of using monthly licensing. This is a 5.2% increase from the base rate.

This suggests that as the market for data grows, there is likely to be more use of subscriptions, which are more adept at extracting consumer surplus. But these facts also point to more provision of high-quality data.

5.4 Slower Sales Growth

The final prediction of the model, H4, gets to the core of the mechanism. The problem a data seller faces is that they need to restrict new data sales to earn monopoly rents. This need to restrict sales is not a problem faced by most other non-data tech firms in the industry because more technologies do not exhibit strategic substitutability. A computer is not less valuable when others buy the same computer. Instead, for most technology products the reverse is true: If everyone purchases and uses Apple products, for example, using non-Apple products becomes less compatible with others and less valuable. Therefore, we test H4 by comparing the growth in customers and revenue of data firms to their

non-data counterparts after controlling for time or venture geographic market trends.

Table 7: Time-Series Trend of Data Value and Potential Customers

	(1) ΔVenture Value	(2) ΔGoogle Trends Index
Data Provider	-0.526**	-0.066*
	(0.216)	(0.035)
Observation-level	Company-Round	Search Term-YearMonth
Observations	54,124	19,741
R-squared	0.024	0.012
Fixed Effects	Year State	Year-Month

Notes. ΔVenture Value, which captures the percentage change of the current financing round from the previous round. Source: Crunchbase and PitchBook. ΔGoogle Trends Index is the percentage change in the Google Trends index from 2018 to 2023 for data and non-data tech firms. The model controls for company age fixed effects. *$<$ 0.1, **$<$ 0.05, ***$<$ 0.01.

Column (1) of Table 7 examines the time-series value dynamic of startup companies, comparing data provider companies with non-data-provider companies in related industries. The data are at the level of company-financing round, and each observation is a financing round of a startup company. The data provider companies are data providers on Datarade that can be matched to Crunchbase and PitchBook. The control sample includes companies in Crunchbase and PitchBook that are in top 5 industries that data providers operate in. The table reports the β estimate from the following model,

$$\Delta VentureValue_{it} = \alpha + \beta \cdot DataProvider_i + \theta_{t \times State} + \varepsilon_{it}.$$

The key dependent variable is ΔVenture Value, which captures the percentage change of the current financing round from the previous round. The key explanatory variable is *DataProvider* indicating if the company is a data provider. We control for the lag from the previous round to this round, control for year-by-state fixed effects, and cluster standard errors at the year and state levels. The thought experiment is that we are comparing two companies, one is a data provider and one is not, that raised funding in the same year-state, with the same gap from the previous round—we see that data providers value

growth (Δ) is significantly lower than the comparable companies.

Column (2) examines the time-series change in search popularity captured using Google Trends. This analysis compares Google Trends index from 2018 to 2023 of two groups of terms: data provider-related terms consisting of all data categories that are extracted from Datarade (i.e., "commercial market data," "business location data"); as control group, we use top breakthrough technologies identified annually by MIT Technology Review during the sample period (i.e., "custom cancer vaccines," "3-D metal printing," "online privacy").[4] For each of these terms, we extract the monthly Google Trend index. The table reports the estimated value of β from the following model,

$$\Delta GoogleTrend_{it} = \alpha + \beta \times DataProvider_i + \theta_t + \varepsilon_{it}.$$

The key dependent variable is ΔGoogleTrend, which captures the percentage change of the current Google Trends from the previous round. The key explanatory variable is *DataProvider* indicating if the term is a data provider-related term or a general technology term. We control for year-month fixed effects and cluster standard errors at the same level. In this thought experiment, we are comparing data terms with other technology terms in terms of their monthly change in search popularity, after controlling for granular time trends using year-month fixed effects.

A potential measurement challenge is selection: Most surviving firms grow their sales over time. However, our approach of taking the difference between data and non-data firms should remove this effect. As long as data firms have a similar selection of surviving firms as their non-data counterparts, this effect should disappear in the difference. However, future work is to investigate whether the failure rates are indeed similar.

6 Quantifying the Welfare Losses

Our model is a stylized one. It surely does not contain all the welfare-relevant tradeoffs one would want for a thorough policy analysis. However, some rough quantification

[4]The MIT Technology Review's annual breakthrough technologies can be accessed at: https://www.technologyreview.com/supertopic/tr10-archive/.

of the model can give us an idea of the magnitude of the losses from the mechanism we describe. Our results, while not comprehensive, suggest that the way in which sales models regulate monopoly power can have a quantitatively significant effect on consumer surplus and on welfare.

6.1 Consumer Surplus: Data Sales vs. Data Subscriptions

Data poses a trade-off for consumer surplus. Firms without data do not have monopoly pricing power. Lower prices benefit consumers. However, data makes firms more efficient. Firms without data have higher marginal costs, which get passed on to consumers as well. That trade-off shows up throughout our comparison of data market structures as well. To quantify the trade-off, we first derive expressions for consumer surplus from the model, in the case of data sales and data licensing.

To compute consumer surplus, we simple substitute in the solutions for the equilibrium price and quantity p_i and q_i for each variety. Recall that each variety has three possible market structures: 1) both firms have data, 2) one has data, the other does not, and 3) neither has data. Then, we multiply each of these surpluses times the fraction of varieties that yield that surplus. This yields a one-period consumer surplus, as a function of data supplier choices n and z. Finally, we substitute in these data producer choices and cumulate up the one-period surpluses to yield total lifetime consumer surplus. Appendix A follows these steps and prove the following result.

Proposition 4. The lifetime consumer surplus, when data is sold is

$$\sum_{t=0}^{\infty} \beta^t u_t^{sale} = \frac{1}{\sigma - 1} \left[\frac{1 + x - 2\left(\frac{\sigma-1}{\sigma}\right)^{\sigma-1} x}{1 - \beta\delta^2} + \frac{\left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1} x}{1-\beta} + \frac{2x\left(1 - \left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1} x\right)}{1 - \beta\delta} \right], \quad (6.1)$$

where $\delta = \frac{1-\sqrt{1-\gamma}}{\gamma}$, and $x = 1 + \frac{1}{\sigma} \frac{1}{2\left(1+\sqrt{1-\gamma}-\beta\right)}$.

The lifetime consumer surplus, when data is licensed as a subscription is

$$\sum_{t=0}^{\infty} \beta^t u_t^{sub} = \frac{1}{1-\beta} \frac{1}{\sigma - 1} \left[\frac{1}{4} \left(1 + \left(\frac{\sigma}{\sigma - 1} \right)^{\sigma-1} \hat{x} + 2\hat{x} \right) \right], \quad (6.2)$$

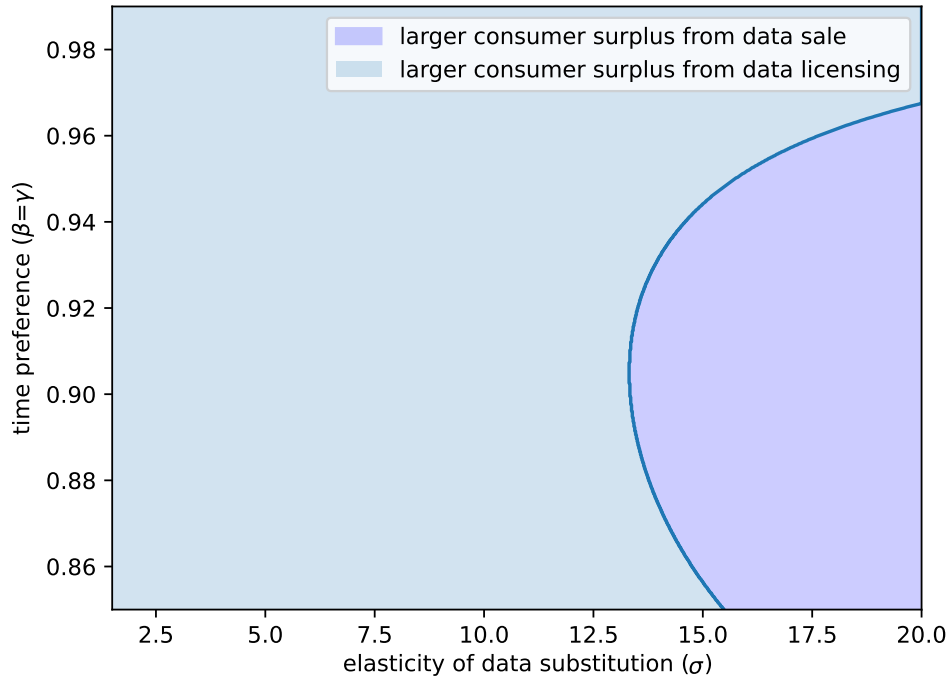
where $\hat{x} = 1 + \frac{1}{\sigma}\frac{1}{2(1-\beta)}$.

In both cases, δ represents the rate at which data sales converse to their lone-run level and x represents the ex-ante data quality choice. It is a monotonic transformation of z.

Even though data sellers can restore their monopoly power using data subscriptions, it is not obvious that this is better for consumers. The two consumer surplus expressions in Proposition 4 are not easily rankable. In some settings, either model could be superior. The reason is that when firms get less profit, they also invest in lower quality data, which also harms consumers.

To see how surplus depends on the model parameters, we choose some values for time preference and elasticity and plot regions where each sales model dominates.

Figure 5: Do Data Sales or Data Subscriptions Maximize Consumer Surplus?



Notes. The darker region indicates combinations of parameters β and σ for which data sale yields higher consumer surplus, i.e., (6.1) $\geq$ (6.2). σ governs the elasticity of substitution between varieties of goods. Depreciation of old data in subscription/licensing model is $\omega \leq 1/z$.

6.2 Calibrating the Model

There are two fundamental parameters of the model, the time preference β and the elasticity of substitution σ. All other variables are choices or functions of these two.

Since there is no aggregate risk in the model, the rate of time preferences should be the inverse of a gross riskless rate $(1 + r)$. The average three-month treasury bill rate since 1954 has been 4.2%.[5] Therefore, we use a rate of time preference of $\beta = 1/(1.042) = 0.96$.

The elasticity of substitution is tougher to estimate. We could use existing estimates of elasticities from the industrial organization literature. However, the substitutability of one data set for another could be quite different from the substitutability of breakfast cereals, video games or ready-mix concrete. Since it is important to get the domain-specific substitutability, we will use our data marketplace data and our model to impute an elasticity value.

Since we have not yet achieved a satisfactory calibration of the elasticity, for now, we show results for a range of elasticities. Future approaches might use goods markups or data seller profits to impute an average elasticity.

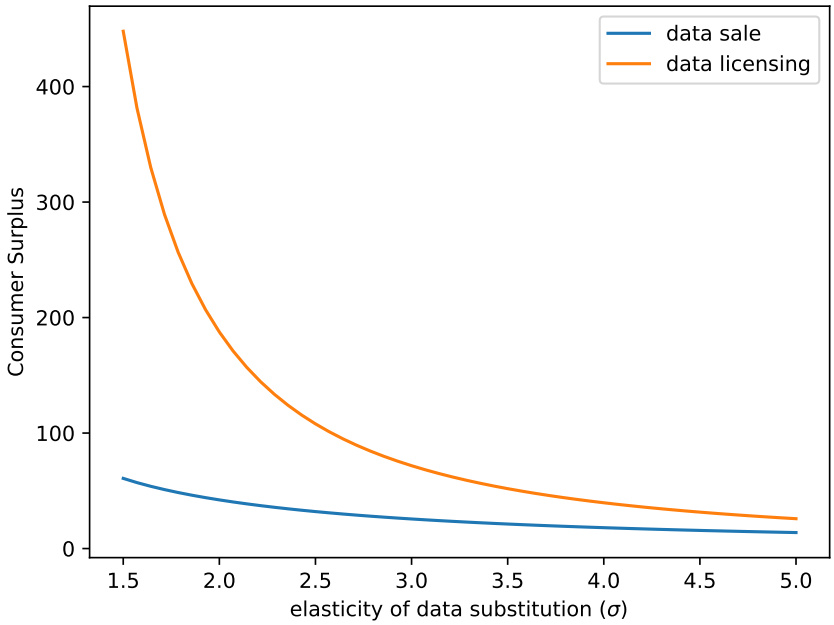
However we calibrate the elasticity, the estimate will be undoubtedly rough. We will use this magnitude to calibrate our model, mostly as a proof of concept. At best, this might give us a sense of the order of magnitude our effect might have.

6.3 Welfare results

Finally, we can use the model to answer the policy relevant question about what sorts of data sales are better for consumers. Figure 6 shows that, unless goods are highly elastic, data licensing or subscription is better for consumers than data sales. At first, that might seem surprising. After all, data sellers make less use of their monopoly power. Monopoly power usually creates deadweight loss that makes consumers worse off. So data sales would seem to be better for consumers. However, because data sellers lose most of their rents from data, they have little incentive to invest in data quality. Since higher quality data makes goods firms more efficient, which in turn, makes good cheaper, consumers

[5] Source: https://fred.stlouisfed.org/series/TB3MS.

Figure 6: Welfare from data sales and data subscriptions



Notes. Welfare is consumer surplus plus producer surplus. Consumer surplus is given by (6.1) for data sales and (6.2) for data licensing or subscriptions. Producer surplus is the value V in (3.2) and (3.3). The elasticity parameter σ is reported on the x axis of the plot. The discount rate is $\beta = 0.96$.

benefit when more data is sold. Only if goods are highly elastic might data sales be better because, in this case, data producers have little incentive to invest in quality data anyway.

The last question we pose to the model is whether firms' incentives to choose data sales or data subscriptions aligns with welfare. Figure 3 revealed that firms choose data subscriptions when they are patient, relative to data buyers. Since we interpret patient firms to mean non-financially-constrained firms, this suggests that firms without financial constraints are more likely to make welfare-maximizing choices about their means of selling data. However, in many cases firms incentives to choose the right data business model will be misaligned with the consumers' interests. This suggests a role for regulation or subsidies to ensure sufficient data production.

7 Conclusion

Many policy makers are concerned about the market power of data sellers. If most data were sold with a one-time fee and data purchasers are reasonably patient, then market power in data markets should not be regulated for its own sake. The inability of data sellers to commit to sell limited copies of data, combined with the fact that data's strategic value declines in the number of users, forces competitive pricing. Even if the seller is a monopolist, the inability to restrict future data sales makes the seller compete with its future self in data provision. Of course, with the loss of monopoly power comes a loss of incentive to produce quality data.

Not only is monopoly power not entirely bad, one might consider protecting it. Just like patent laws protect the monopoly power of innovators to encourage innovation, copyright law could be seen as protection for data to encourage the discovery of new, high-quality data sources.

However, data subscription services are a tool for firms to restore monopoly power. While subscriptions restore monopoly power, they also restore an incentive for data sellers to invest in producing high-quality data. Our quantitative analysis of the model teaches us that, on net, subscriptions benefit consumers for all but the most elastically demanded goods.

Market power in data markets depends on the choice of pricing models data sellers choose to implement. We collected data from one of the largest on-line data marketplaces to investigate data sellers' pricing strategies. We find that, while subscription models are the tool of choice for most firms, finally constrained firms may still choose one-time data sales because it produces more cash flow up-front. Firms selling data with few customers face less cost from adopting one-time fees. Similarly, firms selling data that depreciates rapidly face little effect of future sales. Such firms also choose one-time fees. These findings align with the theory but also give us new insight into the functioning of this rapidly-expanding and politically controversial market.

References

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. 2021. "Too much data: Prices and inefficiencies in data markets." *Forthcoming, American Economic Journal: Microeconomics.*

Arnott, Richard, and Joseph Stiglitz. 1991. *Equilibrium in Competitive Insurance Markets with Moral Hazard.* NBER Working Papers 3588. National Bureau of Economic Research, Inc, January.

———. 1993. "Price Equilibrium, Efficiency, and Decentralizability in Insurance Markets with Moral Hazard." *Working Paper.*

Bergemann, Dirk, and Alessandro Bonatti. 2022. *Data, Competition, and Digital Platforms.* Technical report.

Brunnermeier, Markus K., and Martin Oehmke. 2013. "The Maturity Rat Race." *Journal of Finance* 68, no. 2 (April): 483–521.

Cong, Lin William, Danxia Xie, and Longtian Zhang. 2021. "Knowledge Accumulation, Privacy, and Growth in a Data Economy." *Management Science* 67 (10): 6480–6492. ISSN: 15265501. https://doi.org/10.1287/mnsc.2021.3986.

DeMarzo, Peter, and Zhiguo He. 2021. "Leverage Dynamics without Commitment." *Journal of Finance* 76 (3): 1995–1250.

Eckbo, B. Espen, ed. 2008. *Handbook of Empirical Corporate Finance SET.* Elsevier Monographs 9780444532657. Elsevier.

Farboodi, Maryam, and Laura Veldkamp. 2022. *A Model of the Data Economy.* Working Paper, Working Paper Series 28427. National Bureau of Economic Research. http://www.nber.org/papers/w28427.

Fudenberg, Drew, and Jean Tirole. 1991. *Game Theory.* The MIT Press.

Green, Daniel, and Ernest Liu. 2021. "A Dynamic Theory of Multiple Borrowing." *Journal of Financial Economics* 139 (2): 389–404.

Grossman, Sanford, and Joeseph Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *American Economic Review* 70(3):393–408.

Grossman, Sanford J, and Joseph E Stiglitz. 1980. "On the impossibility of informationally efficient markets." *The American economic review* 70 (3): 393–408.

Hellwig, Christian, Sebastian Kohls, and Laura Veldkamp. 2012. "Information choice technologies." *American Economic Review* 102 (3): 35–40.

Ichihashi, Shota. 2020. "Online Privacy and Information Disclosure by Consumers." *American Economic Review* 110, no. 2 (February): 569–595. https://ideas.repec.org/a/aea/aecrev/v110y2020i2p569-95.html.

Jones, Charles I, and Christopher Tonetti. 2020. "Nonrivalry and the Economics of Data." *American Economic Review* 110 (9): 2819–58.

Kirpalani, Rishabh, and Thomas Philippon. 2020. *Data sharing and market power with two-sided platforms.* Technical report. National Bureau of Economic Research.

Lambrecht, Anja, and Catherine E. Tucker. 2015. *Can Big Data Protect a Firm from Competition?* Technical report. mimeo.

Segal, I. 1999. "Contracting with externalities." *Quarterly Journal of Economics* CXIV (May): 337–388.

Appendix

A Proofs

Proofs of Propositions 1, 2, and 3

Proof. First conjecture $\bar{\pi}(n) = \frac{\xi(a-bn)}{1-\beta}$ and we solve the data producer's problem.

$$V(n) = \max_{\widetilde{n}} \left\{ (\widetilde{n} - n) \frac{(a - b\widetilde{n})}{\sqrt{1-\beta}} + \beta V(\widetilde{n}) \right\}.$$

$$\text{FOC} \quad \frac{(a - bg(n))}{\sqrt{1-\beta}} - \frac{(g(n) - n)\, b}{\sqrt{1-\beta}} + \beta V'(n') = 0$$

$$\text{Envelope} \quad V'(n) = -\frac{(a - bg(n))}{\sqrt{1-\beta}}$$

Substitute the envelope condition into the first-order condition:

$$(a - bg(n)) - (g(n) - n)\, b - \beta \left[a - bg^2(n) \right] = 0$$

Conjecture $a - bg(n) = \delta(a - bn)$, we get

$$\left(2\delta - 1 - \beta\delta^2 \right)(a - bn) = 0$$

Given $\delta \in (0,1)$, the solution is $\delta = \frac{1-\sqrt{1-\beta}}{\beta}$.

We now solve for the data buyer firms' willingness to pay under rational expectation:

$$\begin{aligned} \bar{\pi}(n) &= \sum_{t=1}^{\infty} \beta^{t-1} \left[a - bg^{t-1}(n) \right] \\ &= \frac{a - bn}{1 - \beta\delta} \\ &= \underbrace{\frac{1-\beta}{1-\beta\delta}}_{\xi} \frac{a - bn}{1-\beta} \end{aligned}$$

Substitute $\delta = \frac{1-\sqrt{1-\beta}}{\beta}$, we can simplify $\xi = \sqrt{1-\beta}$.

Using these solutions, we can solve for the data producer's value function:

$$\begin{aligned} V(n) &= \frac{\xi}{1-\beta}\sum_{t=1}^{\infty}\beta^{t-1}\left(g^t(n)-g^{t-1}(n)\right)\left(a-bg^t(n)\right) \\ &= \frac{\xi}{1-\beta}\sum_{t=1}^{\infty}\beta^{t-1}\delta^{t-1}(1-\delta)\left(\frac{a}{b}-n\right)\delta^t(a-bn) \\ &= \frac{\xi}{1-\beta}\sum_{t=1}^{\infty}\beta^{t-1}\delta^{2t-2}\delta(1-\delta)\frac{1}{b}(a-bn)^2 \\ &= \frac{\xi\delta(1-\delta)\frac{1}{b}(a-bn)^2}{(1-\beta)(1-\beta\delta^2)} \\ &= \frac{\delta(1-\delta)b(\bar{n}-n)^2}{(1-\beta\delta)(1-\beta\delta^2)} \end{aligned}$$

where $\bar{n} \equiv a/b$. The value at time 0 is

$$V(0) = \frac{\delta(1-\delta)a^2/b}{(1-\beta\delta)(1-\beta\delta^2)} \tag{A.1}$$

$$= \frac{\delta a^2}{2b\sqrt{1-\beta}} \tag{A.2}$$

$$= \frac{a^2}{2b}\frac{\sqrt{1-\beta}-(1-\beta)}{\beta(1-\beta)} \tag{A.3}$$

The proofs of propositions 2 and 3 follow analogous steps by substituting the respective value functions and then guess-and-verify. □

Proof of Proposition 4 In case 1) where both firms have data, the price is the firms' marginal cost, which is $1/z$. Since demand is $q_i = p_i^{-\sigma}$, substituting these into consumer utility (1.1) yields a one-period consumer surplus of $v_1 = z^{\sigma-1}/(\sigma-1)$.

In case 2) where firms are asymmetric, the price was $z\sigma/(\sigma-1)$, which implies a quantity of $(z\sigma/(\sigma-1))^{-\sigma}$. Substituting price and quantity into consumer utility (1.1) yields a one-period consumer surplus of $v_2 = \frac{1}{\sigma-1}\left(z\frac{\sigma-1}{\sigma}\right)^{\sigma-1}$.

In case 3) where neither firm has data, the price and quantity are both 1. One-period consumer surplus is $v_3 = 1/(\sigma-1)$.

Multiplying each of these three consumer surplus expressions by the fraction of vari-

eties that have each market structure (the probabilities), we can express total consumer surplus as a function of number of copies of data sold n_t and data quality z:

$$u_t = \frac{1}{\sigma - 1}\left[\left(\frac{2 - n_t}{2}\right)^2 + \left(\frac{\sigma}{\sigma - 1}\right)^{\sigma - 1} z\frac{n_t^2}{4} + \frac{2n_t\left(2 - n_t\right)}{4}z\right]$$

Of course, the copies of data sold and the data quality are also endogenous choices of the data provider. The next step is substitute those in.

Given the equilibrium policy function in (1.7): $n_t = (1 - \delta)\frac{a}{b} + \delta n_{t-1}$ with $\frac{a}{b} = 2$, we know

$$2 - n_t = \delta\left(2 - n_{t-1}\right) = 2\delta^t$$

$$\left(2 - n_t\right)^2 = 4\delta^{2t}$$

$$n_t^2 = \left(2 - 2\delta^t\right)^2 = 4 + 4\delta^{2t} - 8\delta^t$$

$$2n_t\left(2 - n_t\right) = 4\left(2 - 2\delta^t\right)\delta^t = 8\delta^t - 8\delta^{2t}$$

$$\sum_{t=0}^{\infty}\beta^t\left(\frac{2 - n_t}{2}\right)^2 = \sum_{t=0}^{\infty}\beta^t\delta^{2t} = \frac{1}{1 - \beta\delta^2}$$

$$\sum_{t=0}^{\infty}\beta^t\frac{n_t^2}{4} = \frac{1}{1 - \beta} + \frac{1}{1 - \beta\delta^2} - \frac{2}{1 - \beta\delta}$$

$$\sum_{t=0}^{\infty}\beta^t\frac{2n_t\left(2 - n_t\right)}{4} = \frac{2}{1 - \beta\delta} - \frac{2}{1 - \beta\delta^2}$$

We can write the consumer's ex-ante surplus as

$$\begin{aligned}
&\sum_{t=0}^{\infty}\beta^t u_t \\
&= \frac{1}{\sigma - 1}\left[\frac{1}{1 - \beta\delta^2} + \left(\frac{\sigma}{\sigma - 1}\right)^{\sigma - 1} x\left(\frac{1}{1 - \beta} + \frac{1}{1 - \beta\delta^2} - \frac{2}{1 - \beta\delta}\right) + x\left(\frac{2}{1 - \beta\delta} - \frac{2}{1 - \beta\delta^2}\right)\right] \\
&= \frac{1}{\sigma - 1}\left[\frac{1 + \left(\frac{\sigma}{\sigma - 1}\right)^{\sigma - 1} x - 2x}{1 - \beta\delta^2} + \left(\frac{\sigma}{\sigma - 1}\right)^{\sigma - 1} x\left(\frac{1}{1 - \beta} - \frac{2}{1 - \beta\delta}\right) + x\frac{2}{1 - \beta\delta}\right] \qquad \text{(A.4)} \\
&= \frac{1}{\sigma - 1}\left[\frac{1 + x - 2\left(\frac{\sigma - 1}{\sigma}\right)^{\sigma - 1} x}{1 - \beta\delta^2} + \frac{\left(\frac{\sigma}{\sigma - 1}\right)^{\sigma - 1} x}{1 - \beta} + \frac{2x\left(1 - \left(\frac{\sigma}{\sigma - 1}\right)^{\sigma - 1} x\right)}{1 - \beta\delta}\right] \qquad \text{(A.5)}
\end{aligned}$$

Under data sales, we have $\delta = \frac{1-\sqrt{1-\gamma}}{\gamma}$, and ex-ante data choice $x = 1 + \frac{1}{\sigma}\frac{1}{2\left(1+\sqrt{1-\gamma}-\beta\right)}$. We can compute consumer surplus by substituting δ and x into (A.5).

Under subscription, $n^* = 1$, $x = 1 + \frac{a^2}{4b(1-\beta)} = 1 + \frac{1}{\sigma}\frac{1}{2(1-\beta)}$. The consumer surplus per period is

$$u_t^{sub} = \frac{1}{\sigma - 1}\left[\frac{1}{4}\left(1 + \left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1} x + 2x\right)\right].$$

The ex-ante consumer surplus is

$$\sum_{t=0}^{\infty} \beta^t u_t^{sub} = \frac{1}{1-\beta}\frac{1}{\sigma-1}\left[\frac{1}{4}\left(1 + \left(\frac{\sigma}{\sigma-1}\right)^{\sigma-1} x + 2x\right)\right].$$

B Market Foundations for Data Externality

Strategic substitutability of data arises in many contexts. Here is a simple one where there is imperfect competition and firms use data to forecast uncertain shocks to their profit.

FIRMS There are n_F firms, indexed by i: $i \in \{1, 2, \ldots, n_F\}$. Each firm chooses the number of units of each good they want to produce, an $N \times 1$ vector $\mathbf{q}_i$, to maximize risk-adjusted profit, where the price of risk is ρ_i.

$$U_i = \mathbf{E}\left[\pi_i | \mathcal{I}_i\right] - \frac{\rho_i}{2}\mathbf{Var}\left[\pi_i | \mathcal{I}_i\right] - g(\chi_c, \widetilde{\boldsymbol{c}}_i). \tag{B.1}$$

This mean-variance objective is consistent with empirical corporate finance evidence on firms' decisions (Eckbo 2008) and is a second-order approximation to a broader class of firm managers' utility functions.

Firm production profit π_i depends on quantities of each good, $\mathbf{q}_i$, the market price of each good, $\mathbf{p}$, and the marginal cost of production of that good, $\boldsymbol{c}_i$:

$$\pi_i = q_i'\left(\mathbf{p} - \boldsymbol{c}_i\right). \tag{B.2}$$

PRODUCTS AND ATTRIBUTES The product space has N attributes, indexed by $j \in \{1, 2, \ldots, N\}$. Goods, indexed by k, are combinations of attributes.

Each good $k \in \{1, 2, \ldots, N\}$ can be represented as an $N \times 1$ vector $\boldsymbol{a}_k$ of weights that good places on each attribute. The jth entry of vector $\boldsymbol{a}_k$ describes how much of attribute j the kth good requires. This collection of weights describes a good's location in the product space. Let the collection of $\boldsymbol{a}_k$'s for each good k be an $N \times N$, full-rank matrix A.

The quantity of attributes that a firm i produces is a vector $\tilde{\mathbf{q}}_i$, with jth element $\tilde{q}_{ij}$. The attribute vector is the vector of firm i's product quantities, q_i, times the inverse attribute matrix A^{-1}:

$$\tilde{\mathbf{q}}_i = A^{-1}\mathbf{q}_i. \tag{B.3}$$

The marginal cost of producing a good is $\boldsymbol{c}_i$. The firm produces each attribute j at a unit cost of $\widetilde{c}_{ij}$. The vector $\widetilde{\boldsymbol{c}}_i$ is the N-by-1 vector of all marginal production costs of firm i for each attribute. The vector $\boldsymbol{c}_i = A'\widetilde{\boldsymbol{c}}_i$ is the vector of firm i's marginal cost for each product. The cost of producing a unit of good k for firm i is therefore $\boldsymbol{c}_i = \boldsymbol{a}_k'\widetilde{\boldsymbol{c}}_i$.

PRICE Our demand system embodies the idea that goods with similar attributes are partial substitutes for each other. Therefore, the price of good i can depend on the amount every firm produces of every good.

The price of each good depends on the attributes of a good. The price of good k is the units of each attribute a_k times the price of each attribute $\tilde{p}$:

$$p_k = \sum_{j=1}^{N} a_{jk}\tilde{p}_j. \tag{B.4}$$

Each attribute j has an average market price that depends on an attribute-specific constant and on the total quantity of that attribute that all firms produce:

$$\tilde{p}_j^M = \bar{p}_j - \frac{1}{\phi}\sum_{i=1}^{n_F}\tilde{q}_{ij}. \tag{B.5}$$

Each firm does not receive the market price for its good, but rather has a firm-specific price that depends on a firm-specific demand shock $\mathbf{b}_i$. The demand shock $\mathbf{b}_i$ is a vector with jth element b_{ij}. This vector is random and unknown to the firm: $\mathbf{b}_i \sim N(0, I)$, which is i.i.d. across firms. The price a firm receives for a unit of attribute j is thus $\tilde{p}_j + b_{ij}$. We

can express firm i's price in vector form as

$$\tilde{p}_i = \left[\tilde{p}_1^M, \tilde{p}_2^M, \ldots, \tilde{p}_N^M\right]' + b_i. \tag{B.6}$$

The price a firm receives for a unit of good k is therefore $p_k + \sum_{j=1}^{N} a_{jk} b_{ij}$.

INFORMATION Each firm generates n_{di} data points. Each data point is a signal about the demands for each attribute: $\tilde{\mathbf{s}}_{i,z} = \boldsymbol{b}_i + \tilde{\boldsymbol{\varepsilon}}_{i,z}$, where $\tilde{\boldsymbol{\varepsilon}}_{i,z} \sim N(\mathbf{0}, \tilde{\Sigma}_e)$ is an $N \times 1$ vector. Signal noises are uncorrelated across attributes and across firms. All firms can observe all the data generated by each firm. Of course, other firms' data is not relevant for inferring b_i. But this allows firms to know what other firms will do.

Because we are interested in how data affects competition, we will take data (n_{di} and $\tilde{\Sigma}_e$) as given. The question will be what happens to market competition and markups when we exogenously change these data conditions of some or all firms. Section **??** explores what aspects of the results change when data is generated as a by-product of economic transactions.

EQUILIBRIUM

1. Each firm chooses a vector of marginal costs $\tilde{\boldsymbol{c}}_i$, taking as given other firms' cost choices. Since the data realizations are unknown in this ex ante investment stage, the objective is the unconditional expectation of the utility in (B.1).

2. After observing the realized data, each firm updates beliefs with Bayes' law and then chooses the vector $\mathbf{q}_i$ of quantities to maximize conditional expected utility in (B.1), taking as given other firms' choices.

3. Prices clear the market for each good.

Substitutability Externality of Information To show substitutability, we now want to consider, what happens when one additional firm gets a signal with one unit of precision, about consumer demand? How does that effect the utility of another firm observing that same amount of information?

We start with the optimal production decision of a firm. Define $\boldsymbol{H}_i = \left(\rho_i \mathbf{Var}\left[\boldsymbol{b}_i|\mathcal{I}_i\right] + \frac{2}{\phi}\boldsymbol{I}_N\right)^{-1}$. Using Bayes law to replace the expectation $\mathbf{E}\left[\boldsymbol{b}_i|\mathcal{I}_i\right]$ with the weighted sum of signals $\boldsymbol{K}_i\boldsymbol{s}_i$, with $\boldsymbol{K}_i = \Sigma_{b_i}(\Sigma_{b_i} + \Sigma_{\epsilon_i})^{-1}$ yields

$$\tilde{\boldsymbol{q}}_i = \boldsymbol{H}_i\left(\bar{\boldsymbol{p}} + \boldsymbol{K}_i\boldsymbol{s}_i - \frac{1}{\phi}\sum_{j=1,j\neq i}^{n_F}\tilde{\boldsymbol{q}}_j - \tilde{\boldsymbol{c}}_i\right). \tag{B.7}$$

We have set the model up so that the only way one firm's information affects another firm is through the level of production. Notice that the firm's output is increasing in H_i, which itself in decreasing in conditional variance. Data reduces conditional variance. Thus, data will increase a firm's expected level of production.

A one unit increase in the precision of data increases conditional precision by one unit. The decrease in conditional variance, the inverse of conditional precision is $-\mathbf{Var}\left[\boldsymbol{b}_i|\mathcal{I}_i\right]^2$.

The size of this effect of data on the own level of production is $\partial q_i/\partial n_{di} = -\partial E[q_i]/\partial H_i \cdot \partial H_i/\partial Var \cdot \mathbf{Var}\left[\boldsymbol{b}_i|\mathcal{I}_i\right]^2$. Since $E[q_i]/\partial H_i > 0$ but $\partial H_i/\partial Var < 0$, the effect of data on own production is positive. This makes sense because a firm with more data faces less uncertainty and produces more aggressively.

The effect of firm i's data on firm i' works through the price level. WHen firm i produces one unit more of attribute j, the price of attribute j falls by $1/\phi$. Thus, $\partial p/\partial n_{di} = -1/\phi\partial q_i/\partial n_{di}$.

Next, we solve for the effect on expected profits. Expected profits can be expressed as:

$$\mathbf{E}\left[\tilde{\boldsymbol{q}}_i'\left(\tilde{\boldsymbol{p}}_i - \tilde{\boldsymbol{c}}_i\right)\right] = \mathbf{E}\left[\tilde{\boldsymbol{q}}_i'\left(\mathbf{E}\left[\tilde{\boldsymbol{p}}_i|\mathcal{I}_i\right] - \tilde{\boldsymbol{c}}_i\right)\right]. \tag{B.8}$$

Notice that the effect of a one unit increase in price of an attribute is an increase of $\tilde{\boldsymbol{q}}_{i'}'$ in profits. Thus, putting these effects together with the chain rule, we find that the marginal effect of an increase in data owned by firm i on firm i's profit is $\tilde{\boldsymbol{q}}_{i'}' \cdot (-1)/\phi\partial q_i/\partial n_{di} < 0$. So one firm's data reduces another firm's profit.

C Foundations for Data Depreciation

To understand why data depreciates and how much it depreciates, we need to model how firms derive competitive advantage from data. Data is information. Big data, used with modern big data techniques is used for prediction. AI and machine learning are, at their core, prediction technologies. So the data we are talking about is information used to make predictions more accurate. More accurate predictions can inform more optimal or efficient actions. The greater efficiency of actions is the source of firms' competitive advantage. Understanding the role of data will allow us to deduce its depreciation rate.

Consider a firm that uses data with normally-distributed noise to forecast some profit-relevant variable that follows an AR(1) process with normal innovations:

$$\theta_{t+1} \;=\; \rho\theta_t + \zeta_{t+1}, \qquad \zeta_{t+1} \sim (0, \sigma_\zeta^2), \tag{C.1}$$

for $0 < \rho < 1$.

Perhaps the cost of production of the firm is related to the distance between an action a_{it} they choose and this state, $(a_{it} - \theta_t)^2$. The optimal choice of action each period would be to choose $a_{it} \;=\; E[\theta_t|\mathcal{I}_{it}]$. This would make the marginal cost the expected squared forecast error $(E[\theta_t|\mathcal{I}_{it}] - \theta_t)^2$, which is the definition of the conditional variance $V[\theta_t|\mathcal{I}_{it}]$.

The prior mean and variance are given by $E[\theta_t|\mathcal{I}_t]$ and $V[\theta_t|\mathcal{I}_t] := \eta_t^{-1}$, where $\mathcal{I}_t$ represents whatever information set the agent has at time t. We define η with the inverse because this lends itself to interpreting η_t as the amount of data. A lower variance estimate or more accurate estimate implies more data about θ_t.

Consider the variance of tomorrow's state, given today's data. Taking the variance of both sides of (C.1), we get $\mathrm{V}[\theta_{t+1}|\mathcal{I}_t] = \rho^2\eta_t^{-1} \;+\; \sigma_\zeta^2$.

This conditional variance is the expected squared forecast error: $V[\theta_{t+1}|\mathcal{I}_t] \equiv E[(\theta_{t+1} - E[\theta_{t+1}|\mathcal{I}_t])^2|\mathcal{I}_t]$. It reveals how inaccurate the firm's prediction is, or how poor or scarce their predictive data is. In Bayesian language, this is a prior variance of θ_{t+1}.

If the data used to forecast θ_{t+1} has normally distributed noise, then according to Bayes' Law, all newly-acquired data can be combined and represented as a signal about

tomorrow's state $s_t = \theta_{t+1} + e_t$, with $e_t \sim N(0, \sigma_e^2/m_{it})$, where m_{it} is the number of new data points firm i observes at time t, each with precision σ_e^{-2}. The $t+1$ information set is equivalent to $\mathcal{I}_{t+1} = \{\mathcal{I}_t, s_t\}$, which is the information available today, plus the signal observed at the end of period t.

According to Bayes' law, combining a normal prior belief with a normal signal yield a posterior precision that is the prior precision (the inverse of equation (C)), plus the precision of the new data $\sigma_e^{-2} m_{it}$:

$$\eta_{t+1} \;=\; (\rho^2 \eta_t^{-1} \;+\; \sigma_\zeta^2)^{-1} \;+\; \sigma_s^{-2}. \tag{C.2}$$

This law of motion for the amount of data says that we take the existing stock of data η_t, depreciate it by transforming it into $(\rho^2 \eta_t^{-1} + \sigma_\zeta^2)^{-1}$ and then add on the precision of newly-acquired data. This is similar to a law of motion for a stock of capital: $k_{t+1} = (1-\psi)k_t + i_t$, where i_t is new investment. For data that predicts a persistent process, the depreciation rate is

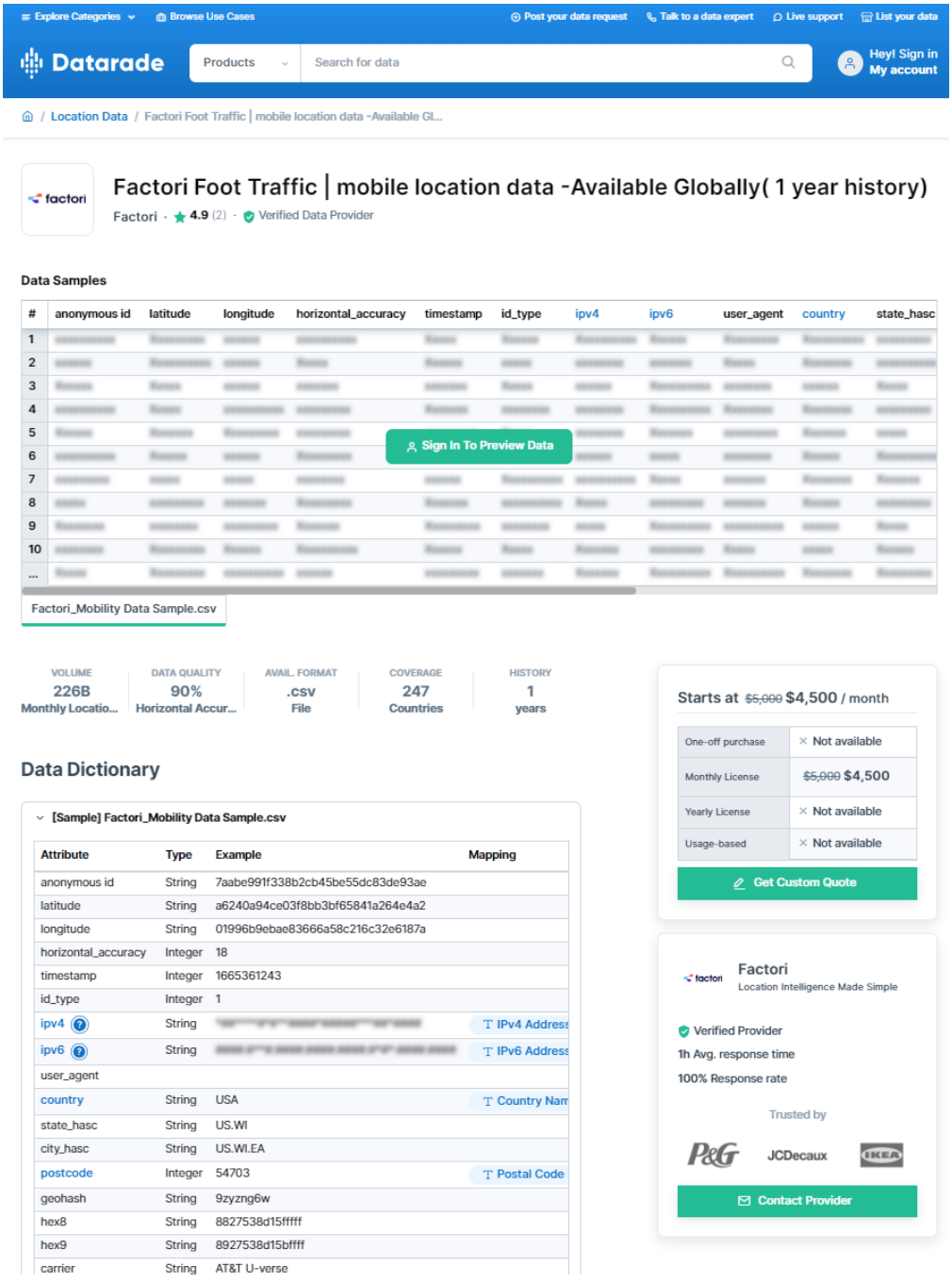
$$\psi_t \;=\; 1 - \frac{1}{\rho^2 + \sigma_\zeta^2 \eta_t}. \tag{C.3}$$

Note that if the AR(1) process is highly volatile (high σ_ζ), then the amount of data will depreciate quickly. Data about yesterday's state is less relevant to today's state because the state is changing quickly. This the the basis for our use of sector volatility as a proxy for data depreciation.

D Data Appendix

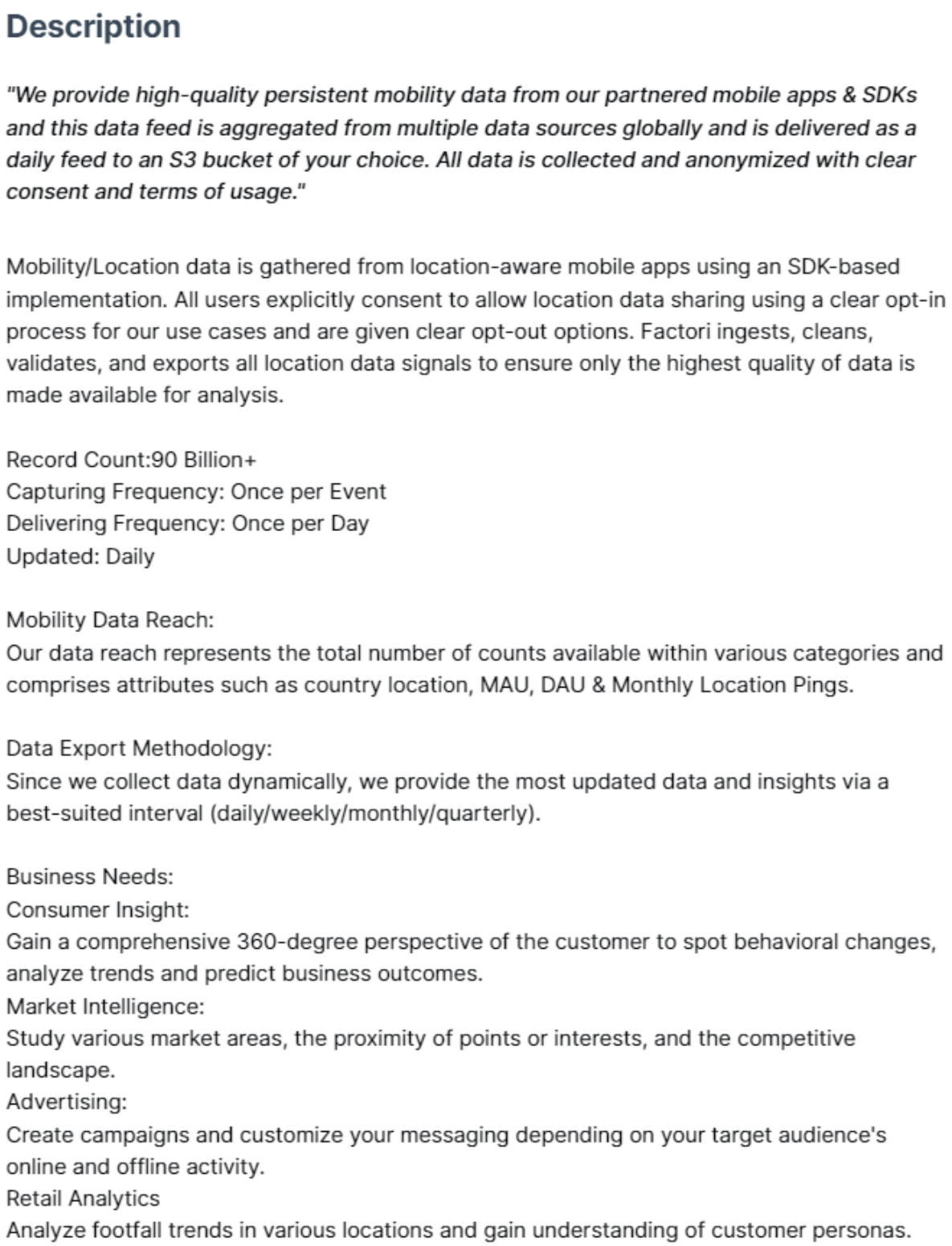
This appendix provides a richer description of our data sets. They give a visual illustration of the data product page, describe the topics the data pertains to, the industry and geographical locations of data providers and the categories of business data.

Figure A.1: Examples of Datarade Product Page



Notes. This is a screenshot of a data product hosted on Datarade, URL address: https://datarade.ai/data-products/lifesight-foot-traffic-data-global-mobile-location-data-2-lifesight.

Figure A.2: Examples of Datarade Product Page: Data Description



Description

"We provide high-quality persistent mobility data from our partnered mobile apps & SDKs and this data feed is aggregated from multiple data sources globally and is delivered as a daily feed to an S3 bucket of your choice. All data is collected and anonymized with clear consent and terms of usage."

Mobility/Location data is gathered from location-aware mobile apps using an SDK-based implementation. All users explicitly consent to allow location data sharing using a clear opt-in process for our use cases and are given clear opt-out options. Factori ingests, cleans, validates, and exports all location data signals to ensure only the highest quality of data is made available for analysis.

Record Count:90 Billion+
Capturing Frequency: Once per Event
Delivering Frequency: Once per Day
Updated: Daily

Mobility Data Reach:
Our data reach represents the total number of counts available within various categories and comprises attributes such as country location, MAU, DAU & Monthly Location Pings.

Data Export Methodology:
Since we collect data dynamically, we provide the most updated data and insights via a best-suited interval (daily/weekly/monthly/quarterly).

Business Needs:
Consumer Insight:
Gain a comprehensive 360-degree perspective of the customer to spot behavioral changes, analyze trends and predict business outcomes.
Market Intelligence:
Study various market areas, the proximity of points or interests, and the competitive landscape.
Advertising:
Create campaigns and customize your messaging depending on your target audience's online and offline activity.
Retail Analytics
Analyze footfall trends in various locations and gain understanding of customer personas.

Notes. This is a screenshot of a data product hosted on Datarade, URL address: https://datarade.ai/data-products/lifesight-foot-traffic-data-global-mobile-location-data-2-lifesight.

Table A.1: Geographical Locations of Data Providers

Headquarter	Count	Percentage
United States	1176	47.84%
United Kingdom	223	9.07%
India	117	4.76%
Germany	115	4.68%
Canada	65	2.64%
France	54	2.20%
Netherlands	40	1.63%
Israel	38	1.55%
China	34	1.38%
Australia	33	1.34%
Other	471	19.16%
Missing	92	3.74%
Total	2458	100.00%

Notes. This table presents the geographical locations of data providers on the Datarade platform.