An empirical study of the value of data

Imanol Arrieta-Ibarra, Tiago Caruso, Diego Jiménez Hernández, E. Glen Weyl

October 30, 2020
(exremely preliminary, do not cite or distribute)

Abstract

There has been a recent call for companies to acknowledge and pay for the data their algorithms use to make profits. In order to achieve this, there needs to be a way to price data. Current efforts have tried to ascribe the value of data to hidden costs such as the loss of privacy. We present a novel approach to estimating an upper bound for the economic value of data in algorithms. Our method does not assume that users have failed to internalize any costs in data production (such as privacy). Here we show that the price of data is in great part determined by the power dynamics present in markets. We apply our method to ride-sharing, simulating a market using data from a large ride-sharing platform (Uber). We estimate that in our scenario, with users having full market power, data would contribute up to 47% of Uber’s revenue. This would translate to average payments to drivers of approximately $30 per day, solely as compensation for the value of the data they generate as drivers, which corresponds to 20 to 40 percent of a average drivers daily earnings.

1 Introduction

Data has been said to be the fuel of the 21st century. The value of many companies have risen by billions of dollars as a consequence of new methods to extract revenue from it. Meanwhile, producers of data have been mostly excluded from the new economic paradigm and are forced to barter their data in exchange for services. While some firms have been using the data they collect to provide better services, others have been profiting from the free and ubiquitous availability of data without any apparent direct impact to consumers. Moreover, some governments have started to consider the extent to which this barter should be taxed and how. However, the true value of data in most of these applications remain, to a large extent, unknown and unmeasured.

There is little consensus on how to measure the value of data. Except for a small set of cases, data is a “missing market”. It is not traded in a regular market, but exchanged for consumption experiences of online services. “Producers” of data consumers of the goods and services consider it a small price to pay in order to get services that were not available in the past. On the other hand, the “consumers” of data find clever ways to amass vast quantities of it. While there have been calls for research on this topic (Arrieta-Ibarra et al., 2018; Lanier, 2014; Posner and Weyl, 2018), the advances on quantifying the effects of big data and artificial intelligence on companies have been slow.

Some governments, in particular inside the European Union (Regulation, 2016), have expressed interest in taxing data revenue and providing data rights. Imagine, for example, a firm that operates in the France but is based in Ireland. This company provides a service in exchange for data in the France
but pays taxes in Ireland. The French government can argue that if there were a market for data, the firm would pay for user’s data and would charge for the service it provides. The firm could pay taxes in Ireland but the revenue resulting from the provision of data would be taxed in France. To understand what would be the taxable income, France and the company would need to agree on the share that data contributes to the company’s income.

In this work we take a rather general view on what data is. Some of the examples we consider are user transactions for credit card companies, GPS location for ride-sharing, Facebook posts, Amazon reviews, genomic code information, labels provided by Mechanical Turk workers¹, Wikipedia articles, YouTube videos, and YouTube comments. Some of these examples correspond to actively generated data (think Facebook posts) while others are passive (GPS location). In this work we consider active and passive data in an undifferentiated way, leaving a heterogeneous treatment as future work.

Our goal is to present a systematic calculation of the economic value of data. The analysis in this work delivers an estimated upper bound on that economic value. By doing so, we provide a range of values which regulators could use as definitions for taxable income, and companies to assess the value that data generates to their different services. The research output of this framework can provide regulators with a basis for assessing tax data revenue. From the perspective of corporations, the analysis is a guard against over-taxation, limiting the amount of revenues that can be attributed to data. Most importantly, these methods would make transparent to data producers how much their data is worth.

We also show how these methods for modeling data can be applied to a ride-sharing application using simulations derived from data provided by Uber. These simulations serve us to exemplify a way with which we could estimate the economic value of data for Uber on a somewhat restrict case. We argue that the simulations we consider are a close approximation for the real Uber case, and provide a close upper bound. However, we highlight that more work, including an experimental design, may be needed to properly give policy advice through this quantification.

This work is structured in four parts. Section 2 gives a summary of prior related work. Section 3 presents a simplified framework through which to estimate what we’ll define as the “value of data”. Section 4 then translates this framework to a particular real world application: ride-sharing. We use Uber data to simulate a revenue model in terms of data, and estimate what the value of data would be under this model. Finally we conclude and present some guidance for future work.

2 Related Work

The work closest to ours is (Bajari2018) where the authors estimate the effect that data has on a series of machine learning algorithms using Amazon proprietary data. However, the authors mostly concentrate on how data impacts algorithms through different dimensions in the data and do not look for how this translates into economic output. They also find diminishing error rates with respect to the use of data, and observe differences depending under which dimension is data perturbed. For example they show that increasing a data-set in terms of number of available Amazon products produces different returns to the performance of algorithms; compared to increasing the amount of information for a fixed number of products. We were able to reproduce both findings. However, for the purpose of our current

¹Amazon’s microtask labor market
work, we pay attention to the homogeneous quantity of data as the variable of interest and do not concern ourselves with other dimensions. Nevertheless, we acknowledge it is of great importance to ask questions about which particular characteristics of data make it valuable and continue this line of research assuming heterogeneity in data.

With a similar spirit (Vincent, Johnson, et al., 2019) analyze how user generated content, in the form of Wikipedia articles, influences the behavior of search engines. They find that a high amount of results in google are consequence of user generated content in Wikipedia, and as such it drives a considerable part of Google’s advertising revenue. Such relationship is suggestive of a “free-riding” externality where Wikipedia editors are not compensated for their work. In subsequent work (Vincent, Hecht, and Sen, 2019), they explore how “data strikes” may affect different firms in the quality of their services. Nagaraj and Stern, 2020, on the other hand, analyze the value that map making provides to the economy. They do not address the matter of the distribution of revenue, however, they make it evident that industries are greatly benefited by the availability of free sources of information, and that ascribing the correct value to them would lead to an increase in their production.

Recent research has tried to ascribe value to data by deriving it from privacy loss. Jones and Tonetti, 2019 develop an economic model where they show that in the current market equilibrium, firms are incentivized to hoard data and have a disregard for privacy. These issues are corrected when data rights are given to consumers. However, their argument relates the cost of data to how much it reveals about a user’s privacy and the user’s valuation of her privacy. While this is an important component of the value of data, we argue that the loss of privacy should not be the only way in which it is valued. We posit that data has intrinsic value and that its current pricing is due to the unequal power distribution in the data market.

Work by Gefen et al., 2019 and Brynjolfsson, Collis, and Eggers, 2019 points out that people do ascribe high value to their data. However, this value depends heavily on the type of applications this data is used for. That said, in online platforms, users lose their ability to monitor how their data is being used and to what purpose. This negates users the possibility of making proper estimations on how to price their data which results in an underpricing of it.

All data is not created equal, and there has been advances in the question of how to distribute payments based in the usefulness of data. Bax, 2019 utilizes an approximation to the Shapley value (Shapley, 1953) and the Owens Value (Owen, 1977) in order to solve this problem. On (Arrieta-Ibarra et al., 2018) we proposed using a Leave-One-Out approximation (Koh and Liang, 2017). However Yona, Ghorbani, and Zou, 2019 show that using the Shapley Value in the context of data produces a more just way of distributing value.

Related to this work there is a rich literature on how data analysts would need to consume data in case this was priced or in the desire to preserve privacy. (Chen, Immorlica, et al., 2017) produce an incentive compatible mechanism with which a data analyst would choose how much data to input in her model. (Chen and Zheng, 2019) analyze the problem of estimating population parameters from users who price their data, with budget constraints.
2.0.1 Ride Sharing

On the side of ride-sharing our work looks closely at Castillo, 2019 where they look to answer the question of who benefits from surge pricing. To answer that question, ibid. develop a model to simulate counterfactual scenarios for different pricing strategies. We base our simulations on their model by reproducing their strategy for demand estimation and rider movement. The main idea is that since experimenting on the real Uber system may be difficult, we can approximate a counterfactual solution by modeling the behaviour of riders and drivers and perturbation the way Uber takes decisions.

Furthermore, ibid. utilizes an extensive literature on demand estimation which was first implemented for the case of Uber by Cohen et al., 2016. We borrow greatly from all this work to make sure our simulations are as close as possible to the state of the art in rider and driver simulation for Uber.

3 Model overview

To fix ideas, we first propose two thought experiments that represent opposites on a spectrum for how to value data. On one extreme, which we call data monopsony, where data providers (e.g., individuals) are price takers and data consumers (e.g., tech companies, insurance companies, banks) form a monopoly on the demand side.

The other extreme we call data monopoly, in which data providers (e.g., individuals) form a monopoly, what we’ll call a “data union”, and “data consumers” (e.g., companies) will be assumed to be price-takers. While this situation departs from the status quo, it shows a very simple way to understand how to think about the value of data. This model shows how, under certain market conditions, even without taking into account the loss of privacy, and even when the cost of producing data is negligible, it is still on the data producers’ best interest to charge a positive amount for the use of their data. We then show how this may be the case even when the firms provide services whose quality is dependent on data.

We define two types of agents in this models: firms and data producers. Firms generate services for consumers, and use data provided by consumers as input. Individuals may enjoy these services and produce data that companies can use for their services.

3.1 Firms

For both cases we assume that firms use data as their only input. They use this input to generate services for individuals. To get the input they need, firms have to pay a per unit cost \( p \) for each unit of data. We assume that there is a function \( F : D \subset \mathcal{R} \to \mathcal{R} \) which maps data to an objective function, that the company tries to maximize. This function \( F \) can be thought of as the relationship between the machine learning automation algorithms and long-term revenue to the company. We assume that this function is monotonous with decreasing returns to scale. We also assume that the company has an outside option where they operate without data. The outside option is represented by \( K \) which is the marginal utility of the first data point (If there is no outside option, then \( K = \infty \)). Finally we assume that there exists an amount of data \( \Delta \) after which the marginal contribution of an additional data point is zero:

\(^2\)For our audience without economic background, a monopsony is a monopoly on the demand side. Firms can be a monopsony in the data market if they are the only ones to which people can sell their data to.
\[ F'(D) \geq 0 \quad \forall D \in \mathbb{R} \]  
\[ F''(D) \geq 0 \quad \forall D \in \mathbb{R} \]  
\[ \lim_{D \to 0} F'(D) = K. \]

### 3.2 Data Monopsony

We begin by reviewing how standard economic theory predicts that monopsonies set prices. Let’s imagine a market where there is only one buyer and multiple sellers, think the Department of Defence for military devices, or a coal mine in a small town that purchases labor. The lack of competition implies that the single buyer company can obtain low prices from the sellers in the market, relative to those that would arise under perfect competition.

The firm’s objective function is to maximize an objective function (long-term revenue for example) minus costs. Let \( D \) represent the amount of data used (devices bought or hours of labor), \( F(D) \) the company’s objective function and \( P \) the price for data. The firm has then the following objective to maximize:

\[
\max_D F(D) - P(D)D,
\]

We explicitly denote \( P(D) \) to show that the price paid for devices depends on the amount of devices bought. By first order conditions we have that

\[
F'(D) = P(D) + P'(D)D.
\]

Economists call the left hand side of this equation the Marginal Revenue Product (MRP) for the devices (or labor). The right hand side is the Marginal Cost (MC) for a device (or for an hour worked). Figure 1 is a standard illustration from economic textbooks of monopsony pricing. The pink line represents the supply of devices; the blue line, the MC; and the green line, the MRP. Now, under perfect competition, the price and number of devices will be determined where supply and demand meet, that is, in point C. However in a monopsony, as we’ve seen, the quantity is determined where MRP and MC intersect, or point A. Then the price of devices purchased is that of point B which is the price that the device company will place on \( q \) devices. The ABC triangle formed between the points is what economist call the loss in terms of welfare.

Figure 1 also evidences the two main issues that monopsonic power causes: lower prices and less devices bought than under perfect competition. In the labor market its even more impacting since this signifies lower wages and less labor bought.

Now, for the case of data, we can imagine that the supply curve \( P(D) \) would be flat and close to zero. On the other hand, we would expect the MRP to drop fast and then move asymptotically towards zero. This means that firms that consume data would pay a smaller price for it and consume a lower quantity of data than in perfect competition. Already in Arrieta-Ibarra et al., 2018 we mentioned the possibility

\(^3\text{Note that because of economic convention, the axes are inverted from what another discipline may normally expect.}\)
Figure 1: The pink line represents the supply of devices, the blue line; the MC, and the green line; the MRP. Now, under perfect competition, the price and number of devices will be determined where supply and demand meet, that is in point C. However in a monopsony, as we’ve seen, the quantity is determined where MRP and MC intersect, or point A. Then the price of devices purchased is that of point B which is the price that the device company will place on q devices.

that firms may be consuming less data than optimal in order not to pay for it. We argued that companies could produce better services by paying for either higher quality or more diverse data.

This, we argue, is the main reason behind the price of data being zero in the status quo. Companies only need to produce services that marginally benefit producers of data in order to offset the small cost of producing data. However, from our previous explanation, it is possible to see that both producers of data and society may be paying a cost for these power dynamics. It gets even worse in the case where the data production costs are really negligible. In the next subsection we’ll make the extreme assumption that the MC is 0, which in this case would make the price of data be immediately 0.

Now, in the next subsection we’ll model data producers as monopolies and show that if this were the case, there would be occasions where data does get exchanged for positive prices, even when the marginal cost of producing data is close to zero. This still causes some market inefficiencies, so it would be in the interest of governments and society to decide what should they prioritize.

3.3 Data Monopoly

Similar to the last section, figure 2 shows a simplified diagram of the consequences of having monopolistic power. In this case, the data union holds all the market power and as such charges a higher price than that of perfect competition. However, opposite to the case of a monopsony, if $MC = 0$, the data union would still charge a positive price. This price is what we’ll use to upper bound the value of data.
Figure 2: The pink line represents the demand for devices, the blue line; the MC, and the green line; the MRP. Now, under perfect competition, the price and number of devices will be determined where supply and demand meet, that is in point C. However in a monopoly, the quantity is determined where MRP and MC intersect in point B. Then the price of devices purchased is that of point A which is the price that the data union will place on q devices.

In this scenario, the firm has then the following objective to maximize:

$$\max_D F(D) - pD,$$

(5)

where \( p \) is the price a unit of data. By first order conditions, the firm will buy data until \( F'(D^*) = p \). Thus we can write \( D^* = D(p) = F' - 1(p) \). Note that contrary to the monopsony, here the firm takes prices as given.

As an example, lets assume that the objective function has the form of a quadratic production function \(^4\) where \( F(D) = A - 2\alpha \Delta D + \alpha D^2 \) for \( D \leq \Delta \), where \( A \) represents all the other factors of production, and \( \alpha < 0 \). We would get that \( D^* = D(p) = \frac{p}{2\alpha} + \Delta \). \( \Delta \) corresponds to the point where the first derivative reaches zero so that we are only interested in the interval \((0,\Delta)\). Now, we concentrate on the case of a quadratic function because we’ll use it to fit \( F(D) \) in our practical application. The choice of the functional form in that case is only to simplify future analysis.

### 3.3.1 Data Union

Here we assume there is a single individual (or a data union) who is a “data monopolist”. This assumption means that the data union can decide the price at which to sell data. Thus, it will incorporate the companies’ decision problem into its own. After defining a price for data, companies would decide how much to collect. Therefore, the data monopolist tries to maximize revenue \((p \cdot D(p))\) by selling data, with the first order condition being:

$$D(p) + D'(p) \cdot p = 0.$$

(6)

\(^4\)There are multiple options when choosing how to model the functional form of a production function. For more on the topic refer to (Griffin, Montgomery, and Rister, 1987). However a quadratic production function allows us to have analytic results and satisfies all the conditions required.
Solving this equation will give us the optimal price individuals would choose for their data, \( p^* \). We refer to \( p^* \) as the “value of data” under no additional gains from the service, since it is the equilibrium price resulting from the data provider possessing full market power. It is important to note that we are assuming the cost of producing data to be zero. Even so, it is on the data provider’s best interest to charge a positive price for her data.

Following the previous subsection’s example under a quadratic production function, where we concluded that \( D(p) = \frac{p}{2a} + \Delta \), we would get that \( p = -\alpha \Delta \). So the value of data for this case would depend both on how the data contributes to the objective function of the company and the amount of data necessary to get marginal benefit of zero.

Now, the reader may think: “well, if the quality of the service is linked to the amount of data used by the company, then for sure it would be on the data provider’s best interest to charge less for their data.” We address this issue by incorporating a quality utility component to the maximization problem of the data provider. Let \( Q : \mathcal{D} \to \mathcal{R} \) represent the value in monetary terms of the service as the function of data used. We assume that:

\[
\begin{align*}
Q'(d) &> 0 \quad (7) \\
Q''(d) &< 0, \quad (8)
\end{align*}
\]

so that individuals value the service provided but the service improvements grow at decreasing rates.

The data provider’s problem then changes to:

\[
\max_p \quad p \cdot d(p) + Q(d(p)), \quad (9)
\]

where the first part summarizes the dollar value of getting paid for the data she provides, whereas the second part summarizes the quality changes that may come from additional data. We can refer to \( c^{**} \) as the value of data when the service is useful to individuals. We can make additional assumptions to guarantee the solution is so that \( c^{**} > 0 \). The intuition is that the individual utility gain from selling more data cannot be too big relative to firms’ revenue gains from more data use, as otherwise we would have to pay firms to use your data in equilibrium. There may also be cases where \( c^{**} \leq 0 \), meaning that the firm may actually be providing a positive externality to society.

### 4 The value of data in ride-sharing

This section will deal with how to apply the framework developed in Section 3 to a practical setting where a machine learning algorithm is learning to improve a product based on user data. The example we examine is within the context of Uber, one of the world’s largest ride-sharing applications, studying the value of data for ride-sharing firms.

The best way to quantify the effects of data in Uber would be to run a series of experiments constraining Uber’s capacity of utilizing user data. However, for our purposes, we’ll concentrate on a simulated

---

5Namely that (i) \( Q'(d) < M \) for all \( d \), where \( M \) is a positive constant and (ii) \( |d(c)/d'(c)| \geq M \) for all \( c \).

6\( c^{**} < 0 \) only if there are extra costs derived from acquiring data, and the service provides extreme benefits to its consumers.
environment derived from Uber data similar to that studied by Castillo, 2019. This simulation intends to reproduce driver and rider behaviour under different decision making processes taken by Uber.

Figure 3 shows a diagram of how we simulate the ride-sharing market. On the left side of the diagram we have Uber’s matching and price estimating algorithms. These algorithms both utilize data from riders and drivers to make decisions. These decisions in turn modify the simulation environment for riders and drivers. By constraining the flow of data we expect the algorithm performance to deteriorate, which in turn should reduce revenue.

The purpose of the simulations is to vary the amount of data that gets imputed to Uber’s algorithms. We will use these simulations to estimate revenue counter-factual for different data amounts - under this simulated market. We’ll make several simplifying assumptions, and are aware that in practice some of these may not hold entirely. However, after performing some stress tests on the assumptions, we find no impact in the quantities we estimate. The following lists the assumptions we are making to simulate the ride-sharing market:

1. Riders open the app and decide whether to take a trip with Uber or not. In case the trip is not taken its because there was a better outside option, as in (ibid.).
2. Drivers move randomly in the map but gravitate towards hexagons with higher multipliers as in (ibid.).
3. Matching is carried out greedily, associating each rider to the closest driver.
4. We model $F$ as a quadratic function. We stress test this assumption by also fitting a cubic and a modified Cobb-Douglas function.
5. When Uber has no data, their revenue is close to zero. The reason being that viable alternatives like broadcasting the requests from riders and letting drivers sort rides by themselves would not
be a competitive product in the market. To stress test this, we check what would happen if there was a no-data alternative that gave Uber 10% or 20% of their current revenue.

4.1 Data

We were given access to two proprietary ride-sharing datasets that cover a span of three months during 2018 in Dallas, Texas. The first dataset, which we’ll call rider data, consists on data for sessions started by drivers. For each session, we have the pickup latitude and longitude, the date and time when the session started, and when the trip started. We also have data on what Uber algorithms recommend as the ride price multiplier and the multiplier that users actually observe in the app. This data covers the full 3 month span. We use this dataset to estimate the demand for rides and to simulate rider behavior.

The second dataset, which we’ll call driver data, corresponds to GPS location of drivers for the span of the first week of April in 2018. This data is used by Uber to match drivers with riders. We’ll use it for the purposes of matching and for simulating driver behavior.

We make use of Uber’s geospatial hierarchical system in order to do all our computations. Just as Uber algorithms do, we don’t work directly with the Euclidean distance, but define distance in terms of hexagons in a map. Figure 4 shows the hexagons covering Dallas, Texas that we limit our analysis to. We only consider hexagons with a high volume of riders and drivers in order to further simplify our analysis.

Figure 4: We work with a subset of hexagons in order to avoid problems of sparsity in the data.

4.2 Uber’s objective function

Even though a company’s objective function in the long run may be the maximization of revenue, in the short run companies may limit themselves to maximizing over an objective function that approximates long-term revenue. In the case of Uber, we view their objective function as the number of rides subject to clearing subsets of the market (Cohen et al., 2016). Clearing the market and raising prices to do so allows for a better distribution of drivers in a map. As such, we’ll be working with the following objective function for Uber:

\[ F(D) = \mathbb{E} \left[ \sum_{u \in U} q(\hat{p}_r(D), t, h, r_u) M_u \right], \quad (10) \]

These are different since the final multiplier gets rounded and smoothed across space and time.
where $q$ is the probability of a ride being completed subject to price $\hat{p}_r(D)$ and beginning in market $(t, h)$ for a distance $r_u$ where the expectation is taken over matchings. $M_u$ is a random variable indicating if there was a matching for a given session or not. The price estimate is

$$
\hat{p}_r(D) = \arg\max_{p_r} \sum_{u \in U} \hat{q}(p_r, t, h, r_u; D),
$$

(11)

where $\hat{q}(p_r, t, h, r_u; D)$ is the estimated demand with data $D$.

### 4.3 Demand Estimation

For demand estimation we refer to Castillo, 2019 and Cohen et al., 2016 who use the external variations between Uber’s recommended prices and the prices users see their app. We follow the literature on discrete choice demand using the micro data to identify rider demand across different markets.

Conditional on opening the app, rider $i$ in market $t$ that wishes to travel a distance $r_i$ has utility from taking a trip based on:

$$
u_i = \alpha(r_i, h, l) - \beta(r_i) \cdot p_i - \gamma(r_i) \cdot w_i + \epsilon_i
$$

(12)

where $p_i$ is the price quoted and $w_i$ is the estimated time until pick-up (ETP) for the ride. The outside option of rider $i$ is not taking the trip for any reason, and is normalized to 0. In this model, $\alpha(r_i, h, l)$ is the utility gain from taking a trip, $\beta(r_i)$ is the sensitivity of utility with respect to the price, while $\gamma(r_i)$ measures the sensitivity with respect to the ETP. We allow $\beta$ and $\gamma$ to vary with the distance of the trip, $r_i$. The error, $\epsilon_i$, is an idiosyncratic utility shock that is observed by riders. We allow the error to be correlated with prices and ETP.

Our goal is to estimate equation (12). However, the error is likely to be correlated with prices and ETPs. If this is the case, identification of $\beta(\cdot)$ and $\gamma(\cdot)$ is not guaranteed. To identify these coefficients, we further decompose the idiosyncratic error into:

$$
\epsilon_i = g(\hat{p}_i, \hat{w}_i, l, h) + \eta_i,
$$

where $g(\cdot)$ is a flexible function intended to capture the endogenous part of prices and ETPs, and $\eta_i$ is an orthogonal iid EV1 error.

There are two consequences of limiting the amount of data for estimating prices. The first impacts Uber’s revenue directly since if the price is too high, Uber looses customers; while, if the price is too low, it looses revenue. The second is that, by estimating prices inaccurately, the multipliers may not be set to their optimal levels, making drivers gravitate towards the wrong directions.

### 4.4 Ride Matching

We do a greedy naive matching by pairing each rider to the closest driver in order of appearance. In Figure 5, lets imagine that a ride begins at the central hexagon. After looking for a driver in that hexagon, we would then look at the first concentric circle and so on until we find a free driver.

This is the same procedure Uber used in their first version of their app. Today the matching algorithm has added complexity, however from conversations with the engineering team the improvements to the
algorithm have not been large enough for the purposes of this work.

Figure 5: We try to match riders with the first driver in the closest concentric circle.

Limiting the amount of data for ride matching would have Uber miss-calculate the position of drivers. This in turn would make riders and drivers not match optimally and reduce the overall number of rides. For example in image 6 we have that two drivers move hexagons from the blue position to the red one. Under the full data regime, rider A would be assigned to car 1 and rider B to car 2. However under a smaller data regime the blue cars would correspond to the best estimation of where the drivers are. Uber would then match rider A to car 2 and rider B to car A. Causing longer arrival times and less trips because of all the wasted time.

Figure 6: Drivers move from the blue position to the red position. Under a smaller data regime, Uber would only know their first location and would predict that the cars are in the blue position. Consequently the matching under the full data regime would be more efficient than that of the smaller data one.

4.5 Quality

For the case of Uber we can divide quality into two components. The first one has to do with drivers and the increase in work they get from a better service. As such we include their share of the earnings.
in their optimization function. Riders, on the other hand, derive utility from having a precise estimation of when the driver will arrive. We will concentrate in this work on the first side of quality since it has a bigger impact on the overall exercise and leave the rider side of quality for future work.

4.6 Simulations

Recall that $F(\cdot)$ and $Q(\cdot)$ summarize the utility to companies and consumers from data. As such we can use simulations to estimate those functions and get back the value of data. Our simulations consist on counterfactual modeling of how riders and drivers behave in the market when different amounts of data are used to train the algorithms.

To estimate Uber’s objective function as a function of the data, we model how riders and drivers appear and interact in the market. We create a test set consisting of the first week of April\(^8\). We then use the remaining data in the rider dataset to estimate the demand function $q(p_r, t, h, r)$ where $p_r$ is the price of a ride, $t$ the time slot, $h$ the hexagon and $r$ the travel distance. We define this function as the probability of a ride being completed conditional on the given covariates. This we do by taking the logistic function of (12). For our simulations we’ll assume that $q(p_r, t, h, r)$ is our data generating process. We can then estimate $\hat{q}(p_r, t, h, r; D)$ using different amounts of data.

We then divide the market in hexagons of a higher resolution than those in Figure 4 (which we use only to model drivers moving across the map). These markets last one minute each and are represented by the tuple $(t, h)$. For each market, we set the price multiplier to 1 whenever there are more drivers than riders. Otherwise we set the multiplier so that it clears the market in expectation. For each market then, we have a price $\hat{p_r}(D)$. If we over-estimate the demand we could put too high a price which would result in a smaller number of completed rides. Underestimating the demand would only have indirect adverse effects in this setup. It results in a sub-optimal distribution of drivers, less drivers would go to the areas that need them most.

4.6.1 Riders

The progression of how riders will enter the system will be the same as that of the first week of April in the Uber rider dataset\(^9\). We don’t consider long term consequences of a customer taking or not taking a ride so each simulation will recreate the same order of events.

4.6.2 Drivers

To simulate drivers we use the driver dataset as the basis of our simulations. Drivers will appear at market $(t, h)$ whenever they appear in the dataset. We know these times and know the amount of time each driver works for. Other than that, drivers will move throughout the map according to the following indications.

- If a driver is unoccupied, she will look at the $k$ closest neighboring hexagons and will average the multipliers. She will then move in the direction of the highest multiplier.

---

\(^8\)This is the same timespan as that of the driver dataset.

\(^9\)This week was not used in training.
• If a driver is occupied, she will disappear from the system and appear back at random somewhere at a distance \( r \) from the departing hexagon. If the resulting hexagon lies outside of the markets we are considering, she’ll move in the direction of the closest market.

5 Results

Figure 7 shows how Uber’s objective function changes as we increase the amount of data used. We fit a quadratic function to these datapoints parametrized as in Section 3 and use this function to estimate the value of data in terms of Uber’s objective function. With that, we get that data accounts of \( 47\% \) of Uber’s long-term revenue. Table 1 summarizes our findings under different model specifications. We would like to stress that the decision of choosing a quadratic function to fit the data is to simplify the analysis. We fitted other functional forms that satisfied our assumptions and found no real difference. If we had access to more computational power, this assumption would not even be needed. We could produce a higher number of simulations and then solve numerically for the empirical \( F \) without any parametric assumption. Fitting a quadratic in this case serves just as a form of smoothing.

Figure 7: Revenue as a function of data. The units for both axis are in percentages in order to comply with privacy requirements from Uber.

Table 1 shows these results for different specifications of functional forms. We observe that for all specifications we get very similar results\(^{10}\). The upper bound price per datapoint we estimate to be 0.01 dollars. This makes the data share be 47% of Uber’s revenue. Because we are talking about a distortion in the market, a concern would be that this change in power dynamics would impact the overall revenue significantly, affecting all parties. Nevertheless, in this fictitious equilibrium, the overall revenue would be 89% of the current status quo. This is, changing the market power dynamics would cost both firm and users around 11% in lost revenue. However, drivers in this case would get on average $30 more per day as data payments which corresponds to 20 to 40 percent of their daily earnings.

The results when varying the outside option are similar in spirit. In general the objective function for Uber would be that of choosing between the revenue from using data, and that from not using data. This means that if Uber’s outside option for data provides is greater than 53% of their revenue, they would chose not to pay for data. However, we find this very unlikely for two reasons. First, having less data would make Uber’s service worse than that of competitors. Second, we find it would be difficult to

\(^{10}\)They are different only at 4 orders of magnitude.
justify a 25% charge\textsuperscript{11} on a service that uses no data.

As such, in this scenario, riders and Uber would be the most affected parties from the change in market power dynamics with drivers would benefit greatly. But again, this should be taken as an upper bound given the extreme nature of the exercise. The overall message should be to show that data can have a considerable price depending on market power dynamics and that the current status quo is, to a considerable degree, a consequence of the monopolistic power that companies have.

<table>
<thead>
<tr>
<th>Functional Form</th>
<th>Price</th>
<th>%Revenue</th>
<th>% Rides</th>
<th>Payments per Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadratic</td>
<td>0.01</td>
<td>47%</td>
<td>89%</td>
<td>30$</td>
</tr>
<tr>
<td>Cubic\textsuperscript{12}</td>
<td>0.01</td>
<td>47%</td>
<td>89%</td>
<td>30$</td>
</tr>
<tr>
<td>Cobb-Douglas with linear correction</td>
<td>0.01</td>
<td>47%</td>
<td>89%</td>
<td>30$</td>
</tr>
</tbody>
</table>

Table 1: The upper bound price per datapoint we estimate to be 0.01 dollars making the data share be 47% of revenue. In this fictitious equilibrium, the overall revenue would be 89% of the current status quo, and drivers would get on average $30 per day as data payments.

6 Discussion

In the previous section we have shown that on a world where data unions control all negotiating power in the market, and under certain structural assumptions, the average driver would expect to earn 30$ per workday on average from their data activity in the ride-sharing market. This corresponds to 20 to 40 percent of their daily salary.

These results are robust to different modeling choices when constructing counterfactuals. That said, an actual experiments with real users would be the gold standard evaluation. However, given this analysis, it is evident that there exists situations (however hypothetical they are) where user’s data can cost high prices.

In our sanity checks we find that the contributions from riders and drivers have different value in terms of the total earnings. This is just an example of the importance of expanding this analysis to cover heterogeneous data providers. Even in the case of drivers, a driver that works in the periphery would provide a different value from those that work in the city center.

To summarize, we’ve presented a new way of measuring the economic value of data that does not rely on data having any cost in being produced. The economic value studied here is separate from any privacy value. This analysis could be useful to companies, governments, and consumers in that it would allow a framework with which to compensate consumers for the data they produce.

We see two main paths ahead. The first one is getting more efficient ways to estimate the objective and quality functions. We think running experiments are the natural next step, however even for this we require some different functional assumptions. Additionally, we think there is plenty of work to be done on the subject of the heterogeneity of data. In this work we considered data as a homogeneous input whereas in reality different data-points can provide different value. This is most obvious for example on the provenance of data in the Uber experiment. Further analysis as that in Yona, Ghorbani, and Zou, 2019 would be useful to differentiate payments between users.

\textsuperscript{11}Which is the approximate amount Uber currently charges riders for the service.
References


