Surge Pricing Solves the Wild Goose Chase*

Juan Camilo Castillo† Dan Knoepfle‡ E. Glen Weyl§

July 2017

Abstract

Ride-hailing apps introduced a more efficient matching technology than traditional taxis (Cramer and Krueger, 2016), with potentially large welfare gains under the appropriate market design. However, we show that when price is too low they fall into a failure mode first pointed out by Arnott (1996) that leads to market collapse. An over-burdened platform is depleted of idle drivers on the streets and is forced to send cars on a *wild goose chase* to pick up distant customers. These chases occupy cars, reducing the number of customers served, earnings and thus effectively removing drivers from the road and exacerbating the problem. We use data from Uber to show that wild goose chases are indeed a problem in the Manhattan market. The effects of wild goose chases dominate more traditional price theoretic considerations and imply that welfare and profits fall dramatically as price falls below a certain threshold and only gradually move in price above this point. A platform forced to charge uniform prices over time will therefore have to set very high prices to avoid catastrophic chases. Dynamic "surge pricing" can avoid these high prices while maintaining the system functioning when demand is high.

Keywords: wild goose chases, ride-hailing, surge pricing, dynamic pricing, hypercongestion
JEL classifications: D42, D45, D47, L91, R41

*We appreciate the helpful comments of Susan Athey, Eduardo Azevedo, Timothy Bresnahan, Liran Einav, Matthew Gentzkow, Ramesh Johari, Jonathan Levin, Andy Skrzypacz, Christopher Snyder, Rory Sutherland and seminar participants at Microsoft Research New York City and Stanford University.

†Department of Economics, Stanford University, 579 Serra Mall, Stanford, CA 94305; jccast@stanford.edu, http://sites.google.com/site/juancamcastillo/.

‡Uber Technologies, 1455 Market Street, San Francisco, CA 94103; knoepfle@uber.edu.

§Microsoft Research, One Memorial Drive, Cambridge, MA 02142 and Department of Economics, Yale University; glenweyl@microsoft.com, http://www.glenweyl.com.

1 Introduction

Ride-hailing applications (apps) like Uber and Lyft introduced a promising new technology to compete with traditional taxis. Cramer and Krueger (2016) show that the fraction of working time that a driver actually spends with a rider in the back seat is roughly 40% higher for Uber than for traditional taxi markets. Ride-hailing, however, is not more efficient than taxis under all circumstances. In this paper we show both theoretically and empirically using data from Uber that, unlike traditional street-hail taxi systems, ride-hailing platforms are prone to a matching failure first anticipated by Arnott (1996). When there are too few drivers relative to demand, drivers are quickly occupied and thus free drivers are spread thinly throughout a city, forcing matches between drivers and passengers that are far away from each other. Cars are thus sent on a *wild goose chase* (WGC) to pick up distant customers, wasting drivers' time and reducing earnings. This reduces the number of available cars both directly by occupying cars and indirectly as cars exit in the face of reduced earnings, exacerbating the problem. This harmful feedback cycle can lead the system to collapse, but can be avoided by using prices to ration demand when it is high. This may help explain why these platforms have relied so heavily on "surge" pricing, in contrast to traditional taxi markets.

Because he was focused on optimal allocations, Arnott discounted WGCs as Pareto-dominated and thus just a theoretical curiosity. However, we show that at times of high demand, if prices do not appropriately adjust, all equilibria of the market are WGCs when using a first-dispatch protocol, in which an idle driver is immediately dispatched every time a rider requests a trip (as many ride-hailing services have committed to). This suggests two ways in which pricing can avoid WGCs. First, one might set a single high price all the time, sufficiently high to avoid WGCs even at peak-demand periods. This design has the drawback that prices will be unnecessarily high, and thus demand inefficiently suppressed, at times of low demand. A more elaborate mechanism is to use dynamic "surge pricing" that responds to market conditions. Such a system was introduced by Uber early in its development. Prices are set high during peak-loads, but can fall when demand is more normal. Thus, against the common perception, surge pricing allows ride-hailing apps to *reduce* prices from the baseline of static pricing instead of increasing them.

Our analysis starts with a theoretical model in a homogeneous spatial region that highlights the phenomenon of WGCs. The main components of our model are demand for trips, labor supply, and a matching technology that defines how labor supply translates into supply of trips. The characteristic feature of WGCs is that the supply of trips given a fixed number of drivers is a non-monotonic function of pickup times due to two opposing effects. An increase in pickup times requires fewer idle drivers, which frees up drivers that can serve more customers. But as pickup times go up, drivers spend a larger fraction of their time picking up passengers instead of driving them to their destination. WGCs occur at high pickup times, when the latter effect dominates over the former. High demand puts the system under stress by reducing the number

of idle drivers on the road and increasing pickup times. This inefficient use of driver time results in a lower number of trips in equilibrium.

Despite this novelty, WGCs are similar to "hypercongestion", a related phenomenon in transportation economics (Walters, 1961; Vickrey, 1987). When enough cars enter a road, speeds of all cars on the road fall sufficiently that the total throughput of the road actually falls, causing traffic jams.[1] However, the effects of WGCs may be much more severe than those of hypercongestion because the supply of drivers is endogenous, and may collapse in reaction to the fall in earnings due to less trips being completed. We show that under WGCs a decrease in prices leads to sharp decreases in welfare, number of trips, platform revenue, and drivers' surplus. We also show that WGCs can always be avoided by increasing prices.

We back our theoretical findings with empirical evidence of WGCs using data of Uber trips in Manhattan between December 2016 and February 2017. We find that the number of trips given a fixed number of drivers indeed exhibits the nonmonotonicity in pickup times predicted by our theory. Beyond this, our theory has a fairly sharp prediction that when a theoretically-derived (intuitive but non-obvious) measure of system slack falls below a specified threshold for WGCs that can be derived purely from data on traffic and matching flow, the system should experience rapid and catastrophic failure. We verify this prediction by showing that a variety of market performance measures degrade drastically when the market falls below this threshold: pickup times, trip cancellation rates, and the fraction of unserved customers rises steeply, while the fraction of people who request a trip plummets.

We then calibrate our theoretical model to the data in order to make a detailed quantitative analysis of the welfare effects of surge pricing. WGCs dominate more traditional price theoretic considerations. Consistent with the main results of our theory section, welfare and revenue fall dramatically as price falls below a certain threshold and the market enters a WGC. On the other hand, welfare and revenue only gradually move in price above this point. Thus, the main concern for a ride-hailing platform when deciding how to price is to avoid WGCs.

We analyze the behavior of a welfare maximizing platform that serves more than one market, as defined by different times of the day. We first compute the optimal prices with surge pricing, where the platform sets different prices for each individual market. Then we analyze the behavior if it is constrained to set a single price for all markets. The only way to avoid the drastic loss in welfare from WGCs is to set prices close to the highest one under surge pricing. In a simple two market analysis, where the platform serves the times of highest and lowest demand during weekdays, the platform would set prices at more than 97% of the surge pricing level at high demand, and 66% higher than the level at low demand. Thus, surge pricing only leads to

[1]While this possibility was largely dismissed in the early years of the transportation economics literature (Arnott and Inci, 2010), empirical evidence from the engineering literature has clearly shown that hypercongestion occurs in practice (Muñoz and Daganzo, 2002). Hall (2016) highlights that the existence of hypercongestion dramatically strengthens the case for the pricing of roads, just as we argue that wild goose chases may be the reason that dynamic pricing is widely used in ride-hailing but not elsewhere.

very modest increases in prices at times of high demand, whereas it allows drastic reductions in low demand times. This goes against the perception among the public and regulators that surge pricing is a form of price gouging. For example, the splash page on competitor Gett's home page on June 27, 2017 stated "The only time we surge is never o'clock" and many cities in the developing world have banned or otherwise forced Uber to desist from surge pricing.

Pricing is not the only tool ride-hailing apps can use to avoid WGCs. We discuss two alternative approaches. First, rationing rides when demand is high avoids over-burdening the market and WGCs. However, this makes the service unreliable, eliminating one key advantage of ride-hailing over traditional taxis. Second, setting a small maximum dispatch radius also avoids WGCs, but it creates passenger queues. Passengers then have to wait without being matched to a driver and without knowing how long they will have to wait to be picked up. A maximum dispatch radius is thus in tension with a user interface feature of current ride-hailing apps—that riders know immediately upon request the location and trajectory of a car driving towards them. This feature is considered very appealing to riders and our internal interviews suggest product leaders at Uber would be loath to compromise that element of the rider experience. Hence, although surge pricing is not the only way to avoid WGCs, alternative approaches have drawbacks that limit their appeal to ride-hailing platforms.

Our analysis begins in the next section with our theoretical model with elements that are similar to Arnott. In Section 3 we describe how WGCs arise, we show the catastrophic effects WGCs have on welfare and revenue, and we show how increasing prices avoid WGCs. In Section 4 we show empirical evidence of WGCs in the Uber market in Manhattan. Then in Section 5 we calibrate our model to our data and analyze the effects of a ban on surge pricing, as well as the redistributive effects, assuming that Uber only caters to two markets, rush hour and non-rush hour. In section 6 we discuss some alternative solutions to WGCs, and why we believe surge pricing is the best option for platforms. We also discuss ride-sharing or "pooling" in Section 7, and show that WGC are also present and might even be worse than without pooling. In the next draft of this paper we will also include a more realistic welfare analysis that in which instead of facing a small number of markets, Uber faces a large number of markets during every time of the week, each one of them with different primitives.

2 Model

We consider a static, steady-state model of a ride-hailing service. Dynamics are critical to a variety of aspects of the model and to the concept of surge pricing, but we reduce short-term dynamics to a static steady-state analysis and model dynamics over longer periods of time as allowing or prohibiting differential pricing based on market conditions.

2.1 Demand for trips

Let λ be the density of arrival of users (measured, for instance, in users per minute per square kilometer). These are the users that might potentially request a ride if the price and the pickup time are good enough for them. We assume that users will request a ride when they are willing to pay the associated price and are able to wait the associated pickup time. Demand is then given by a function $D(T, p) \leqslant \lambda$, where T is average pickup time and p is price.[2] We now list the main assumptions on this function:

Assumption 1. $D(T, p)$ *satisfies the following:*

1. *It is bounded above*
2. *It is continuously differentiable in* (T, p) *and decreasing both in pickup time and prices.*
3. $\lim_{T\to\infty} D(T, p) = 0$ *for all* $p \geqslant 0$ *and* $\lim_{p\to\infty} D(T, p) = 0$ *for all* $T \geqslant 0$.
4. *For all* p, *the distribution of the maximum willingness-to-pay has finite mean.*

Part 1 is motivated by the fact that even with zero pickup time and with prize zero a bounded number of people λ are in need of transportation. Part 2 is standard for demand functions. Part 3 just states the fact that nobody is willing to pay an infinite price nor wait infinite time to get a ride. Part 4 assumes that the distribution of willingness-to-wait is not too fat tailed. Note that for much of our analysis we consider equilibrium *holding the price fixed*. The clearing variable, instead, will be pickup times, so $D(T, p)$ can be thought of as a decreasing demand function, where T plays the role of prices, and p is an exogenous demand shifter.

2.2 Labor supply

Individual drivers decide whether to work based on expected hourly earnings e, and this results in a supply of drivers $l(e)$ (measured in drivers per square kilometer, for instance), where we assume that $l(\cdot)$ is increasing and continuously differentiable. To find an expression for e, let τ be the fraction of the price charged to passengers that the platform takes as revenue. If Q is the equilibrium density of rides per unit of time, and the price is p, total earnings per unit of time per unit area are $(1-\tau)pQ$. The average earnings per unit of time for an individual driver are $e = (1-\tau)p\frac{Q}{L}$. Labor supply then satisfies $L = l\left((1-\tau)p\frac{Q}{L}\right)$ in equilibrium. We make the following assumptions on the function l:

Assumption 2. $l(e)$ *is continuously differentiable and increasing, and* $l(0) = 0$.

[2]Demand actually depends not on average pickup time but on the realizations of pickup time. So from a primitive demand function $\tilde{D}(\tilde{T}, p)$ that depends on realized pickup time $\tilde{T}$, demand would be $\int \tilde{D}(\tilde{T}, p)dF(t)$, where F is the distribution of $\tilde{T}$. We will later show that the distribution depends on I, the density of idle drivers, which has a one to one mapping with average pickup times. So, to be precise, what we describe here is $D(T, p) = \int \tilde{D}(\tilde{T}, p)dF(\tilde{T}; I(T))$

These are all standard properties for a supply function. A straightforward consequence of this assumption is that L, as defined implicitly by $L = l\left((1-\tau)p\frac{Q}{L}\right)$, is increasing in total earnings $(1-\tau)pQ$.

2.3 Matching technology and supply of trips

Our demand function measures something quite different than the labor supply of the previous subsection. Whereas demand is the number of trips requested, supply is the number of drivers working. We thus need the third main component of our model, the matching technology, in order to translate the number of drivers working into the number of trips supplied. The goal of this section is then to obtain a supply function of the form $S(T, L)$ that gives the number of trips that can be served by L drivers when pickup times are T. The reason why T is relevant for supply is that it is also the time drivers have to spend picking up passengers.

At any given moment working drivers are in one of three states: idle (waiting to be matched to a rider), *en route* (on their way to pick up a passenger), or driving a passenger to her destination. The total number of drivers working thus has to be equal to the sum of drivers in each one of these states. We defined I to be the number of idle drivers. In equilibrium, tQ drivers are driving a rider, where t is the average trip duration. This is the product of the number of trips per unit time and the average time it takes to pick up a rider. By a similar reasoning, TQ drivers are on their way to pick up a rider (*en route* drivers) in equilibrium, since T is the time it takes on average for a driver to pick up a passenger.

Based on the previous expressions for the number of drivers in each state, the following identity accounts for the total density of drivers in equilibrium:

$$L = \underbrace{I}_{\text{Idle}} + \underbrace{tQ}_{\text{Driving}} + \underbrace{TQ}_{\textit{En route}} . \tag{1}$$

The expression TQ for *en route* drivers shows an essential feature of dispatch systems: high pickup times are bad both because they make riders wait longer and because drivers have to spend more time not taking passengers to their destination, which as we will see reduces the number of trips the whole market is able to serve. And this is essentially different from street-hail taxi markets, where drivers pick up riders while they are idle on the street. Therefore, there is no pickup time and the total density of drivers is accounted for by $L = I + tQ$.

The average pickup time is $T(I)$, a decreasing function of the density of idle drivers: if there are a lot of idle drivers, a new arriving rider will on average be matched to a driver that is closer to him, so he will have to wait less time before being picked up. This pickup time function is the only primitive of the supply function. We will assume a simple geometry with no inefficiencies beyond pickup time and a uniform distribution of drivers, thus abstracting from the important systematic differences in supply compared to demand at different points in space studied by

Buchholz (2016) and treating these differences only through our analysis of separate markets that are treated as entirely segmented. Given this segmentation assumption it may be easier to interpret our markets as representing different times, as in the analysis of Frechette et al. (2016), rather than different places within a city; in either case our static model that leaves out substitution and complementarity across markets is an important modeling simplification.

In our calibration we will assume a specific functional for for $T(I)$ that fits the data closely (Appendix A). For now, we will simply make the following assumptions:

Assumption 3. $T(I)$ *is continuously differentiable, decreasing, and convex. It also satisfies* $\lim_{I\to\infty} T(I) = 0$ *and* $\lim_{I\to 0} T(I) = \infty$.

The fact that $T(I)$ is decreasing reflects the fact that more drivers decrease expected distance to the closest driver and therefore pickup times. Convexity means diminishing marginal returns of additional idle drivers. The first limit condition simply means that with an infinity of drivers pickup times would go down to zero, and the second one means that riders would have to wait infinitely long with zero idle drivers. All these conditions are satisfied by the empirically motivated functional form we later assume in our calibration, as well as by simple functional forms that can be theoretically motivated. For instance, if *en route* drivers drove in a straight line at a constant speed in an n-dimensional space, $T(I) \propto I^{-\frac{1}{n}}$,[3] which satisfies all these properties.

Since $T(I)$ is a decreasing function, we can define its inverse, $I(T)$, which will turn out to be more convenient for our model. It can be interpreted as the density of idle drivers that is needed to ensure an average pickup time T. As direct consequences of 3, $I(T)$ is continuously differentiable, decreasing, and convex, $\lim_{T\to\infty} I(T) = 0$, and $\lim_{T\to 0} I(T) = \infty$.

Isolating Q from 1 and substituting in $I(T)$ gives us the expression we wanted for the supply of trips:

$$S(T, L) = \frac{L - I(T)}{t + T} \tag{2}$$

The functional form for this expression is intuitive. The numerator is the number of busy drivers (those that are not idle). These are the drivers that could potentially be taking a passenger to her destination. The denominator is the average busy time it takes to complete a trip, which is the sum of the time it takes to pick up the passenger and then drive her to her destination. Dividing the number of busy drivers by the time per trips gives the number of trips that can be completed per unit time.

This functional form has increasing returns to scale since $S(T, bL) > bS(T, L)$ for $b > 1$. Thicker markets can thus achieve lower pickup times while holding the number of trips per driver constant, or increase the number of trips per driver while holding pickup times constant.

[3]The choice of $n = 2$ is not entirely obvious, since some places like Manhattan are somewhere in between one and two dimensional, and speed depends on the length of travel. This is why we use a functional form below that is flexible enough to fit observed traffic flow data.

3 Wild Goose Chases

We now use the model of the previous section to highlight the key forces driving our analysis.

3.1 Normal and wild goose chase matching equilibria

We analyze in detail the form of the trip supply function $S(T, L)$. Its main properties are summarized in the following lemma:

Lemma 1. *Supply $S(T, L)$ is continuously differentiable. Given some fixed number of drivers L, supply for $T = T(L)$ is $S(T(L), L) = 0$. There exists some positive pickup time $\hat{T}(L)$ such that $S(T, L)$ is increasing in T for $T(L) < T < \hat{T}(L)$ and decreasing in T for $\hat{T}(L) < L$. Finally, $\lim_{T\to\infty} S(T, L) = 0$*

Proof. See Appendix B.1. □

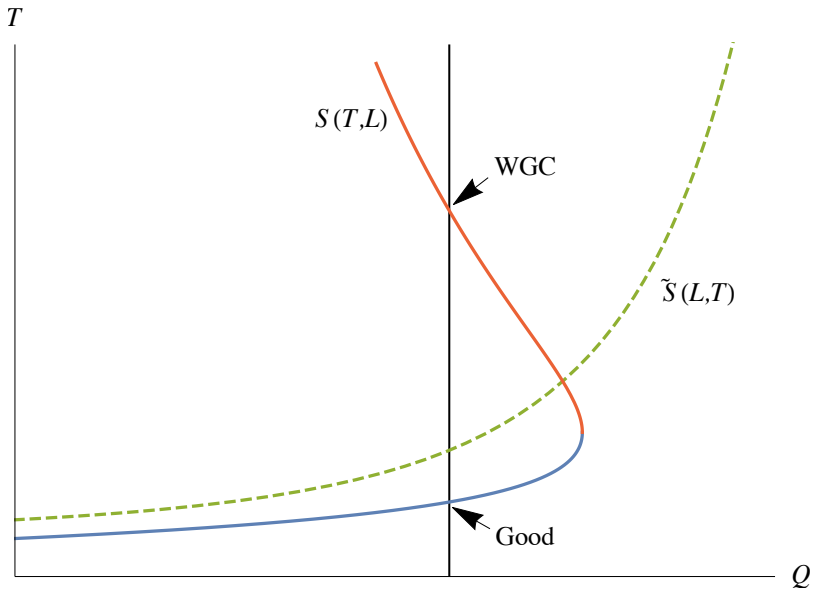


Figure 1: Supply of trips. This figure illustrates the backward-bending supply curve in a ride-hailing market, with the WGC and the good regions in red and blue, respectively. For comparison, the green line shows how supply looks for a street-hailing taxi market. Given the results in Cramer and Krueger (2016), the taxi technology is less efficient for good outcomes, which means higher pickup times.

The main features in this lemma are illustrated in Figure 1. For $T < T(L)$ the value of the functional form for $S(T, L)$ is negative; the current number of drivers cannot achieve such low pickup times, even if all of them were idle. Supply is then increasing in T, like a traditional functional form for supply, until $\hat{T}(L)$. After that point, supply decreases, converging to zero as pickup time goes to infinity. This backwards bending supply curve is very different from a traditional supply function, and it is the main driver of our results.

The intuition for the nonmonotonicity of $S(T, L) = \frac{L-I(T)}{d+T}$ is as follows. The numerator is an increasing function; to understand why, suppose that the platform wants to achieve a lower pickup time. In order to do so, it needs more idle drivers on the streets, which decreases the number of drivers that are busy and reduces the potential capacity of the market. This effect is the main driving force in the blue region of the supply function in Figure 1.

As pickup times increase and supply approaches its maximum, a second effect starts to kick in. With higher pickup times the denominator becomes larger and larger, since drivers have to spend a significant portion of their time picking up passengers that are far away. So, despite the fact that higher pickup times requires less idle drivers, thus freeing some of them to drive passengers, the total time it takes to complete a trip becomes longer, and after some point capacity starts to decrease. This second effect dominates in the red region.

Figure 1 illustrates that there are two ways to supply the same number of trips for a fixed number of rides. The first one, which we call a good outcome, has a low pickup time (in the blue region of the supply curve). The one above the maximum, plotted in red, has a longer pickup time. It is evident that the latter situation is inefficient, as it achieves the same number of trips with the same number of drivers, but with higher pickup times for passengers. We call these situations *wild goose chases* (WGCs). In colloquial English, wild goose chases refer to extended, wasteful and ultimately vain pursuits of an unattainable objective. By analogy, in this bad situation, the ride-hailing system, by trying to serve beyond its capacity, must send drivers to distant locations that ultimately reduce the number of rides it can effectively provide. And clearly the system should never be in this situation.

Our supply curve bears some similarity to a Laffer curve in tax theory (or the revenue curve in monopoly theory): governments face a tradeoff between high tax rates and high tax revenue. But beyond a certain point taxes are so high that revenues also decrease, and the tax rate should never be beyond this point. In our model, if the platform has to choose an equilibrium in some point along the supply curve, it will face a tradeoff between quantity and quality (in terms of pickup time) whenever it is in a good outcome. But if it is in a WGC, it no longer faces a tradeoff since moving upwards along the curve decreases the number of trips served while at the same time increasing pickup times. It thus becomes evident that the platform would like to move down along the supply curve to get back to a good outcome.

WGCs can be easily diagnosed in the data. Note that when they happen the derivative of $S(T, L)$ with respect to T is negative. If we rewrite it as $S(T(I), L) = \frac{L-I}{t+T(I)}$, this is equivalent to the derivative of $S(T(I), L)$ being positive. After some simple algebra steps this can be written as $I < -\epsilon_I^T T(I) Q$, where ϵ_I^T is the elasticity of pickup time with respect to the density of idle drivers. If we define *slack* to be $s = \frac{I}{T(I)Q}$, the ratio of idle drivers to en route drivers, this simply means that WGCs occur when slack is less than ϵ_I^T.

This makes the the theory straightforward to test. The number of idle drivers and the number of en route drivers are directly observable in the data, so slack can be easily computed. Although

ϵ_I^T is not directly observable, we will show in Section 5.1 that $I(T)$ can be easily fit to the data closely, from which we obtain an estimate of ϵ_I^T that is somewhere between -0.25 and -0.5.

WGCs are unique to ride-hailing markets. For comparison's sake, we will now analyze the equivalent supply curve for a traditional street-hailing market. By this we mean a market in which taxis can only be hailed by standing in the street and not by calling them.[4] In this kind of market there is no pickup time, since whenever a taxi is "matched" to a rider the trip starts immediately. There is also a decreasing function $\tilde{T}(I)$ that maps idle drivers to pickup times: the more idle drivers in the street the less time a rider should expect to wait before a taxi shows up. Because of the greater efficiency of ride-hailing, $\tilde{T} > T$; we do calibrate this efficiency gain precisely in our numeric section, and in our illustration in Figure 1 we represent a magnitude that roughly matches a gap between these similar to that found by Cramer and Krueger (2016).

Let $\tilde{I}(T)$ be the corresponding inverse function. The driver identity is then $L = \tilde{I}(T) + tQ$. The same exercise as before leads to a supply function $\tilde{S}(T, L) = \frac{L - \tilde{I}(T)}{t}$. This is an increasing function, as illustrated in Figure 1. The intuition is the same as in good outcomes for ride-hailing markets: decreasing pickup times requires more idle taxis on the streets, which reduces the number of drivers available to take passengers to their destination. But the second effect of drivers wasting more time picking up drivers is no longer present, so supply is no longer backward bending.

3.2 Equilibrium

We will now proceed to put together the three elements in our model: demand for trips, labor supply, and trip supply.

The first condition for equilibrium is that trip supply and demand must be equal:

$$Q = D(T, p) = S(T, L) \tag{3}$$

Figure 2 illustrates this condition. In general the solution in the market for trips can be a good outcome or a WGC, as illustrated in the figure. In Section 3.4 we will show that price is a key determinant of the region in which it falls.

Another condition that must be satisfied in equilibrium is that the number of drivers working has to be equal to labor supply:

$$L = l\left((1-\tau)p\frac{Q}{L}\right) \tag{4}$$

An equilibrium is a joint solution in (T, L, Q) of equations (3) and (4).

In order to simplify our analysis of equilibria, we define $\hat{Q}(L; p)$ to be the solution to equations (3) for a given number of drivers and $\hat{L}(Q; p, \tau)$ to be the solution to equation (4). It is straightforward to see that (4) has a unique solution for every Q, so $\hat{L}$ is well defined. Furthermore, $\hat{L}$ is increasing and continuous in Q and $\hat{L}(0; p, \tau) = 0$. For equation (3), we can

[4] If riders are also allowed to call taxis, the market would then share some features with a ride-hailing market.

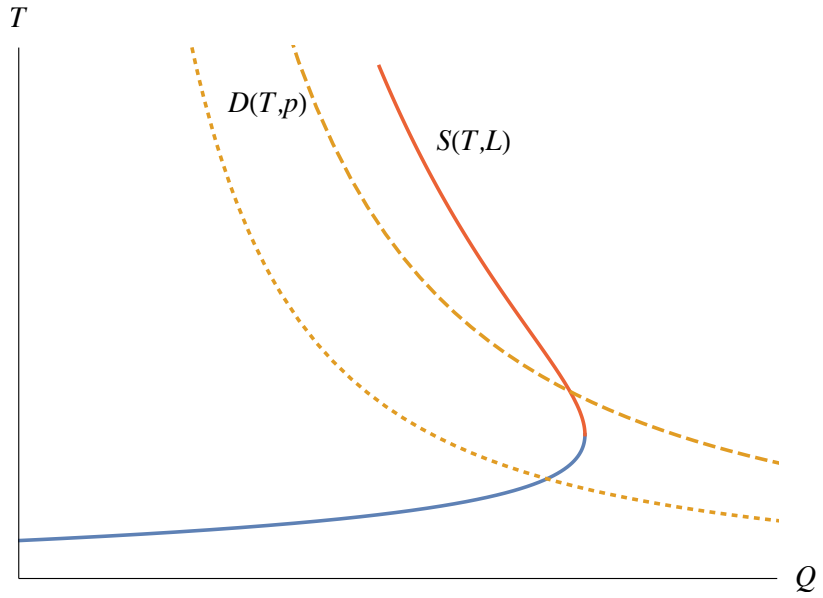
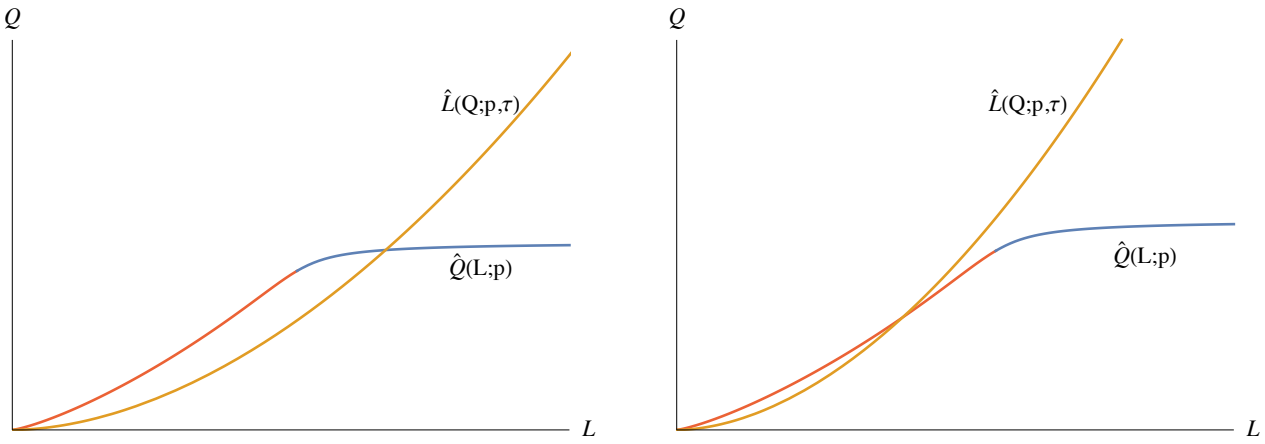


Figure 2: Trip supply and demand with a fixed number of drivers.

also see easily that for $L = 0$ the unique solution is $Q = 0$. On the other hand, we cannot guarantee a unique solution. In fact, multiple solutions arise with fairly simple and reasonable functional forms for supply and demand. In order to deal with this, we pick the highest solution (i.e., the one with greatest Q) whenever there are multiple solutions. The following lemma guarantees the existence of at least one solution, which means that $\hat{Q}(L;p)$ is well defined:

Lemma 2. $D(T,p) = S(T,L)$ *has at least one solution in* T *for all* (p,L). *For the highest solution,* Q *is increasing in* L.

Proof. See Sectionppendix B.2 □



(a) Good outcome

(b) Wild goose chase

Figure 3: Equilibrium. The red part of $\hat{Q}(L;p)$ represents situations with a WGC. These plots show how the equilibrium can be in a good outcome or in a WGC.

An equilibrium can then be characterized as a solution to the following two equations:

$$Q = \hat{Q}(L; p) \qquad L = \hat{L}(Q; p, \tau) \tag{5}$$

An example of how these equations look is shown in Figure 3. Since $\hat{Q}(0; p) = 0$ and $\hat{L}(0; p, \tau) = 0$, there always exists an equilibrium at the origin, although it is sometimes unstable.[5] Whenever the equilibrium at the origin is unstable, there is at least one stable interior equilibrium.[6] We cannot rule out multiple equilibria. Indeed, it is not difficult to construct functions that lead to multiple equilibria. In the case of multiple solutions, we select the highest equilibrium. However, we do not face any situation with multiple equilibria in our calibrations.

3.3 Revenue, welfare, and surplus

Platform revenue is straightforward to define. At an equilibrium with price p and Q trips, the platform gets revenue $\tau p Q$.

We need some additional assumptions to define welfare, riders' surplus, and drivers' surplus. First, let social cost of drivers $C(L)$ be the integral of the inverse supply curve, so that $C'(L) = l^{-1}(L)$ for all L. To pin down the function exactly, let $C(0) = 0$. This is a standard cost function which is increasing and convex. Drivers' surplus is then what they get paid minus their cost:

$$DS(Q, L, p) = (1 - \tau) p Q - C(L) \tag{6}$$

In order to define welfare and consumer surplus, let $U(Q, L, T)$ be gross utility. In our calibration we make much more specific assumptions on its functional form, but for now we will use a general functional form. In general Q and T are not enough to specify gross utility, as it is unclear which passengers are being served. So with U we have in mind the gross utility that would be achieved in an equilibrium with some unspecified price that resulted in Q trips with pickup time T. This reasoning leads to the following assumptions:

Assumption 4. *Gross utility $U(Q, T)$ is continuously differentiable in (Q, T), and it is decreasing in T and increasing in Q with $U_Q(Q, T) = p$.*

Gross utility is decreasing in T because if the same people are served with lower pickup times their utility should be greater. Additionally, an equilibrium with lower waiting times requires a higher price, so customers with higher willingness to pay would be served, which is an additional effect in the same direction.

For the sign of U_Q, note that if pickup time is fixed and the number of people served increased, it might be the case that gross utility decreased if the new people getting a ride have

[5] Stability means that in the (L, Q) plane both functions cross from above. In terms of derivatives, $\hat{Q}'\hat{L}' < 1$.

[6] Since D is bounded above, so is $\hat{Q}$. This implies that if there exists some L such that $\hat{L}(\hat{Q}(L)) > L$ (which is the case when the solution at the origin is unstable), then there exists some $L' > L$ such that $\hat{L}(\hat{Q}(L') = L'$: $\hat{L}(\hat{Q}(L))$ is a bounded increasing function, and the left hand side is an unbounded, continuous, and increasing function.

lower willingness to wait than T. But this cannot happen in equilibrium, as every rider would be willing to wait, justifying the assumption that U is increasing in Q. Furthermore, a change in Q in equilibrium could only happen by a change in prices, through which every rider that now decides not to take a trip was a marginal rider whose utility from a trip is p, and therefore $U_Q = p$.

With this definition of gross utility in mind, we define riders' surplus as

$$RS(Q, T) = U(Q, T) - pQ \tag{7}$$

Finally, welfare is gross utility minus social cost:

$$W(Q, L, T) = U(Q, T) - C(L) \tag{8}$$

Alternatively, welfare is the sum of riders' surplus, drivers' surplus, and revenues, but since payments are just transfers, they cancel out to obtain the same expression.

3.4 Pricing and Wild Goose Chases

We will now analyze how pricing affects the equilibrium in this market. Following the basic intuition from standard microeconomics, we should expect there to be some price that maximizes welfare and one that maximizes revenue, the latter higher than the former. We will see in our calibration that this is indeed the case. In this section we prove some stronger results related to WGCs. We end up with two main conclusions: First, WGCs can always be avoided by increasing prices. Second, lowering prices during WGCs leads the market towards market collapse and very sharp decreases in the number of trips, welfare, and revenue.

To simplify our notation, let $\epsilon_Y^X = \left|\frac{Y}{X}\frac{\partial X}{\partial Y}\right|$ denote the elasticity of X with respect to Y, and let $\sigma = \text{sgn}(\epsilon_T^S)$. The characteristic feature of WGCs is then $\sigma < 0$, whereas $\sigma > 0$ in good equilibria. Also let ϵ_l be the elasticity of $l(\cdot)$. Note, from equation 4, that the labor supply elasticity to a change in prices, given a fixed number of trips, is not given by ϵ_l. An increase in earnings leads to an increase in labor, which spreads out earnings more thinly across drivers, and this effectively means that supply is less elastic. Instead, the labor supply elasticity is given by $\epsilon_L = \frac{\epsilon_l}{1+\epsilon_l} \in [0, 1]$, which will be a key parameter in our model.

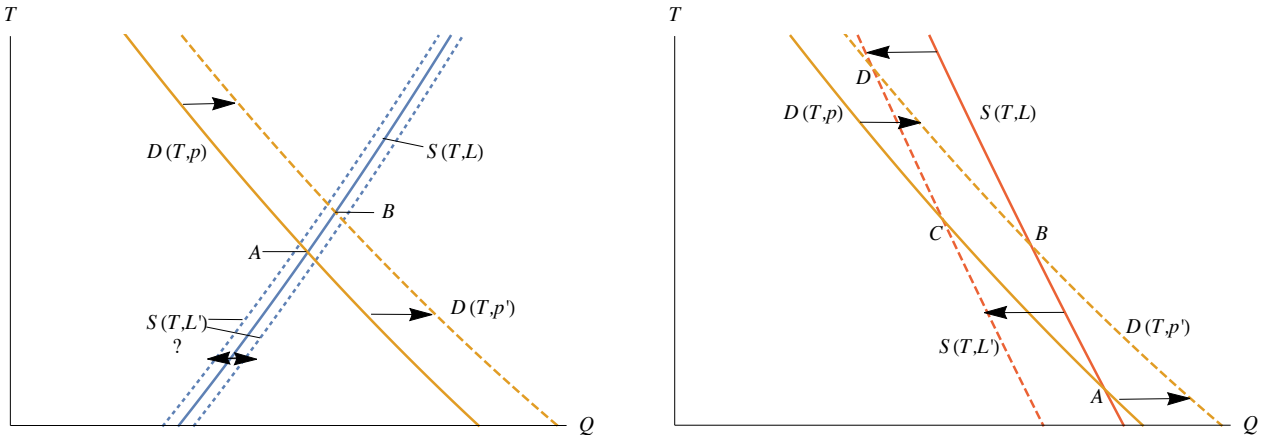
We start with the following proposition:

Proposition 1. *The price elasticities of equilibrium number of trips and drivers and pickup time are*

$$\epsilon_p^Q = \frac{1}{\Delta}\frac{-\sigma\epsilon_T^S\epsilon_p^D + \epsilon_L\epsilon_T^D\epsilon_L^S}{\epsilon_T^D + \sigma\epsilon_T^S} \qquad \epsilon_p^L = \frac{\epsilon_L}{\Delta}\left(1 - \frac{\sigma\epsilon_T^S\epsilon_p^D}{\epsilon_T^D + \sigma\epsilon_T^S}\right) \qquad \epsilon_p^T = -\frac{\epsilon_p^Q + \epsilon_p^D}{\epsilon_T^D} \tag{9}$$

where $\Delta = 1 + \epsilon_L\frac{\epsilon_T^D\epsilon_L^S}{\epsilon_T^D - \sigma\epsilon_T^S} > 0$. *In a good equilibrium the signs of all three elasticities are ambiguous. In WGCs, on the other hand,* $\epsilon_p^Q > 0$, $\epsilon_p^L > 0$, *and* $\epsilon_p^T < 0$.

Proof. See Appendix B.3. □



(a) Good outcome

(b) Wild goose chase

Figure 4: Response to a decrease in prices. These plots should be interpreted as a zoom in on the equilibria in the plots in Figure 2.

To understand this proposition, we will analyze good outcomes and WGCs separately. We start with good outcomes, where the intuitions resemble standard markets. Consider the market response to a price decrease, which is illustrated in Figure 4a. Let (p, L) be the original price and equilibrium number of drivers, and (p', L') be the final price and equilibrium number of drivers. This shifts demand outwards from $D(T, p)$ to $D(T, p')$, and the effect before considering the labor supply response is to move from point A to point B, with an increase in quantities and pickup times. However, the response of the labor market is ambiguous: on the one hand, the price decrease reduces earnings, but on the other hand more trips mean more earnings, so the direction of movement of the supply curve is ambiguous. If L' is the equilibrium labor supply, it is not clear whether $S(T, L')$ is to the right or to the left of $S(T, L)$.

For WGCs, on the other hand, it is clear in which direction the effect on all variables goes. Consider again a price decrease (Figure 4b). The effect of prices before considering labor supply changes is to move from A to B. This implies a decrease in quantities, despite demand shifting to the right. This counterintuitive phenomenon is because of the backwards bending supply curve, and follows the main intuition for what happens during a WGC: the market is overburdened and is beyond its maximum capacity. Further attempts to get more trips from the market result in less idle drivers and even longer pickup times, thus reducing the number of trips that drivers are able to serve. This effect is further reinforced by the effect of labor supply. As both prices and the number of trips decrease, drivers' earnings go down and the supply curve shifts to the left. The equilibrium thus moves further towards the upper left corner, with a decrease in the number of trips and an increase in pickup times.

This analysis explains the signs of these elasticities. If we analyze the expressions in Proposition 1 we can also say something about magnitudes. This leads to a result that will be

present throughout this paper. In WGCs any decrease in prices leads to sharp decreases in number of trips and drivers, revenue, and welfare. This comes from combining three separate effects that make the magnitude of these elasticities greater in a WGC than in a good equilibrium. First of all, with WGCs all terms in all three expressions are positive. This reflects the fact that there are no countervailing effects, which makes the effect of prices stronger.

The second effect is a self-reinforcing feedback cycle between supply and demand for trips when they are both decreasing. In the expressions in Proposition 1, this can be seen by noting that many of the terms have $\epsilon_T^D + \sigma\epsilon_T^S$ in the denominator, which is small when supply is decreasing as in a WGC. In a traditional market with increasing supply, as in Figure 4a, there is a balancing feedback cycle between supply and demand. Without taking into account the supply of drivers, the effect on quantities of a horizontal shift in demand of magnitude $dQ > 0$ (moving from A to B) has magnitude $-\frac{\sigma\epsilon_T^S}{\epsilon_T^D + \sigma\epsilon_T^S} dQ$, which is smaller than dQ because the increase in supply is mitigated by an increase in pickup time. But with a decreasing supply curve, the original effect is reversed and it might be magnified: as ϵ_T^D gets close to ϵ_T^S, the demand shift results in a decrease in quantities, which is then magnified by a further decrease in prices. In Figure 4b, the decrease in demand from A to B might be greater than the original demand shift when demand and supply cross at a small angle. The same feedback cycle affects horizontal supply shifts, such as one moving the equilibrium from B to D. This shift gets magnified by a factor of $\frac{\epsilon_T^D}{\epsilon_T^D + \sigma\epsilon_T^S} > 1$. Again, the fact that both curves cross at a small angle leads to larger changes in quantities and pickup times.

A third effect comes from positive feedback between supply and demand. Consider a decrease in prices. In a WGC, this leads to a decrease in the equilibrium quantity and a decrease in earnings, thus reducing labor supply. The decrease in labor shifts the supply curve to the left, further reducing quantities. For a good equilibrium, on the other hand, it isn't clear whether a quantity increases or decreases, so this feedback cycle might not even exist. But for a WGC this cycle unambiguously leads to positive feedback. The strength of this effect is represented by Δ in the denominator. This term is the determinant of the Jacobian of the matrix in the implicit function theorem, and is greater than one for good equilibria but less than one for WGCs.

Putting together all three effects, it is not surprising to see that the number of trips, welfare, and revenues all decrease very quickly to zero as prices go down in WGCs. The fact that all effects go in the same direction gets compounded by both feedback cycles, leading to a quick market collapse. The mirror effect is that an increase in price increases all these quantities very quickly. The next proposition states that an increase in prices eventually takes the market out of the undesirable state of a WGC.

Proposition 2. *Suppose that the highest equilibrium of the market at price* p *is in a WGC. Then there exists some price* $p' > p$ *at which the highest equilibrium of the market is no longer in a WGC.*

Proof. See Appendix B.4. □

The key to this proposition is the fact that demand is bounded and it goes to zero for all pickup times as price goes to infinity. As the whole demand curve shifts downwards, price will eventually reach some point in which maximum supply is greater than maximum demand. In this case there has to be at least one good outcome in equilibrium, which is evident from a plot like Figure 2.

Although it is true that increasing prices takes the market out of a WGC, it is not necessarily true that decreasing prices takes the market into a WGC.[7] When supply is very high relative to demand, for instance, it might be the case that the highest equilibrium for every price is in a good equilibrium. Consider, for instance, a demand function such that $D(T, 0)$ is always below the curve that joins the loci of maxima of $S(T, L)$ for different values of L. Then for every (p, L) there exists a crossing between supply and demand that takes place at a WGC. This might have been the case, for instance, with taxi telephone dispatch markets. Given how slow it took to get a taxi, demand was limited to a few niche users, such as people wanting to go the airport. This might explain why these markets probably functioned smoothly without WGCs.

The following proposition analyzes the effect of prices on revenue, welfare, and drivers' surplus:

Proposition 3. *The effect of prices on welfare is given by*

$$\frac{dW}{dp} = U_T \frac{dT}{dp} + \frac{Q}{\Delta}\left[(1-(1-\tau)\epsilon_L)\frac{-\sigma\epsilon_T^S\epsilon_p^D}{\epsilon_T^D+\sigma\epsilon_T^S} - \epsilon_L(1-\tau) + \epsilon_L\frac{\epsilon_T^D\epsilon_L^S}{\epsilon_T^D+\sigma\epsilon_T^S}\right]. \tag{10}$$

This derivative is positive in a WGC.

An increase in prices increases revenue and drivers' surplus in a WGC. The effect on riders' surplus is given by

$$\frac{dRS}{dp} = \epsilon_p^T\epsilon_T^U\frac{U}{p} - Q \tag{11}$$

which is positive if and only if $\epsilon_p^T\epsilon_T^U\frac{U}{pQ} > 1$.

Proof. See Appendix B.5. □

This proposition breaks down welfare effects during a WGC into an unambiguous benefit to drivers and the platform (as the earnings of both rise) and a more ambiguous effect on riders. Gross rider utility unambiguously increases, but greater payments might lead to a decrease in their surplus. Marginal riders' payments exactly offset their gross utility, so what matters in the end is inframarginal riders' utility: whether the decrease in pickup times is enough to compensate the increase in prices. In order to pin down whether this is the case, we would have to make additional assumptions on the way pickup times affect their utility (or more precisely, on the way gross utility depends on pickup times). If riders' utility is sensitive to waiting times

[7] It might even be the case that a price decrease takes the market out of a WGC. But this requires a very inelastic demand, so it is a pathological case than a realistic possibility.

with an elasticity on the order of $\frac{pQ}{U}$,[8] then what matters is whether ϵ_p^T is greater or less than one. And from our previous analysis this quantity is very likely to be high due to the reinforcement effects between the supply and demand curves and between the market for trips and the market for labor. So if inframarginal riders' utility does depend on time, riders' surplus is likely to increase with prices.

Even though the effect of prices on riders' surplus might be ambiguous, the total effect on welfare is unambiguous in WGCs. All transfers offset each other, so what really matters is the effect on gross utility and social cost. The direct effect of an additional trip increases gross utility, but it also causes an increase in labor that increases social cost. The first effect has magnitude p, whereas the latter effect has magnitude $(1-\tau)\epsilon_L p$: the wedge introduced by the platform ensures that not too many drivers enter the market, and the fact that drivers split revenues among themselves further magnifies this effect. So the increase in gross utility is greater than the increase in social cost. The direct effect of price on labor is also positive: a fixed percentage increase in drivers shifts supply upwards by a larger percent due to increasing returns to scale, and this is further magnified by the feedback between supply and demand, so the net effect is also an increase in welfare. Both channels get magnified by feedback between the labor and trip markets, which leads to a net increase in welfare.

3.5 Optimal pricing

So far we have worked under the assumption that τ is fixed. Uber has typically maintained a fixed value as time goes by, although it has taken different fractions for different drivers. Furthermore, they have not tried to change it across different times as they surge. Therefore we believe this is a reasonable assumption. For completeness, in this section we analyze how prices would look if a platform were willing to change τ. The rest of the paper will again treat τ as fixed.

We now use the insulating tariff approach as in Weyl (2010). Instead of maximizing directly in prices, we will maximize in the number of drivers working and the number of trips. First, note that setting (p, τ) is equivalent to setting p and $p' = (1-\tau)p$, the effective price for drivers. Furthermore, we can reparameterize the space (p, p') into the two dimensional space (Q, L), which makes the whole analysis much less burdensome.

Our first result gives expressions for optimal prices under welfare and revenue maximization:

Proposition 4. *Welfare maximizing prices are given by*

$$p = \bar{u}_T T \epsilon_Q^T \qquad p' = \bar{u}_T T \epsilon_L^T, \tag{12}$$

where $\bar{u}_T = \frac{U_T}{Q}$ *is the average change in utility of inframarginal users caused by a change in waiting time.*

[8] A crude accounting leads to this kind of conclusion: suppose that $\tilde{U}$ measures rider's utility in units of pickup time. Then $\epsilon_T^{\tilde{U}} = \frac{TQ}{\tilde{U}}$.

Revenue maximizing prices are given by

$$p = \frac{1}{1-\frac{1}{\epsilon_p^D}}\tilde{u}_T T \epsilon_Q^T \qquad p' = \frac{1}{1+\frac{1}{\epsilon_l}}\tilde{u}_T T \epsilon_L^T, \tag{13}$$

where $\tilde{u}_T = -\frac{Q_T}{Q_p}$ *is the average change in utility of* marginal *users caused by a change in waiting time.*

Proof. See Appendix B.6. □

As usual in multi-sided markets, revenue maximizing prices have two distortions compared with welfare maximizing prices (Weyl, 2010). First, there is a Spence (1975)-Sheshinski (1976) distortion: first order conditions only take into account the utility of price-marginal riders and not the surplus of the price-average riders.[9] This distortion biases both prices downwards. Second, there is a markup term that biases passengers' price upwards and drivers' price downwards, since a profit maximizer wants to widen the gap between both prices. The net effect is that drivers' price unambiguously decreases, whereas there is an ambiguous effect on passengers' price (the mark-up raises the price, but the Spence distortion lowers it).

Note that increasing returns to scale implies that $-\epsilon_Q^T + \epsilon_L^T > 0$: as the number of drivers and trips increase by the same proportion, waiting times go down. An immediate consequence of this and of the expressions for welfare maximization is the following:

Proposition 5. *Welfare maximization requires a subsidy, i.e.,* $\tau < 0$.

This is the main point in Arnott (1996). It can be understood as follows: increasing returns to scale mean that increasing the market size yields greater welfare. Thus, there exists and implicit externality from every additional driver and passenger, which means that the market requires a subsidy for optimality.

4 Empirical Evidence of Wild Goose Chases

In this section we show descriptive empirical evidence that WGCs are indeed a problem in actual markets. One could have thought that since Uber's surge pricing algorithm is meant to avoid bad market situations, then it should be able to detect the situations in which WGCs would have occurred. If that was the case we would see no evidence of WGC. We will show that Uber's algorithm seems to be good at avoiding WGCs most times, but there are still a few times (around 1.5% of the time) during which WGCs still occur.

[9] See Bulow and Klemperer (2012) for a general analysis of the harms created by the tendency of random rationing systems to neglect this surplus.

4.1 Data

We use Uber data from Manhattan between December 1st, 2016 and February 28, 2017. We look at data both from UberX and from UberPool because the set of drivers working for both products is the same. As we show in Section 7, WGC are also a problem with UberPool, so our qualitative results for UberX also hold for UberPool. Furthermore, less than a third of the trips in our sample are UberPool trips, which means that our quantitative analysis is mostly driven by UberX.

We aggregate all of Manhattan, which means that our data has a time series format. We do not disaggregate the data into smaller regions because the spatial nature of this market causes complications that we don't consider in our model.[10] We also aggregate the data into half-hour periods, despite the fact that we observe what happens at a higher resolution, because this resembles more closely the steady state we analyze in our model. A high number of ride requests during one minute, for instance, would cause a small number of idle drivers in the next period, but this is all because of transient dynamics.

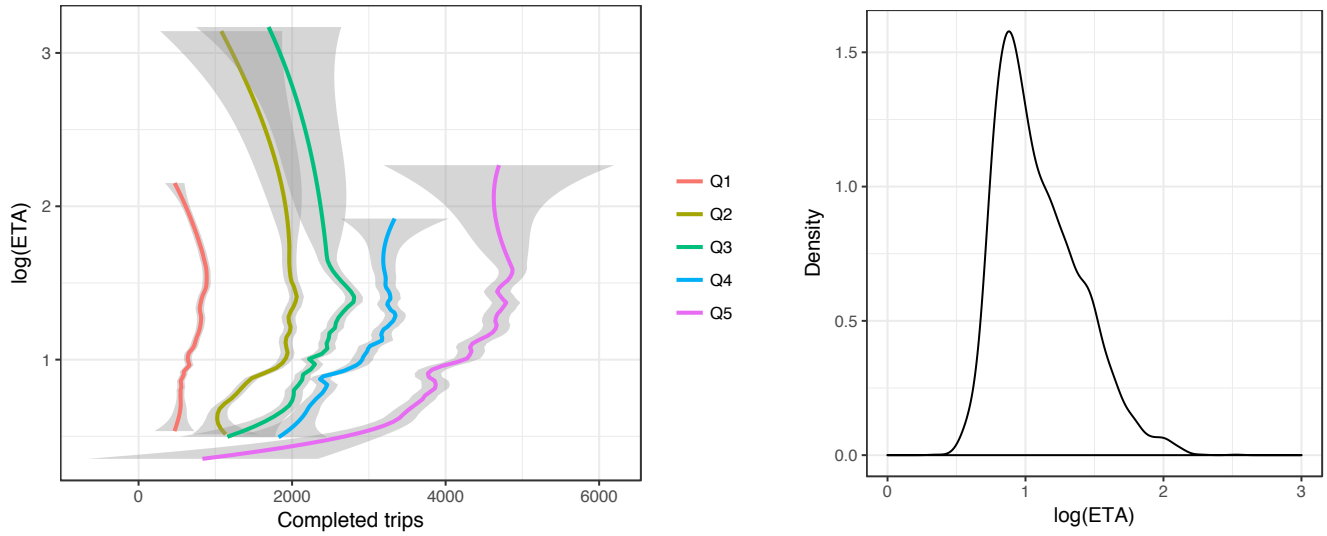
On the supply side, we observe the total number of driver-minutes spent in each one of the three states we consider in our model (idle, en route, and driving a passenger). On the demand side we observe how many riders open the app and look at the UberX or UberPool product page. We also observe all the trips requested during the period we analyze, and the number of trips that were eventually completed and those that were not. For those that were not completed, we can observe the reason why it was not: the driver cancelled, the rider cancelled, or the rider could not be matched because there were no nearby drivers or none of them accepted the trip.

We also have data on surge pricing. We observe the surge multiplier in each one of the small geofences used by the surge algorithm, which updates prices every two minutes. For the analysis in this section we average it as an unweighted mean over all geofences and all two minute period during each hour.

4.2 Supply of trips

In this section we show an empirical analogue of Figure 1, showing the nonmonotonic relation of supply as a function mean pickup times. This nonmonotonicity holds when the number of drivers is fixed, but the actual number of drivers varies substantially in the data. Thus, we cannot simply plot pickup times drivers and the number of trips in the data, as our data would also be affected by how the supply curve shifts to the left and to the right. In order to deal with this, we split the sample into five quintiles of the number of drivers working, and we show that the resulting fit exhibits the characteristic backwards bending supply curve that leads to WGCs.

[10] For instance, we see that there are times in which a very high number of rides in a particular small area are served with essentially zero idle drivers, which opposes our findings from figure 1 as this would imply infinite time. The reason for this behavior is that a huge local demand spike (such as the end of a concert or sports event) was served with the idle drivers from nearby locations.



(a) Pickup times and number of trips

(b) Distribution of pickup times

Figure 5: Subfigure 5a plots the logarithm of expected pickup time when matched, measured as the log of the expected time to arrival, against number of trips completed. Fit lines are locally weighted quadratic regressions using the 25% of the data that is closest to every point, using tricubic weights. The grey shaded regions represent 95% confidence intervals for the pointwise mean. Subfigure 5b shows a kernel density estimate of the distrubution of ETAs.

We will show that one characteristic feature of WGCs are high cancellation rates. Thus, if we used actual pickup times we would have a truncated distribution, as people with longer pickup times would be more likely to cancel trips. Instead, we use the ETA shown in the app immediately after the rider is matched to a driver as our measure of T. Figure 5 shows supply as a function of $\log(T)$. The function exhibits decreasing behavior for high pickup times, just as our model would predict.

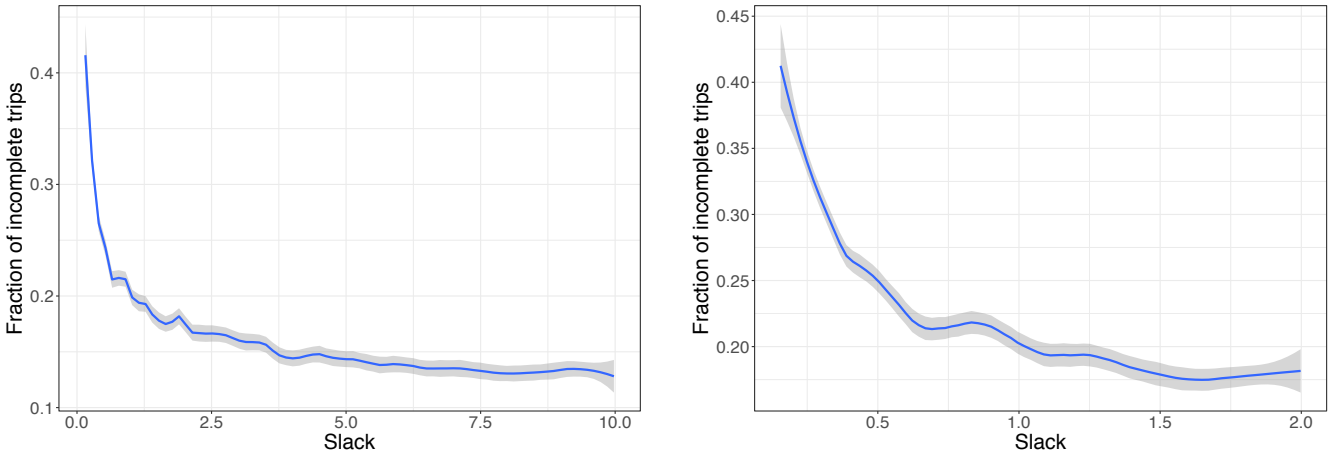
The point at which WGCs start to happen is somewhere around 1.6. The histogram of $\log(T)$ shows that this is relatively rare, which suggests that Uber's surge pricing algorithm indeed avoids getting into the very worst situations in which waiting times become very high. However, there do exist times in which the decreasing trend is clear, which means that there is still some room for improvement.

Note that the WGC behavior is especially clear for quintiles 1-3, at times when the number of drivers working is low. This is evidence that Uber is better able to avoid WGCs at the busiest times, but less so during times of low demand and supply. Even during busy times the leftmost end of the plot seems to be flat, which means that although WGCs are avoided, any change in policy that would decrease prices by only a bit would lead the market to the WGC region. Also note that although all observations in these plots are weighted equally, it is much more important for welfare to avoid WGCs during busy times because the number of passengers and drivers benefitting from the platform is larger. Thus, Uber seems to have calibrated their model well during the most important times, but there seems to be some room for improvement at

some less important times. One likely reason for this to happen is the fact that at these times the market is much thinner, meaning that Uber has been able to collect much less data to calibrate their surge models.

4.3 Performance measures and slack

A separate way to see whether WGCs take place is to look at some performance measures' behavior as a function of slack. We know that WGCs take place when slack is less than some value between $\frac{1}{4}$ and $\frac{1}{2}$, where the exact value depends on the current conditions like traffic speed and the thickness of the market. This means that for the aggregate data we are looking at we should see a sharp decline in these performance measures at these values. For reference, slack is below $\frac{1}{2}$ for 11.4% of observations, and below $\frac{1}{4}$ for 1.36% of observations.



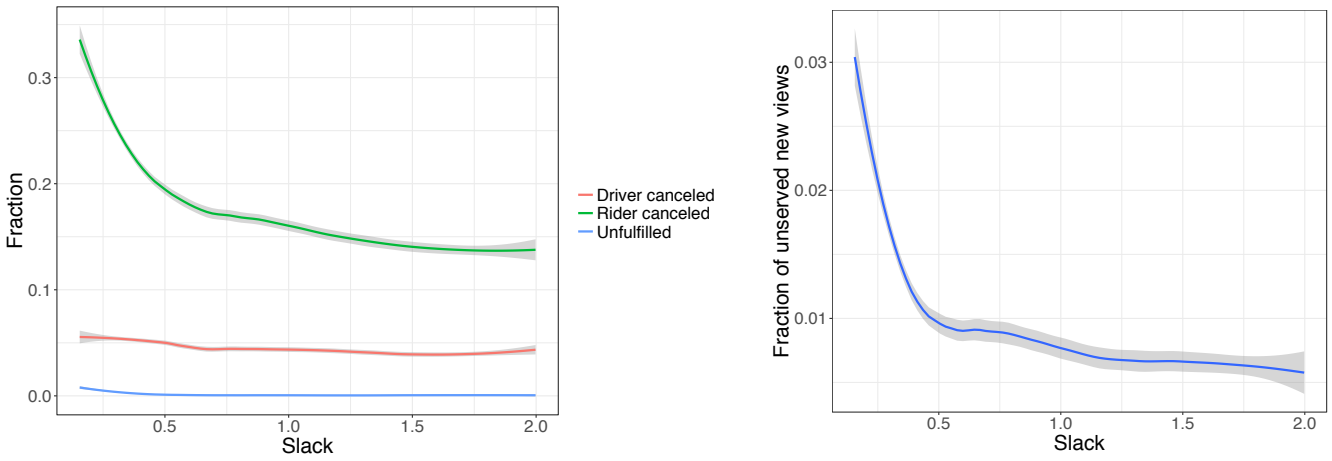
(a) $s < 10$, which includes 96% of the data

(b) Focus on $s < 2$.

Figure 6: Fraction of trips that are not completed as a function of slack. Fit lines are locally weighted quadratic regressions using tricubic weights. The figure to the left uses 5% of the data that is closest to every point, whereas the one to the right uses 25%. The grey shaded regions represent 95% confidence intervals for the pointwise mean.

The first performance measure we use is the fraction of ride requests that are eventually completed. Figure 7 shows how it behaves as a function of slack. The figure to the left shows almost the full range of values of slack, except for a few outliers with $s > 10$. It is very stable for $s > 5$, with values between 0.1 and 0.15. We focus on $s < 2$ in the figure to the right. The fact that the plot is increasing is not surprising, since it is a mechanical relation that times with low completion rates are times with few idle drivers. However, the fact that the slope has a sudden change approaching 0.5 suggests that WGCs might be starting to take place there.

In Figure 7a we disaggregate the trips that were not completed to get an understanding of what happens when completion rates are low. The main cause of non-completed trips are rider cancellations, and it is the main subgroup that varies with slack. This suggests that during WGC it is often the case that passengers decide not to wait for the driver they were matched to. It



(a) Disaggregation of trips that are not completed

(b) New views without a driver nearby

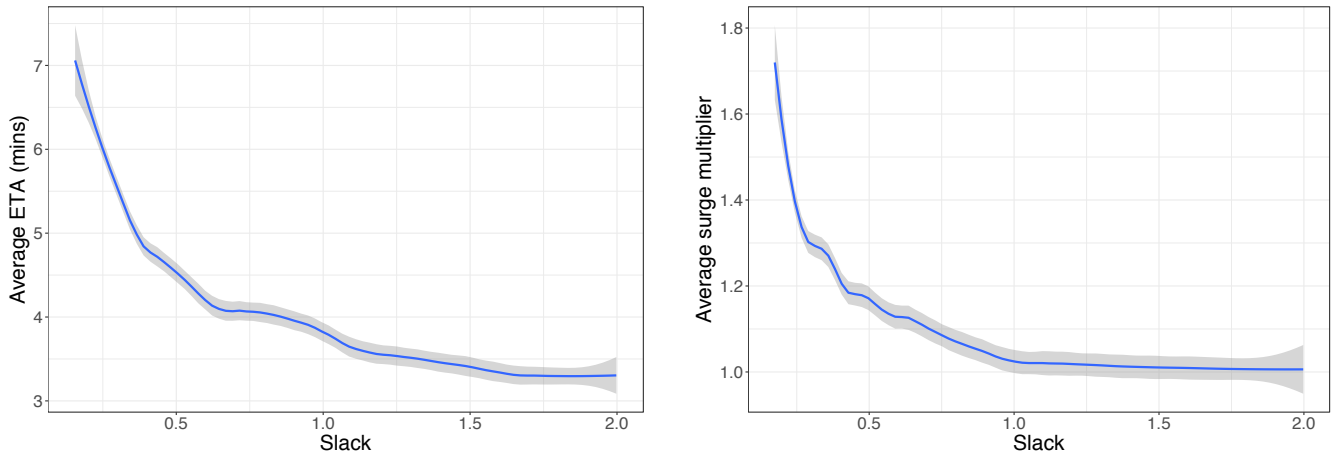
Figure 7: Disaggregation of trips that are not completed and fraction of new views that have no driver nearby as a function of slack. Fit lines are locally weighted quadratic regressions using tricubic weights. Both figures use 25% of the data that is closest to every point. The grey shaded regions represent 95% confidence intervals for the pointwise mean.

might seem surprising that the number of driver cancellations only sees a slight uptick to the left. The main reason for this is that Uber has a system of incentives to avoid drivers cancelling rides. Finally, the number of unfulfilled trips is extremely small compared with cancellations. This is mostly due to the fact that whenever no driver is available within some radius, the app displays a message telling riders that there are no available drivers in their vicinity.[11] This suggests another performance measure we can use: the number of views that are shown a message of no nearby drivers. We analyze this in Figure 7b, and we also see a sudden deterioration in service when slack goes beyond 0.5.

The final performance measure that we analyze is ETAs, in Figure 8a. There is again a mechanical relation between slack and ETA, but, once again, we observe a spike when the market approaches the WGC region.

In Figure 8b we analyze how surge pricing changes with slack. We see that the surge multiplier very rarely goes above 1 when slack is above 1. On the other hand, the surge multiplier becomes larger as slack goes down, and we see a strong increase especially below 0.3. This means that Uber reacts to slack, or at least reacts to measures that are closely connected to slack. However, given that the sudden degradation in performance measures takes place at a higher value of slack than the spike in surge multiplier, Uber might benefit from reacting more strongly when slack is between 0.5 and 0.3.

[11] Decreasing this radius can be used as a tool to manage WGCs. However, Uber uses it mostly as a tool to avoid extreme events in which people are matched to someone very far away, and not as a tool to manage scarcity. In future versions of this project we will compare this mechanism to the surge pricing solution used by Uber.



(a) ETAs

(b) Surge multiplier

Figure 8: ETAs and surge multiplier as a function of slack. Fit lines are locally weighted quadratic regressions using tricubic weights. Both figures use 25% of the data that is closest to every point. The grey shaded regions represent 95% confidence intervals for the pointwise mean.

5 Surge Pricing

In this section we calibrate our model and apply it to quantitatively analyze optimal pricing and in particular the effects of allowing versus prohibiting surge pricing. We begin by discussing our calibration.

5.1 Calibration

In order to calibrate our model we need to make additional functional form assumptions on our model. We also need to fit a few parameters so that the model matches some moments of the data.

Demand

Let $r(p)$ be the fraction of potential riders that are willing to pay a price p, and let $g(T)$ be the fraction that are willing to wait a time T. We assume that willingness to pay and wait are independent, so that demand is $D(T,p) = \lambda g(w) r(p)$[12]. This assumption might be bad if both decisions are positively or negatively correlated. However, we do not have strong reasons to believe they go either way. An example of them being negatively correlated is a businessman that is late for a meeting, and who is willing to pay a lot but is not willing to wait. On the other hand, an example of them being positively correlated is an old man that needs to visit his

[12]This demand function assumes that pickup time is the same for every driver, or that riders only respond to mean pickup times. A more realistic expression would be $D = \int \lambda g(T) r(p) dF(T)$, where F is the distribution of T. We stick to the simpler demand function to avoid computing distributions of waiting times.

daughter but cannot drive: he is willing to pay a high price because of the lack of an outside option, and he is in no rush and willing to wait.

In order to compute gross utility we would have to make assumptions on the way that utility depends on time. Instead of making this kind of assumption, we simply assume that utility does not depend directly on pickup time, although it does depend indirectly through the number of trips requested. We showed in our theoretical section that taking into account the disutility of waiting only makes the effects of WGCs even more striking. We therefore take the most conservative approach, which makes our results less sharp than if we did take it into account.

Gross utility is thus the gross utility per rider willing to wait times the number of riders willing to wait:

$$U(p, T) = \lambda g(T) \left[\int_p^\infty r(p')dp' + pr(p) \right] \tag{14}$$

We assume that willingness to pay has a double Pareto lognormal (Reed, 2003; Reed and Jorgensen, 2004) distribution with parameters $\alpha = 3$, $\beta = 1.43$, $\mu = 1.1$, and $\sigma = 0.45$. The parameters α, β, and σ are chosen so that the distribution has the same shape as the US income distribution, as in Fabinger and Weyl (2016). The parameter μ, which is simply a horizontal rescaling of the distribution, is chosen to fit the elasticities in Cohen et al. (2016), who estimate willingness to pay of riders on the platform Uber. The function $r(p)$ arises from this distribution, where p is the surge multiplier. We also assume that the ability to wait has a lognormal distribution with mode 5 minutes and variance such that the elasticity of the corresponding function $g(T)$ agrees with the value from Cohen et al. (2016).

Labor supply

For labor supply we assume a constant elasticity functional form, $l(e) = A \left(\frac{e}{1+\frac{1}{\epsilon_l}} \right)^{\epsilon_l}$. This results in a cost function $C(L) = A \left(\frac{L}{A} \right)^{1+\frac{1}{\epsilon_l}}$. We assume an elasticity of 1.2 based on Angrist and Caldwell (2017), where they estimate a medium-term elasticity from experiments measuring drivers' supply under contracts with different payment schemes. Very short-term elasticities, for unexpected demand shocks, are likely to be lower and very long-term elasticities, for secular changes in earnings on the platform, are likely to be higher. Since we observe the number of drivers and trips, as well as the average surge multiplier, we can compute the expected hourly earnings and back out the value of A.

Supply of trips

The only primitive that determines the functional form of $S(T, L)$ is the functional form of $T(I)$. We fit it by using data on the average pickup time as a function of distance to the matched driver, which we denote by $\check{T}(x)$. In a simple, homogeneous space, $\check{T}(x)$ is simply a linear

function, $\frac{x}{v}$, where v is the speed. However, matters are considerably more subtle in practice. The pattern of roads in some cities has one-way streets every other block, and in others follows radial rather than axis-aligned coordinates. Furthermore, speeds are greater when traveling longer distances since drivers are able to take larger streets or highways. This implies that the appropriate formula for $\check{T}(x)$ in practice will vary from city to city.

We take a function of the form $\check{T}(x) = a(1 - e^{-bx}) + cx$. The first term captures the fact that cities' street patterns cause inefficiencies when traveling short distances. The second term means that speed eventually reaches some terminal value c, which is the speed once drivers take a main street. This functional form fits very well the data for trips in Manhattan obtained from Uber, as shown in Appendix A.

Once we fit $\check{T}(x)$, we obtain an expression for $T(I)$ as follows. In two dimensional space, the density of drivers at a distance x from an arbitrary point is $2\pi Ix$, (a measure to be integrated with respect to x) which is the hazard function of the nearest driver. The CDF of the distance to the nearest driver $G(x; I)$ is then given by the differential equation $\frac{dG}{dx} = 2\pi Ix(1 - G)$, whose solution, which corresponds to a Weibull distribution, is $G(x; I) = 1 - e^{-\pi Ix^2}$. If the average pickup time as a function of distance is $t(x)$, then $T(I) = \int_0^\infty \check{T}(x) dG(x; I)$. Given the functional form assumption for $\check{T}$, the resulting expression for expected pickup time is $T(I) = \frac{1}{\sqrt{4I}} \left(c + 2ab \exp\left(\frac{b^2}{4\pi I}\right) \Phi\left(\frac{b}{\sqrt{2\pi I}}\right)\right)$ where Φ is the CDF of a standard normal distribution.

Note that under this functional form assumption, $\lim_{I \to 0} -\epsilon_I^w = \frac{1}{2}$, but for larger values of I (about as large as could reasonably be expected in practice), $-\epsilon_I^w$ reaches an interior minimum at a value of about 0.26.[13] That is, in cities with a very dense coverage of drivers, fewer idle drivers relative to those picking up riders are needed to avoid WGCs. This is intuitive because when drivers are very dense, additional idle drivers do not rapidly reduce pickup times. It is therefore not problematic for drivers to spend a greater fraction of their time on "dead miles". Taken to an extreme, as I grows large it is natural that more time is spent picking up passengers relative to being idle, as most drivers must drive around the block to get a nearby rider; only if so many drivers can be made available that one is directly in front of every potential rider's house can this small friction be eliminated. When there are fewer available drivers, on the other hand, increasing driver density is more beneficial and thus more idle drivers relative to those picking up riders are needed to avoid WGCs as each additional driver "fills in" an important part of the city grid.

Calibration to different markets

We calibrate the parameters of our model by using aggregate data from the same dataset for Uber in Manhattan we used in our empirical section, between December 1, 2016 and February

[13] Eventually, however, as $I \to \infty$, it again becomes $\frac{1}{2}$. This makes sense because the inefficiencies of going around the block eventually level off once there are so many cars that pickup time is determined by driving straight down the block.

Market	λ (sessions/h $\cdot$ km^2)	R (trips/h $\cdot$ km^2)	D (drivers/km^2)	A (drivers/h $\cdot$ km^2)
Mean	223.4	97.1	50.6	203.6
Strong	354.3	146.8	71.5	246.2
Weak	147.7	63.4	44.5	281.5

Table 1: Observables and parameters for the mean, weak, and strong market.

28, 2017. We exclude December 15-January 7 since these are atypical days because of holidays. We focus on weekdays between 7 am and midnight. The only parameters that remain to be input in the model are λ, A, and a $r(1)$ (since even with surge multiplier 1x and waiting time zero not every person who opens the app requests a trip). We observe λ directly in the model. We back out A and $r(1)$ as the values that lead to an equilibrium with the observed number of trips and drivers.

For the main calibration we use average values over the whole sample. Thus, this can be thought of as the "average" behavior of the Manhattan market. This is the main specification we use. In a separate specification, we model two different markets, the one between 11 am and noon, which we call the weak market, and the one 6 and 7 pm, which we call the strong market. We assume that for these two markets all the model primitives stay the same as for the average market, except for λ and A. Table 1 compares the average number of drivers, sessions, and trips, as well as the calibrated parameter A, for the weak, strong, and average market. The number of sessions, trips, and drivers are greatest for the strong market and the least for the weak market. The supply shifter A follows a different pattern:[14] Supply is highest in the weak market, in the middle of the workday. It is also higher than average in the strong market, probably because many people work a few hours after their full time job.

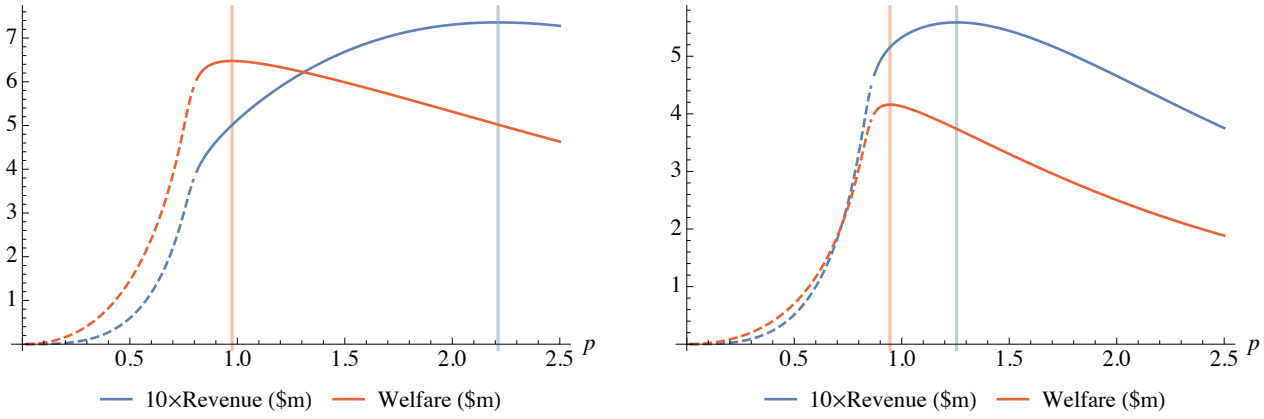
5.2 Quantitative analysis of pricing

Figure 9 shows how revenue, welfare, and rides behave as a function of passengers' price for fixed $\tau = 0.238$, which corresponds to the average value used by Uber in Manhattan.[15] The left region with dashed lines represents prices at which WGC occur.

Subfigure 9a uses the very inelastic estimates for demand from Cohen et al. (2016). The main thing to note is the asymmetry of the welfare function around its maximum. There is a drastic drop in welfare to the left of the WGC threshold. This is evidence that WCG equilibria can lead to dramatic welfare losses and are "Pareto dominated" in the sense that WGCs in aggregate hurt all of drivers, riders and the platform (though they may slightly benefit some price marginal riders who are willing to wait a long time). To the right of the threshold, any price increase

[14] A has the same units of D. Its interpretation is that it is the number of drivers who would be willing to work if their hourly earnings were equivalent to working with no time spent being idle or picking up passengers, with surge multiplier $1 + \frac{1}{\epsilon_D}$.

[15] The exact value varies from driver to driver, depending on the time at which they entered the platform.



(a) Cohen et al. (2016) elasticities.

(b) Twice the Cohen et al. (2016) elasticites.

Figure 9: Revenue and welfare for the Manhattan market as a function of price for passengers. Dashed lines represent points to the left of the WGC threshold. The vertical lines represent the optimal prices for the function with the corresponding color.

benefits some group (typically drivers and the platform) and hurts others (typically passengers), and since there is a tradeoff, changes in welfare are not too large: a 20% increase in prices from the optimum only decreases welfare by less than 8%. On the other hand, a 30% decrease in prices from the optimum leads to a 50% decrease in welfare. This goes in line with our analysis from Section 3: feedback between all parts of the model lead to a quick market collapse as prices go down. Finally, note that optimal prices are just a bit above the price where WGCs start. Any further increase in prices results in social welfare losses due to too many drivers working and a waste of time.

The main implication is that in order to maximize welfare it is much worse to err by setting prices too low than by setting them too high. Thus, in the face of uncertainty, platforms would like to set prices with some margin above the threshold in order to avoid WGC from ever happening.

For this calibration the threshold is in the inelastic part of $r(p)$. By the usual intuition from the analysis of a monopolist, the revenue maximizing price is in the elastic region, which starts at around price 2.2. Even in this case, the revenue function has a kink at the threshold, which means that there is a dramatic deterioration of revenue once WGC start to take place. Furthermore, the effect on welfare of setting the very high revenue maximizing price is mild compared with the potential effect of a WGC. This corresponds to a 130% price increase from the welfare optimum that decreases welfare by 28%, which is the same decrease that would be caused by a 22% price decrease from the optimum.

The elasticity estimates from Cohen et al. (2016) are based on studying the effects of price increases that last only a few minutes typically on ride requests. They are thus unlikely to reflect what would happen if the platform consistently set prices as high as 2.2. Subfigure 9b shows the same calibration, assuming that elasticities are twice those in Cohen et al., i.e., around

0.8-1.2 for prices between 1 and 2. We believe this to be a much better illustration of the way the actual market behaves when prices are predictable and medium-to-long-term adjustments (e.g. switching to another ride-hailing platform or driving to work) are made to these prices by riders. Note first that the general form of the welfare function does not change much. The elastic region starts at 1.2, which is the revenue maximizing price. Revenue and welfare maximizing prices are now close to each other, and more importantly, changes in welfare and revenue are not substantial for prices between them. On the other hand, both revenue and welfare drop dramatically after entering the WGC threshold. Thus, welfare and profit changes between both optima are second order when compared to the changes when getting below the threshold.

This implies that revenue and welfare are relatively well-aligned. Unless elasticities are as low as in Cohen et al. (2016), the main concern both of a revenue and a welfare maximizer is to avoid WGC. Whereas a welfare maximizer might be tempted to set prices close to the threshold, this would mean risking huge welfare losses given the uncertainty of the market, and maximizing expected welfare would imply setting a higher price very close to the profit maximizing one.

Given this, from now on we will analyze the social benefits of surge pricing using the original elasticities measured by Cohen et al.. We also assume that the platform maximizes welfare. We believe this to be a good approximation since these platforms' main concern is long run profit instead of short run revenue, and they are thus very concerned about consumer satisfaction. By adjusting incentives directly (by assuming welfare maximization) we correct for the tendency of the platform to lower prices to account for longer-term platform growth while maintaining realistic degrees of responsiveness to price changes to determine the effects of pricing on system engineering.

By surge pricing we mean the ability of the platform to change prices at different times. We still assume that τ is fixed. In order to do this analysis, we focus on a setup similar to the one in Aguirre et al. (2010) to analyze the welfare effects of price discrimination. As we mentioned in the calibration section, we analyze the market between 11 am and noon, which we call the weak market, and the market between 6 and 7 pm, which we call the strong market. These are, on average, the one hour intervals in our database with the highest and lowest demand. We first require the platform to have the same price for both markets, which is similar to what happens, for instance, with Gett, which does not have surge pricing. In the second setup we do allow the platform to set different prices for each market.

Figure 10 shows the results of this analysis. The constrained price is extremely close to the unconstrained price for the strong market. The reason for this is that welfare drops much more sharply to the left of the optimum than to the right. Another way to put this is that if the platform is constrained, it has little freedom to set prices below the strong market unconstrained optimum because it gets close to the WGC threshold, under which welfare in the strong market declines very abruptly. This means that allowing price discrimination leads to a significant

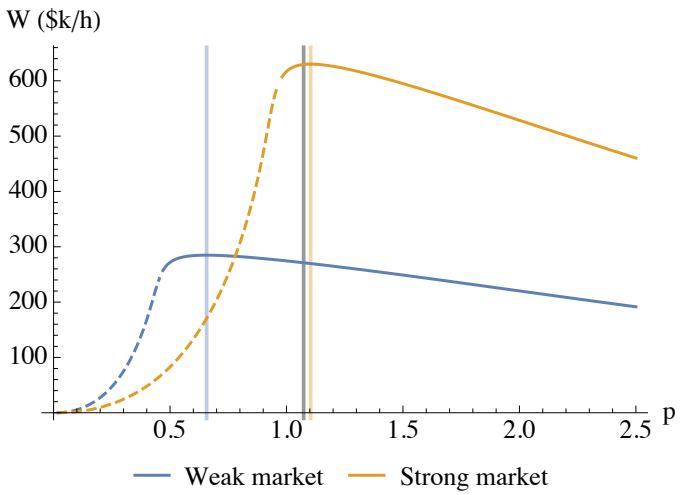


Figure 10: Price discrimination with fixed τ. The gray vertical line represents the optimal price without surge pricing. The blue and yellow vertical lines represent the optimal prices with discrimination for the weak and strong market, respectively.

reduction of prices in weak markets, whereas it only leads to modest increases in prices for strong markets, as we highlighted in the introduction.

This is very different from a common perception across the media and some regulators, that surge pricing hurts consumers. The main idea is that without surge pricing platforms' prices would always be at their base fare, whereas surge pricing allows them to engage in price gouging and extract large rents from the market, especially hurting riders. Our results show a very different story. Without surge pricing the platform would set prices at a high level to avoid ever getting in WGCs, which are very bad both for welfare and for their revenues. And allowing surge pricing would benefit consumers, potentially at the expense of drivers. Thus, drivers are the ones that should be most concerned about surge pricing, which reduces their welfare at times of low demand.

We find that the welfare maximizing price is around 1.1 in the strong market. One might think that this is at odds with the fact that multipliers often go above 1.5. However, there is substantial spatial variation, as well as between days of the week, and there is a high degree of unpredictability which often leads to high demand and scarcity of drivers. None of these sources of variation is captured in the way we calibrated this strong market, so it is not surprising to find a multiplier only slightly above one.

Our results also explain the fact that ride hailing platforms typically change prices upwards but not downwards. The consequences are not too bad if the ideal price was 0.7 but the actual price is constrained to be 1, whereas welfare decreases by a lot if the ideal price is 1.3 and the platform is constrained to 1. Even despite this fact, one might wonder why platforms have not decided to set prices below 1. The main reason is because of reputational pressures: they constantly face criticism for drivers not being paid well, and for predatory pricing trying to

avoid new entrants.

6 Alternative Solutions for Wild Goose Chases

We have focused on pricing as the main tool to avoid WGCs. However, there are many alternative ways in which platforms can avoid them. We will now discuss them and argue that they are not as desirable as surge pricing, either because they are less efficient or because they have practical or public relations limitations that make them infeasible. This is an informal analysis, but we will formalize these results in future versions of this paper.

A naive way to solve WGCs is to randomly deny trips to some riders. This is equivalent to an inwards shift of the demand curve, which moves the equilibrium away from a WGCs by the analysis from Figure 4. However, such random denial is inefficient compared to an increase in prices that achieves the same number of trips in equilibrium: an increase in prices is equivalent to denying trips to riders with the lowest willingness to pay, and therefore results in a more efficient allocation of trips.

A more nuanced solution is to deny trips to people whose realized pickup time is large. This still has the disadvantage that some people with high value will end up not receiving a trip, also resulting in an inefficient allocation. However, it has an advantage over a pure price mechanism because it avoids those trips with highest pickup times, in which a rider causes the greatest externality on other riders by removing an idle driver from the streets for the longest time. As a matter of fact, Uber does set a maximum dispatch radius on the order of twenty minutes, which means that it implements a similar kind of policy. However, it is only set in order to avoid the most extreme cases, thus truncating the right tail of the distribution of pickup times, and the policy is not modified dynamically to manage demand and supply imbalances.

Despite its apparent desirability, there is one important reason why expanding the use of a dispatch radius and making it the main tool used to avoid WGCs is not feasible in practice. One of the main reasons why ride hailing platforms have become successful is the guarantee of a reliable service: sometimes the rider might have to pay a high price, but in times of need a ride will always be available. Using smaller dispatch radiuses goes against this guarantee, so although it might avoid the problem of WGCs, it essentially means that at times of high demand there exists some probability of being denied a trip. Ride hailing platforms are therefore unwilling to make it their main tool manage supply and demand.

There exists a variant of a maximum dispatch radius that somewhat diminishes the problem of unreliability. Instead of simply denying a trip, the platform might make the passenger wait until some driver drops off a passenger in the rider's vicinity or some idle driver moves close to the rider. Essentially, this implies creating passenger queues in equilibrium, increasing thickness and improving match quality at the expense of waiting time, a common theme in the matching literature (Akbarpour et al., 2016).

However, it is unlikely that a riders' only option is to wait. Besides other modes of transportation, there are various competitors in the ride hailing market, so the user in many cases would just leave the platform and open the competing app. Furthermore, passenger queues are in tension with a user interface feature of current ride-hailing apps—that riders know immediately upon request the location and trajectory of a car driving towards them. This feature is considered very appealing to riders and our internal interviews suggest product leaders at Uber would be loath to compromise that element of the rider experience. Thus, although this solution might seem attractive from a social perspective, in practice it is unlikely that a private platform would be willing to follow it given the fierce competition between ride-hailing companies.

Besides these alternatives to solve WGCs, there also exist tweaks to their mechanism design that, although they may not completely avoid WGCs, they might be able to mitigate them, and if coupled with surge pricing might allow price changes to be milder. Some of these have been implemented to some extent by ride-hailing apps. The first one is to match passengers to drivers who are about to finish a trip nearby. This effectively increases the density of idle drivers, thus reducing pickup times and allowing drivers to spend more time driving passengers. But this does not entirely avoid WGCs unless the number of trips ending in some area is very high.

Another mitigating solution is to rematch drivers and passengers. WGCs can be understood as a dynamic inefficiency: sometimes drivers and passengers could have been matched better if the platform had waited long enough, but the fact that matching takes place greedily leads to suboptimal matching. In those cases two driver-rider pairs could be rematched to obtain a lower pickup time for both pairs. This is, again, a solution that effectively increases the density of idle drivers. But it has some important limitations. First, it creates uncertainty on the part of passengers about the arrival time of their driver, since rematching might drastically reduce it. Second, once a rider sees an ETA for the driver, if after rematching the ETA is reduced, it would not be possible to force the rider to meet the driver at an earlier time. Thus, this solution would only be helpful if riders are available before the initial ETA shown and if they are paying attention to see the new ETA.

One final mitigating solution that has nice theoretical properties is to charge for pickup time and not only for the segment from the origin and the destination. This effectively means pricing according to the externality caused on other riders by depleting streets of idle drivers. This solution, however, has the problem that riders would like to know the ETA before requesting the trip, which would require tentative matching with drivers before deciding to request. Besides this complication, there is a public relations issue that might arise: unlucky passengers that are matched to someone far from them would not only have to wait a long time; they would also have to pay a high price. This would be problematic since people expect platforms to provide a low pickup time and would feel that they would be charged for the platform's inability to offer low pickup times.

7 Ride-Sharing

Pooling services such as UberPool and Lyft Line, which allow trips to be shared by multiple riders, have increasingly become a widespread alternative to standard ride hailing products. In this section we extend our analysis to pooling. Given the much greater complexity of these services, we make stronger simplifying assumptions. Our main finding is that WGCs are also a problem, and most of our previous analysis still holds.

7.1 Model

We assume a simple model where demand and supply take the same form as in our previous model, and in which passengers can only take pooling trips. The difference between ride sharing and ride hailing is the matching technology. We will now show that it also results in a backwards bending supply function, which the leads to all of the main results in our paper.

Drivers can now be in one of five states. They can be idle, I, with one rider, B_1, with two passengers, B_2, picking up a rider while empty, K_1, and picking up a rider while driving one rider, K_2. Thus, at any given time the total number of drivers satisfies the following equation:

$$L = I + B_1 + B_2 + K_1 + K_2 \tag{15}$$

If a new rider requests a ride, he is matched to the nearest driver among those that are idle and those with one rider that go in a similar direction. Let q be the probability that some driver is taking a rider in a similar direction. We assume that this is independent of the state of the system. It is a quantity that depends crucially on how willing is the platform is to deviate a driver that is taking a rider to his destination. The rider that requests a ride thus sees an effective density of drivers $I + qB_1$, which is the density of drivers that could pick him up if he requested a ride. The pick-up time is therefore $T(I + qB_1)$. With this in mind, in equilibrium the total number of passengers picking up passengers $K_1 + K_2$ is equal to the rate of ride requests times the pickup time wR, which means that $L = I + B_1 + B_2 + T(I + qB_1)Q$.

We also assume that if a driver with a rider is deviated to pick up another rider, the trip time of the rider in the car increases by the time it takes to pick up the new rider. This amounts to assuming that on average the pick up location of the new rider is neither closer nor farther away from the final destination of the first rider. With this in mind, the total time of trips (without counting the pick up time) is equal to tQ, which must be equal to the time spent by drivers with passengers. The time spent driving two passengers counts twice, so this means that $tQ = B_1 + 2B_2$.

The number of drivers driving two passengers and one rider are related by the rate at which those with one rider are dispatched to pick up a second rider and the rate at which those with two passengers finish their trip. The rate at which they finish trips is twice the inverse average

length of a trip, $\frac{2}{t}$. The rate at which drivers get a second ride can be written as $\frac{qQ}{I+qB_1}$: since the effective density of available drivers is $I + qB_1$, the region for which the closest driver is any given driver is the inverse of this density, $\frac{1}{I+qB_1}$. Since the density of trip request rate is Q, the arrival rate to this area is $\frac{Q}{I+qB_1}$, and the probability that the arriving rider goes in the same direction as the old rider is q, which multiplies this rate. Therefore, $B_2 = \frac{t}{2}\frac{qQ}{I+qB_1}B_1$.

7.2 Wild goose chases

From the previous analysis, supply is given by the solution in (Q, B_1, B_2) to the following system of equations:

$$L = I + B_1 + B_2 + T(I + qB_1)Q \tag{16}$$

$$tQ = B_1 + 2B_2 \tag{17}$$

$$B_2 = \frac{tQ}{2}\frac{qB_1}{I + qB_1} \tag{18}$$

In order to make sense of these equations, fix the number of idle numbers. Solving equations (17) and (18) for B_1 and B_2 tells the proportion of busy time that drivers spend with one or two passengers. Equation (17) simply states that the total time spent with passengers by drivers has to be such that all the requested rides are completed. Equation (18) says that if the number of available drivers with one rider qB_1 is large compared with the total number of available drivers $I + qB_1$, then the balance tilts towards more rides being served by ride shares.

Solving this system of equations, given values of q, L, and T, results in a solution with the following properties:

Proposition 6. *There is at most one solution to equations* (16)-(18) *with positive* Q. *Let the solution with highest* Q *be* $S(T, L)$. *This solution satisfies the following properties*

1. *There exists some pickup time* $\check{T}(L)$ *such that* $S(T, L) < 0$ *for* $T < \check{T}(L)$ *and* $S(T, L) > 0$ *for* $T > \check{T}(L)$.

2. $\lim_{t\to\infty} S(T, L) = 0$ *for all* T.

Proof. See Appendix B.7. □

We did not prove that supply has the simple form it has for a simple ride-hailing system, where it is single peaked. However, we did show that it is initially increasing and at the end it is decreasing. It might have more than one local maximum, but the main behavior from WGCs is still present. Thus, most of our analysis still carries through. In particular, all our results about WGCs still hold whenever supply and demand cross at a point where supply is decreasing.

The intuition for this result is very similar to a traditional ride hailing service. When demand is high, the relevant density of drivers for pickup time $I + qB_1$ is very low. The reason is that a

higher demand leaves only a low number of idle drivers, and forces most busy drivers to serve two customers. This low $I + qB_1$ leads to high pickup times. And just as before, this is wasteful because drivers have to spend a lot of time picking up passengers, either while having no rider, or while driving the first rider. And this ends up reducing the capacity of the market.

8 Conclusion

In this paper we analyze the motivations behind surge pricing in ride-hailing apps. We find that it is an essential part of their success, since otherwise they would not be able to charge the low prices at low demand times that have made them a more desirable alternative to traditional taxis. The main reason is that, despite a better matching technology that can potentially increase the efficiency of the market, they are prone to a catastrophic matching failure which we call *wild goose chases* (WGC), in which high demand depletes the streets of idle drivers. This forces matches between drivers and passengers that are very far from each other, and drivers end up spending most of their time on a futile search for passengers far away from them instead of taking passengers to their destination. This also reduces their earnings, reducing the number of drivers working and amplifying the problem. Dynamically varying pricing allows platforms to avoid WGCs at times of high demand by increasing prices, while lowering prices at times of low demand. Without surge pricing, the only alternative would be to set high prices at all times to avoid the sharp decrease in welfare and revenue that would arise from WGCs at times of high demand.

Our main model is that of a market that clears through pickup times instead of prices. Whereas demand is a traditional decreasing curve, supply of trips is initially increasing but then it bends backwards. This behavior arises because of two opposing effects: higher pickup times require less idle drivers, thus freeing up more of them to take passengers to their destination, but it also takes up more of their busy time, reducing their ability to serve riders. WGCs occur when the market equilibrium is in the backward bending part of supply. We show that the market is very sensitive to subtle price changes when that is the case, and price decreases lead quickly to market collapse. On the other hand, an increase in prices decreases demand and increases supply, moving the equilibrium out of a WGC.

We corroborate empirically the validity of our results using data for Uber in Manhattan. First, we show descriptive evidence that supply does exhibit the backwards bending form that our theory predicts. We also observe a drastic degradation in performance measures when the market conditions are such that a WGC should take place according to our theoretical analysis, which is evidence of the welfare drop predicted by our model. We then also use the data to calibrate our model and quantify the effects of banning surge pricing. We find that in order to avoid WGCs the platform would have to set consistently high prices, corresponding to 97% of the price at times of high demand. These results oppose the notion that ride-hailing platforms

use surge as predatory pricing, and that without it they would always set low prices around their current base prices to the benefit of passengers.

We discuss a variety of measures that ride hailing platforms can take to in order to avoid WGCs or mitigate their effects. For a variety of practical and public relations reasons, we conclude that surge pricing is the best option. We also analyze the behavior for pooling products like UberPool and Lyft Line, and we also find a backwards bending supply curve, which means that all of our main theoretical results are still valid.

In future versions, we hope to make more detailed use of the data we obtained. We want to analyze additional performance measures like ETAs, and we want to make a more detailed breakdown of the times when WGCs take place. We will also quantify the importance of surge pricing by computing the welfare decrease if dynamic pricing is not allowed. We would like to make two separate computations. First, how valuable is surge pricing when responding to predictable market characteristics, such as rush hour or the day of the week. Second, how valuable it is to respond to unpredictable changes, such as rain, concerts, or random fluctuations from the steady state.

References

Iñaki Aguirre, Simon George Cowan, and John Vickers. 2010. Monopoly Price Discrimination and Demand Curvature. *American Economic Review* 100, 4 (2010), 1601–1615.

Mohammad Akbarpour, Shengwu Li, and Shayan Oveis Gharan. 2016. Thickness and information in dynamic matching markets. *Working Paper* (2016).

Joshua D. Angrist and Sydnee Caldwell. 2017. Uber vs. Taxi: A Driver's Eye View. *Working Paper* (2017).

Richard Arnott. 1996. Taxi Travel Should Be Subsidized. *Journal of Urban Economics* 40, 3 (1996), 316–333.

Richard J. Arnott and Eren Inci. 2010. The Stability of Downtown Parking and Traffic Congestion. *Journal of Urban Economics* 68, 3 (2010), 260–276.

Nicholas Buchholz. 2016. Spatial Equilibrium, Search Frictions and Efficient Regulation in the Taxi Industry. (2016). http://scholar.princeton.edu/sites/default/files/nbuchholz/files/taxi_draft.pdf.

Jeremy Bulow and Paul Klemperer. 2012. Regulated Prices, Rent-Seeking and Consumer Surplus. *Journal of Political Economy* 120, 1 (2012), 160–186.

Peter Cohen, Robert Hahn, Jonathan Hall, Steven Levitt, and Robert Metcalf. 2016. Using Big Data to Estimate Consumer Surplus: The Case of Uber. (2016). http://www.nber.org/papers/w22627.

Judd Cramer and Alan B. Krueger. 2016. Disruptive Change in the Taxi Business: The Case of Uber. *American Economic Review* 106, 5 (May 2016), 177–82. DOI:http://dx.doi.org/10.1257/aer.p20161002

Michal Fabinger and E. Glen Weyl. 2016. The Average-Marginal Relationship and Tractable Equilibrium Forms. (2016). https://ssrn.com/abstract=2194855.

Guillaume Frechette, Alessandro Lizzeri, and Tobias Salz. 2016. Frictions in a Competitive, Regulated Market: Evidence from Taxis. (2016). http://www.columbia.edu/ ts3035/websitefiles/frechette_lizzeri_salz.pdf.

Jonathan D. Hall. 2016. Pareto Improvements from Lexus Lanes: The Effects of Pricing a Portion of the Lanes on Congested Highways. (2016). http://individual.utoronto.ca/jhall/documents/PI_from_LL.pdf.

Juan Carlos Muñoz and Carlos F. Daganzo. 2002. The Bottleneck Mechanism of a Freeway Diverge. *Transportation Research Part A: Policy and Practice* 36, 6 (2002), 483–505.

William J. Reed. 2003. The Pareto Law of Incomes – an Explanation and an Extension. *Physica A* 319 (2003), 469–486.

William J. Reed and Murray Jorgensen. 2004. The Double Pareto-Lognormal Distribution – A New Parametric Model for Size Distributions. *Communicatoins in Statistics – Theory and Methods* 33, 8 (2004), 1733–1753.

Eytan Sheshinski. 1976. Price, Quality and Quantity Regulation in Monopoly Situations. *Economica* 43, 170 (1976), 127–137.

A. Michael Spence. 1975. Monopoly, Quality, and Regulation. *Bell Journal of Economics* 6, 2 (1975), 417–429.

William S. Vickrey. 1987. Marignal and Average Cost Pricing. In *The New Palgrave Dictionary of Economics*, Steven N. Durlauf and Lawrence E. Blume (Eds.). Basingstoke, UK: Palgrave Macmillan.

Alan A. Walters. 1961. The Theory and Measurement of Private and Social Cost of Highway Congestion. *Econometrica* 29, 4 (1961), 676–699.

E. Glen Weyl. 2010. A Price Theory of Multi-Sided Platforms. *American Economic Review* 100, 4 (2010), 1642–1672.

Appendix A Fit of pickup time function

See Figure 11.

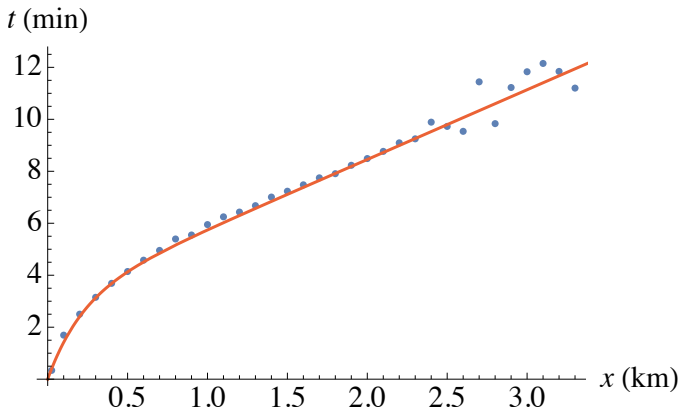


Figure 11: Average pickup time as a function of distance from matched driver, as well as a fit of the form $t(x) = a(1 - e^{-bx}) + cx$. There are very few trips with distance greater than 2.5 km, which explains the high variability in the data.

Appendix B Proofs

B.1 Proof of lemma 1

Proof. Fix L. From the original function $T(I)$, $I(T(L)) = L$. Since the denominator in equation (2) is positive for positive T and $I(T)$ is decreasing, it is clear that $S(\underline{T}(L), L) = 0$, $S(T, L) < 0$ for $0 < T < \underline{T}(L)$ and $S(T, L) > 0$ for $T > \underline{T}(L)$.

Note that $\frac{\partial S}{\partial T} = \frac{1}{d+T}\left[-I'(T) - \frac{L-I(T)}{d+T}\right] = \frac{1}{d+T}\left[-I'(T) - S(T, L)\right]$. Since S is continuously differentiable, starting at $T = \underline{T}(L)$, it is increasing in T until it reaches $-I'$ at some point $\hat{T}(L)$, at which its derivative is zero so it attains a local maximum and becomes greater than $-I'$. Thus, its derivative then becomes negative. And note that at no point greater than $\hat{T}(L)$ can $S(T, L) = -I'(T)$: that would imply $\frac{\partial S}{\partial T} = 0$, which is a contradiction since $-I'$ is decreasing so for S to cross it from above its derivative would have to be negative. So S is decreasing in T for all $T > \hat{T}(L)$.

The numerator in equation (2) is bounded by L, and the denominator goes to infinity as T goes to infinity, so $\lim_{T\to\infty} S(T, L) = 0$. □

B.2 Proof of lemma 2

Proof. This is equivalent to saying that $D(T, p)$ and $S(T, L)$ intercept at least once for all L.

Fix L. We start by showing that for any $\epsilon > 0$ there exists $\bar{T}$ such that $S(T, L) \geqslant \frac{L-\epsilon}{T}$ for all $T > \bar{T}$. Fix ϵ. Then $S(T, L) - \frac{L-\epsilon}{T} = \frac{L-I(T)}{d+T} - \frac{L-\epsilon}{T} = -\frac{I(T)}{d+T} - \frac{Ld}{T(d+T)} + \frac{\epsilon}{d+T}$. The first two term decay

faster than the third, which is the only positive term, so for all $T > \bar{T}$ the third term dominates and $S(T, L) - \frac{L-\epsilon}{T} \geqslant 0$.

This implies that for fixed L $S(T, L) = O(T^{-1})$ as $T \to \infty$. This would be the same behavior of demand if the upper tail of willingness to pay was as in a Pareto distribution with $\alpha = 1$, so the fact that the tail is thinner implies that $D(T, p) < S(T, L)$ for sufficiently high T. Additionally, note that demand is an increasing function, whereas supply is increasing for $T < \bar{T}(L)$ but decreasing for $T > \bar{T}(L)$, and it is zero for $T = T(L)$. By the mean value theorem, $D(T, p)$ and $S(T, L)$ intercept at least once for all L.

Given that $D(T, p)$ is a decreasing function, the solution with the highest Q is also the one with the lowest T. Since $D > S$ for low enough T, at the highest solution $\frac{\partial D}{\partial T} < \frac{\partial S}{\partial T}$. From the implicit function theorem and the chain rule, $\frac{dQ}{dL} = \frac{D_T S_L}{D_T - S_T}$, which is positive and $S(T, L)$ is increasing in L. It might also be the case that a change in L leads to a new solution with a higher value of Q due to two new intersections of D and S at a lower value of T. This would lead to a discontinuity in $\hat{Q}$, which would be a positive jump. Note that an increase in L never leads to supply and demand no longer crossing at the previous highest solution, as it would mean that $S > D$ for all T above the current solution, which contradicts $D(T(L)) = 0$. □

B.3 Proof of proposition 1

Proof. We find the comparative statics of equilibria from the implicit function theorem. The total differential of equations (3) and (4) looks as follows:

$$\begin{pmatrix} 1 & -\frac{\epsilon_T^D \epsilon_L^S}{\epsilon_T^D + \sigma\epsilon_T^S} \\ -\epsilon_L & 1 \end{pmatrix} \begin{pmatrix} d\log Q \\ d\log L \end{pmatrix} = \begin{pmatrix} -\frac{\sigma\epsilon_T^S \epsilon_p^D}{\epsilon_T^D + \sigma\epsilon_T^S} \\ \epsilon_L \end{pmatrix} d\log p, \tag{19}$$

Some simple algebra shows that this is equivalent to the first two expression in the proposition. To get the third expression, substitute in the total differential of $D(T, p)$, $\epsilon_T^D d\log T = d\log Q - \epsilon_p^D d\log p$.

In a WGC ϵ_T^S is negative. Furthermore, the highest solution of (3) (and any stable solution) has $\epsilon_T^S > \epsilon_T^D$. This implies that all three numerators are positive, so the sign of the elasticities of Q and L is the sign of the denominator, whereas the elasticity of T has the opposite sign.

For any stable solution $\hat{Q}'\hat{L}' < 1$. From the total differential, $\hat{Q}'\hat{L}' = -\epsilon_L \frac{\sigma\epsilon_T^D \epsilon_L^S}{\epsilon_T^D - \sigma\epsilon_T^S}$, and Δ is positive whenever this is less than one. So in the highest solution, which is stable, the determinant is positive and both the number of drivers and trips increases with prices. □

B.4 Proof of proposition 2

Proof. Starting from price p, an increase in prices also increases L as long as the equilibrium is still in a WGC, which shifts the supply curve upwards. This also shifts the demand curve

downwards. This can continue until either (a) the equilibrium is no longer a WGC, or (b) $D(0,p) \leqslant \max_T S(T,L)$. But once $D(0,p) \leqslant \max_T S(T,L)$ it has to be the case that supply and demand cross in the good region, which proves that the equilibrium is eventually in the good region. □

B.5 Proof of proposition 3

Proof. The total differential of welfare is $dW = U_T + U_Q dQ - C'dL$. Note first that $U_Q = p + \tilde{u}_Q$, where $\tilde{u}$ is the derivative of the utility of inframarginal passengers. Also $C' = p(1-\tau)\frac{Q}{L}$. Plugging in these expressions and the expressions for the elasticities of Q and L yields equation (10) after a few algebra steps.

In order to see that this is positive, note that in a WGC $\epsilon_T^S < 0$. Also $\epsilon_L \in [0,1]$, so $(1-(1-\tau)\epsilon_L) > 0$. This means that the first term in parentheses is positive. For the second term, $\epsilon_T^S > 1$ since the matching technology has increasing returns to scale, so $\frac{\epsilon_T^D \epsilon_L^S}{\epsilon_T^D + \sigma\epsilon_T^S} - (1-\tau) \geqslant \epsilon_L^S - (1-\tau) > 0$. Thus, the sum of the second and third terms is also positive in a WGC. We also showed that $\frac{dT}{dp}$ is negative, and U_T is negative, which means that welfare always increases with price increases in WGCs.

For revenue, note that $\frac{dR}{dp} = \tau\left(Q + p\frac{\partial Q}{\partial p}\right)$. This is positive whenever $\frac{\partial Q}{\partial p} > 0$, which is the case in a WGC by Proposition 1. For drivers' surplus, note that given drivers' equilibrium equation $C'(L) = (1-\tau)pQ$, we can write $RS = LC'(L) - C(L)$. Differentiating this with respect to prices yields $\frac{dDS}{dp} = LC''(L)\frac{dL}{dp}$, which is positive in a WGC.

For riders' surplus, $\frac{dDS}{dp} = U_T\frac{dT}{dp} - (U_Q - p)\frac{dQ}{dp} - Q$. The term in the middle cancels out, which yields the expression in the proposition. □

B.6 Proof of Proposition 4

Proof. Welfare maximization can be written as $\max U(Q, T(Q,L)) - C(L)$. The first order conditions are $U_Q + U_T T_Q = 0$ and $U_T T_L - C'(L) = 0$. Noting that $U_Q = p$ and $C'(L) = p'\frac{Q}{L}$ and substituting elasticities for derivatives yields the desired expressions.

For revenue maximization, we want to solve $\max Q(p(Q,L) - p'(Q,L))$. The first order conditions are then $p - p' + Qp_Q = 0$ and $p_L = p'_L$. Note from $dQ = Q_p dp + Q_T dT$ that $\frac{dp}{dQ} = \frac{1}{Q_p} - \tilde{u}_T T_Q$ and $\frac{dp}{dL} = -\tilde{u}_T T_L$. Also note, from the total differential of $LC'(L) = p'Q$, that $\frac{dp'}{dQ} = -\frac{p'}{Q}$ and $\frac{dp}{dL} = \left(1 + \frac{1}{\epsilon_l}\right)\frac{p'}{L}$. Substituting these expressions in the FOCs and some algebra steps yield the desired expressions. □

B.7 Proof of proposition 6

Proof. First, let $f(x) = T^{-1}(x)$ (which we can no longer call I because the effective density is $I + qB_1$). Note that $I = f(T) - qB_1$. Some algebra from (16)-(18) leads to the following solution

for Q from a quadratic equation:

$$Q = \frac{qL - \frac{t+T}{t}f(T) \pm \sqrt{(1-2q)f(T)^2 + L^2q^2 + \frac{2f(T)}{t}((1-2q)f(T) + qL)T + \frac{T^2}{t^2}f(T)^2}}{q(t+2T)} \tag{20}$$

The highest solution is evidently the one with the positive sign for the square root. Let this solution be $S(T, L)$.

Note that, as $T \to 0$, $f(T) \to \infty$ and $Tf(T) \to 0$. This implies that $\frac{1}{S(T,L)} \frac{-f(T)+\sqrt{(1-2q)f(T)^2}}{q(t+2T)} \to 1$, which means that $S(T, L)$ is negative for low enough T.

Not also that, as $T \to \infty$, $f(T) \to 0$ and $Tf(T) \to \infty$. This implies that $\frac{1}{S(T,L)} \frac{2L}{t+2T} \to 1$, which means that $S(T, L)$ is positive for high enough T, and this proves the second part of the proposition.

Continuity of $S(T, L)$ implies that there is at least one T such that $S(T, L) = 0$. By plugging in the system of equations, it is evident that it takes place at $I = L$, $B_1 = B_2 = 0$, which implies $T = T(L)$. This means that there can only be one solution with zero Q, and this implies the first part of the proposition. And since there is only one root, it cannot be that the lower solution ever becomes positive, which proves that there is at most one positive solution.

□