Surge Pricing Solves the Wild Goose Chase*

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

March 2018

Abstract

Ride-hailing apps usually match more efficiently than taxis, but they can enter a failure mode anticipated by Arnott (1996) that we call *wild goose chases*. High demand depletes the platform of idle drivers, so cars must be sent to pick up distant customers. Time wasted on pick-ups decreases drivers' earnings, leading to exit and exacerbating the problem. Raising prices, either by keeping them consistently high or "surge" pricing only at high demand times, brings demand back under control and avoids these catastrophic failures. Banning surge pricing would thus likely result in always-high prices. Alternative solutions would undermine ride-hailing's brand promise.

Keywords: wild goose chases, ride-hailing, surge pricing, dynamic pricing, hypercongestion

JEL classifications: D42, D45, D47, L91, R41

*We appreciate the helpful comments of Mohammad Akbarpour, Susan Athey, Eduardo Azevedo, Dirk Bergemann, Timothy Bresnahan, Liran Einav, Matthew Gentzkow, Ramesh Johari, Brad Larsen, Jonathan Levin, Paul Milgrom, Andy Skrzypacz, and Rory Sutherland. We would also like to thank seminar participants at Microsoft Research New York City, Stanford University, 2017 Uber Economics Seminar, 2017 ASSA, 2017 NBER Urban Economics and IT and Digitization Summer Meetings, 2017 NBER Market Design Meeting, and 2017 Columbia/Duke/MIT/Northwestern IO Theory Conference, especially our discussants Richard Arnott, Chiara Farronato, and Cristopher Snyder. We are grateful to the Alfred P. Sloan Foundation for their financial support through Weyl's fellowship.

†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 promising new technologies that improve upon traditional taxis. Cramer and Krueger (2016) show that the fraction of working time that a driver 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 that, unlike traditional street-hail taxi systems, ride-hailing platforms are prone to a matching failure first anticipated by Arnott (1996). We argue that surge pricing is crucial to addressing this failure while allowing low prices at most times.

When drivers are scarce relative to demand, drivers are quickly occupied and thus idle drivers are spread thinly throughout a geographic region, forcing matches between drivers and passengers that are on average 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 decreases 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 with a steep decline in welfare. Street-hail taxis, on the other hand, do not fall into WGCs. Taxis and passengers are next to each other when they are matched, so drivers never waste time picking up passengers.[1]

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 relative to supply all equilibria of the market are WGCs under the commonly-used "first-dispatch" protocol in which every rider requesting a ride is immediately matched, if possible, to the nearest idle driver. This undesirable outcome can be avoided by increasing

[1]Radio taxis do need to spend time picking up passengers after being matched. They are thus prone to WGCs, which we believe are frequent, as anyone that has tried to call a taxi during rush hour on a rainy day can testify.

prices, which suggests two ways to avoid WGCs.[2]

First, one might set a sufficiently high price all the time 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 market design is to use dynamic "surge pricing" that responds to market conditions. Prices are set high during peak-loads, but 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. This suggests that, despite the framing as "surge prices" to make the platform attractive to drivers, in fact dynamic pricing is more similar to a discount. This helps explain why these platforms have relied so heavily on surge pricing, which allows them to undercut taxi markets at most times while still having the flexibility to increase prices and avoid WGCs.

Our analysis starts with a highly-simplified static/steady-state theoretical model in a homogeneous spatial region that is nonetheless able to capture the key features WGCs with sufficient precision to fit the most striking features of the data we consider below. Rider demand depends on waiting time and price, drivers' labor supply is driven by per-unit-time earnings and matching is by first-dispatch. The characteristic feature that gives rise to WGCs is that the supply of trips given a fixed number of drivers is a non-monotone 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 rise, drivers spend more of their time picking up passengers. WGCs occur when pickup times are high and the latter effect dominates.

In that case reducing demand (even without increasing price) benefits drivers, as it mitigates strain on the system, allowing them to complete more rides and eventually eliminating the WGC. This stands in contrast to standard price theory, where a fall in demand harms suppliers. These unusual results resemble many features of "hypercongestion", a related

[2]Changing dispatch protocols may also help, as we discuss below.

phenonomenon in transportation economics (Walters, 1961; Vickrey, 1987): when too many cars enter a road, speed decreases to an extent that the total throughput of the road falls. Furthermore, we show that as in the literature on hypercongestion, WGCs occur "sharply". In particular, when the ratio of idle drivers to those on the way to pick up a customer, what we call "slack", falls below a critical threshold (between $\frac{1}{4}$ and $\frac{1}{2}$ depending on traffic density, the number of one-way roads and other traffic flow conditions).

Sharp predictions raise the chances of extracting a signal of theoretical validity from inevitably noisy data. This motivates our empirical analysis of Uber trips in Manhattan between December 2016 and May 2017. As predicted, trip supply exhibits the nonmonotonicity in pickup times. Even more striking, Figure 1 bears out our prediction of sharp decay in the market. It shows that the rider cancellation rate, a simple measure of market performance, increases drastically as the market enters a WGC. It also shows that these are precisely the times at which Uber sets high surge multipliers, suggesting that it uses surge pricing as a tool to avoid WGCs and that slack closely tracks whatever proprietary metrics it uses to set surge prices. We obtain very similar results with other market performance measures.

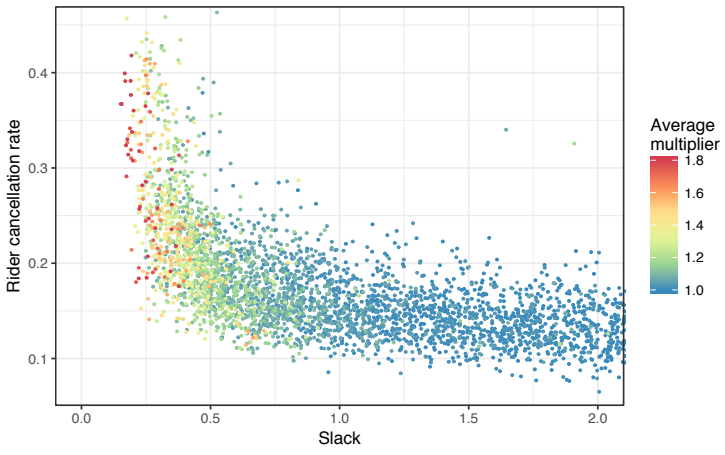


Figure 1: Fraction of trips cancelled by riders as a function of slack (the number of idle drivers divided by the number of drivers on their way to pick up a rider).

Having thus validated the rough empirical relevance of the model, we then calibrate it to these data in order to analyze the welfare effects of surge pricing. Unsurprisingly given our theoretical results, 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, regardless of whether it maximizes revenue or welfare, 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 fixed price for all markets. The only way to avoid the drastic loss in welfare from WGCs without surge pricing is to set the fixed price close to the highest prices under surge pricing. The optimal fixed price is at the 92nd percentile of the price distribution if it is allowed to set different prices for each hour of the week. Thus, surge pricing only leads to very modest increases in prices at times of high demand, whereas it allows drastic reductions at low demand times. This belies the perception among the public and some regulators that surge pricing is a form of price gouging.[3]

Pricing is not the only tool ride-hailing apps can use to avoid WGCs. One appealing option is to match drivers and riders only when they are closer than a certain *maximum dispatch radius* (MDR), avoiding especially wasteful trips. We show that, starting from a WGC, decreasing the MDR increases welfare and eventually avoids the WGC. However, strong reliance on MDRs requires frequently denying trips to requesting riders, which contrasts with the "brand promise" of ride-hailing apps to almost always match a rider with a car near-immediately so the rider can plan for the

[3]For instance, the splash page on competitor Gett's home page on June 27, 2017 stated "The only time we surge is never o'clock", and the government of Manila banned surge pricing during early 2017.

driver's arrival.

More interestingly, therefore, we show that setting a fixed, relatively high MDR mitigates the welfare losses of WGCs significantly, while denying no trips at times when the market performs well. This helps explain Uber's policy to set a high MDR that serves as a circuit breaker, avoiding the worst market outcomes, instead of using it to actively manage supply and demand imbalances. We briefly discuss alternative approaches, such as randomly denying trips and creating passenger queues, that can solve WGCs but strongly conflict with ride-hailing's brand promise.

Our analysis begins in section 2 with our theoretical model that has elements that are similar to Arnott's. In Section 3 we describe how WGCs arise and how increasing prices avoids WGCs. In Section 4 we show empirical evidence of WGCs. Then in Section 5 we calibrate our model to our data and analyze the effects of a ban on surge pricing. In section 6 we discuss alternative solutions to WGCs.

Related Work

Our paper is part of a large and growing literature on the economics and spatial dynamics of ride-hailing. A pioneering paper that set the stage for much of this work was Cramer and Krueger (2016), who find that Uber is substantially more efficient than traditional taxis. They attribute the improvement to better matching, returns to scale, less regulation, and more flexible labor supply.

The crucial distinguishing feature of our analysis is our focus on WGCs, which we argue can undermine these efficiency gains dramatically, and on the simplest model sufficient to quantitatively capture and analyze this phenomenon and potential solutions. To our knowledge, the only precedent for analyzing this feature is Arnott (1996), which we repeatedly discuss. Our model is very similar to his, though somewhat more general and most importantly, we analyze WGCs in detail while he only note the possibility as being Pareto-dominated and thus not part of the optimal

solution. Other papers have analyzed taxi markets, both theoretically and empirically (Lagos, 2003; Frechette et al., 2016; Buchholz, 2017), and in some cases they also note but do not analyze in detail the possibility of phenomena similar to WGCs, though most focus on street-hailing where this issue does not arise.

The next most closely related set of papers are those that analyze other aspects of surge pricing, such as the welfare analysis in Cachon et al. (2017). Other papers about ride hailing consider profit-maximizing pricing (Afeche et al., 2017; Bimpikis et al., 2016), incentive compatibility issues for drivers (Ma et al., 2018), queuing without explicit spatial dynamics (Galichon and Hsieh, 2016; Banerjee et al., 2015), economics of density (Shapiro, 2017), and complementarities and substitutabilities with public transportation (Hall et al., 2017a) . More distantly related is work on driver labor supply (Chen and Sheldon, 2015; Hall and Krueger, 2016; Hall et al., 2017b; Chen et al., 2017; Angrist and Caldwell, 2017; Cook et al., 2018) and rider demand (Cohen et al., 2016; Lam and Liu, 2017), which we use to calibrate our model.

We also draw on the literature on hypercongestion in road traffic flow (Walters, 1961; Muñoz and Daganzo, 2002; Hall, 2018), which WGCs resemble and which we draw on the analogy to, and the literature on monopolies in platform markets (Rochet and Tirole, 2003; Armstrong, 2006; Weyl, 2010), which in form our analysis of optimal pricing.

2 Model

We consider a static, steady-state model of a ride-hailing service in homogeneous space.

2.1 Trip demand and labor supply

Let demand be $D(T, p)$, the number of people who request a trip given expected pickup time T and price p.[4] It is measured in trip requests per unit area and per unit time (which we henceforth without loss of generality refer to as "hourly"). We make the following assumption:

Assumption 1. *$D(T, p)$ is bounded and is differentiable and decreasing in both arguments. For all $T, p \geqslant 0$ respectively, $\lim_{T\to\infty} D(T, p) = 0$ and $\lim_{p\to\infty} D(T, p) = 0$. For all p, the distribution of the willingness-to-wait has finite mean.*

This states that there is a finite pool of potential riders, both higher prices and higher waiting times are undesirable, nobody is willing to pay or wait infinitely and, more strongly, that the distribution of willingness-to-wait is not too fat tailed.

Drivers decide whether to work based on expected hourly earnings e, which results in a supply of drivers $L = l(e)$, measured in drivers per unit area, and which we assume is differentiable and increases from 0 at 0 (no drivers are willing to drive for free). Let Q be the equilibrium number of hourly trips (per unit area), and τ be the fraction charged by the platform. Earnings are then given by $e = (1-\tau)p\frac{Q}{L}$.

2.2 Trip supply given labor supply

Demand measures the number of trips requested, whereas labor supply is the number of drivers working. We thus need a trip supply function that gives the number of trips that can be served by L drivers. We assume that the platform uses a first-dispatch protocol. That is, whenever a rider requests a trip, he is matched immediately to the closest idle driver. We consider departures from this assumption in section 6.

[4] To be precise, demand depends on *realized* pickup time. Consider a primitive demand function $\tilde{D}(\tilde{T}, p)$ that depends on realized pickup time $\tilde{T}$. The distribution of realized pickup time has distribution $F(\cdot; T)$, which, as we will later show, can be parameterized by average pickup time. Thus, demand is $D(T, p) = \int \tilde{D}(\tilde{T}, p)dF(t; T)$.

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 thus has to be equal to the sum of drivers in each one of these states. Let I be the number of idle drivers per unit area. In steady state, tQ drivers are driving a rider, where t is the average trip duration. This is the number of hourly trips times the average time (in hours) it takes to complete a trip. By a similar reasoning, TQ drivers are on their way to pick up a rider in equilibrium, since T is the time it takes on average for a driver to pick up a passenger. This last expression captures a defining feature of ride-hailing markets: pickup times affect both riders and drivers. Whenever a rider waits a long time to be picked up, the driver spends that same amount of time *en route*.

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

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

The average pickup time $T(I)$, which captures the matching technology, is a decreasing function of the number 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. We assume that T continuously differentiably and convexly decreases from ∞ when there are no idle drivers to 0 if there are infinite idle drivers. Convexity means diminishing marginal returns of 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.[5]

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

[5]For instance, if *en route* drivers drive in a straight line at a constant speed in n-dimensional space, $T(I) \propto I^{-\frac{1}{n}}$, which satisfies all these properties.

T. It is easily seen that $I(T)$ has the same shape properties as $T(I)$.

Isolating Q from 1 and substituting $I(T)$ 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). The denominator is the average "trip time" (busy time it takes to complete a trip), the sum of pick-up time and drop-off time. Dividing the number of busy drivers by the trip time yields the number of hourly trips that can be completed. This functional form has increasing returns to scale since $S(T, bL) > bS(T, L)$ for $b > 1$.

3 Wild Goose Chases

3.1 Good and wild goose chase matching equilibria

In our analysis we consider equilibrium *holding the price fixed*, since price is a parameter that can be chosen by the platform. The clearing variable, instead, is pickup time, 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 (Figure 2). Following this idea for our analysis, trip supply $S(T, L)$ can be thought of as a continuous function of T where L is a supply shifter. Its behavior is summarized in the following lemma:

Proposition 1. *Fix labor supply* L. *No pickup time below* $T(L)$ *is achievable. Trip supply is zero at* $T(L)$*, then increases in pickup time until it achieves its maximum, and then decreases until zero asymptotically.*

A supply function that satisfies this proposition is illustrated in Figure 2. The most remarkable feature is the nonmonotonicity of trip supply, which is the driver of our main results. The intuition for this unusual characteristic of $S(T, L) = \frac{L - I(T)}{t + T}$ is as follows. The numerator is increasing. If the platform

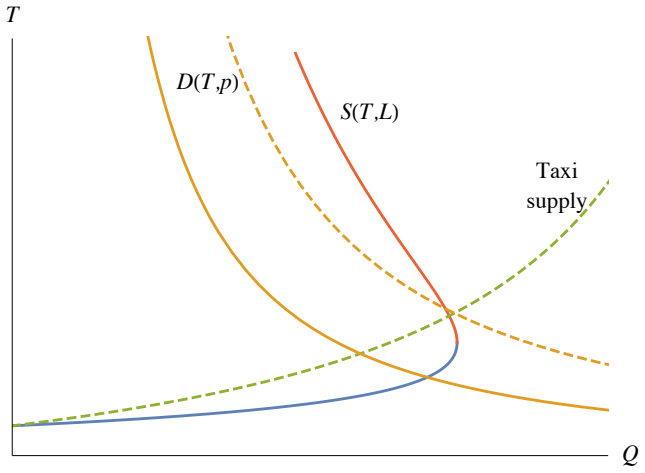


Figure 2: Trip supply and demand with a fixed number or drivers. Supply bends backwards in the ride-hailing market. Taxi supply is shown in green.

wants to achieve a lower pickup time, it needs more idle drivers, which decreases the number of busy drivers and reduces the potential capacity of the market. This effect is the main driving force in the blue region.

As pickup times increase and supply approaches its maximum, a second effect starts to kick in. The denominator grows as drivers spend a significant portion of their time picking up passengers that are far away. So, despite the fact that higher pickup times require less idle drivers thus freeing some of them to drive passengers, the total time it takes to complete a trip becomes longer, and eventually decreases. This second effect dominates in the red region.

Nonmonotonicity means that there are two ways to supply the same number of trips for a fixed number of drivers. We call the one below the maximum of supply a *good outcome* (in the blue region of the supply curve). The one above the maximum, plotted in red, has a longer pickup time. The latter situation is evidently inefficient, as it achieves the same number of trips with the same number of drivers, but with higher pickup times. We call these situations *wild goose chases* (WGCs). 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.

Our supply curve bears some similarity to a Laffer curve in tax theory or the revenue curve in monopoly theory. If the platform has to choose an equilibrium at some point along the supply curve, it faces a tradeoff between number of trips and 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. Let *slack* be $s = \frac{I}{TQ}$, the ratio of idle drivers to en route drivers, an adimensional measure of the availability of drivers. Also let ϵ_I^T be the elasticity of pickup time with respect to the density of idle drivers.WGCs take place when the derivative of trip supply is negative, which, after some simple algebra steps, can be rewritten as $s < -\epsilon_I^T$.[6] 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, $I(T)$ can be easily fit to the data closely, and in Section 5.1 we obtain an estimate of ϵ_I^T that is somewhere between -0.25 and -0.5.

Remarkably, WGCs do not happen with street-hail taxis, which can only be hailed by standing on the street and not by calling them.[7] There is no pickup time, since whenever a taxi is "matched" to a rider the trip starts immediately. The driver identity is then $L = I + tQ$; pickup times do not affect supply directly, since drivers spend no time picking up. There is a decreasing function $\tilde{T}(I)$ that tells how long passengers must wait before an idle taxi drives by: the more idle drivers on the street the less time a rider should expect to wait. Isolating Q leads to an increasing supply function $\tilde{S}(T, L) = \frac{L - \tilde{I}(T)}{t}$ (Figure 2).[8] The intuition is the same as in good outcomes

[6] After rewriting trip supply in terms of $T(I)$, the condition on the derivative can be written as $I < -\epsilon_I^T T(I)Q$.

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

[8] Because of the greater efficiency of ride-hailing, $\tilde{T} > T$; we calibrate this efficiency gain precisely in our numeric section. In Figure 2 we represent a magnitude that roughly matches a gap between these similar to the one found by Cramer and Krueger (2016).

for ride-hailing markets: decreasing pickup times requires more idle taxis, 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:

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

In Figure 2, this condition means that supply and demand cross in equilibrium. 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.

The second 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). The following proposition guarantees the existence of stable equilibria:

Proposition 2. *A stable equilibrium with* $(Q, L) \geqslant 0$ *always exists, possibly with* $Q = L = 0$*. If there are multiple equilibria, ordering them by* Q *or* L *is equivalent. The highest equilibrium is always stable.*

The idea behind the proof of this proposition (with full details in appendix A.2) is that the highest solution to equation 3 yields Q as a function of L, whereas equation (4) gives L as a function of Q. They can thus be thought of as loci in the (Q, L) plane with equilibria at their intersections. Our formal definition of stable equilibria, which is in the appendix, re-

quires further notation, but, informally, they are equilibria such that iterated responses from small perturbations lead back to the equilibrium.

It is not difficult to construct reasonable functional forms that lead to multiple equilibria, but none of our calibrations have this feature and for the purposes of our theoretical analysis, we select the highest and thus best equilibrium.

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 τpQ.

We need some additional assumptions to define welfare, riders' surplus, and drivers' surplus. First, let the 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 earnings minus cost:

$$DS(Q, L, p) = (1 - \tau)pQ - C(L) \tag{5}$$

In order to define welfare and consumer surplus, let $U(Q, T)$ be gross utility. It satisfies the following assumptions:

Assumption 2. *Gross utility* $U(Q, T)$ *is continuously differentiable in* (Q, T)*. It is decreasing in* T *and increasing in* Q *with* $U_Q(Q, T) = p$*.*

Gross utility is decreasing in T through two mechanisms. First, if the same people are served with lower pickup times, their utility increases. Additionally, an equilibrium with lower waiting times requires a higher price, so trips are reallocated to customers with higher willingness to pay. A change in Q in equilibrium with T fixed can only happen through a change in prices, and every rider that now decides not to take a trip is a marginal rider whose utility from a trip is p. Thus, $U_Q = p$.

We define riders' surplus as

$$RS(Q, T) = U(Q, T) - pQ \tag{6}$$

Finally, welfare is gross utility minus social cost:[9]

$$W(Q, L, T) = U(Q, T) - C(L) \tag{7}$$

3.4 Pricing and Wild Goose Chases

We will now analyze how pricing affects the equilibrium in this market. For now we hold the platform's take rate τ fixed as has historically been the case. Deviations from this policy are in theory desirable, as we show in section 3.5, but have only been introduced recently because of regulatory concerns.

We first look at the local effect of pricing under WGCs:

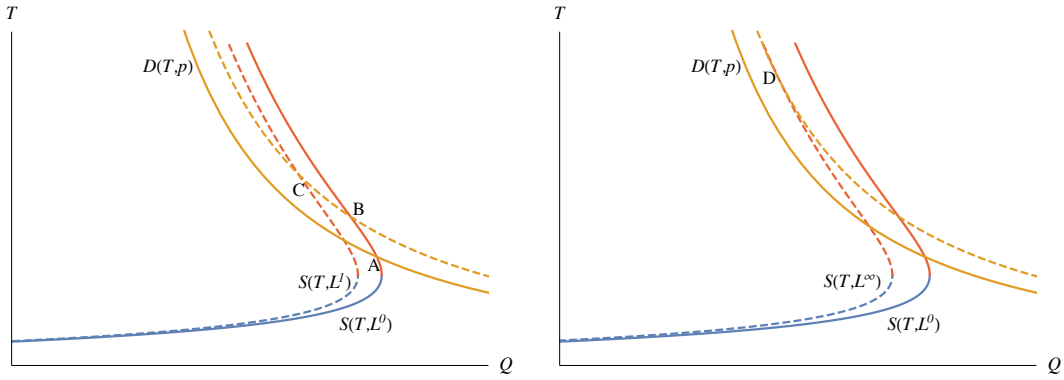
Proposition 3. *Starting from a WGC equilibrium, a price decrease increases pickup times while decreasing the number of trips, the number of drivers working, drivers' surplus, revenue, and welfare. The effect on riders' surplus is ambiguous.*

This proposition highlights that it is evidently good to increase prices whenever the market is in a WGCs, although it is not quite a Pareto improvement, since it might be the case that riders' surplus decrease.

This result might not be too surprising when comparing it with a more traditional market. If prices start from a very low base, one might expect a further decrease to have the effects in this proposition. The key feature of WGCs for our purposes here is that they can be diagnosed and lead to sharper effects than in other markets because of a feedback cycle, as we will now argue.

Consider a WGC equilibrium like point A in figure 3a with labor supply L^0, and suppose there is a small price *decrease*. Consider first the effect with fixed supply, as would be the case with self-driving cars. Demand shifts outwards until the equilibrium is at point B, with less trips than at point A. Earnings decrease through two channels: lower prices and less trips. Now consider the first order labor response to such decrease in earnings, to a lower labor supply level L^1. This moves the equilibrium to point C,

[9] Equivalently, welfare is the sum of riders' surplus, drivers' surplus, and revenues.



(a) First- and second-order response

(b) New equilibrium

Figure 3: Response to a decrease in prices starting from a WGC equilibrium.

which is substantially to the left of point B. The reason for this is that both supply and demand are decreasing, causing a feedback cycle in which a lower number of trips and higher pickup times reinforce each other.

A second feedback cycle further reinforces these effects. The decrease in trips from B to C further reduces earnings, so labor supply further decreases to a new level L^2. By the same mechanism as before, the number of trips also goes further down, reducing earnings and further decreasing labor supply. The cycle operates until the market equilibrium settles down at a point like D, with labor supply L^∞. This iteration process converges because the equilibrium is stable. All these repeated decreases in the number of trips and increases in pickup times mean that the market eventually ends up with substantially less trips and substantially higher waiting times than at point A, despite the fact that the original change in prices was small. This whole analysis is formalized in the appendix, where we find expressions for derivatives of the number of trips, labor supply, and pickup times that highlight these feedback cycles.

The previous analysis clarifies the effect of prices on the number of trips, waiting times, and labor supply from Proposition 3. We will now explain the rest of the effects, which are proven formally in Appendix A.3. After the price decrease from figure 3a, drivers' surplus decreases, since more drivers working means a movement down the labor supply curve.

In order to analyze the effect on riders' surplus, note that the effect on marginal drivers' surplus is zero, so only the effect on inframarginal riders matters. The issue thus becomes whether the increase in pickup times is justified by the reduction in price. This crucially depends on the functional form for utility, so we cannot make any further statements without stronger assumptions. Even though the effect of prices on riders' surplus is ambiguous, total welfare unambiguously decreases. Gross utility decreases both through less trips and higher pickup times. The cost of drivers decreases by $(1-\tau)p\frac{\epsilon_l}{1+\epsilon_l}$, where ϵ_l is the elasticity of supply, but this is more than offset by the decrease in gross utility of p through less trips.[10]

The following proposition shows that prices can be used to avoid WGCs:

Proposition 4. *Suppose that the highest equilibrium of the market at price* p *is in a WGC. Then there exists some higher price* $p' > p$ *at which the highest equilibrium of the market is no longer in a WGC.*

This proposition arises from the fact that demand is bounded and goes to zero as price goes to infinity. As the whole curve shifts downwards, maximum demand eventually becomes less than maximum supply. In this case there has to be at least one good equilibrium.

Decreasing prices does not necessarily take the market into a WGC—although it always does in our calibrations. 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.[11] This might have been the case with some radio taxi markets. Given how long 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.

[10] Drivers' cost is multiplied by $\frac{\epsilon_l}{1+\epsilon_l}$ because in response to an increase in earnings, for instance, labor supply increases, thus diluting earnings among a larger number of drivers and partially offsetting the initial response.

[11] 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.

3.5 Optimal pricing

While in most of the paper we assume a fixed τ, we feel our analysis would be incomplete without an analysis of optimal pricing from a social and private perspective. To simplify our notation, let $p' = (1-\tau)p$, the price on drivers' side. Also let $\bar{u}_T = \frac{U_T}{Q}$ be the average change in utility of inframarginal users from an increase in pickup time, and $\bar{u}_T = -\frac{Q_T}{Q_p}$ be the average change in utility of *marginal* users. The following proposition gives expressions for optimal prices under welfare and revenue maximization:

Proposition 5. *Welfare maximizing prices are given by*

$$p = \bar{u}_T T \epsilon_Q^T \qquad p' = \bar{u}_T T \epsilon_L^T, \tag{8}$$

which requires a subsidy ($p > p'$). Revenue maximizing prices are given by

$$p = \frac{1}{1 - \frac{1}{\epsilon_p^D}} \tilde{u}_T T \epsilon_Q^T \qquad p' = \frac{1}{1 + \frac{1}{\epsilon_\iota}} \tilde{u}_T T \epsilon_L^T, \tag{9}$$

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) distortion: first order conditions only take into account the utility of price-marginal riders and not the surplus of the price-average riders.[12] 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.

Increasing returns to scale implies that $-\epsilon_Q^T + \epsilon_L^T > 0$, which is the reason why welfare maximization requires a subsidy. This is the main point in Arnott (1996). Increasing the market size yields greater welfare. Thus, there exists an implicit positive externality from every additional driver and passenger, so the market requires a subsidy for optimality. While analyzing

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

competition is beyond the scope of this paper, these results suggest that there are significant trade-offs in introducing competition because, while market power may distort prices upwards, increasing returns to scale make a monopolistic market structure desirable. A regulatory model in the spirit of public utilities may be more appropriate.

4 Empirical Evidence of Wild Goose Chases

4.1 Data

We use Uber data from Manhattan between December 1st, 2016 and May 14, 2017. We look at both UberX and UberPool because the set of drivers working for both products is the same. As we show in Appendix D, WGC are also a problem with UberPool, so our qualitative results for UberX also hold for UberPool. 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 our data as a single time series for the full city because spatial heterogeneity and flows across neighborhoods complicate our model.[13] We aggregate the data into half-hour periods, because this captures the steady state we analyze in our model and abstracts away from transient dynamics.

On the supply side, we observe the total number of driver-minutes spent in each one of the three states 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 trips requested and whether they were completed. For those that were not, we observe why: the driver canceled, the rider canceled, or the rider could not be matched to a driver.

Finally, we observe data on surge pricing. A surge multiplier, which multiplies the base fare to obtain the final fare, is set every two minutes in

[13] For instance, we see small areas with zero drivers and a large number of trips: a large demand spike, like a concert, was served by the idle drivers from nearby locations.

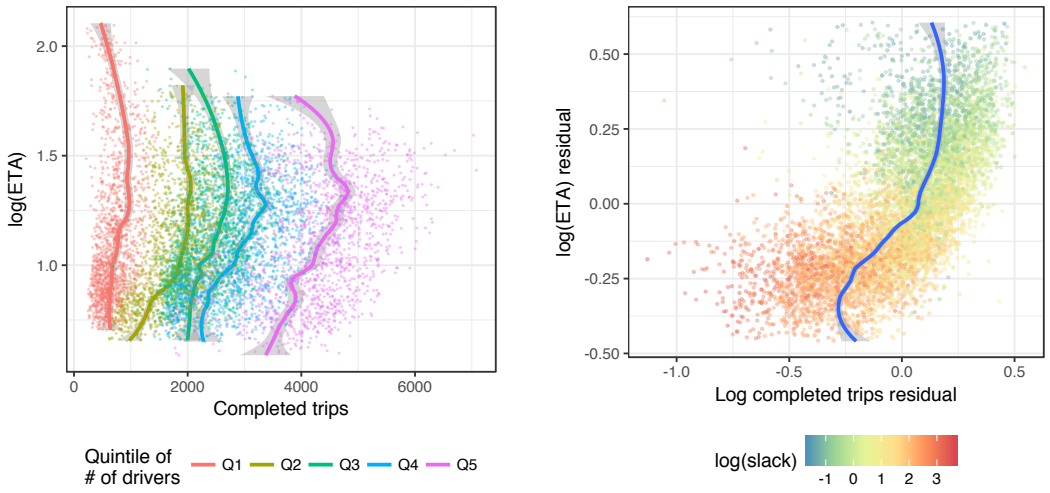
small hexagon-shaped locations that are either 600 (before January 15) or 350 meters (after Januray 15) across. We average surge multipliers over all of Manhattan in every half-hour period to match the structure of our data. All this aggregation should dull the observability of our key predictions, so the still relative sharpness of our findings in line with our theory is all the more striking.

4.2 Supply of trips

In this section we show an empirical analogue of Figure 2, which exhibits the backward bending trip supply curve that gives rise to WGCs. We first illustrate this with some descriptive statistics, and then move on to a more nuanced regression methodology.

Trip supply $S(T, L)$ is nonmonotonic given a fixed number of drivers, but data shows a lot of variation in the number of drivers. Thus we first split the sample into five quintiles of the number of drivers working, and we show the behavior of the number of trips for each quintile. The result of this exercise is Figure 4a. The horizontal axis represents the number of trips completed, and the vertical axis is the log of the ETA shown to riders as soon as they request a trip. We use ETA instead of realized pickup time because riders with a high ETA cancel a higher number of trips before being picked up, so we would end up with a biased sample. All quintiles show the backward bending relation we would expect from our theory.

There are some factors that affect trip supply beyond those we model, most importantly traffic speed and average trip length. To account for this, we plot in Figure 4b the residuals of regressions of $\log(\mathrm{ETA})$ and $\log Q$ on average traffic speed, average trip length, the number of drivers working, and the square of each one of these variables. This shows dependence of the log of pickup time on the log of the number of trips when controlling flexibly for confounders. Again we observe a backward bending curve, but now only around 5% of the observations are in the decreasing part of the curve—in a WGCs. The figure also shows that this is happens when slack



(a) By quintile

(b) Residuals for whole dataset

Figure 4: 4a plots log(ETA) against the number of trips completed. Fit lines are locally weighted quadratic regressions using the 25% of data closest to every point, using tricubic weights. Shaded regions represent 95% confidence intervals. Outliers, with ETA in the top or bottom percentile by quintile, are excluded. 4b shows residuals of regressions of log(ETA) and the log of completed trips. Outliers are also excluded.

goes roughly below 0.35, which is consistent with WGCs starting below a threshold between $\frac{1}{4}$ and $\frac{1}{2}$.

Our data show that WGCs occur roughly 5% of the time. This suggests there are times when the platform fails to set prices high enough to avoid the worst market outcomes. We show in Appendix B that WGCs take place predominantly during the morning and evening rush hours, especially on Mondays and Fridays, which are also the times with highest average surge multiplier.

To interpret the plots in Figure 4 as supply curves, the main source of variation that traces them out must be demand shifts or market conditions like traffic speed, instead of supply shifts. This concern is partially addressed by controlling for the main determinants of the supply curve. We now complement it with a regression analysis that uses demand shifts as

instrumental variables. We run regressions of the following form:

$$Q_t = \alpha T_t \times a_t + \beta T_t \times (1 - a_t) + a_t P(L_t, v_t, l_t; \gamma) + (1 - a_t) P(L_t, v_t, l_t; \delta) + \epsilon_t, \tag{10}$$

In this regression t is the time period, Q_t the number of trips completed, T_t the average ETA, L_t the number of drivers working, v_t the average trip speed, and l_t the average trip length. In each regression we run, we split the sample in two, according to whether T_t is greater or less than some threshold T^{th}. a_t is a dummy variable that is one if pickup time is above the threshold ($T_t > T^{th}$) and zero otherwise, and $P(L_t, v_t, l_t; \gamma)$ is a second order polynomial in (L_t, v_t, l_t).

Table 1: Nonmonotonicity of supply in pickup time

	Dependent variable:							
	Number of trips completed							
	$T^{th} = \infty$	$T^{th} = 8$	$T^{th} = 7$	$T^{th} = 6$	$T^{th} = 5$	$T^{th} = 4$	$T^{th} = 3$	$T^{th} = 2$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: OLS							
ETA × above		−0.798*	−0.525*	−0.560***	−0.650***	−0.589***	0.344*	3.860***
		(0.423)	(0.300)	(0.199)	(0.143)	(0.120)	(0.184)	(0.410)
ETA × below	4.099	5.208***	5.876***	6.768***	8.526***	12.051***	17.563***	27.785**
	(0.425)	(0.152)	(0.148)	(0.143)	(0.146)	(0.191)	(0.458)	(14.038)
	Panel B: 2SLS							
ETA × above		−0.972***	−0.235	−0.889	2.137	6.975***	12.744***	15.384***
		(0.072)	(0.559)	(0.677)	(2.675)	(1.695)	(1.487)	(0.560)
ETA × below	15.845***	15.764***	16.521***	18.282***	20.509***	27.061***	57.522***	−1.929
	(0.568)	(0.554)	(0.554)	(0.510)	(0.509)	(0.630)	(2.172)	(150.503)
Obs. above threshold	0	29	98	199	504	1,481	3,749	7,680
Obs. below threshold	7,870	7,841	7,772	7,671	7,366	6,389	4,121	190
Observations	7,870	7,870	7,870	7,870	7,870	7,870	7,870	7,870
Note:								Robust standard errors in parentheses: *p<0.1; **p<0.05; ***p<0.01

Panel A in Table 1 shows the OLS estimates for α and β. Each column uses a different value of T^{th}. Column (1) uses a threshold of infinity, so it reports the result of a regression on the full dataset. In the other columns we see that the coefficient below the threshold is always positive, and the coefficient above is negative for higher thresholds.

Pickup time T_t is endogenous, since it is determined in equilibrium. We instrument it with the number of people who open the app λ_t. The exclusion restriction is that market outcomes in no way influence whether people open the app or not, which seems reasonable since people only

observe prices and ETAs *after* they have opened the app. Panel B in Table 1 shows 2SLS estimates, with a pattern that is consistent with our OLS estimates. The main difference is that the coefficient of ETA above the threshold changes sign at a higher value of ETA, and our coefficients are somewhat less precisely estimated.

Our results thus support our theoretical prediction that supply is non-monotonic in pickup times. A similar exercise is to run regression where a_t is defined by whether slack is above or below some threshold. The results are shown in appendix C. They show evidence that trip supply is increasing in pickup time for high slack, but is decreasing for low slack, just as our theory would predict.

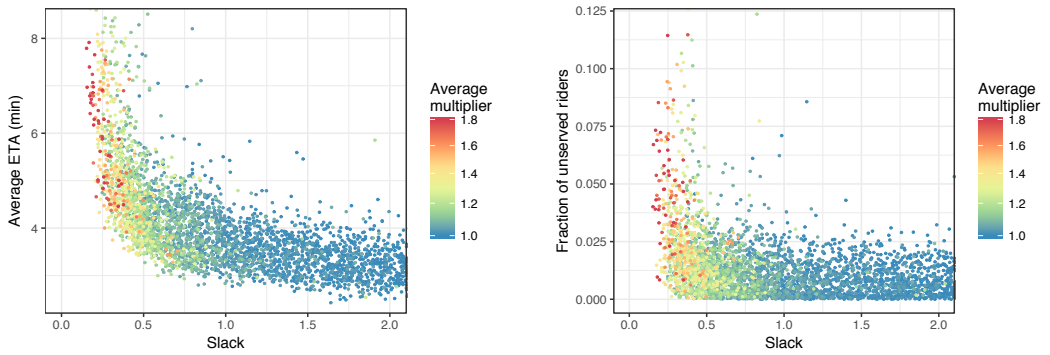
4.3 Performance measures and slack

We will now show evidence that market performance deteriorates drastically when the market reaches a WGC, as defined by slack crossing some threshold between $\frac{1}{4}$ and $\frac{1}{2}$. Slack is below $\frac{1}{2}$ for 11.5% of observations, below $\frac{3}{8}$ for 5.5% of observations, and below $\frac{1}{4}$ for 0.95% of observations.

Figure 1 in the introduction plots the fraction of requested trips that are subsequently cancelled by riders. The cancellation rate increases as slack goes down, which is not entirely surprising: lower slack means low rider availability, which harms rider experience and should cause more cancellations. What is remarkable is how a mild relation suddenly spikes as soon as slack reaches values below 0.5. Figure 5 shows similar results for two additional performance measures, average ETA and the fraction of riders that see no riders in their vicinity. Both variables have the same behavior as trip cancellations, with a sudden spike below 0.5.

The sudden deterioration is related to our analysis from Section 3.3, which shows how sensitive the market is to small price deviations that greatly reduce the number of trips and increase pickup times. We will show in Section 5.1 that it is also consistent with our calibration.

These figures show that the surge multiplier very rarely goes above 1



(a) Pickup times

(b) Unserved riders

Figure 5: Performance measures as a function of slack.

when slack is above 1. On the other hand, the surge multiplier becomes larger as slack goes down, and we see especially high multiplier values for slack below 0.5. This is exactly what one would expect if Uber reacts to slack, or at least to closely related measures. Despite this behavior, WGCs do take place, as evidenced by observations with low slack and bad performance measures. This indicates that Uber might benefit from increasing prices even more aggressively at troublesome times.

5 Surge Pricing

We now calibrate our model and apply it to quantitatively analyze pricing and the effects of allowing versus prohibiting surge pricing.

5.1 Calibration

Let λ be the arrival rate of potential riders. Let $r(p)$ be the fraction of those 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 demand is $D(T, p) = \lambda g(T) r(p)$. This assumption would be violated if both decisions are strongly correlated, but it is not entirely clear

whether the correlation is positive or negative.[14]

As a simplifying assumption, we assume that utility does not depend directly on pickup time, although it does depend indirectly through the number of trips requested. Taking into account the disutility of waiting would only make the effects of WGCs stronger, since high pickup times hurt consumers even more. We therefore take the most conservative approach. Gross utility is gross utility per rider willing to wait times the number of riders willing to wait, $U(p,T) = \lambda g(T)\left[\int_p^\infty r(p')dp' + pr(p)\right]$.

We assume that willingness to pay has a double Pareto lognormal distribution (Reed and Jorgensen, 2004) with parameters $\alpha = 3$, $\beta = 1.43$, $\mu = 1.1$, and $\sigma = 0.45$. We choose parameters so that the distribution has the same shape as the US income distribution (Fabinger and Weyl, 2016). We choose the remaining parameter μ, a rescaling of the distribution, to fit the elasticities in Cohen et al. (2016). The form of $r(p)$ arises from this distribution, where p is the surge multiplier. We assume that willingness to wait has a lognormal distribution with mode 5 minutes and variance such that the elasticity of $g(T)$ agrees with the value from Cohen et al. (2016).

For labor supply, we assume a constant elasticity functional form, $l(e) = A\left(\frac{e}{1+\frac{1}{\epsilon_l}}\right)^{\epsilon_l}$. We assume an elasticity of 1.2 based on Angrist and Caldwell (2017). 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.

We fit the matching technology, described by $T(I)$, using data on average pickup time as a function of distance to the matched driver, which we denote by $\check{T}(x)$. We take a function of the form $\check{T}(x) = a(1 - e^{-bx}) + cx$. The first term captures the fact that cities' (e.g. one-way) 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

[14]For instance, a businessman that is late for a meeting and is willing to pay a lot suggests a negative correlation, whereas an elderly person that cannot drive but is in no rush suggests a positive correlation.

trips in Manhattan obtained from Uber, as shown in Appendix B.

Once we fit $\check{T}(x)$, we obtain an expression for $T(I)$ by integrating $\check{T}(x)$ over the distribution of the distance to the nearest driver as a function of the density of idle drivers.[15] 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.

Under this functional form $\lim_{I\to 0} -\epsilon_I^w = \frac{1}{2}$, 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.[16] These are the values we use throughout this paper to diagnose WGCs. This means that with a high density of drivers lower slack is needed to avoid WGCs.[17]

Calibration to different markets

We calibrate the remaining parameters of our model to the data. We exclude December 15-January 7 since these are atypical days because of holidays, and we focus on weekdays between 7 am and midnight. We set $\tau = 0.238$, the average take rate of Uber.[18] 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. In a separate specification, we model two different markets, the weak market between 11 am and noon, and the strong market between 6 and 7 pm. These are the one hour intervals with the highest and lowest demand. We assume that all the model primitives stay the same as for the average market, except

[15] The density of drivers at a distance x is $2\pi Ix$. The CDF of the distance to the nearest driver $G(x;I)$ is given by $\frac{dG}{dx} = 2\pi Ix(1-G)$, whose solution corresponds to a Weibull distribution $G(x;I) = 1 - e^{-\pi Ix^2}$.

[16] However, as $I \to \infty$, it again becomes $\frac{1}{2}$. The inefficiencies of going around the block eventually level off once pickup time is determined by driving straight down the block.

[17] This is intuitive because when drivers are very dense, additional idle drivers do not rapidly reduce pickup times because of the inefficiency of driving around the block. Lower slack values are thus natural since more time is spent picking up passengers relative to being idle

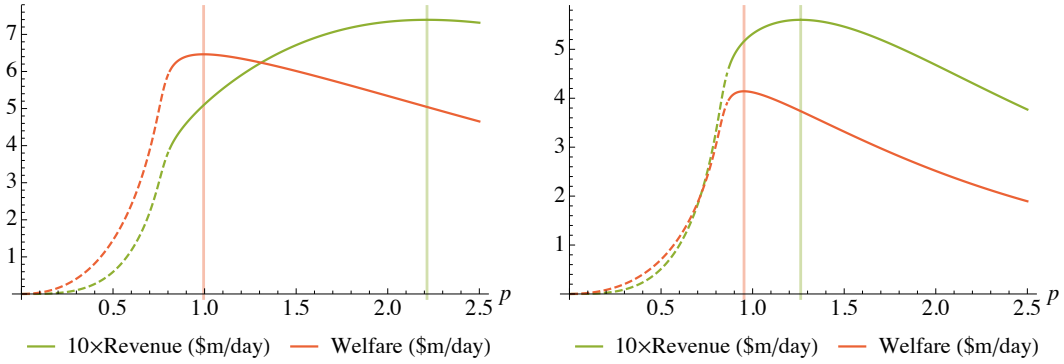
[18] The value of τ varies between drivers, depending on when they started at Uber.

Market	λ (sessions/h $\cdot$ km^2)	Q (trips/h $\cdot$ km^2)	L (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 2: Observables and parameters for calibrated markets

for λ and A. Table 2 compares aggregate numbers as well as parameters for the average, weak, and strong markets. 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: 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. In a final specification, with details in Appendix B, we calibrate the parameters separately for every hour of the week.

5.2 Quantitative analysis of pricing



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

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

Figure 6: Revenue and welfare as a function of price. Dashed lines represent WGCs. Vertical lines represent optimal prices.

Figure 6 shows how revenue and welfare behave as a function of price. The left region with dashed lines represents prices at which WGCs occur. Subfigure 6a uses demand estimates from Cohen et al. (2016). 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 in aggregate hurt all of drivers, riders and the platform (although they may slightly benefit some price marginal riders who are willing to wait a long time).

To the right of the threshold, on the other hand, any price increase typically benefits drivers and the platform while hurting passengers. Since there is a tradeoff, changes in welfare are not too large: a 30% increase in prices from the optimum only decreases welfare by less than 4%. On the other hand, a 30% decrease in prices from the optimum leads to a 41% decrease in welfare.

For this calibration WGCs start 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.

The elasticity estimates (around 0.4-0.6 for prices between 1 and 2) from Cohen et al. (2016) are based on very short-run price changes, which explain why they are so low. Subfigure 6b shows the same calibration, assuming elasticities are twice as large. 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 region. 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 WGCs.

We now obtain similar results for street-hail taxis. We assume that average waiting time has the form $\check{T}(I) = \frac{\theta}{I}$, where θ is a parameter we calibrate to the data.[19] We take the average number of active taxis in

[19]This functional form corresponds to the arrival rate of an idle taxi being a Poisson process with rate that is proportional to the number of idle drivers.

Manhattan from Frechette et al. (2016) and assume that it results in an average waiting time of one minute. Demand and supply stay the same.

Figure 7 compares welfare between ride-hailing and traditional street-hail taxis. Although there also is a sharper decrease in welfare to the left than to the right of the maximum for taxis, the asymmetry is much less than for ride-haling. Ride-hailing leads to higher welfare with the right price, but there exists a region with WGCs in which the taxi market performs better. Higher welfare for Uber might seem at odds with the casual observation that waiting times are lower for taxis in Manhattan, but this arises because or increasing returns to scale: the average number of taxi drivers is almost four times the average number of Uber drivers in our data, based on data from Frechette et al. (2016).

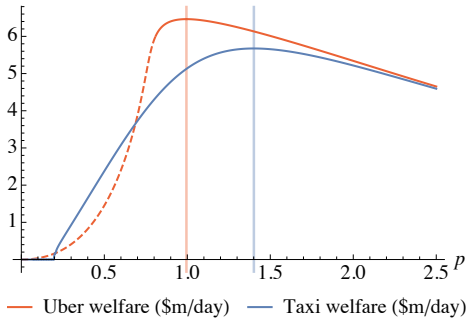
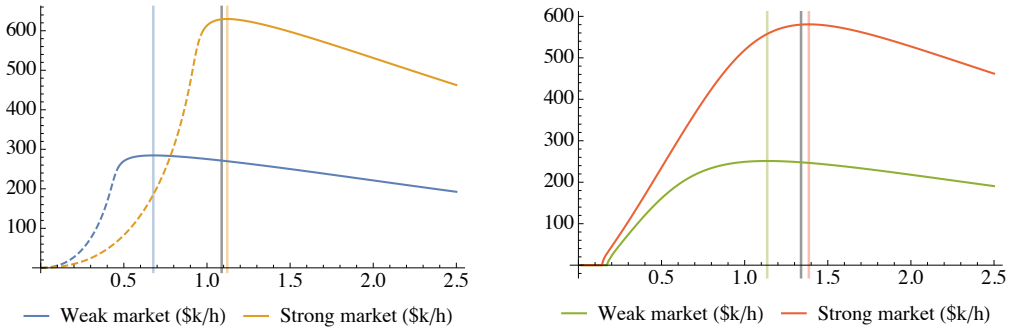


Figure 7: Comparison of welfare with ride-hailing and traditional street-hail taxis. Vertical lines represent welfare maximizing prices.

5.2.1 Two-market pricing

We now analyze the social benefits of surge pricing using the original elasticities measured by Cohen et al. (2016). We assume that the platform maximizes welfare since their main concern is long run profit instead of short run revenue. We first analyze a simple setting in which the platform only faces the weak and the strong market. In our first setup, static pricing, we require the platform to have the same price for both markets. In the second setup, surge pricing, we allow the platform to set different prices.



(a) Ride-hailing

(b) Traditional street-hail taxis

Figure 8: Two-market pricing with fixed τ. The gray vertical line represents the optimal price with static pricing. Colored vertical lines represent optimal prices with surge pricing.

Figure 8 shows the results of this analysis. For ride-hailing (Subfigure 8a), the static price is extremely close to surge pricing for the strong market. Whereas the surge price for the strong market is only 3% above the static price, the surge price for the weak market is 38% below. The reason for this extreme asymmetry is the sudden drop in welfare to the left as the market enters a WGC. There is some asymmetry for traditional street-hail taxi markets (Subfigure 8b), but it is of a different order of magnitude. Furthermore, optima are much closer together with street-hail taxis, which implies that surge pricing is less important than with ride-hailing.

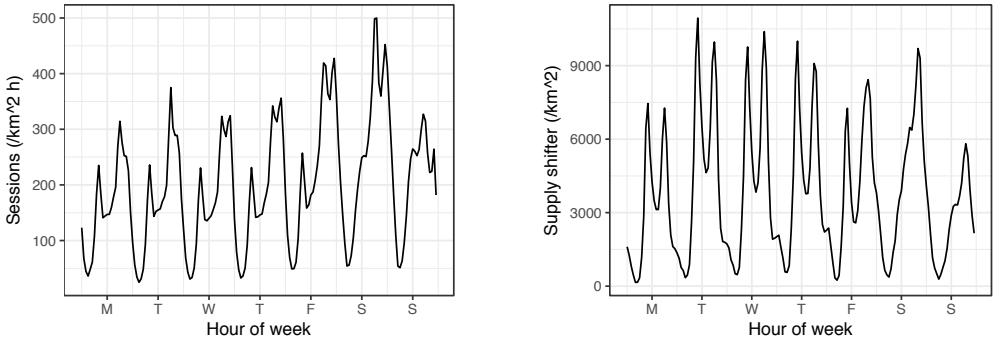
If the platform is constrained to static pricing, it has little freedom to set prices below the strong market surge optimum because it would enter a WGC with a substantial welfare drop. This means that allowing dynamic pricing leads to a significant reduction of prices in weak markets, whereas it only leads to modest increases in prices for strong markets. This drastically differs from the perception that surge pricing hurts consumers. If anything, drivers should be the ones worried that surge pricing does not allow inefficiently high prices at times of low demand.

Our results also suggest an explanation for 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. They might not want to go below 1 because they face criticism for drivers not being paid well, and for predatory pricing to avoid new entrants.

5.2.2 Pricing by hour of the week

We now analyze a platform that faces one separate market for each hour of the week. Besides λ and A, we also calibrate the average trip time t and the average speed v. We assume that average waiting time has the form $T(I, v) = \frac{v_{avg}}{v} T(I, v_{avg})$, where v_{avg} and $T(I, v_{avg})$ are speed and pickup time for the average market.

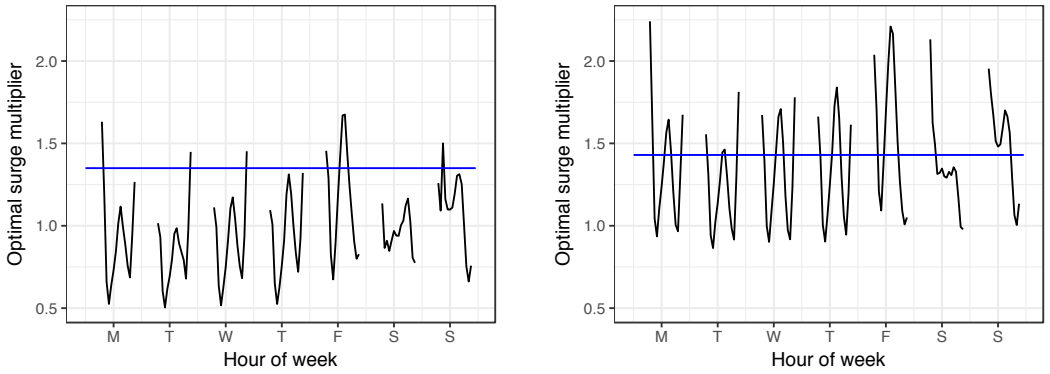


(a) Demand shifter (λ). (b) Supply shifter (A).

Figure 9: Value of main parameters of the model for different times of the week. Labels for the day of the week represent noon.

Figure 9 shows the behavior of supply and demand shifters. For demand, we see a small peak during the morning rush hour and a higher and longer peak for the afternoon rush hour. It is at its lowest late at night and early in the morning. Saturday has the highest demand, with no dip in between rush hours, and Sunday has a relatively low demand. For supply, we see the highest values in the early morning and late afternoon. This is most likely due to people with part time jobs or flexible schedules that are able to work before or after their main job. Appendix B shows the behavior of all other variables.

Figure 10 shows welfare maximizing prices. We focus on times between 7 am and 10 pm; at other times traffic conditions are so different that we do not believe our main fit of $T(I, v_{avg})$ is appropriate. The pattern for Uber (Subfigure 10a) is similar Monday through Thursday, with the highest prices in the early morning and at night, which are times with low supply. Morning and afternoon rush hour have high supply, which explains why they have somewhat lower prices despite high demand. Surge is very high during Friday afternoon and evening. Saturdays have low prices, mostly because higher traffic speed requires less drivers to satisfy high demand. Sunday have the highest prices because of low supply.



(a) Uber matching

(b) Taxi matching

Figure 10: Welfare maximizing prices with surge (black) and static pricing (blue) for different times of the week. Labels for the day of the week represent noon.

Separate patterns can be seen during Fridays, Saturdays, and Sundays. Surge must be high during Friday afternoon: demand is very high, whereas supply is essentially the same as other weekdays. Prices are low during Saturday. Despite demand being high throughout the whole day, traffic speed is low in the morning, and supply is high during the afternoon. Sunday has the highest prices, essentially because supply is low. All these patterns roughly follow the actual observed prices (see appendix B), except for the fact that surge pricing is actually high during Saturdays.

The optimal static price is higher than the optimal surge price at all but 7

hours of the week, consistent with our finding that a high price is necessary to avoid WGCs. The optimal static price is at the 92nd percentile of the distribution of optimal dynamic prices. A similar calculation using a taxi technology shows that the optimal static price is only at the 58th percentile (Subfigure 10b). Despite the fact that Uber would have substantially lower prices with surge pricing, it would have to set static prices at essentially the same level as with taxis.

6 Alternative Solutions to Wild Goose Chases

6.1 Maximum Dispatch Radius

A simple way to avoid WGCs is to randomly deny trips with some probability $1-q$ to shift the supply curve inwards, towards the good region. A better way to do it that specifically targets the externality caused by drivers driving too far to pick up riders is to deny trips in which riders would be matched to drivers at a distance greater than a *maximum dispatch radius* (MDR) R.[20] This is a solution that Uber has implemented to some extent, since they have a MDR for every location. As opposed to surge pricing, which varies dynamically, MDRs mainly serve as circuit breakers that avoid the worst market outcomes. In this subsection we model MDRs and suggest an explanation for this design choice.

Demand and labor supply are the same as in our main model. A rider who requests a trip is now matched with the probability $q(I, R)$ that there is an idle driver within a distance R. We assume this probability increases in both parameters, is zero if either parameter is zero, and converges to one if either parameter goes to infinity.

Expected pickup time conditional on being matched is now a function $T(I, R)$ of both the number of idle drivers and of the MDR that is decreasing in I and increasing in R. We make some additional assumptions that we list in the appendix, all of which are consistent with a functional form

[20] Or, similarly, a maximum dispatch ETA, which is equivalent in our model.

microfounded by a matching technology based on a pickup time $\tilde{T}(x)$ as in our calibration. We can also define an inverse function $I(T, R)$, which gives rise to a trip supply function $S(T, L, R) = \frac{L - I(T,R)}{t+T}$. Trip supply is again backward bending. It shifts to the right as R decreases: the dispatch radius avoids long pickups, thus increasing the number of trips that the same number of drivers can serve. The market equilibrium with an MDR is defined by

$$Q = q(I(T, R), R)D(T, p) = S(T, L, R) \qquad L = l\left((1 - \tau)p\frac{Q}{L}\right) \tag{11}$$

Proposition 6. *An equilibrium with $R \geqslant 0$ always exists (possibly with $Q = 0$, $L = 0$). The highest equilibrium is always stable and satisfies the following:*

1. *Starting from a WGC, a decrease in R leads to a decrease in pickup time and an increase in the number of trips, welfare, riders' surplus, and revenue.*

2. *If the highest equilibrium with R is a WGC, there exists some $R' < R$ such that the highest equilibrium is no longer a WGC.*

Starting from a WGC equilibrium, reducing R has the same properties as increasing prices: it improves market outcomes and eventually leads the equilibrium out of a WGC. To understand this, think first of the first order response, a shift to the left in supply, as more trips are denied, and a shift to the right in the supply curve. This then leads to the feedback cycles illustrated in Figure 3, which eventually lead the market out of the WGC.

Figure 11 shows how welfare changes with the introduction of a MDR, using the parameters from our main calibration. The surge multiplier is 0.75, which induces a WGC in the strong market but not in the weak or average market. In the strong market, once the MDR kicks in (at around 1 km) welfare starts to increase as trips with the strongest externality are avoided. The WGC is eventually avoided, at around 0.25 km. An excessively small MDR denies too many trips and welfare drops down to zero. The plot for the average market shows that even markets without WGCs might

benefit from a MDR. The plot for the weak market shows that there is no observable benefit of a MDR in markets that are far from being in a WGC.

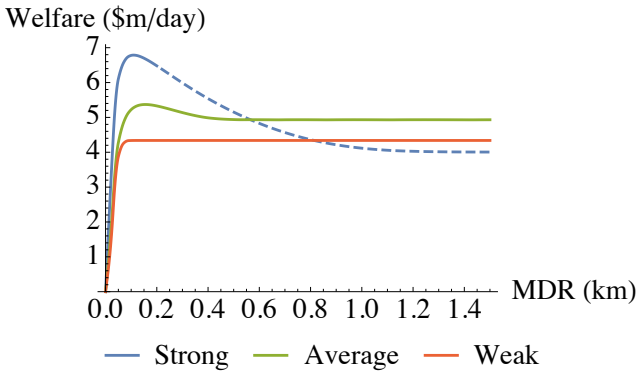
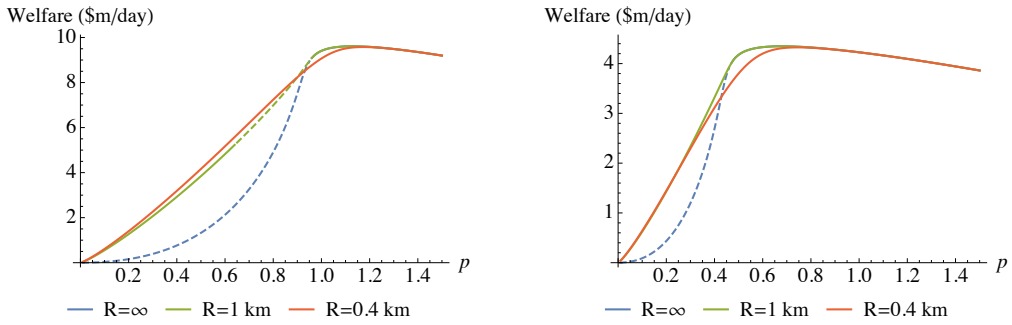


Figure 11: Welfare as a function of the MDR with a surge multiplier of 0.75.

Platforms could in principle manage WGCs by varying the MDR. However, the key for their success is the guarantee of a reliable service. Uber's mission statement is "Transportation that is as reliable as running water," and Lyft states in its signup website "A ride whenever you need one" (as of February 14, 2018). Using smaller MDRs goes against this guarantee. Ride hailing platforms are therefore unwilling to make it their main tool to manage supply and demand.

However, platforms have used a fixed MDR in combination with dynamic pricing. For the period we obseve, Uber fixed a maximum ETA between 10 and 25 (which roughly corresponds to a MDR between 2.5 and 6 km) minutes that was set in advance . The effects of this kind of policy are illustrated in Figure 12. A MDR of 1 km in practice dominates having no MDR: it mitigates the welfare loss during WGCs, at the cost of a decrease in welfare for high prices that is too small to observe in the figures. A smaller MDR, like the one shown in the plots (0.4 km) can further limit the harm of WGCs, but it starts denying trips at times when it is not necessary, thus causing a decrease in welfare when there are no WGCs. This is especially clear for the weak market. Most importantly, introducing a MDR removes the sharp decrease in welfare with low prices due to WGCs.



(a) Strong market

(b) Weak market

Figure 12: Welfare as a function of price for different dispatch radii. Equilibria with WGCs are shown with a dashed line.

The MDR works as a circuit breaker that limits the harm caused by WGCs. It does not affect behavior when the market is working well, since in that case very few passengers would be matched to a driver beyond the MDR. But it substantially increases welfare when prices are mistakenly set too low. Our results suggest that Uber could possibly benefit from a more aggressive MDR than the one they currently set.

A similar alternative to MDRs is to make passengers wait if there is no rider within R. This creates passenger queues in equilibrium, increasing thickness and improving match quality at the expense of waiting time. However, it is unlikely that riders' only option is to wait. There are other modes of transportation, as well as several competitors in the ride hailing market, so the user could just leave the platform and open the competing app. Losing customers to the competition is a major concern for these platforms, so although it might be socially optimal to make riders wait, in practice platforms are unlikely to do so.

6.2 Other strategies that attenuate WGCs

A variety of market design elements may mitigate WGCs, although they do not completely avoid them. Platforms can match passengers to drivers who are about to finish a trip nearby. This effectively increases the density

of idle drivers, reducing pickup times. Platforms can also rematch two rider-driver pairs if everyone benefits. Both approaches are improvements in the pickup time function $T(I)$, which can take the market out of WGCs in certain circumstances, but pickup time is still nonmonotonic and thus prone to WGCs.

A increasingly important share of the market are ride-sharing trips, with product such as UberPool or Lyft Line. Riders get a cheaper ride in which they are matched, if possible, to other riders going in the same direction. This is a more efficient usage of driver time, which results in a trip supply function that is further to the left, so it might avoid WGCs. However, as we show in Appendix D, the supply curve is also backwards bending. High demand depletes the market of idle drivers. This means, first, that idle drivers have to spend longer picking up passengers. But it also means that riders have to go through longer detours in order to pick up other riders. Thus, WGCs also occur.

The platform can also charge for pickup time and distance. This means pricing according to the externality caused on other riders by depleting streets of idle drivers and should theoretically restore efficiency. Howerver, unlucky passengers that are matched to someone far from them would have to pay a high price, besides having to wait a long time. People expect platforms to provide a low pickup time and would feel they are being charged for the platform's inability to offer a good service. Platforms are thus unwilling to implement this kind of pricing because of public relations.

7 Conclusion

Ride-hailing apps have improved transportation with market design innovations in matching and pricing. Whereas public opinion overwhelmingly supports their novelties in matching, there are mixed feelings about surge pricing, especially because of worries that it is a form of price gouging that hurts consumers. We have argued that matching and pricing are fundamentally tied together: although matching can perform better than

with incumbent technologies, this is only the case under the right pricing. Otherwise it enters wild goose chases, a market failure in which drivers spend too much time picking up far away costumers. Surge pricing is thus a fundamental component of ride hailing apps, which allows them to serve markets efficiently and reliably while having lower prices than older technologies when demand is low.

There are many interesting potential directions for future research, some of which (related to richer spatio-temporal dynamics) one of us is currently working on. However, a fascinating and policy-relevant dimension that has been entirely neglected thus far is the role of platform competition. While there are a variety of rich questions to be treated, our analysis highlights a surprising one: competition may be undesirable because idle drivers are a common pool resource. Most idle drivers have many competing applications open simultaneously. If one platform rations drivers by carefully managing its market, these drivers may be diverted to other platforms, destroying its ability and incentive to manage WGCs. Another unusual downside of competition will arise only once ride-hailing platforms grow to be dominant on the roads: a monopoly platform would then have an incentive to manage road congestion, while competitors would not. Both of these suggest that utility regulation may be more appropriate than competition policy in correcting the harms of market power.

References

Afeche, Philipp, Zhe Liu, and Costis Maglaras, "Ride-Hailing Networks with Strategic Drivers: The Impact of Platform Control Capabilities on Performance," *Working paper*, 2017.

Angrist, Joshua D. and Sydnee Caldwell, "Uber vs. Taxi: A Driver's Eye View," *Working paper*, 2017.

Armstrong, Mark, "Competition in two-sided markets," *The RAND Journal of Economics*, 2006, *37* (3), 668–691.

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

Banerjee, Siddhartha, Carlos Riquelme, and Ramesh Johari, "Pricing in ride-share platforms: A queueing-theoretic approach," *Working paper*, 2015.

Bimpikis, Kostas, Ozan Candogan, and Saban Daniela, "Spatial pricing in ride-sharing networks," *Working paper*, 2016.

Buchholz, Nicholas, "Spatial Equilibrium, Search Frictions and Efficient Regulation in the Taxi Industry," *Working paper*, 2017.

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

Cachon, Gérard P., Kaitlin M. Daniels, and Ruben Lobel, "The Role of Surge Pricing on a Service Platform with Self-Scheduling Capacity," *Manufacturing & Service Operations Management*, 2017, *19* (3), 368–384.

Chen, M. Keith and Michael Sheldon, "Dynamic Pricing in a Labor Market: Surge Pricing and Flexible Work on the Uber Platform," *Working paper*, 2015.

Chen, M Keith, Judith A Chevalier, Peter E Rossi, and Emily Oehlsen, "The value of flexible work: Evidence from Uber drivers," Technical Report, Working paper 2017.

Cohen, Peter, Robert Hahn, Jonathan V. Hall, Steven Levitt, and Robert Metcalfe, "Using big data to estimate consumer surplus: The case of uber," *Working paper*, 2016.

Cook, Cody, Rebecca Diamond, Jonathan Hall, John A List, and Paul Oyer, "The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers *," *Working paper*, 2018.

Cramer, Judd and Alan B. Krueger, "Disruptive Change in the Taxi Business: The Case of Uber," *American Economic Review*, May 2016, *106* (5), 177–82.

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

Frechette, Guillaume, Alessandro Lizzeri, and Tobias Salz, "Frictions in a Competitive, Regulated Market: Evidence from Taxis," *Working paper*, 2016.

Galichon, Alfred and Yu-Wei Hsieh, "A Theory of Decentralized Matching Markets Without Transfers, with an Application to Surge Pricing," *Working paper*, 2016.

Hall, Jonathan D., "Pareto improvements from Lexus Lanes: The effects of pricing a portion of the lanes on congested highways," *Journal of Public Economics*, 2018, *158*, 113 – 125.

_ , Craig Palsson, and Joseph Price, "Is Uber a substitute or complement for public transit?," *Working paper*, 2017.

Hall, Jonathan V. and Alan B Krueger, "An analysis of the labor market for Uber's driver-partners in the United States," *Working paper*, 2016.

_ , John Horton, and Dan Knoepfle, "Labor Market Equilibration: Evidence from Uber," *Working paper*, 2017.

Lagos, Ricardo, "An Analysis of the Market for Taxicab Rides in New York City," *International Economic Review*, 2003, *44* (2), 423–434.

Lam, Chungsang Tom and Meng Liu, "Demand and Consumer Surplus in the On-Demand Economy: The Case of Ride Sharing," *Working paper*, 2017.

Ma, Hongyao, Fei Fang, and David C Parkes, "Spatio-Temporal Pricing for Ridesharing Platforms," *arXiv preprint arXiv:1801.04015*, 2018.

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

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

Rochet, Jean-Charles and Jean Tirole, "Platform Competition in Two-Sided Markets," *Journal of the European Economic Association*, 2003, *1* (4), 990–1029.

Shapiro, Matthew H., "Density of Demand and the Benefit of Uber," *Working paper*, 2017.

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

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

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

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

Appendix

A Proofs

A.1 Proof of proposition 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 $\lim_{T\to\infty} S(T, L) = 0$. □

A.2 Proof of proposition 2

In order to simplify our analysis of equilibria, we define $\hat{Q}(L; p)$ to be the highest solution to equations (3) for a given number of drivers and $\hat{L}(Q; p, \tau)$ to be the solution to equation (4). The following lemma states that $\hat{Q}(L; p)$ is well defined:

Lemma 1. $D(T, p) = S(T, L)$ *has at least one solution in* T *for all* (p, L). *For the highest solution,* Q *is increasing in* L.

Proof. 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 $D(T(L)) = 0$. □

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 also see easily that for $L = 0$ the unique solution is $Q = 0$.

An equilibrium can then be characterized as a solution to $Q = \hat{Q}(L; p)$ and $L = \hat{L}(Q; p, \tau)$. We now define stability as the condition that both functions cross from above:

Proof. An equilibrium is stable if $\hat{Q}'\hat{L}' < 1$. □

We are now in a position to prove proposition 2:

Proof. The equilibrium at the origin always exists because $\hat{Q}(0;p) = 0$ and $\hat{L}(0;p,\tau) = 0$. All equilibria can be ordered because both $Q = \hat{Q}(L;p)$ and $L = \hat{L}(Q;p,\tau)$ are increasing. 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. Thus, the highest equilibrium is always stable. □

A.3 Proof of proposition 3

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 $\sigma = \operatorname{sgn}(\epsilon_T^S)$. The characteristic feature of WGCs is then $\sigma < 0$, whereas $\sigma > 0$ in good equilibria. Also let ϵ_l be the elasticity of $l(\cdot)$ and let $\epsilon_L = \frac{\epsilon_l}{1+\epsilon_l} \in [0,1]$.

Lemma 2. *The price elasticities of equilibrium number of trips and drivers and pickup time are*

$$\epsilon_p^Q = \frac{1}{\Delta}\frac{-\sigma\epsilon_T^S\epsilon_p^D + \epsilon_L\epsilon_T^D\epsilon_L^S}{\epsilon_T^D + \sigma\epsilon_T^S} \qquad \epsilon_p^L = \frac{\epsilon_L}{\Delta}\left(1 - \frac{\sigma\epsilon_T^S\epsilon_p^D}{\epsilon_T^D + \sigma\epsilon_T^S}\right) \qquad \epsilon_p^T = -\frac{\epsilon_p^Q + \epsilon_p^D}{\epsilon_T^D} \tag{12}$$

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

$$\begin{pmatrix} 1 & -\frac{\epsilon_T^D\epsilon_L^S}{\epsilon_T^D + \sigma\epsilon_T^S} \\ -\epsilon_L & 1 \end{pmatrix}\begin{pmatrix} d\log Q \\ d\log L \end{pmatrix} = \begin{pmatrix} -\frac{\sigma\epsilon_T^S\epsilon_p^D}{\epsilon_T^D + \sigma\epsilon_T^S} \\ \epsilon_L \end{pmatrix} d\log p, \tag{13}$$

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. □

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

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

$$\frac{dW}{dp} = U_T \frac{dT}{dp} + \frac{Q}{\Delta}\left[(1-(1-\tau)\epsilon_L)\frac{-\sigma\epsilon_T^S\epsilon_p^D}{\epsilon_T^D + \sigma\epsilon_T^S} - \epsilon_L(1-\tau) + \epsilon_L\frac{\epsilon_T^D\epsilon_L^S}{\epsilon_T^D + \sigma\epsilon_T^S}\right]. \tag{14}$$

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

$$\frac{dRS}{dp} = \epsilon_p^T \epsilon_T^U \frac{U}{p} - Q \tag{15}$$

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

Proof. The total differential of welfare is $dW = U_T + U_Q dQ - C'dL$. Note first that $U_Q = p + \bar{u}_Q$, where $\bar{u}$ is the derivative of the utility of infra-marginal 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 (14) 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 2. 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. □

A.4 Proof of proposition 4

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. □

A.5 Proof of Proposition 5

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. □

A.6 Formal assumptions and proof of proposition 6

Assumption 3. *Functions* $q(I, R)$ *and* $T(I, R)$ *satisfy the following properties:*

1. $q(I, R)$ *is increasing in both* I *and* R
2. $q(I, 0) = 0$ *and* $\lim_{R\to\infty} q(I, R) = 1$ *for all* I
3. $q(0, R) = 0$ *and* $\lim_{I\to\infty} q(I, R) = 1$ *for all* R
4. $T(I, R)$ *is decreasing in both* I *and* R*, and it is concave in* I
5. $T(I, 0) = 0$ *and* $\lim_{R\to\infty} T(I, R) = T(I)$ *for all* I
6. $\lim_{I\to 0} T(I, R) = \check{T}(R)$ *where* $\check{T}(R)$ *is bounded and increasing in* R*, and* $\lim_{I\to\infty} T(I, R) = 0$ *for all* R

The behavior of the original model is recovered as $R \to \infty$. The inverse function $I(T, R)$ satisfies the following:

Lemma 4. *Let* $I(T, R)$ *be the inverse of* $T = T(I, R)$ *when holding* R *fixed. It satisfies the following:*

1. *It is defined for all positive* R *and for all* $T \in [0, \check{T}(R)]$
2. $\lim_{T\to 0} I(0, R) = \infty$ *and* $I(\check{T}(R), R) = 0$ *for all* R

Supply is then $S(T, L, R) = \frac{L - I(T,R)}{t+T}$. It satisfies the properties for $S(T, L)$ specified in lemma 1, except for the fact that it is undefined for $T > \check{T}(R)$. The labor supply equation behaves exactly as in the main model, resulting in an increasing supply curve $\hat{L}(Q; p, \tau)$. Following the arguments of lemma 1, there is also a bounded, increasing equation $\hat{Q}(L; p, R)$. By the same arguments as in Appendix A.2, there is at least one solution, and the highest one is stable. The proofs for comparative statics follow essentially the same arguments as the proof in Appendix A.3.

B Additional figures

Figure 13 shows our fit of pickup time as distance from matched driver.

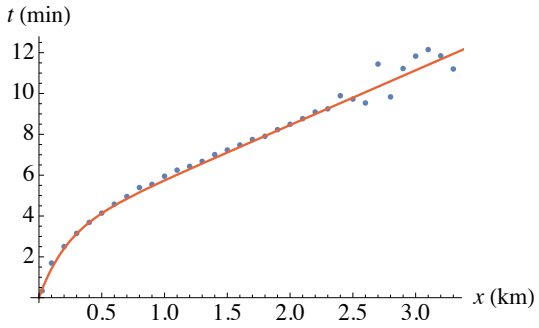


Figure 13: Average pickup time as a function of distance from matched driver, as well as a fit of the form $t(x) = a(1 - e^{-bx}) + cx$. There are very few trips with distance greater than 2.5 km, which explains the high variability in the data.

Figure 14 shows the behavior of the observable market quantities used to fit the model in section 5.2.2.

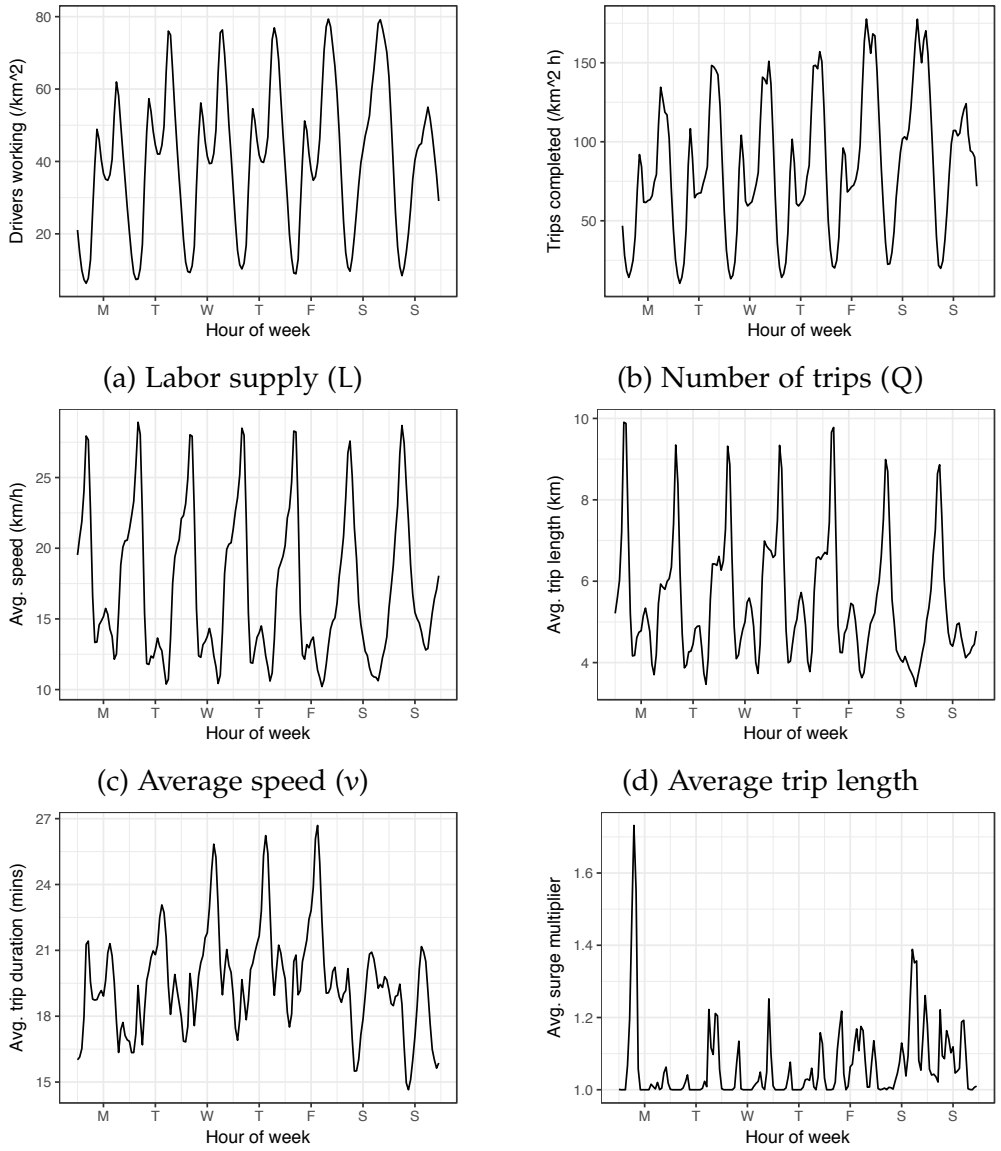
Figure 15 shows what are the times of the week during which WGCs take place, defined as slack below 0.35, as well as the average multiplier.

C Trip supply by slack

Table shows the results of equation (4.2) where a_t is defined by whether slack is above or below some threshold.

D Pooling

In this section we extend our analysis to pooling services, such as UberPool or Lyft Line. Given the much greater complexity of these services, we make stronger simplifying assumptions. Demand and supply take the same form as in our previous model, and in which passengers can only take pooling trips.



(a) Labor supply (L)

(b) Number of trips (Q)

(c) Average speed (v)

(d) Average trip length

(e) Average trip duration (t)

(f) Average surge multiplier (p)

Figure 14: Market characteristics across different times of the week. Labels for the day of the week represent noon for any given day.

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

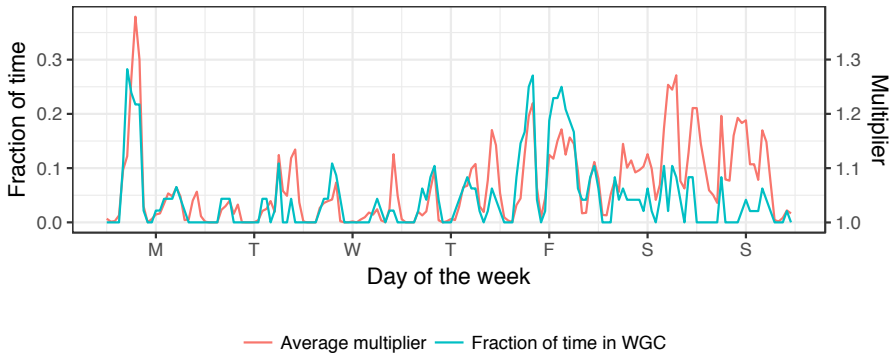


Figure 15: Frequency of WGCs and average multipliers at different times of the week.

Table 3: Behavior of supply with high and low slack

	Dependent variable:							
	Number of trips completed							
	$s^{th} = 0$	$s^{th} = 0.2$	$s^{th} = 0.3$	$s^{th} = 0.4$	$s^{th} = 0.5$	$s^{th} = 0.6$	$s^{th} = 0.8$	$s^{th} = 1$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: OLS							
ETA × above	4.731***	4.793***	5.086***	5.592***	5.964***	6.074***	5.928***	5.678***
	(0.496)	(0.516)	(0.608)	(0.818)	(1.078)	(1.296)	(1.582)	(1.867)
ETA × below		0.134	−0.554*	−0.674***	−0.474***	−0.311**	−0.092	0.025
		(0.810)	(0.325)	(0.175)	(0.133)	(0.128)	(0.107)	(0.094)
	Panel B: 2SLS							
ETA × above	13.686***	14.479***	15.900***	18.053***	20.305***	22.160***	25.140***	28.750***
	(0.385)	(0.357)	(0.368)	(0.387)	(0.446)	(0.528)	(0.694)	(0.888)
ETA × below		−22.450	−1.465	1.148	2.599***	3.599***	4.895***	5.114***
		(80.867)	(1.206)	(0.751)	(0.951)	(1.033)	(1.098)	(0.936)
Obs. above threshold	7,822	7,801	7,632	7,297	6,911	6,593	6,110	5,723
Obs. below threshold	0	21	190	525	911	1,229	1,712	2,099
Observations	7,822	7,822	7,822	7,822	7,822	7,822	7,822	7,822
Note:								Robust standard errors in parentheses: *p<0.1; **p<0.05; ***p<0.01

total number of drivers satisfies $L = I + B_1 + B_2 + K_1 + K_2$

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. 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, $B_2 = \frac{t}{2}\frac{qQ}{I+qB_1}B_1$.

Supply is given by the solution in (Q, B_1, B_2) to the following system of equations:

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

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

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

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

Proposition 7. *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. 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:

$$Q = \frac{qL - \frac{t+T}{t}f(T) \pm \sqrt{(1-2q)f(T)^2 + L^2q^2 + \frac{2f(T)}{t}((1-2q)f(T) + qL)T + \frac{T^2}{t^2}f(T)^2}}{q(t+2T)} \tag{19}$$

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.

□

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.