Surge Pricing Solves the Wild Goose Chase*

Juan Camilo Castillo[†] E. Glen Weyl[‡]

December 2016

Abstract

Why is dynamic pricing more prevalent in ride-hailing apps than movies and restaurants? Arnott (1996) observed that an over-burdened taxi dispatch system may be forced to send cars on a *wild goose chase* to pick up distant customers when few taxis are free. These chases occupy taxis and reduce earnings, effectively removing cars from the road and exacerbating the problem. While Arnott dismissed this outcome as a Pareto-dominated equilibrium, we show that when prices are too low relative to demand it is the unique equilibrium of a system that uses a first-dispatch protocol (as many ride-hailing services have committed to). This effect dominates more traditional price theoretic considerations and implies that welfare and profits fall dramatically as price falls below a certain threshold and then decline only gradually move in price above this point. A platform forced to charge uniform prices over time will therefore have to set very high prices to avoid catastrophic chases. Dynamic "surge pricing" can avoid these high prices while maintaining system functioning when demand is high. We show that pooling can complicate and exacerbate these problems.

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

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

This paper is preliminary and incomplete, as described below. Comments are very, and citations somewhat, welcome.

[*] The name of one of the co-authors of this paper has been removed from it to preserve the anonymity of the ride-hailing platform which provided data for this paper. We appreciate the helpful comments of Eduardo Azevedo, Timothy Bresnahan, Liran Einav, Matthew Gentzkow, Ramesh Johari, Jonathan Levin, Andy Skrzypacz and seminar participants at Microsoft Research New York City and Stanford University.

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

[‡] Microsoft Research New York City, 641 Avenue of the Americas, New York, NY 10011 and Department of Economics, Yale University; glenweyl@microsoft.com, http://www.glenweyl.com.

1 Introduction

The prices of films and restaurants vary at most modestly and rigidly across time despite dramatic fluctuations in the opportunity cost of capacity. Instead, queues and reservations are used to ration these resources. This pattern puzzled many economists (Becker, 1991) and generated broad enthusiasm for the recent rise of real-time dynamic consumer pricing in ride-hailing applications, such as Lyft and Uber. Was this simply a social taboo that innovative ride-hailing apps were willing to break? Or does "surge pricing" play a unique role in the context of ride-hailing? We argue for the later conclusion. In particular, we highlight a crucial technical feature of existing ride-hailing systems, that they are susceptible to a phenomenon we refer to as the *wild goose chase*, which is analogous to "hypercongestion" in transportation engineering, that would make them extremely technically inefficient in the absence of dynamic pricing. This implies that surge pricing is critical to maintaining system capacity and not simply an alternative means of allocating that capacity.

To understand the system failures that would occur in the absence of surge pricing it is important to note that cars in the system may be in one of three states: idle, picking up a rider they have been dispatched to or delivering a rider to her destination.[1] In typical ride-haling systems, 1) a car is immediately dispatched to any rider requesting a ride and 2) drivers are only paid for the time they are actually servicing a customer. Given these features, when the demand for rides is high relative to the number of idle cars, cars will often have to be dispatched on a wild goose chase (WGC) to pick up a rider at a distant location. As a result, the cars will spend a long time picking up their passengers and thus will become idle infrequently. This will reinforce the scarcity of idle drivers, closing this negative feedback loop.

Thus, it may be that in times of high demand the total number of rides completed per unit time may actually be lower than when demand is weaker. To make matters worse, the fraction of time working during which drivers are not paid (what drives call "dead miles") rises when WGCs occur, thereby lowering their earnings relative to periods of lower demand. This perversely discourages drivers from offering services during these times when their services are most needed. When combined, these factors can lead a ride-hailing service with these rules and without surge pricing to grind to a halt at precisely the times when it is most needed.

Surge pricing solves both of these problems. First, by reducing the flow of demand below the volume that creates WGCs, surge pricing avoids the erosion of effective capacity for a fixed number of drivers supplying services. Second, by restoring or boosting earnings during high demand times, surge pricing makes earnings at least as high during peak loads as during normal times. Absent surge pricing or a change to the system's engineering, ride-hailing platforms would have to charge very high uniform prices to ensure WGCs occurs only when demand is

[1] This differs from street taxis, which may only be either idle or delivering a rider. However it is similar to dispatch taxis.

exceptionally heavy, a practice that would harm the welfare of all participants relative to surge pricing.

Note that this mechanism, while hardly unique to taxi dispatch systems does not arise in many seemingly-related contexts. It is not true in fixed-capacity systems like public transit, entertainment facilities or restaurants, as the number of customers who can be served by those systems is independent of the way in which those systems are utilized. It does not apply to street-hail taxis, which only have local pick-ups and thus cannot be sent on WGCs. While central dispatch potentially offers large matching efficiencies over street hailing (Frechette, Lizzeri and Salz, 2016) because they manage to keep drivers employed in some form constantly, these potential efficiencies also make it potentially fragile to employing drivers in unproductive WGCs rather than in the useful depositing of riders. This may be one reason why such systems were largely unsuccessful until the advent of technologies that made dynamic pricing feasible.

Despite this novelty, a related phenomenon has been observed in the literature on transportation economics. In that context, it is called "hypercongestion" (Walters, 1961; Vickrey, 1987) and refers to the fact that when enough cars enter a road to cause what non-specialists would refer to as a "traffic jam", speeds of all cars on the road fall sufficiently that the total throughput of the road actually falls.[2] A similar phenomenon occurs in purely physical systems: if you try to pass a volume of a dry good (like rice) through a funnel, the fastest transmission is possible by a steady pour that avoids clogging the funnel rather than by simply dumping the full volume in. However, the effects of WGCs may be much more severe than those of clogging in these other systems because the supply of drivers is endogenous, and may collapses in reaction to the fall in earnings created by WGCs.

The possibility of WGCs in ride hailing was foreshadowed by Arnott (1996), who considered the optimal design of a centralized and omniscient taxi dispatch system prior to the existence of technology that would enable such a system to be constructed. In his analysis he noted that a Pareto-suboptimal equilibrium could arise "analogous to that for a stable, hypercongested equilibrium in traffic flow theory". However, because he was concerned with an optimal system he "assumed that when there are multiple equilibria, the market settles in the Pareto efficient equilibrium." As shown in Section 3, however, when prices are rigid and riders are free to call and have immediately dispatched to them a ride, the equilibrium will involve WGCs at times of high demand relative to capacity and price.[3]

While Arnott's analysis was astonishingly far-sighted, his vision has largely been implemented over the course of the last half decade. Founded in 2009 and 2012 respectively, ride-

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

[3] There might be multiple equilibria, but in that case all equilibria will involve WGC's at times of high demand.

hailing services Uber and Lyft have become a dominant mode of transportation in many urban areas. Both operate on a first-come-first-serve dispatch basis and, since 2010, both have used dynamic pricing (which Uber labels *surge pricing*) to manage demand. While surge pricing has generated significant excitement among economists, it has been controversial among users of the services and regulators. For example, the splash page on competitor Gett's home page on November 20, 2016 stated "The only time we surge is never o'clock" and many cities in the developing world have banned or otherwise forced Uber to desist from surge pricing.

Our analysis suggests that, absent basic changes to the engineering of ride hailing systems, self- or externally-imposed limitations on surge pricing are likely to have large allocative costs. In an example in Section 4 calibrated to data from a large ride-hailing platform's market in Manhattan, we show that socially optimal prices if surging is prohibited are more than 97% of their level with surge pricing at the highest demand hour of the day and are more than 47% higher than the level during the lowest demand hour when surging is allowed. This quantitatively reinforces our qualitative conclusion that, absent surge pricing or engineering changes, prices would be very close those that prevail with surge pricing at peak demand periods.

None of this is to argue, however, that surge pricing is the only reasonable solution to WGCs. We are currently exploring and in a future draft will include an analysis of how holding a queue of riders, rather than immediately dispatching the next available car, could also help resolve WGCs. Thus surge pricing should not be viewed as the exclusive or necessary response to the possibility of WGCs, but only as the most natural solution that requires the least dramatic reorganization of the engineering and consumer commitments the platforms make.

Our analysis begins in the next section with a model that builds closely on Arnott's model but extends it to endogenize ride requests, driver labor supply and pricing and to allow more realistic matching between drivers and passengers. In Section 3 we describe how WGCs arises in this model and why the unique equilibrium involves WGCs when pricing is too low; we also show the extreme effects WGCs have on all welfare variables and how it causes profits and social welfare to closely align in many cases. Then in Section 4 we calibrate our model to moments supplied by a large ride-hailing platform and quantitatively analyze the effects of a ban on surge pricing. We find that without surge pricing platforms should set prices corresponding to times of highest demand so that WGC never happen. This means that if ride hailing apps like Uber or Lyft did not use dynamic pricing, the alternative would not be to set prices at their base fare at all times, and not even to set prices at their average fares. Instead, prices would be closer to the highest fares that are currently observed.

In recent years, ride-sharing or "pooling" has gained an increasing share of the ride-hailing market. We therefore, in Section 5, discuss a model allowing pooling. This model is much richer and thus we treat it quite superficially at present. More broadly, this paper is a preliminary and primarily theoretical analysis. We have obtained access to detailed microlevel data from a large ride-hailing platform that we plan to use to estimate parameters of an test our model. We thus conclude by discussing the empirical analysis as well as additional theoretical results we plan to add to the paper in a later, more complete draft than the present one.

2 Model

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

2.1 Demand

Let λ be the density of arrival of users (measured, for instance, in users per minute per square kilometer). These are the users that might potentially request a ride if the price and the waiting time are good enough for them. We assume that users will request a ride exactly when they are willing to pay the associated price and are able to wait the associated wait time. We assume these two motives are independent and that there is no lost utility of waiting other than the inability to accept the ride. These assumptions simplify our model, but all of our central results can be derived in a setting that relaxes them. Let $r(p)$ be the fraction of users that are willing to pay for a ride at price p, and let $g(w)$ be the fraction of users that are willing to wait if the time before pickup is w. The number of ride requests is then $R = \lambda g(w) r(p)$.

2.2 Supply

Let D be the number of working drivers per unit area. This causes a total cost $C(D)$ (per unit area per unit of time). Drivers decide whether to work or not by comparing their per-unit time earnings e with their marginal cost, so $C'(D) = e$. To find an expression for e, let τ be the fraction of the price charged to passengers that the platform takes as revenue. Given the total density of rides per unit of time R and the price p, total earnings per unit of time per unit area are $(1-\tau)pR$. The average earnings per unit of time of individual drivers are $(1-\tau)p\frac{R}{D}$. Assuming symmetry among drivers in their expectations and rationality on average, drivers' optimal decision is then given by $C'(D) = (1-\tau)p\frac{R}{D}$.

2.3 Matching

At any given moment drivers are in one of three states. Some of them are idle (waiting to be matched to a rider), which we denote by I. In equilibrium, $\frac{l}{v}R$ drivers are driving a rider, where l is the average trip length and v is the average speed when driving passengers. Finally, wR

drivers are on their way to pick up a rider. Thus, the following identity accounts for the total density of drivers: $D = \frac{l}{v}R + l + wR$.

The average waiting time $w(I)$ is inversely related to the density of idle drivers: if there are a lot of idle drivers, a new arriving rider will on average be matched to a driver that is closer to him, so he will have to wait less time before being picked up. We will assume a simple geometry with no inefficiencies beyond waiting time and a uniform distribution of drivers, thus abstracting from the important systematic differences in supply compared to demand at different points in space studied by Buchholz (2016) and treating these differences only through our analysis of separate markets that are treated as entirely segmented. Given this segmentation assumption it may be easier to interpret our markets as representing different times, as in the analysis of Frechette, Lizzeri and Salz (2016), rather than different places within a city; in either case our static model that leaves out substitution and complementarity across markets is an important modeling simplification.

However, it allows an analytic expression for $w(I)$ as follows. In two dimensional space, the density of drivers at a distance x from an arbitrary point is $2\pi Ix$, (a measure to be integrated with respect to x) which is the hazard function of the nearest driver. The CDF of the distance to the nearest driver $G(x; I)$ is then given by the differential equation $\frac{dG}{dx} = 2\pi Ix(1 - G)$, whose solution, which corresponds to a Weibull distribution, is $G(x; I) = 1 - e^{-\pi Ix^2}$. If the average waiting time as a function of distance is $t(x)$, then $w(I) = \int_0^\infty t(x) dG(x; I)$.

In a simple, homogeneous space, $t(x)$ is simply a linear function, $\frac{x}{v}$, where v is the speed. However, matters are considerably more subtle in practice. The pattern of roads in some cities has one-way streets every other block, and in others follows radial rather than axis-aligned coordinates. Furthermore, speeds are greater when traveling longer distances since drivers are able to take larger streets or highways. This implies that the appropriate formula for $t(x)$ in practice will vary from city to city. We will take a function of the form $t(x) = a(1 - e^{-bx}) + cx$. The first term captures the fact that cities' street patterns cause inefficiencies when traveling short distances. The second term means that speed eventually reaches some terminal value c, which is the speed once drivers take a main street. This functional form fits very well the data for trips in Manhattan obtained from a large ride-hailing platform, as shown in Figure 1. The resulting expression for expected waiting time is $w(I) = \frac{1}{\sqrt{4I}}\left(c + 2ab \exp\left(\frac{b^2}{4\pi I}\right) \Phi\left(\frac{b}{\sqrt{2\pi I}}\right)\right)$.

3 Wild Good Chases

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

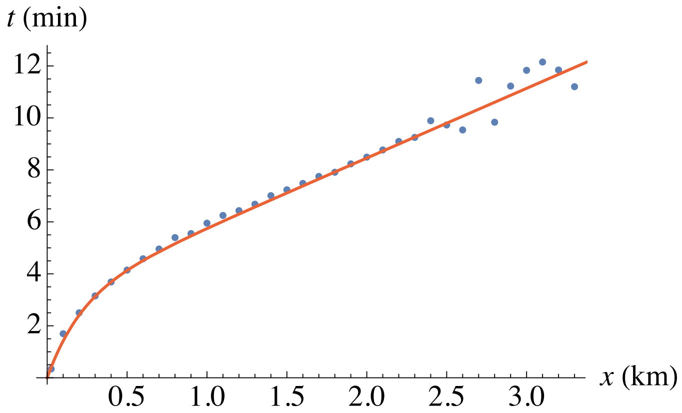


Figure 1: Average waiting time as a function of distance from matched driver, as well as a fit of the form $t(x) = a(1 - e^{-bx}) + cx$. There are very few trips with distance greater than 2.5 km, which explains the high variability in the data.

3.1 Normal and wild goose chase matching equilibria

The identity for the density of drivers is, as derived by Arnott,

$$D = \underbrace{\frac{l}{v}R}_{\text{Driving}} + \underbrace{I}_{\text{Idle}} + \underbrace{w(I)R}_{\text{Picking up}} . \tag{1}$$

We now use this equation to find a solution for $R = Q(D, I) = \frac{D-I}{\frac{l}{v}+w(I)}$. Here $Q(D, I)$ can be interpreted as the capacity of the market: the total number of rides the market is able to serve when there are D drivers and I of them are idle.

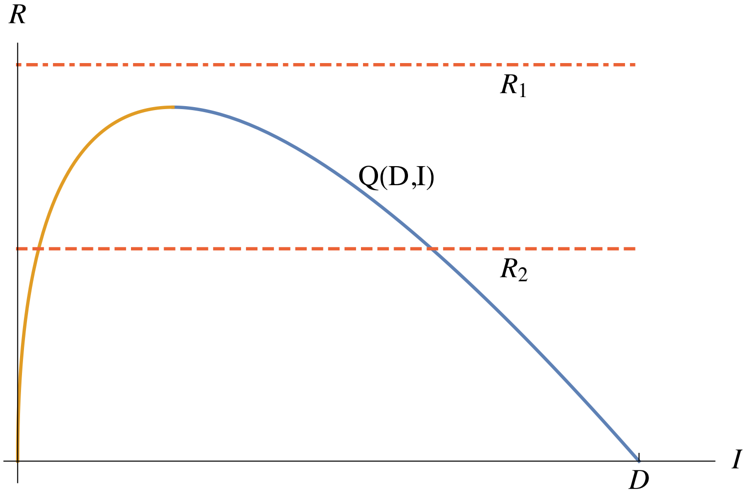


Figure 2: Solutions for number of idle drivers as a function of drivers and ride requests.

The solution to this equation is shown in Figure 2. Note that $w(I)$ is decreasing, convex, $\lim_{\rho\to\infty} w(\rho) = 0$, and $\lim_{\rho\to 0} w(\rho) = \infty$. This causes the inverted-U shape of $Q(D, I)$. The intuition behind it is as follows. When the number of idle drivers I is high, i.e., close to D, very few drivers are available to drive passengers to their destination, and the capacity of the system is low. This explains the behavior of $Q(D, I)$ in the blue region, to the right of its maximum. The reason behind its behavior in the yellow region is more subtle. When there are very few idle drivers, waiting times become very high, since $\lim_{\rho \to 0} w(\rho) = \infty$. Thus, by remembering identity (1), drivers spend most of their time on their way to pick up passengers, and very little time remains for them to drive passengers to their destination. This results in the market having a low capacity.

When finding a solution for $R = Q(D, I)$, two different situations can take place. Let $\bar{R}(D)$ be the maximum of $Q(D, I)$. For $R > \bar{R}(D)$, the number of riders is beyond the maximum capacity given the number of drivers (as with R_1 in 2). The two solutions with $R < \bar{R}(D)$ (R_2 in 2) correspond to two possible levels of idle drivers.[4] In one case there is a high density of idle drivers, which then leads to short pick-up times. The other solution is a perverse equilibrium in which there is a low density of idle drivers, but pick-up times are high and therefore a large number of drivers spend time picking up passengers that are far away. Arnott (1996) pointed out the existence of these two solutions, and noted that the bad solution is clearly Pareto inefficient: the first solution leads to lower waiting times and more passengers getting a ride. He was looking for the social optimum and therefore simply discarded the inefficient solution. But we will show that analyzing this bad solution is essential to understanding how to set prices dynamically.

We call the situation in the bad equilibrium *wild goose chases* (WGCs). In colloquial English, wild goose chases refer to extended, wasteful and ultimately vain pursuits of an unattainable objective. By analogy, in this bad situation, the ride-hailing system, by trying to serve beyond its capacity, must send drivers to distant locations that ultimately reduce the number of rides it can effectively provide. An easy way to diagnose WGCs comes from noting that the derivative of the left hand side of $Q(D, I)$ with respect to I is positive. One way to write this is that WGCs happens when $I < -\epsilon_I^w w(I) R$, where ϵ_I^w is the elasticity of waiting time with respect to the density of idle drivers. This inequality is easy to interpret: the number of idle drivers being less than $-\epsilon_I^w$ times the number of drivers picking up passengers is a red flag for WGCs.

Note that under the functional form we use for $t(x)$ (which, as highlighted above, appears to be a close fit to the data), $\lim_{I \to 0} -\epsilon_I^w = \frac{1}{2}$, 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 .26.[5] That is, in cities with a very dense coverage of drivers, fewer idle drivers relative to those picking up riders are needed to avoid WGCs. This is intuitive because when drivers are very dense, increased numbers of idle drivers do not rapidly reduce waiting times. It is therefore not problematic for drivers to spend a greater fraction of their time on "dead miles". Taken to an

[4] In the knife edge case $R = \bar{R}(D)$ there is a unique solution.

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

extreme, as I grows large it is natural that more time is spent picking up passengers relative to being idle, as most drivers must drive around the block to get a nearby rider; only if so many drivers can be made available so that one is directly in front of every potential rider's house can this small friction be eliminated. When there are fewer available drivers, on the other hand, increasing driver density is more beneficial and thus more idle drivers relative to those picking up riders are needed to avoid WGCs as each additional driver "fills in" an important part of the city grid.

Let $I^g(R, D)$ denote the good equilibrium, and $I^h(R, D)$ denote the WGC solution. These two solutions also lead to good and WGC solutions for waiting time and fraction of passengers able to wait, which we denote by $w^g(R, D) = w(I^g(R, D))$, $w^h(R, D) = w(I^h(R, D))$, $g^g(R, D) = g(w(I^g(R, D)))$, and $g^h(R, D) = g(w(I^h(R, D)))$, with a slight abuse of notation. The functions for the fraction of passengers able to wait has the following characteristics:

Lemma 1. *Assuming continuity of* $w(I)$ *and* $g(w)$*, functions* $g^g(R, D)$ *and* $g^h(R, D)$ *are continuous and satisfy the following:*

- $\frac{\partial g^g}{\partial D} > 0$ *and* $\frac{\partial g^g}{\partial R} < 0$
- $\frac{\partial g^h}{\partial D} < 0$ *and* $\frac{\partial g^h}{\partial R} > 0$
- $g^h(0, D) = 0$ *and* $g^g(\bar{R}(D), D) = g^h(\bar{R}(D), D)$
- $\lim_{R \to \bar{R}(D)} \frac{\partial g^g(R)}{\partial R} = -\infty$ *and* $\lim_{R \to \bar{R}(D)} \frac{\partial g^h(R)}{\partial R} = \infty$

Proof. By the implicit function theorem, $\frac{\partial I}{\partial D} = \frac{1}{1+w'R}$ and $\frac{\partial I}{\partial R} = \frac{1+w}{1+w'R}$. In the stable solution $1 + w'R > 0$, whereas in the WGC solution $1 + w'R < 0$. Also $g_w < 0$ and $w_I < 0$, which proves the first two points. The WGC solution with $R = 0$ simply has $I = 0$ and $w(I) \to \infty$, which leads to $g = 0$. $\bar{R}(D)$ is defined by the level such that $D - \frac{l}{v}R - I$ and $w(I)R$ are tangent to each other so there is one unique solution, which means that $g(\bar{R}(D), D) = g^h(\bar{R}(D), D)$. Finally, $1 - w'R \to 0$ as $R \to \bar{R}(D)$, which means that $\frac{\partial I}{\partial D} = \frac{1+w}{1+w'R}$ goes either to $+\infty$ or $-\infty$ since the numerator is always positive. □

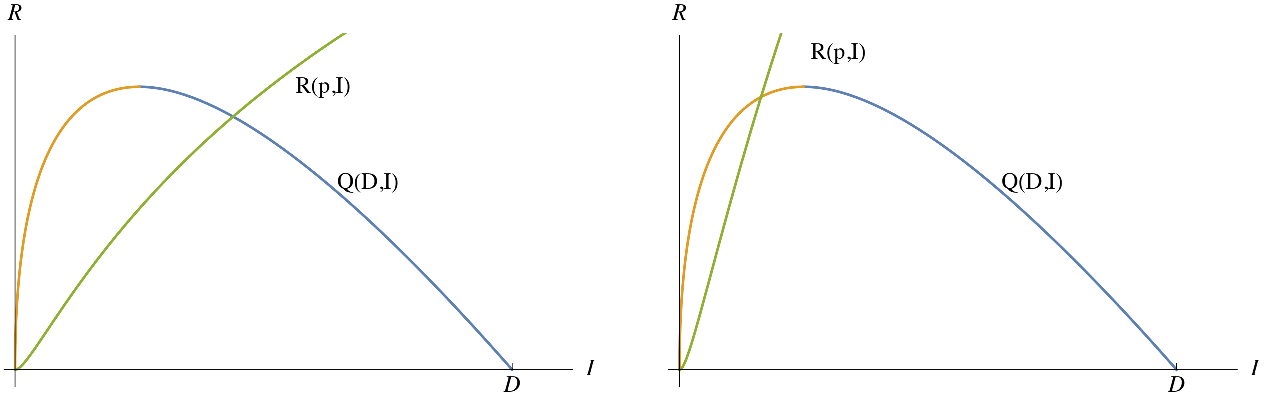
3.2 Equilibrium given prices

So far we have talked about matching equilibria, given a number of ride requests and drivers. But both of these quantities are endogenous once we take into account agents decisions. Our next task is to find the equilibrium given agents decisions.

In order to see this, start with passengers' decisions. Substituting $w(I)$ into the number of ride requests yields $R(p, I) = \lambda g(w(I))r(p)$. This is a demand equation which also depends on the number of idle drivers, since more idle drivers lower waiting times and increase the number of ride requests. In equilibrium the demand of rides must equal the number of rides provided $Q(D, I)$:

$$R(p, I) = Q(D, I) \tag{2}$$

In order to find a solution, start with Figure 2 and add $R(p, I)$, as in Figure 3. Note that two very different behaviors are possible. Let $\bar{R}(D)$ be the maximum value of $Q(D, I)$ and $\bar{I}$ the value that maximizes it. Also let $\bar{p}$ be the price such that $R(\bar{p}, \bar{I}) = \bar{R}(D)$, which we call the *WGC threshold*. If $p > \bar{p}$, there is a unique solution to equation (2) with a good equilibrium (Figure 3a). There could also be solutions with WGCs, but in this case we will restrict our analysis to the good equilibrium.[6] On the other hand, if $p < \bar{p}$, the lower price shifts the whole curve to the right and there is no good solution, but there is at least one solution with WGCs (Figure 3b). It is thus clear that one way to avoid WGCs is by always setting the price high enough.



(a) Good solution

(b) Solution with WGC

Figure 3: Solutions to the demand equation

In case of multiple equilibria, we will restrict our analysis to the solution with the greatest R. With these restrictions, (2) implicitly defines a demand function $\hat{R}(D, p)$.[7] This function is

[6] Intuitively these may occur especially when driver supply is very elastic. In that case, even if the market might potentially be healthy, there are also self-reinforcing equilibria where the lack of other drivers creates WGCs, lower earnings and further reduce the number of drivers.

[7] The highest solution is stable. In order to see this, note that in equilibrium the number of people requesting a ride per unit time times the average time a ride takes must be equal to the number of busy drivers. The equilibrium equation is thus $D - I = (1 + w(I))\lambda g(w(I))r(p)$. The equilibrium is stable if the left hand side crosses the right hand side from below. In order to see that suppose that the right hand side is too large. Then the number of ride requests is higher than in equilibrium, whereas the number of busy people and thus the number of new idle drivers is lower than in equilibrium. These are both balancing forces.

There is at least one stable solution as long as $\lim_{w \to \infty} wg(w) = 0$, since this ensures at least one crossing from below. This is the case if the distribution of willingness to wait has a right tail that is thinner than a Pareto distribution with $\alpha = 1$, which in turn is the same as saying that the mean willingness to wait is finite.

With a distribution of waiting times, the equivalent condition is that the expected value of $w(I)g(w(I))$ converge to zero as $I \to 0$, which can be written as $\int wg(w)h(w|I)dw$. Here $h(w|I)$ is the pdf of waiting time given some density of idle drivers. A sufficient condition for convergence (assuming $\lim_{w \to \infty} wg(w) = 0$) is that for all δ and for all W there exists I such that $\int_0^W h(w|I)dw < \delta$. In that case, for any $\delta > 0$, choose W such that $wg(w) < \frac{\delta}{2}$ for all $w \geqslant W$, and choose I such that $\max wg(w) \int_0^M h(w|I) < \frac{\delta}{2}$. Then the integral is the sum of two terms that are less than $\frac{\delta}{2}$, so it is less than δ.

continuous, increasing in D, and decreasing in p. An equivalent form to define $\hat{R}(D,p)$, which is useful to show the next results, is as the highest solution to

$$R = \lambda g^{e}(R,D)r(p), \tag{3}$$

where $e \in \{g,h\}$.

The number of working drivers depends on the number of ride requests as well as on prices, as can be seen from the following equation that equates marginal cost to hourly earnings:

$$DC'(D) = (1-\tau)pR \tag{4}$$

The implicit solution to this equation defines a supply function $\hat{D}(R,p)$, which is concave if we assume an increasing elasticity of supply.[8]

Given p, τ, and λ, an equilibrium is a joint solution to (2) and (4). It can be seen graphically as the intersection in the (R,D) plane between the two loci $\hat{R}(D,p)$ and $\hat{D}(R)$ (see Figure 4).[9]

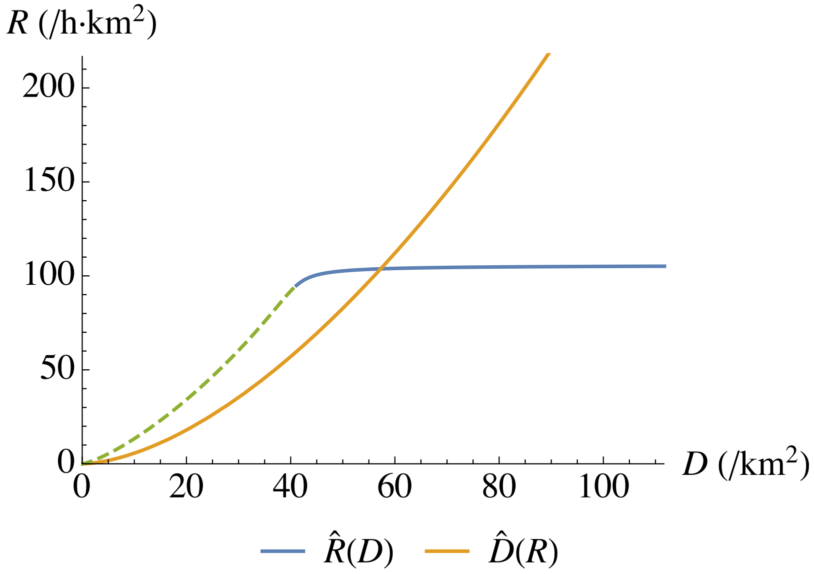


Figure 4: Equilibrium. The green dashed line represents WGC equilibria on the passengers' side.

[8] Since $\frac{\partial \hat{D}}{\partial R} = \frac{(1-\tau)p}{C'(D)} \frac{1}{1+\frac{1}{\epsilon_D}}$. $C'(D)$ is increasing in D, and ϵ_D is decreasing as long as the elasticity of supply is increasing, so $\frac{\partial \hat{D}}{\partial R}$ is decreasing.

[9] There is at least one stable solution if the distribution of willingness to pay has a thinner tail than a Pareto distribution with $\frac{\alpha}{\alpha-1} = 1 + \frac{1}{\epsilon_D(0)}$ (i.e., $\alpha = \epsilon_D(0)+1$). In order to see that, note that the supply equation implies that $\frac{\partial \log \hat{D}}{\partial \log R} = \frac{1}{1+\frac{1}{\epsilon_D}}$, which means that the inverse supply curve has $\frac{\partial \log R}{\partial \log \hat{D}} = 1 + \frac{1}{\epsilon_D}$. On the other hand, as $D \to 0$ $I \to 0$ since $I < D$, so $w(I) \to \infty$ and $w = \frac{D}{R}$. Thus, the demand equation in this limit can be written as $R = \lambda g(D/R)r(p)$, which implies that $\frac{\partial \log \hat{R}}{\partial \log D} = \frac{\epsilon_w^g}{\epsilon_w^g - 1}$. Under the previous condition on the tail of the willingness to wait, $\frac{\epsilon_w^g}{\epsilon_w^g - 1}$ converges to a value less than $1 + \frac{1}{\epsilon_D}$, so the demand equation is above the inverse supply equation for $D <<$. Also the inverse supply equation increases without bound, whereas the demand equation is bounded above at $\lambda r(p)$, which means that they must cross at least once in the right direction.

3.3 Pricing

Suppose first that prices are set to maximize welfare, given by

$$W = \lambda g(R, D) r(p) \bar{u}(p) - C(D) \tag{5}$$

where $\bar{u}(p) = \frac{1}{r(p)} \int_p^\infty r(p')\, dp' + p$ is the average gross utility of those passengers that get a ride. This choice variables are p and τ, but this can be reparameterized as a choice of r and D, as in Weyl (2010): $W = \lambda \tilde{g}(r, D) r \bar{u}(p(r)) - C(D)$.[10] The first order conditions for this problem can be written as:

$$p^{\star\star} = -\epsilon_r^{\tilde{g}} \bar{u}(r) \qquad p^{\star\star}\,(1 - \tau^{\star\star}) = \epsilon_D^{\tilde{g}} \bar{u}(r) \tag{6}$$

Here and in the rest of the paper, we denote the elasticity of X with respect to Y by ϵ_Y^X.

These first order conditions have an intuitive explanation: the price charged to passengers should be $-\epsilon_r^{\tilde{g}} \bar{u}(r)$, the externality they cause on other riders by increasing their waiting time and reducing the likelihood that they will get a ride. This clearly represents the intuition behind the optimal pricing of capacity for a facility like a movie theatre or a restaurant: the cost users should be charged is the opportunity cost of capacity diverted from other potential users. While driver costs do not explicitly appear in the rider price, they are there implicitly in the same way they would appear in the theatre setting: the decision to build capacity (here raise driver wages) is driven by the expected value this yields. In particular, the price paid to drivers is the positive externality they cause on passengers by increasing the density of drivers and decreasing waiting times. The price to riders thus is based on capacity pricing, while optimal choice of capacity determines the price to drivers. Note too that the optimal price to riders is based on the *average* gross utility rather than on the gross utility of the marginal rider (the price itself).

The elasticities can be rewritten as $\epsilon_D^{\tilde{g}} = \frac{\epsilon_D^g}{1 - \epsilon_R^g}$ and $\epsilon_r^{\tilde{g}} = \frac{\epsilon_R^g}{1 - \epsilon_R^g}$, which means that $\tau^{\star\star} = \frac{\epsilon_R^g + \epsilon_D^g}{\epsilon_R^g}$. Let η be the elasticity of scale of waiting times $\frac{\partial \log w(aR, aD)}{\partial \log a} = \epsilon_R^w + \epsilon_D^w$. A matching technology has increasing returns to scale if $\eta < 0$ so that waiting times fall when both sides are proportionally increased. This in turn implies that $\epsilon_R^g + \epsilon_D^g = \epsilon_w^g(\epsilon_R^w + \epsilon_D^w) > 0$, in which case the optimal value of τ is negative as $\epsilon_R^g < 0$. This implies that in the social optimum there should be a subsidy because of the increasing returns to scale, a fact derived by Arnott (1996) and in a less micro-founded model before him by Douglas (1972) who in turn built on the related model of bus transport by Mohring (1972). These increasing returns in the matching function arise from the economies of density inherent to spatial transportation: a space covered more densely by riders and drivers will result in shorter pick up times and thus more efficient transit.

By using the same reparameterization as above, the profit maximization problem can be written as

$$\Pi = \lambda \tilde{g}(r, D) r \left[p_R(r, D) - p_D(r, D)\right] \tag{7}$$

[10] $\tilde{g}(r, D)$ is the implicit solution in R of $R = \lambda g(R, D) r$ divided by λr.

The FOCs for this problem can be written as

$$-\frac{\epsilon_r^{\tilde{g}}}{1-\frac{1}{\epsilon_r}} = 1 \qquad p_D = \frac{\epsilon_D^{\tilde{g}}}{1+\frac{1}{\epsilon_D}} p_R \tag{8}$$

where p_R and p_D denote the price for passengers and drivers, respectively. Since $p_R = \tilde{u}$, where $\tilde{u}$ is the utility of marginal drivers, the previous two equations can be rewritten as

$$p_R = -\frac{\epsilon_r^{\tilde{g}}}{1-\frac{1}{\epsilon_r}} \tilde{u} \qquad p_D = \frac{\epsilon_D^{\tilde{g}}}{1+\frac{1}{\epsilon_D}} \tilde{u}, \tag{9}$$

which we can directly compare with the FOC for welfare maximization. As usual in multi-sided markets, profit maximizing prices have two distortions compared with welfare maximizing prices (Weyl, 2010). First, there is a Spence (1975)-Sheshinski (1976) distortion: first order conditions only take into account the utility of price-marginal riders and not the surplus of the price-average riders.[11] This distortion biases both prices downwards. Second, there is a markup term that biases passengers' price upwards and drivers' price downwards, since a profit maximizer wants to widen the gap between both prices. The net effect is that drivers' price unambiguously decreases, whereas there is an ambiguous effect on passengers' price (the mark-up raises the price, but the Spence distortion lowers it). We return to the implications of these results in Subsection **??** below.

We now fix τ, which more closely resembles the day to day problem of surge pricing given that most ride hailing apps do not dynamically adjust their proportional extraction dynamically. We will consider three problems: welfare maximization, profit maximization, and ride number maximization. We did not analyze the unconstrained ride number maximization problem because in that case the optimal number of drivers is unbounded.

Starting with the equilibrium conditions (3) and (4), the comparative statics give the following equation:

$$\begin{pmatrix} 1-\epsilon_R^g & -\epsilon_D^g \\ -1 & 1+\frac{1}{\epsilon_D} \end{pmatrix} \begin{pmatrix} d\log R \\ d\log D \end{pmatrix} = \begin{pmatrix} -\epsilon_r \\ 1 \end{pmatrix} d\log p, \tag{10}$$

after which the result is:

$$\frac{d\log R}{d\log p} = \frac{-\epsilon_r(1+\frac{1}{\epsilon_D}) + \epsilon_D^g}{\Delta} \qquad \frac{d\log D}{d\log p} = \frac{1-\epsilon_r-\epsilon_R^g}{\Delta}, \tag{11}$$

where $\Delta = (1-\epsilon_R^g)(1+\frac{1}{\epsilon_D}) - \epsilon_D^g$ is the determinant of the matrix. This result leads to the following lemma:

Lemma 2. *The optima for constrained welfare, profit, and ride number maximization are not in the WGC*

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

region. The optima for unconstrained profit and welfare maximization are not in the WGC region.

Proof. Note that in the WGC region $\epsilon^g_R > 1$, $\epsilon^g_D < 0$, and $\epsilon^g_R + \epsilon^g_D < 0$. So the determinant is negative, and the number passengers increases as long as the elasticities of g are large enough. Thus, a price increase always increases both profits and the number of rides.

For welfare, there is a tradeoff: average utility and rides increase, which means that gross utility increases, but cost also increases since the number of drivers increases. In order to look at which effect dominates, note that the change in welfare can be written as $\frac{dW}{dp} = R\frac{d\bar{u}}{dp} + \bar{u}\frac{dR}{dp} - C'(D)\frac{dD}{dp}$. Note first that $\frac{d\bar{u}}{dp} = \epsilon_r\left(\frac{\bar{u}}{p} - 1\right)$. Plugging in the previous results from the comparative statics yields an expression whose numerator is $\frac{\bar{u}}{p}(\epsilon_r((1-\epsilon^g_R)(1+\frac{1}{\epsilon_D}) - \epsilon^g_D) + \epsilon^g_D - \epsilon_r(1+\frac{1}{\epsilon_D})) - \epsilon_r((1-\epsilon^g_R)(1+\frac{1}{\epsilon_D}) - \epsilon^g_D) - (1-\tau)(1-\epsilon_r-\epsilon^g_R)$. $\frac{\bar{u}}{p} > 1$ and $-\epsilon_r(1+\frac{1}{\epsilon_D})) > \epsilon^g_R$ yields the desired result: $(\frac{\bar{u}}{p} - 1)\Delta < 0$, $\frac{\bar{u}}{p}\epsilon^g_D - (1-\tau)\epsilon^g_R < 0$ and $-\epsilon_r(1+\frac{1}{\epsilon_D}) - (1-\tau)(1-\epsilon_r) < 0$.

Since no constrained optimum is in the WGC region, the unconstrained problem is never in the WGC region. The only remaining case is unconstrained welfare maximization, in which case τ might be negative, so the previous proof does not work. But in this case there is a Pareto improvement by giving away free rides to get to the good equilibrium with higher number of rides, which means that the optimum cannot be in the WGC region. □

Lemma 3. *The optimal price for constrained profit maximization is above the optimal price for constrained welfare and ride number maximization.*

Proof. Profit maximization is equivalent to maximizing $R(p)p$, which is the number of rides times an increasing function of prices. Thus, its maximum is above the maximum for ride number maximization.

Welfare maximization can be written as $\max R(p)p\frac{\bar{u}(p)}{p} - f(pR(p))$, where f is an increasing function, since the number of riders is an increasing function of the total hourly earnings, which is $\frac{1-\tau}{\tau}$ times the profits. $\frac{\bar{u}(p)}{p}$ is a decreasing function,[12] so without the final term this function would be maximized at a lower price than profits. And for prices below the optimal price for profits the final term is decreasing, which implies that the optimal price is even lower. □

The last two lemmas mean that the optimal prices for constrained maximization of welfare, profits, and ride numbers are bounded by the lowest price that leads to WGCs and the optimal price of profits. We will now look at how far these two bounds are from each other.

The profit maximization problem is $\max \Pi(p) = \tau R(p)p$, with first order condition $\tau R(p) + \tau p R'(p) = 0$. After substituting the expression for the elasticity of R with respect to prices, we obtain an expression with numerator $1 - \epsilon_r - \epsilon^g_R$, where $\epsilon_r = \frac{\partial \log r}{\partial \log p}$. This is the same numerator as in the elasticity of drivers, which is no coincidence: the number of riders is an increasing function of the total hourly earnings, which is $\frac{1-\tau}{\tau}$ times the profits.

[12] It is easy to check that it is constant for a constant elasticity of demand, and for any given price a function with increasing elasticity has a lower value of $\bar{u}(p)$ than the constant elasticity supply curve with the same elasticity at p.

Analyzing the expression $1 - \epsilon_r - \epsilon_R^g$ leads to some insights. First of all, as prices converge to the WGC threshold from the right, ϵ_R^g converges to negative infinity, which means that increasing price increases profits. Further price increases lead to a quick decrease in $-\epsilon_R^g = -\epsilon_w^g \epsilon_I^w \epsilon_R^I$: $\epsilon_I^w = \frac{1}{2}$, ϵ_I^w decreases with price assuming Myerson (1981) regularity of willingness to wait, and the marginal effect of additional drivers on the density of idle drivers very quickly reaches a small value. This means that, as long as $1 - \epsilon_r < 0$, the FOC $1 - \epsilon_r - \epsilon_R^g = 0$ is satisfied very close to the WGC threshold. On the other hand, with $1 - \epsilon_r > 0$ the price has to go up until some point in which demand is elastic, which might be far from the WGC threshold.

3.4 Discussion

The possibility of WGCs aligns social and private incentives in the sense that both the planner and a monopolist wish to keep prices (constrained or unconstrained) above the level leading to WGCs. However, it does not perfectly align them: a monopolist will still set (constrained or unconstrained) higher prices than will a planner, which would (unconstrained) want to subsidize travel. In what follows we will quantitatively explore the relative size of these effects. Before turning to this, it is worth briefly considering, however, the mechanism that drive WGC and the harms it creates, because these make WGCs potentially more harmful in our setting than hypercongestion in the traffic flow literature.

In the traffic context, hypercongestion reduces the capacity of a fixed roadway to serve cars, lengthening travel times. While this is not fully self-correcting, travel times cannot increase too dramatically as travelers will either choose not to travel or find a different route. A more severe failure is possible in the context of ride hailing. As WGCs lengthens wait times, it may discourage drivers more rapidly than it drives off passengers. If so, the system may enter a downward spiral: as wait times lengthen, more cars exit the road because of reduced earnings than passengers are discouraged from requesting rides. This worsens WGCs and perpetuates the vicious cycle. Even if this cycle does not cause complete market collapse, it can create a feedback loop that makes ride hailing systems highly sensitive to WGCS, well beyond what occurs in traffic flow with hypercongestion. We will explore this dynamic quantitatively in the next section.

4 Surge Pricing

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

4.1 Calibration

We calibrate the parameters of our model by using aggregate data from a large ride-hailing platform in Manhattan for the week between October 3 and 7, 2016. We focus on weekdays between 8 am and midnight. For each one hour period we observe the number of trips, the average trip distance, the average trip time, the number of sessions (i.e., the number of users who opened the app and saw the screen to request a ride), the total number of hours that drivers spent working, and the average surge multiplier. We calibrate the primitives of the model based on this data.

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

For drivers' cost function we assume a constant elasticity supply: $C(D) = A\left(\frac{D}{A}\right)^{1+\frac{1}{\epsilon_D}}$. We assume an elasticity of 1.5 based on research that is presently not publicly available but will soon be released using data from the ride-hailing platform to which we calibrate. This represents a medium-term elasticity of driver supply across different hours of the day that are anticipated to have different demand levels, as this corresponds to the counter-factual we focus on below. Very short-term elasticities, for unexpected demand shocks, are likely to be lower and very long-term elasticities, for secular changes in earnings on the platform, are likely to be higher. Since we observe the number of drivers and trips, as well as the average surge multiplier, we can compute the expected hourly earnings and back out the value of A. Finally, we also observe a database with the average waiting time as a function of the distance to the matched driver for batches of 100 m. We fit the waiting time to this data as shown in Figure 1.

We use average values over the whole period we observe to calibrate the primitives of the model. Thus, this can be thought of as the “average" behavior of the Manhattan market. This is the main specification we use. In a separate specification, we model two different markets, the one between 11 am and noon, which we call the weak market, and the one 6 and 7 pm, which we call the strong market. We assume that for these two markets all the model primitives stay the same as for the average market, except for λ and A. Table 1 compares the average number of drivers, sessions, and trips, as well as the calibrated parameter A, for the weak, strong, and average market. The number of sessions, trips, and drivers are greatest for the strong market

Market	λ (sessions/h $\cdot$ km^2)	R (trips/h $\cdot$ km^2)	D (drivers/km^2)	A (drivers/h $\cdot$ km^2)
Mean	191.9	86.1	55.9	326.7
Strong	265.8	130.8	73.6	314.8
Weak	136.9	61.6	46.7	393.9

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

and the least for the weak market. The supply shifter A goes in the opposite direction, although the differences are not very large.[13] This makes sense since more people would like to work at normal working hours like 11 am instead of during hours when most drivers would prefer to be on the way home or with their families.

4.2 Quantitative analysis of pricing

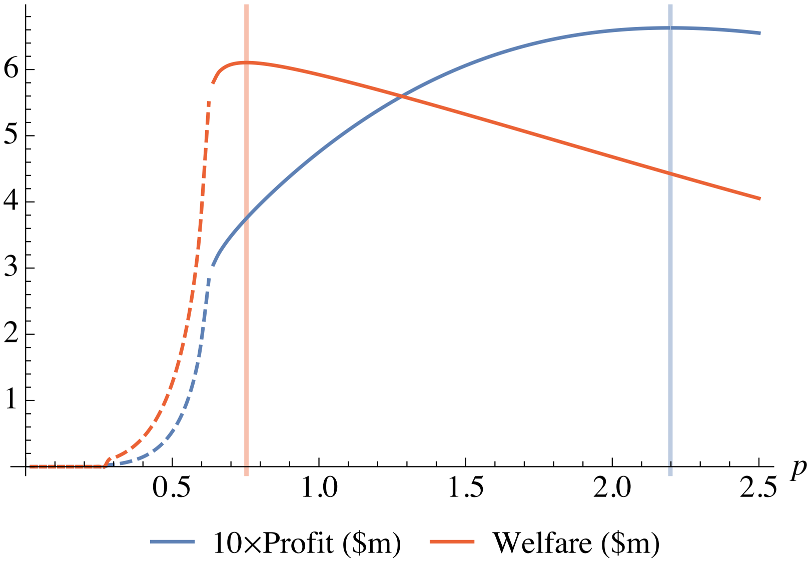


Figure 5: Profits and welfare for the Manhattan market as a function of price for passengers. Dashed lines represent points to the left of the WGC threshold. The vertical lines represent the optimal prices for the function with the corresponding color.

Figure 5 shows how profits, welfare, and rides behave as a function of passengers' price for fixed $\tau = 0.238$, which corresponds to the average value used by our ride-hailing platform in Manhattan.[14] The left region with dashed lines represents prices at which WGC occur. The main thing to note is the asymmetry of the welfare function around its maximum. There is a drastic drop in welfare to the left of the WGC threshold. This is evidence that WCG equilibria can lead to dramatic welfare losses and are "Pareto dominated" in the sense that WGCs in aggregate hurt

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

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

all of drivers, riders and the platform (though they may slightly benefit some price marginal riders who are willing to wait a long time). To the right of the threshold, any price increase benefits some group (typically drivers and the platform) and hurts others (typically passengers), and since there is a tradeoff changes in welfare are not too large: a 20% increase in prices from the optimum only decreases welfare by less than 5%. On the other hand, price decreases in the WGC region hurt everyone, which explains why a 20% decrease in prices from the optimum leads to a 50% decrease in welfare. Finally, note that optimal prices are just a bit above the hypercongestion threshold. Any further increase in prices results in social welfare losses due to too many drivers working and a waste of time.

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

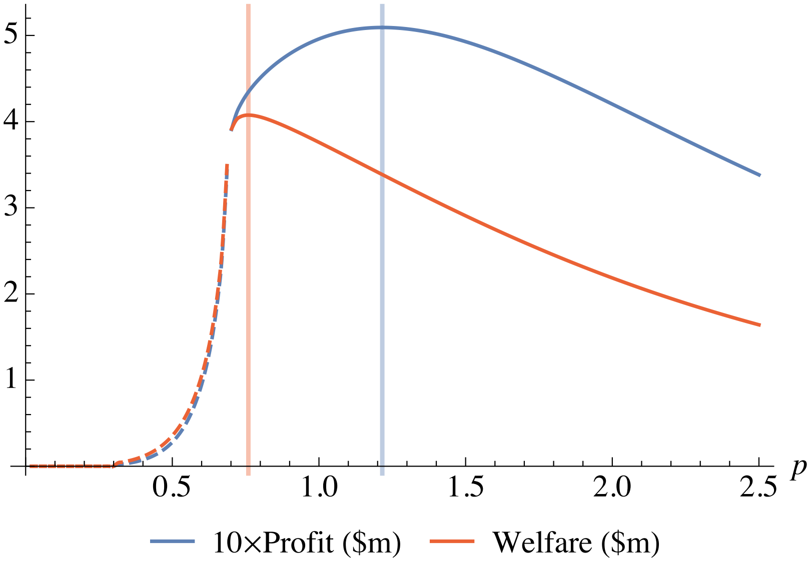


Figure 6: Profits and welfare for the Manhattan market assuming elasticity twice the one in Cohen et al. (2016). Dashed lines represent points to the left of the WGC threshold. The vertical lines represent the optimal prices for the function with the corresponding color.

For this calibration the threshold is in the inelastic part of $r(p)$. By the reasoning in section 3.3, the profit maximizing price is in the elastic region, which starts at around price 2.4. Even in this case, the profit function has a kink at the threshold, which means that there is a dramatic deterioration of profits once WGC start to take place. Furthermore, the effect on welfare of setting the very high profit maximizing price is mild compared with the potential effect of a WGC. This corresponds to a 190% price increase from the welfare optimum that decreases welfare by 18%, which is the same decrease that would be caused by a 17% price decrease from the optimum.

The elasticity estimates from Cohen et al. (2016) are based on studying the effects of price increases that last only a few minutes typically on ride requests. They are thus unlikely to reflect what would happen if the platform consistently set prices as high as 2.4. Figure 6 shows the same calibration, assuming that elasticities are twice those in Cohen et al., i.e., around 0.8-1.2 for prices between 1 and 2. We believe this to be a much better illustration of the way the actual market behaves when prices are predictable and medium-to-long-term adjustments (e.g. switching to another ride-hailing platform or driving to work) are made to these prices by riders. Note first that the general form of the welfare function does not change much. The elastic region starts at 1.2, which is the profit maximizing price. Profit and welfare maximizing prices are now close to each other, and more importantly, changes in welfare and profits are not substantial for prices between them. On the other hand, both profits and welfare drop dramatically after entering the WGC threshold. Thus, welfare and profit changes between both optima are second order when compared to the changes below the threshold.

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

Given this, from now on we will now analyze the social benefits of surge pricing assuming that the platform maximizes welfare but using the elasticities measured by Cohen et al.. We use this as our central specification because these elasticities are likely to be more correct in terms of the response of riders to relatively short-term price fluctuations in terms of the effects they have on system engineering, but do a poor job capturing platform incentives. By adjusting incentives directly (by assuming welfare maximization) we correct for the tendency of the platform to lower prices to account for longer-term platform growth while maintaining realistic degrees of responsiveness to price changes to determine the effects of pricing on system engineering.

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

Figure 7 shows the results of this analysis. The constrained price is extremely close to the unconstrained price for the strong market. The reason for this is that profits drop much more sharply to the left of the optimum than to the right. Another way to put this is that if the

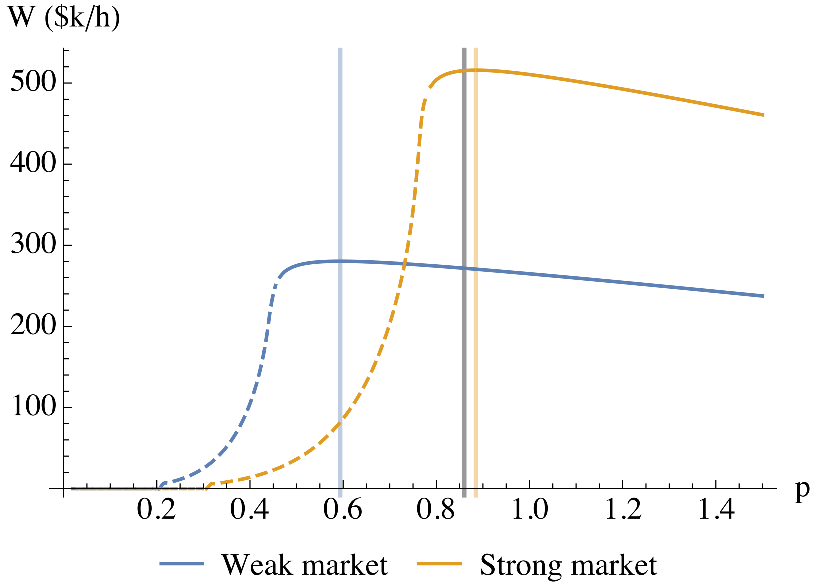


Figure 7: Price discrimination with fixed τ. The gray vertical line represents the optimal price without surge pricing. The blue and yellow vertical lines represent the optimal prices with discrimination for the weak and strong market, respectively.

platform is constrained, it has little freedom to set prices below the strong market unconstrained optimum because it gets close to the WGC threshold, under which welfare in the strong market declines very abruptly. This means that allowing price discrimination leads to a significant reduction of prices in weak markets, whereas it only leads to modest increases in prices for strong markets, as we highlighted in the introduction.

The welfare maximizing price is never above 1, even in the strong market. This is consistent with the data, since this platform only surges 28% of the time between 6 and 7 pm. One might then think that the platform should never surge. The reason there should be surge pricing is because there is substantial spatial variation, as well as between days of the week, and there is a high degree of unpredictability which often leads to high demand and scarcity of drivers. Without surge pricing the platform would have to set prices above 1 in order to avoid the catastrophic consequences of the market being in a WGC at these times of high demand.

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

	Cohen et al. (2016) elasticities					
	W (\$k/h)	Π (\$k/h)	RS (\$k/h)	DS (\$k/h)	ΔU	ΔMLD
Dynamic	398.08	25.32	340.27	32.49	0.005391	0.001076
Static	393.49	27.45	330.84	35.21	0.005158	0.001132
	2× Cohen et al. (2016) elasticities					
	W (\$k/h)	Π (\$k/h)	RS (\$k/h)	DS (\$k/h)	ΔU	ΔMLD
Dynamic	264.23	28.57	198.99	36.67	0.003285	0.001752
Static	256.10	30.43	186.62	39.05	0.003029	0.001788

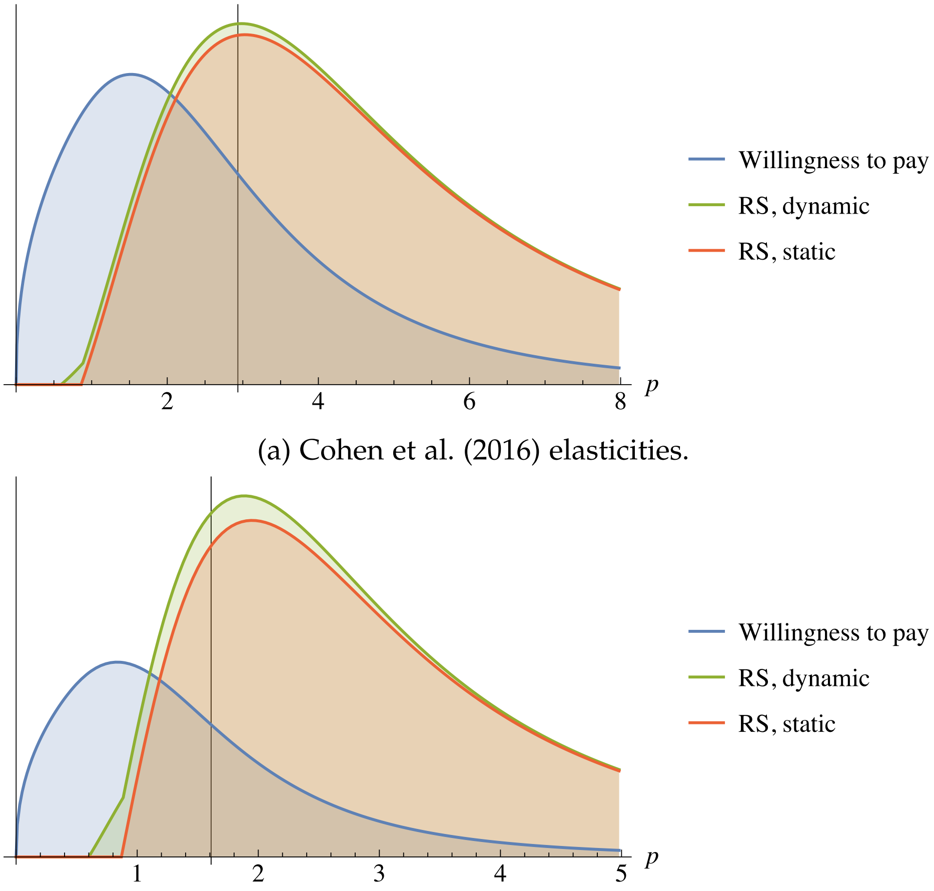
Table 2: Redistribution of welfare with static and dynamic pricing.

4.3 Redistribution of welfare

Surge pricing obviously leads to an increase in welfare relative to static pricing since it is an unconstrained problem. The main question now becomes how it transfers welfare among riders, drivers and the short-run profits of the platform. Table 2 summarizes these results. The first thing to note is that static pricing increases short run profits. This is not surprising given previous results: short-run profit maximization requires bringing prices all the way up to the elastic region of demand, and switching from dynamic to static pricing has almost no effect on price in the strong market whereas it leads to a substantial price increase in the weak market. Static pricing also benefits drivers, which is also due to the fact that prices mostly increase. Passengers' surplus, on the other hand, is higher with dynamic pricing, also due to the fact that static prices are on average higher, which hurts passengers. Thus, dynamic pricing leads to an increase in welfare since it causes a substantial increase in passengers' surplus, but at the cost of decreasing drivers' surplus and profits.

Is the redistribution from dynamic pricing egalitarian? First of all, it decreases the utility of the platform's shareholders, who are most likely concentrated at the upper extreme of the income distribution. It also decreases the utility of drivers, who tend to be towards the lower part of the income distribution. And it increases the utility of passengers, who are relatively wealthy. However, a more detailed analysis shows that the passengers who benefit the most from surge pricing are those who are willing to pay the least, that is, quantities between the weak market prices with dynamic and static pricing.

To see this, Figure 8a shows the distribution of willingness to pay, as well as the distribution of passengers' surplus, assuming the elasticities in Cohen et al. (2016). Clearly the passengers that benefit the most from dynamic pricing are those with willingness to pay between the weak market price of 0.62 and around 4 since the price decrease in the weak market causes an important increase in their surplus, whereas passengers willing to pay a lot only see a small percentage increase in their surplus and therefore surplus almost does not change in the upper tail. The majority of the benefit is below the mean of willingness to pay, and it thus means lower inequality. This, however, is less clear if we assume higher elasticities as in Figure 8b: an important part of the change in consumer surplus takes place above the mean.



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

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

Figure 8: Distribution of income and riders' surplus both for static and dynamic pricing. The black vertical line is the mean of the income distribution.

In order to quantify this, we make the assumption that passengers' willingness to pay is proportional to their income (more precisely, to fix magnitudes, we assume it is 1%).[15]. We then compute the mean log deviation (Theil index) of the original income distribution, as well as the mean log deviation of a new income distribution assuming that each passenger's income increases by their surplus. The changes in mean log deviation are shown in the last column of table 2. The first thing to note is that both dynamic and static pricing increase inequality: they mostly benefit rich people. With the Cohen et al. (2016) elasticities dynamic pricing is less inegalitarian, for reasons we highlighted above. The same happens if we double elasticities, but the effect is less clear, which reflects the intuition from Figure 8.[16]

As a final measure of welfare, we compute the average utility gain of passengers under the

[15] This number only changes the magnitude of our measurements, but not their relative sizes.

[16] As a further note on the dependence on elasticity, if we double once more elasticities, we get to a point in which the change in log deviation with dynamic pricing is 0.001752, whereas the one with static pricing is 0.001788. Thus, in this case static pricing is more egalitarian because in that case most people willing to pay the actual prices are above the mean and improving their welfare increases inequality.

assumption that their utility is the logarithm of their consumption. We also assume that their willingness to pay is 1% of their income. For the Cohen et al. (2016) elasticities, the mean utility without the ride-hailing market is 0.8548, and as shown in table 2 the gain in utility is about 5% larger with dynamic pricing. Using twice these elasticities the initial mean utility is 0.4151, and again the gain in utility is about 5% greater with dynamic pricing. The gain from dynamic pricing is greater with higher elasticities as the high prices necessitated by static pricing limit demand more when demand is highly elastic.

5 Ride-Sharing

In this section we extend our analysis to allow trips to be shared by multiple riders as in the UberPool and Lyft Lines services.

5.1 Model

The main difference between ride sharing and ride hailing is the matching technology. Drivers can now be in one of five states. They can be idle, I, with one rider, B_1, with two passengers, B_2, picking up a rider while empty, K_1, and picking up a rider while driving one rider, K_2. Thus, at any given time the total number of drivers gives the following equation:

$$D = I + B_1 + B_2 + K_1 + K_2 \tag{12}$$

If a new rider requests a ride, he is matched to the nearest driver among those that are idle and those with one rider that go in a similar direction. Let q be the probability that some driver is taking a rider in a similar direction. We assume that this is independent of the state of the system. It is a quantity that depends crucially on how willing is the platform is to deviate a driver that is taking a rider to his destination. The rider that requests a ride thus sees an effective density of drivers $I + qB_1$, which is the density of drivers that could pick him up if he requested a ride. The pick-up time is therefore $w(I + qB_1)$, where the function w satisfies the same properties as in the original model. 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 waiting time wR, which means that $D = I + B_1 + B_2 + w(I + qB_1)R$.

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 $\frac{1}{v}R$, which must be equal to the time spent by drivers with passengers. The time spent driving two passengers counts twice, so this means that $\frac{1}{v}R = B_1 + 2B_2$.

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, $\frac{2\nu}{l}$. The rate at which drivers get a second ride can be written as $\frac{qR}{I+qB_1}$: 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 arrival rate is R, the arrival rate to this area is $\frac{R}{I+qB_1}$, and the probability that the arriving rider goes in the same direction as the old rider is q, which multiplies this rate. Therefore, $B_2 = \frac{qR}{I+qB_1}B_1$.

The supply side is almost the same as with ride hailing, with the only difference that earnings per hour are greater by a factor of $\gamma = \frac{B_1+2B_2}{B_1+B_2}$, since drivers are paid twice when carrying two passengers. The equilibrium condition is then

$$DC'(D) = (1-\tau)pR\gamma(R,D) \tag{13}$$

where $\gamma(R,D)$ arises from the engineering equilibrium, to which we now turn.

5.2 Wild goose chases

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

$$D = I + B_1 + B_2 + w(I + qB_1)R \tag{14}$$

$$\frac{l}{\nu}R = B_1 + 2B_2 \tag{15}$$

$$B_2 = \frac{l}{2\nu}R\frac{qB_1}{I + qB_1} \tag{16}$$

In order to make sense of these equations, fix the number of idle numbers. Solving equations (15) and (16) for B_1 and B_2 tells the proportion of busy time that drivers spend with one or two passengers. Equation (15) simply states that the total time spent with passengers by drivers has to be such that all the requested rides are completed. Equation (16) says that if the number of available drivers with one rider qB_1 is large compared with the total number of available drivers $I + qB_1$, then the balance tilts towards more rides being served by ride shares. Call the solution to these two equations $B_1(I,R)$[17]. This function is continuous, increasing, concave, $B_1(0,R) = 0$, $\lim_{I\to\infty} B_1(I,R) = \frac{l}{\nu}R$, and $\lim_{I\to 0} \frac{\partial B_1(I,R)}{\partial I} = \infty$. Note that we have not made any assumptions about functional forms so far. If we plug this in equation (14) and substitute $B_2 = \frac{1}{2}\left(\frac{l}{\nu}R - B_1\right)$

[17] The closed form solution is $B_1(I,R) = I\frac{\sqrt{1+\frac{4lqR}{\nu I}}-1}{2q}$. It is easier to understand its properties from the inverse function $I(B_1,R) = \frac{B_1^2 q}{\frac{l}{\nu}R - B_1}$

we obtain

$$D = I + \frac{1}{2}\left(\frac{l}{v}R + B_1(I,R)\right) + w(I + qB_1(I,R))R \tag{17}$$

In order to find a solution, we would have liked to do an analogue of Figure 2, but that would require isolating R from equation (17), which is not easy to solve. Instead, we show a straightforward counterpart of what we did before. In Figure 2 we plotted $Q(D,I)$, the number of rides that could be served with D drivers, I of which are idle. Now, in Figure 9, we show $M(R,I)$, the number of drivers needed to serve R riders when there are I idle drivers, which can be thought of as the dual problem to $Q(D,I)$. We assume that this function is quasiconcave,[18] as in Figure 9. This has the same intuition as before: With a very high number of idle drivers, an even greater number of them is required to serve any fixed amount of rides, which explains the behavior in the good region drawn in blue. With a low number of idle drivers, drivers will have to spend most of their time picking up passengers, and therefore a high number of drivers is needed to serve R rides.

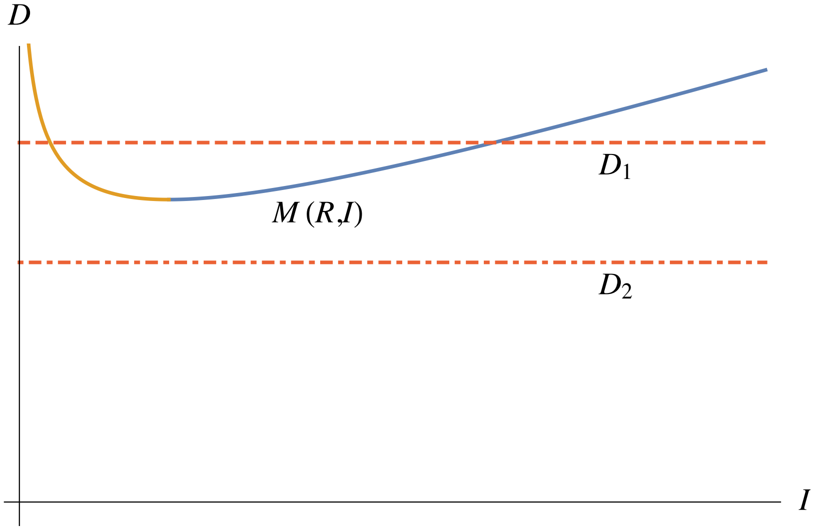


Figure 9: Solutions for the number of idle drivers as a function of the number of drivers and ride requests.

From this equation we can see that a condition for WGCs is:

$$I + qB_1 < \frac{1 + qB_1'}{1 + \frac{1}{2}B_1'}\epsilon_\rho^w wR \tag{18}$$

[18] This is the case, for instance, if waiting time has constant elasticity $|\epsilon_\rho^w| \geqslant \frac{1}{20}$ (ϵ_ρ^w is negative) and $q \geqslant \frac{1}{50}$. If we want to drop the assumption of constant elasticity, this is the case if $|\epsilon_\rho^w| \geqslant \frac{1}{5}$ for all values and $q \geqslant 0.11$. To see this, note that one sufficient condition for there to be at most two solutions is that equating the derivatives of both sides $-w'(I(B_1,R) + qB_1) = \frac{1}{R}\frac{I'(B_1,R)+\frac{1}{2}}{I'(B_1,R)+q}$ must have a unique solution: both B_1 and w are continuously differentiable, so the derivatives of both sides have to be equal at three or more points for there to be four or more solutions. Equating both derivatives gives the previous expression after some straightforward algebra. With constant elasticity, this equation can be written as $a\epsilon_\rho^w(\frac{x}{1-x})^{-1-\epsilon_\rho^w} = \frac{1}{2q} - \frac{1-2q}{q}x + \frac{1-2q}{2q}x^2$ where $x = \frac{v}{l}\frac{B_1}{R}$, and it is not hard to check that for this range or parameters there is a unique solution in the relevant range, $x \in [0,1]$.

The outcome from this matching technology is, given the number of ride requests R and the number of drivers D, what is the equilibrium waiting time $w(R, D)$. It can be computed by finding the value of I that solves 17, and then substituting the value in $w(I + qB_1(I, R))$. There is no closed form solution, but it can be computed numerically. Note that this function depends on q, despite the fact that we do not denote this dependence explicitly.

One key performance measure to compute is the average trip time. For ride hailing this quantity is simply $\frac{l}{v}$, but for ride sharing deviating to pick up or drop off another rider might lengthen the trip. An expression for the average trip time is $T = \frac{l}{v}\frac{B_1+2B_2+K_2}{B_1+2B_2} = \frac{l}{v}\frac{B_1+2B_2+\frac{qB_1}{I+qB_1}wR}{B_1+2B_2}$. The fraction in the first expression is the total time passengers spend in a car divided by the total time they spend while going in the right direction. We assume that passengers request a ride if the sum of the waiting time plus the additional time they spend in a car picking up someone is less or equal to their willingness to wait. Therefore, the fraction of ride requests is $g(w + T - \frac{l}{v})$. From this, we can see that demand is given by

$$R = \lambda g^P(R, D)r(p) \tag{19}$$

where $g^P(R, D) = g(w(R, D) + T(R, D) - \frac{l}{v})$.

6 Conclusion

In this paper we analyze the motivations behind surge pricing in ride-hailing apps. We find that in this context surge pricing is much more important than in apparently similar markets, such as restaurants or films. The main reason is that when prices are too low a perverse equilibrium which we call a *wild goose chase* arises. In this kind of equilibrium a low number of idle drivers leads to deficient matching and long pickup times. Drivers spend too much time picking up passengers instead of driving them or waiting to be matched, resulting in a low number of idle drivers and thus completing a vicious circle. Surge pricing is then a natural tool to avoid WGCs, which are catastrophic for welfare, since they vastly decrease the capacity of the market, thus reducing drivers' and passengers' surplus as well as profits. Absent surge pricing or an engineering solution, uniformly high prices would have to be used which would reduce demand and especially harm riders.

There are many things we plan to add to this paper, both in terms of analysis and measurement in future drafts. On the analytic side, our two primary goals are to calibrate our ride-sharing model to the Manhattan market, so that we can make quantitative analyses like the ones in section 4, and to analyze the alternative mechanisms platforms may use to avoid WGC, such as setting a maximum matching distance or holding a priority queue on the riders' side of the market to maintain a sufficient density of idle drivers to avoid WGCs or periodically rematching passengers and drivers. There are also many results that we derived in the project but have not yet had time to exposit about whether competition (implying lower revenue extraction but lower economies of density) is desirable, optimal regulation and the nature and extent of distortions to pricing on the two sides of the market that we hope to include in the next draft.

However, more importantly, we hope to make more detailed use of micro data from our ride-hailing partner to measure the key predictions of our model. We will first analyze the data in detail to find situations in which we believe WGC might be occurring. We expect performance measures of the market, such as waiting time and the fraction of requested trips served, to drop down significantly in these situations and for this effect to occur steeply over a small range of prices, as in our results from Subsection 4.2. This would be compelling evidence that the phenomenon we are describing is a real issue with important welfare consequences. We will also quantify the importance of surge pricing by computing the welfare decrease if dynamic pricing is not allowed. We would like to make two separate computations. First, how valuable is surge pricing when responding to predictable market characteristics, such as rush hour or the day of the week. Second, how valuable it is to respond to unpredictable changes, such as rain, concerts, or random fluctuations from the steady state.

References

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

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

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

Becker, Gary S. 1991. "A Note on Restaurant Pricing and Other Examples of Social Influences on Price." *Journal of Political Economy*, 99(5): 1109–1116.

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

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

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

Douglas, George W. 1972. "Price Regulation and Optimal Service Standards: The Taxicab Industry." *Journal of Transport Economics and Policy*, 6(2): 116–127.

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

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

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

Mohring, Herbert. 1972. "Optimization and Scale Economies in Urban Bus Transportation." *American Economc Reveiew*, 62(4): 591–604.

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

Myerson, Roger B. 1981. "Optimal Auction Design." *Mathematics of Operations Research*, 6(1): 58–73.

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

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

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

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

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

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

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