PARETO IMPROVEMENTS FROM LEXUS LANES: THE EFFECTS OF PRICING A PORTION OF THE LANES ON CONGESTED HIGHWAYS

JONATHAN D. HALL

Abstract. This paper shows that a judiciously designed toll applied to a portion of the lanes of a highway can be a Pareto improvement even before the revenue is spent. I achieve this new result by extending a standard dynamic congestion model to reflect an important additional traffic externality which transportation engineers have recently identified: additional traffic does not simply increase travel times, but can also introduce additional frictions that reduce throughput. By using a time varying toll to smooth the rate that people depart for work it is possible to avoid these frictions, increasing speed and throughput. Increasing throughput shortens rush hour, which directly helps all road users. However, adding tolls changes the currency used to pay for use of the highway during rush hour from time to money. This change hurts the inflexible poor and most of the time will outweigh the benefit they reap from having a shorter rush hour. We can avoid hurting the inflexible poor by only adding tolls to a portion of the lanes. Doing so preserves their ability to pay with time instead of money. When there are two families of agents, one rich and the other poor, then as long as some rich drivers use the highway at the peak of rush hour then adding tolls to a portion of the lanes is a Pareto improvement. To confirm the real world relevance of this theoretical possibility I use survey and travel time data to estimate the effects of adding optimal time varying tolls. I find that adding tolls to a fourth of the lanes is a Pareto improvement, and that social welfare gains of doing so are over a thousand dollars per road user per year.

1. Introduction

In the ninety years since Arthur C. Pigou introduced the idea that tolls could be used to alleviate traffic congestion, carts have given way to automobiles and congestion has grown to consume 38 hours per commuter annually, nearly an entire work week (Schrank et al., 2012). In addition to the 5.5 billion hours drivers lost to additional travel time in 2011, congestion wasted 2.9 billion gallons of fuel (Schrank et al., 2012), releasing an additional 26 million metric tons of carbon dioxide into the atmosphere,[1] as well as a host of other pollutants. This additional pollution amounts to more than six times the annual emissions

Date: October 29, 2013.
Email: jhall@northwestern.edu.
I am especially grateful for the guidance and support Gary Becker and Eric Budish have given me. This paper is much improved due to discussions with William Hubbard and Ethan Lieber. I am also grateful for helpful feedback from Ian Fillmore, Brent Hickman, Devin Pope, Mark Phillips, Chad Syverson, and George Tolley as well as seminar audiences at the University of Chicago, Northwestern University, RSAI, and Kumho-Nectar.
[1]The Environmental Protection Agency (2011) estimates that the average gallon of gasoline releases 8,887 grams of carbon dioxide.

saved by the current fleet of hybrid and electric vehicles,[2] and is responsible for up to 8,600 preterm births a year (Currie and Walker, 2011). Congestion also retards economic growth; cutting congestion delay in half would raise employment growth by 1% per year (Hymel, 2009).

Despite the significance of these costs, the vast majority of roads remain unpriced. A major barrier to implementing congestion pricing is the received wisdom among economists, policy makers, and the public that congestion pricing makes many, if not most, road users worse off.[3] Lindsey and Verhoef (2008) suggest that "most likely, these losses are the root of the longstanding opposition to congestion tolling in road transport," a view echoed throughout the literature.[4] Most politicians view it as politically toxic[5] and polls almost always find widespread opposition to congestion pricing.[6] As one voter put it, "Turkeys don't vote for Christmas–and motorists won't vote for more taxes to drive."[7] If the median voter believes they will be worse off under congestion pricing, then perhaps it is no surprise congestion pricing hasn't been widely implemented.

Of course the toll revenue isn't lost, and since the received wisdom is that congestion pricing is a Kaldor-Hicks improvement, meaning the winners gain more than the losers

[2]The Environmental Protection Agency (2011) estimates that the typical passenger vehicle emits 5.1 metric tons of carbon dioxide a year, so the additional pollution is equivalent to that of 5 million vehicles. Samaras and Meisterling (2008) estimate that a plug-in hybrid reduces greenhouse gas emissions by 32% compared to conventional vehicles while traditional hybrids reduce greenhouse gas emissions by 29%. The U.S. Energy Information Administration (2013, table 58) estimates that by the end of 2013 there will be 2.73 million hybrid and electric cars and trucks, of which 98.7% are traditional hybrids. While Samaras and Meisterling (2008) estimate the reduction in greenhouse gas emissions over the vehicle lifetime, so these numbers are not perfectly comparable, these numbers imply the current fleet of electric and hybrid cars is equivalent to removing 0.8 million conventional vehicles from the road. Dividing 5 million by 0.8 million gives us the result that we need 6.25 times the current number of electric and hybrid vehicles to counteract the additional pollution due to congestion.

[3]Other barriers include the belief that it is unfair to let some pay with money to get faster travel times and concern that since tolling brings in more revenue for the government it will lead to increased government spending. For an example of the first belief see Malady, Matthew. 2013. "Want to Save Civilization? Get in Line," New York Times, May 31, 2013. When drivers in Southern California were asked why they oppose congestion pricing, in the specific form of allowing solo drivers to pay to travel in carpool only lanes, 40% responded that either the government will waste the money or it will increase government bureaucracy (Fall 1999 Commuter Survey from Sullivan (1999)).

[4]For example, see Starkie (1986); Cohen (1987); Giuliano (1992); Arnott et al. (1994); Lave (1994); Small et al. (2005) and Small and Verhoef (2007).

[5]Ison (2000, 276) finds that in the United Kingdom 80% of local politicians with responsibility for transportation issues, academics who studied such issues, and transport interest groups "view urban road pricing as being publicly unacceptable."

[6] For example, Jones (1991) reports on twelve polls in the United Kingdom and finds widespread opposition to congestion pricing; Harrington et al. (2001) cites a number of surveys in the United States finding opposition to congestion pricing as well as finding 57% of their survey respondents oppose congestion pricing; and Podgorski and Kockelman (2006) find that 70% of Texans oppose pricing existing roads. The notable counterexample is that after congestion pricing has been implemented it generally finds widespread support. For example, in Stockholm they voted to keep congestion pricing after a seven month trial (Hårsman and Quigley, 2010).

[7]Sturcke, James. 2008. "Manchester Says No to Congestion Charging," Guardian, December 12, 2008.

lose, then there are transfers that will make all road users better off, thereby making congestion pricing a Pareto improvement.[8]

Unfortunately, even when we can design transfers that make a policy a Pareto improvement, it can still be difficult to implement. Stiglitz (1998) points out that it may not be enough to identify Pareto improving transfers because the transfers are transparent, and thus harder to defend, and the government cannot commit to maintaining the transfers. This makes policies that naturally generate a Pareto improvement all the more valuable.

The main result of this paper is that, contrary to the received wisdom, a carefully designed toll on a portion of the lanes of a highway can be a Pareto improvement, even before the toll revenue is spent. This means that all road users will be better off regardless of what is done with the revenue.

Before moving on to how I am able to get this new result and the intuition for it, let me first make two clarifying points. First, to price a portion of the lanes we take a highway and split it into two routes using some barrier, often pylons or simply a painted line, and price one of the routes using electronic tolling as envisioned by Vickrey (1963) and seen today in the form of E-ZPass in the Northeastern United States and similar systems elsewhere. This practice is often called value pricing, since drivers have the option of paying more for something of greater value. The priced lanes are called HOT Lanes when solo drivers can pay to access high occupancy vehicle lanes. The acronym stands for High Occupancy/Toll. The lanes have also been given the epitaph of Lexus Lanes to convey the accusation that only those who can afford a Lexus can afford to drive in them.[9]

Second, obtaining a Pareto improvement comes at a cost. We are not pricing all of the lanes and therefore leave some of the potential Kaldor-Hicks efficiency gains unrealized. However, obtaining a Pareto improvement should make congestion pricing more acceptable to the public and politicians and so allow for widespread adoption. Because of this, what we are really doing is trading potential, but unrealized, efficiency gains for actual efficiency gains.

I am able to get this new result by extending the bottleneck congestion model of Vickrey (1969) and Arnott et al. (1990, 1993) to reflect an important additional traffic externality that has been identified by the transportation engineering literature but that has largely been ignored in the economics literature. Not only does each additional vehicle slow others down, but in heavy enough traffic additional vehicles can create additional frictions which reduce throughput. That is, the road will produce fewer trips per unit time.

To understand the two externalities better, consider a two-lane highway that merges down to one lane at some point. The point where the lanes merge is a bottleneck. As a standard highway lane has a capacity of roughly 40 vehicles per minute (Council, 2000)

[8]See Small (1983, 1992) for practical proposals of how to use the revenue to get close to a Pareto improvement.
[9]The empirical evidence is that those of all income levels use the priced lanes, though the rich use them more frequently (Sullivan and Harake, 1998; Sullivan, 2002).

let's assume that at the start of rush hour more than 40 vehicles per minute start down our hypothetical two lane highway. When these vehicles reach the bottleneck they will find that they cannot all get through at once and so they must wait for their turn to use the scarce road space; a queue will form. An additional car that travels during rush hour *lengthens the queue*, increasing the travel time of all those behind him by 1.5 seconds, the amount of time it takes him to go through the bottleneck. This lengthening of the queue is the standard externality. However, what this simple externality fails to capture is the fact that a queue creates additional frictions that *reduce throughput*. Rather than moving 40 vehicles per minute the bottleneck will only move 36, or even fewer, vehicles per minute. What this means is that each car on the highway imposes two externalities, both of which lead to longer travel times: they *lengthen the queue* and *reduce throughput*.[10]

Over fifty years ago economists conjectured that this second externality existed, that too many cars on the road could reduce throughput (e.g., Walters, 1961; Johnson, 1964). Vickrey (1987) even gave it a name, hypercongestion. This was before the transportation engineering literature had identified the causal mechanisms and the economics literature rejected the claim (e.g., Newell, 1988; Verhoef, 1999, 2001; Small and Chu, 2003).[11]

Tolls can prevent throughput from falling by smoothing the rate at which vehicles get on the highway so that the queue stays below the length where throughput falls. Figure 1 gives a stylized example of how this can work.

When the road is unpriced drivers depart from home at rate $r(t)$. At the start of rush hour 48 vehicles per minute depart from home, but the highway's maximum throughput is only 40 vehicles per minute, so a queue forms and travel times start climbing. As the queue gets longer the second externality takes effect and highway throughput falls to just 32 vehicles per minute. As we approach 8:30 the number of vehicles on the highway and travel times climb to their peak. Eventually rush hour will end, so there must be a period of time where more vehicles are getting off the highway than are getting on it. Starting at 8:30 only 8 vehicles per minute depart from home, allowing the length of the queue, and thus travel times, to start falling, until eventually everyone has reached work and rush hour ends at 9:20. This can be an equilibrium because drivers face a trade-off in leaving

[10]This is a bit of a simplification, as when there are just a few cars on the road adding an additional vehicle can reduce speeds while increasing throughput, but will hold exactly in my model. An alternative way of viewing the two externalities that is more accurate but doesn't separate the two externalities as cleanly is to look at the elasticity of speed with respect the number of vehicles on the road, or density. First note that throughput (vehicles/hour) is the product of speed (miles/hour) times density (vehicles/mile); $T = S \times D$. The standard externality is that $\frac{\partial S}{\partial D} < 0$. As long as the elasticity of speed with respect to density, $\epsilon_{S,D} = -\frac{\partial S}{\partial D}\frac{D}{S}$, is less than one, throughput will be increasing in density. However, when $\epsilon_{S,D} > 1$ the additional externality is in force and additional vehicles will reduce throughput.

[11]There is research arguing hypercongestion is possible for urban centers, which is the context in which Vickrey (1987) defined it. See Small and Chu (2003); Arnott and Inci (2010) as well as Fosgerau and Small (2013).

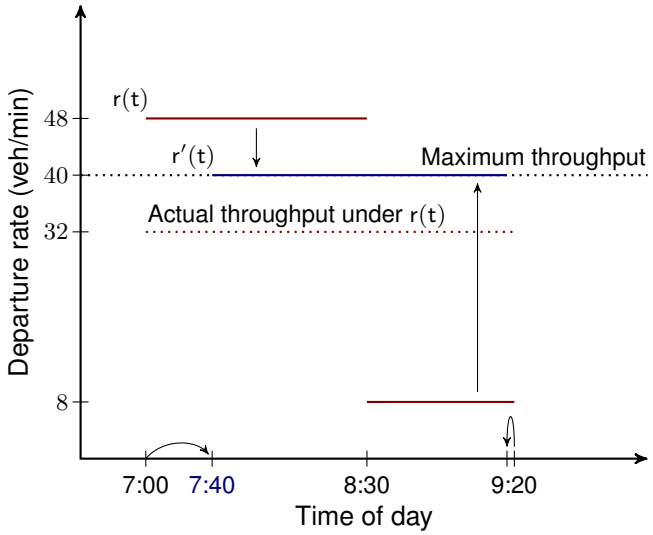


FIGURE 1. Tolls can change when drivers depart and increase throughput

early to avoid traffic but getting to work earlier than desired versus leaving so as to arrive right on-time but enduring a long commute in bad traffic.

Using time varying tolls we can induce drivers to depart at rate $r'(t)$, reducing the departure rate before 8:30 and increasing the rate thereafter. By preventing the queue from forming we eliminate both externalities; there is no queue and throughput remains high at 40 vehicles per minute. Since throughput is higher, rush hour does not need to be as long. In our stylized example, rush hour can start 40 minutes later and end 4 minutes earlier.

By considering the effect of pricing on the first driver to depart in the morning we can take our first look at the welfare impacts of congestion pricing. When the road is free this driver doesn't face any congestion, but will get to work very early. When the road is priced he will not need to leave as early, since rush hour is shorter, still doesn't face any congestion, and for reasons we will see later, pays no toll. This first driver is better off since he has reduced how early he arrives at work.

If all drivers are identical then the first driver to depart being better off means all drivers must be better off; we have obtained a Pareto improvement before spending the revenue.

Once we allow drivers to be heterogeneous then we can no longer use changes to the first driver's welfare as a sufficient statistic for all drivers' welfare, and the analysis gets more complicated. When we add tolls to roads, we change the currency drivers pay with from time to money. However, not all drivers value their time at the same amount, and changing the currency used can hurt some drivers. Consider, for example, a poor driver who traveled at the peak when the road was free. He now faces three choices: he can travel at the same time and pay the toll; travel at a different time when the toll is lower,

increasing how early or late he is to work but avoiding most of the toll; or not travel on this road at all. There are no guarantees he will be better off.

However, we likely can still price some of the lanes and make all road users better off. Pricing some of the lanes allows us to increase total highway throughput, reducing the length of rush hour. Reducing the length of rush hour directly helps all road users. Leaving some of the lanes unpriced allows drivers who prefer to pay with their time to do so. This allows us to avoid harming the poor. As long as some rich drivers are using the highway at the peak of rush hour then we can price some of the highway while still leaving enough unpriced capacity for all of the poor drivers who had been arriving during the peak to still do so.

My main theoretical contribution is showing that it is possible for pricing a portion of the lanes to make all road users better off, even before using the toll revenue. Specifically, I find the set of parameter values for which pricing some or all of the lanes is a Pareto improvement when there are two families of agents. A family is a set of agents with the same value of time and inflexibility (a measure of drivers willingness to trade time early at their destination for travel time) but members of the same family can differ in their desired arrival time. I find an intuitive sufficient condition for value pricing to be a Pareto improvement: as long as some rich drivers use the highway at the peak of rush hour then adding tolls to a portion of the lanes is a Pareto improvement.

My main empirical contribution is to confirm the real world relevance of this theoretical possibility. I allow for a continuum of families and estimate the joint distribution of agents value of time, desired arrival time, and inflexibility for road users on California State Route 91. To the best of my knowledge this is the first time the distribution of inflexibility has been estimated, despite its importance in dynamic congestion models, as well as the first time this joint distribution has been measured. I then use these estimates to evaluate the effects of congestion pricing. I find that the welfare gains from congestion pricing are large. Pricing all of the road increases social welfare by $2,300 per road user per year, but at the cost of hurting some road users by more than $3,000 per year. However, by pricing just a fourth of the lanes we obtain a Pareto improvement while still capturing 43% of the social welfare gains.

I make three additional contributions. First, I introduce to the economics literature the evidence on why throughput falls when roads are congested. Second, I show how the bottleneck model's implicit assumption that throughput is unaffected by pricing explains the differences between the welfare effects of congestion pricing in the bottleneck model relative to other models; differences that have been attributed to the bottleneck model being dynamic. Finally, I extend the bottleneck model to allow for a continuum of desired arrival times. This feature, with otherwise homogeneous agents, was in the initial papers

using the bottleneck model (Vickrey, 1969; Hendrickson and Kocur, 1981), but was subsequently dropped as it didn't affect equilibrium outcomes.[12] However, once agents are heterogeneous along other dimensions then allowing for agents' desired arrival time to be continuously distributed has significant effects on equilibrium outcomes and is vital for matching the model to the data.

In the following sections I will explain the evidence on why throughput falls when roads are congested from the transportation engineering literature (Section 2), extend the standard bottleneck model to allow for the second externality, a continuum of desired arrival times, and multiple routes (Section 3), and work through the model with homogeneous agents both to build intuition for how the model works and to highlight how the welfare effects of congestion pricing depend crucially on whether tolling increases or decreases throughput (Section 5). Then, in order to better address the concerns about the distributional impacts of congestion pricing, I show when pricing part or all of the highway is a Pareto improvement when there are two families of agents (Section 6). I then solve the model with a continuum of families (Section 7) and use those results to to estimate the joint distribution of agent preferences (section 8) and evaluate what would happen were we to price all or part of the road (Section 9).

2. Understanding traffic congestion

This paper depends critically on the claim that tolls can be used to increase highway throughput. To understand this we must first understand the frictions that can occur at bottlenecks that reduce highway throughput. Transportation engineers have identified two causal mechanisms for why a queue behind a bottleneck can reduce highway throughput.

A bottleneck can occur at any place the capacity of a highway decreases, generally because of a reduction in lanes. While the most noticeable bottlenecks are generally the result of lane closures due to construction or an accident, far more common are bottlenecks due to on-ramps. The typical on-ramp is a bottleneck since it is a lane that joins the highway and then ends; it adds vehicles but not capacity.

2.1. **Queue spillovers.**

The first additional friction occurs when the queue behind a bottleneck grows long enough that it blocks upstream exits. Drivers who wish to use these now blocked exits are forced to wait in the line to use the bottleneck, even though they have no need to pass through the bottleneck. Similarly, a queue can grow at a busy off-ramp, spilling onto the mainline of the freeway and blocking through traffic. Vickrey (1969) labeled the first situation a trigger neck and transportation engineers call both situations a queue spillover.

[12]The one other paper to consider agents with a continuum of desired arrival times who are heterogeneous in other dimensions is de Palma and Lindsey (2002), who numerically solve for equilibrium when there are no tolls.

Queue spillovers are the reason that beltways or ring roads that go around major cities, such as I-495, which encircles Washington D.C., and Boulevard Périphérique, which encircles Paris, are especially prone to crippling congestion (Vickrey, 1969; Daganzo, 1996).[13] Muñoz and Daganzo (2002) find that queue spillovers frequently reduce throughput by 25% where I-238 diverges from I-880N outside of San Francisco.

2.2. **Throughput drop at bottlenecks.** In addition, throughput at the bottleneck itself can fall once a queue forms, so that instead of 40 vehicles per minute getting through the bottleneck only 36 or even fewer vehicles get through per minute. On our two lane highway the vehicles in the right lane will need to change lanes before getting to the bottleneck. When traffic is light this is easy, but when traffic is heavy it becomes difficult. At some point a vehicle will come to a stop before merging and force its way over. This is what transportation engineers call a destructive lane change. We can see the damage in two ways. First, the vehicle that forced its way over will be moving very slowly and so go through the bottleneck at a slow speed. Equivalently, he will open up a gap in front of him; this will be a period of time that the bottleneck, the scarce resource on the highway, is not being used. Notice how this contrasts with most queues; while a long line at the grocery store means you will have to wait a while, it does not affect the rate at which customers are served. In fact, most queues increase throughput by avoiding wasted time.

There is a large transportation engineering literature documenting that throughput at bottlenecks drops once a queue forms, which they refer to as the two-capacity hypothesis.[14] All of the papers in the literature have found evidence for the two-capacity hypothesis, though Banks (1991) found at two sites that it only affected the merging lanes, not the entire road. The estimates for the size of the drop range from 2–16%, and are presented in Table 1.[15] This phenomenon has also been modeled in the physics literature in Helbing and Treiber (1998) and Treiber et al. (2000).

All of these papers follow the same basic procedure; they measure the capacity of the bottleneck by identifying bottlenecks that are not constrained by a downstream bottleneck and then measuring flows immediately before a queue forms and while there is a queue. These papers' data come either from video cameras or loop detectors. When using loop detectors they identify when a bottleneck is active and has a queue using either sharp increases in speed or sharp declines in occupancy, the fraction of time a loop detector has a vehicle over it, between consecutive detectors.

[13]There is a animation on Daganzo's website that illustrates this nicely at www.its.berkeley.edu/volvocenter/gridlock/.

[14]The name, "two-capacity hypothesis," refers to the idea that a road has one capacity, or throughput, when there is no queue and a different capacity when there is a queue.

[15]There are papers not expressly testing the two-capacity hypothesis that present results that can be interpreted as evidence for or against the throughput drop at bottlenecks. These includes Hurdle and Datta (1983) who find no capacity drop and Elefteriadou et al. (1995) and Leclercq et al. (2011) who do.

TABLE 1. Findings of transportation engineering literature on throughput drop at bottlenecks

Paper	Throughput drop (%)	Location
Banks (1990)	2.8	I-8, San Diego
Hall and Agyemang-Duah (1991)	5.8	Queen Elizabeth Way, Toronto
Banks (1991)	−1.2–3.2	4 sites in San Diego
Persaud et al. (1998)	10.6–15.3	3 site/time pairs on Toronto Highway 401
Cassidy and Bertini (1999)	7.4–8.7	2 sites in Toronto
Bertini and Malik (2004)	4	US-169, Minneapolis
Zhang and Levinson (2004)	2–11	27 sites in Minneapolis–St. Paul
Bertini and Leal (2005)	9.7	M4, London
	12	I-494, Minneapolis
Cassidy and Rudjanakanoknad (2005)	11.7	I-805, San Diego
Rudjanakanoknad (2005)	13.2	SR-22, Orange County, California
Chung et al. (2007)	12.3	I-805, San Diego
	6.2	SR-24, San Francisco
	5.8	Gardiner Expressway, Toronto
Guan et al. (2009)	15	Fourth Ring Road, Beijing
Oh and Yeo (2012)	8.9–16.3	16 sites in California
Srivastava and Geroliminis (2013)	15	US-169, Minneapolis

Figure 2 shows the relationship between speed and throughput for a bottleneck on I-805N in San Diego. It was created using data from Cassidy and Rudjanakanoknad (2005), who used video recordings of morning traffic to extract second-by-second throughput for each lane at four locations, each 120 meters apart; as well as to measure how long it took to traverse the entire 360 meters, which was measured every five seconds. From the video they are able to verify that vehicle flows through the bottleneck are not constrained by traffic further downstream.

What this means is that Figure 2 maps out a production possibility frontier (PPF). Notice that the PPF bends backwards, that we can have a throughput of 9,000 vehicles per hour at either 28 miles per hour, or 50, and that we can even have a throughput of 10,500 vehicles per hour at a speed of fifty miles per hour.

This PPF bends backwards because throughput falls when a queue forms at the bottleneck. As a measure of whether a queue has formed we can look at the number of vehicles in the rightmost lane, which is represented in Figure 2 by the color of each dot. The dot

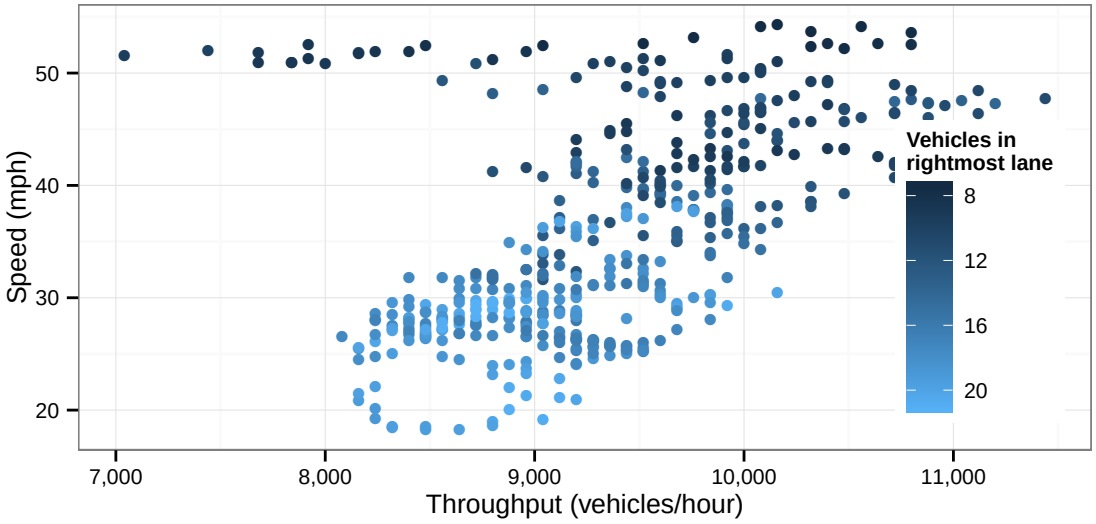


FIGURE 2. Production possibilities frontier implied by observations from I-805N at Palm Avenue on September 17, 2002, 6:08–6:28 a.m. and September 18, 2002, 6:10-6:30 a.m. I take the 45 second moving average of 5 second level data. Data from Cassidy and Rudjanakanoknad (2005).

is red when there are a large number of vehicles in the rightmost lane, and those points invariably are those with low speed and throughput.

3. MODEL

It is by rearranging when people travel within rush hour that tolls can decrease travel times and increase throughput. Because the primary margin on which drivers will adjust is their choice of when to travel, we need to use a model that explicitly includes this choice. The standard model used in the economics literature that does so is the bottleneck model of Vickrey (1969), which was formalized by Arnott et al. (1990, 1993).[16]

This model corresponds to the transportation engineers much beloved hydrodynamic theory of traffic flow (Lighthill and Whitham, 1955; Richards, 1956) when the left side of the density-flow curve is linear, as well as the widely used cell transmission model (Daganzo, 1994). The key difference is that the bottleneck model doesn't consider what happens within the queue or how long the queue grows, while the transportation engineers' models deal explicitly with those issues at the cost of greater complexity. The models all make the same predictions for travel times (and so average speed) as a function of the history of departures and so for the questions I will address in this paper the simplifications the bottleneck model makes are innocuous.

[16]See Arnott et al. (1999) for a survey of the literature which uses this model.

I will make three important modifications to the model. The first is to add the second externality by allowing throughput to fall when a queue forms. This is a natural way to model the throughput drop at bottlenecks and serves as shorthand for the effects of queue spillovers.[17] The second modification is to allow the social planner to choose the fraction of the lanes that are priced; this goes beyond existing work, such as van den Berg and Verhoef (2011), which considers the welfare implications of pricing a fixed portion of the lanes. The final change is to allow agents' desired arrival time at work to be continuously distributed, as in Vickrey (1969); Hendrickson and Kocur (1981); de Palma and Lindsey (2002).

Allowing for a continuum of desired arrival times is important because it allows drivers to be inframarginal, meaning that if the cost of their chosen arrival time increases, they do not change when they arrive. This matters primarily because it is necessary to match the data. The evolution of travel times across the day suggests that the marginal driver at any point it time is quite willing to change when they arrive in order to save just a little travel time, which means the marginal driver cannot be a shift worker. A model that does not allow for inframarginal drivers must either make ridiculous predictions about the evolution of travel times or not contain agents with very inflexible schedules.

3.1. **Congestion technology.** There is a single road connecting where people live to where they work; this road has two types of routes: tolled and free. The social planner chooses the relative size of each route as well as a time varying toll schedule to maximize social welfare taking total road capacity and consumer preferences as given. Let λ_{toll} and λ_{free} denote the fraction of capacity devoted to each route, where $\lambda_{\text{toll}} + \lambda_{\text{free}} = 1$.[18] Travel along this road is uncongested, except for a single bottleneck through which at most s^* cars can pass per unit time. Letting r denote the route and t the time of departure from home, when the departure rate on a route, $r_r(t)$, exceeds its capacity, $\lambda_r \cdot s^*$, a queue develops. Once the queue is more than ϵ vehicles long the throughput of the bottleneck for that route falls to $\lambda_r \cdot s = \lambda_r \cdot s^*$, where $s \leq s^*$. Therefore, queue length, measured as the number of vehicles in the queue, evolves according to

(1)
$$\frac{\partial Q_r(t)}{\partial t} = \begin{cases} 0 & \text{if } Q_r(t) = 0 \text{ and } r_r(t) \leq \lambda_r \cdot s^*, \\ r_r(t) - \lambda_r \cdot s^* & \text{if } Q_r(t) \leq \epsilon \text{ and } r_r(t) > \lambda_r \cdot s^*, \\ r_r(t) - \lambda_r \cdot s & \text{if } Q_r(t) > \epsilon; \end{cases} \qquad r \in \{\text{free}, \text{toll}\}.$$

[17] Under some very specific assumptions about the structure of the road network (Y-shaped network) and distribution of destinations (constant over time) a model of queue spillovers maps exactly into this model.

[18] Implicit in this is the assumption it is costless to split the road into two routes. In reality some separation between the priced and unpriced lanes is required. The Federal Highway Administration recommends a three to four foot buffer when a pylon barrier is used (Perez and Sciara, 2003, p. 39-40) and on I-394 in Minnesota there is a two foot buffer without any barrier (Halvorson and Buckeye, 2006, p. 246). As federal standards call for twelve foot lanes on interstates (AASHTO, 2005, p. 3), splitting the road into two routes could cost as much as a third of a lane. This space can come from narrowing the existing lanes at the cost of reducing the design speed of the highway or the highway could be widened by a few feet.

We then simplify by taking the limit as $\epsilon \to 0$, so throughput on a congested route is constant.[19]

It is in allowing $s < s^*$ that we add in the empirical finding of a throughput drop at bottlenecks, and allowing λ_{toll} to be any number between zero and one, rather than just zero or one, allows us to consider pricing a portion of the lanes, rather than just all or none.

Travel time along route r for an agent departing at t is

$$T_{d,r}(t) = T^f + T^v_{d,r}(t) \qquad r \in \{\text{free}, \text{toll}\}, \tag{2}$$

where T^f is fixed travel time, the amount of time it takes to travel the road absent any congestion, and $T^v_{d,r}(t)$ is variable travel time for route r. Variable travel time is only due to queuing and is the length of the queue divided by the rate at which cars leave the queue

$$T^v_{d,r}(t) = \frac{Q_r(t)}{\lambda_r \cdot s}. \tag{3}$$

For simplicity, and without loss of generality, we will assume that $T^f = 0$. Throughout the rest of this paper when we discuss travel time we are only referring to the variable, congestion related, travel time.

It will be simpler to focus on arrival times instead of departure times, so define $T_r(t)$ as the travel time on route r that an agent *arriving* at t. Because this model is deterministic there is a one-to-one mapping between departure times and arrival times, and so doing so is innocuous.[20]

The production possibilities frontier (PPF) of the bottleneck model is shown in Figure 3. The solid line is the PPF, while the dotted line shows speed-flow combinations that are possible even though they are not on the PPF. The PPF is horizontal up to s^* because up till the point which the bottleneck is binding there is no congestion. Once the bottleneck is binding, throughput falls to s and travel times climb as the queue grows. Since travel time is simply total distance divided by average speed, this means average speed is falling. For different queue lengths there will be different average speeds, all of which have throughput of s. Thus the dotted line is vertical.

[19]This small bit of mathematical jiu jitsu allows us to keep the model simple while avoiding existence of equilibrium problems which can occur when given that the route will be congested, the equilibrium departure rate is too low to create congestion, but when the route is uncongested the equilibrium departure is high enough to create congestion.

[20]If an agent departs at t_d then he will arrive at $t = t_d + T_d(t_d)$, and so $f(t_d) = t_d + T_d(t_d)$ is the function that maps between departure and arrival times. As long as $dT_d/dt_d \neq -1$ then f will be one-to-one and so invertible.

For $dT_d/dt_d = -1$ we would need there to be an interval where a queue exists but no agents depart. This will not happen in equilibrium in this model, though could with more general preferences. However, even were this to happen there is still a unique cost-minimizing departure time for each arrival time, as leaving at the end of the interval yields the same arrival time as leaving at any point within it, but with lower travel time. Should the departure rate have a point mass there will not be a one-to-one mapping between departure times and arrival times. However, we will shortly make some assumptions that rule this out.

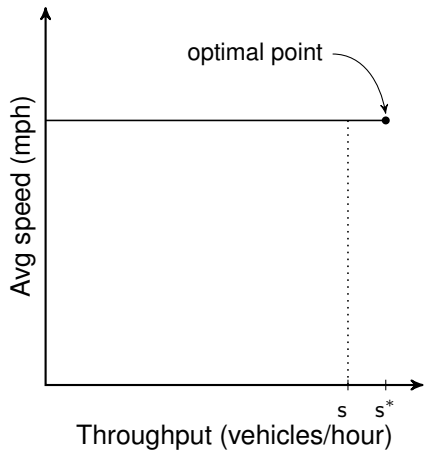


FIGURE 3. Production possibility frontier for bottleneck model

Notice that there is a single point which maximizes speed and throughput, labeled on Figure 3 as the optimal point. The bottleneck model does not have the traditional trade-off between throughput and speed, where increasing one requires decreasing the other (cf. Pigou, 1912; Knight, 1924; Walters, 1961).

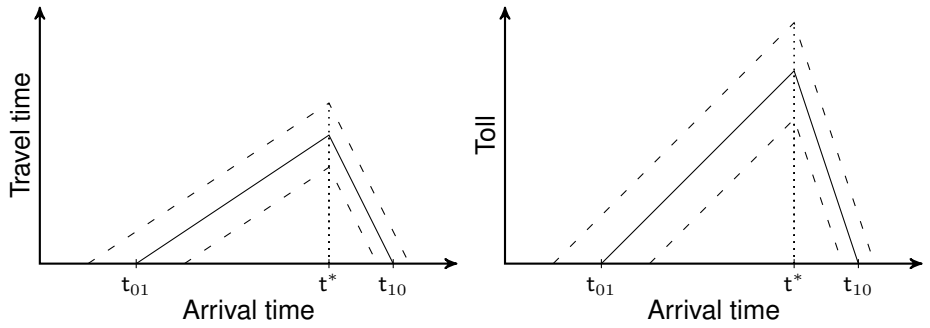
3.2. **Agent preferences.** There are G families of agents; let $i \in \{1, \ldots, G\} = \mathcal{G}$ denote an arbitrary family. Agents choose when to arrive at work and which route to take to minimize the cost of traveling. Agents dislike three aspects of traveling: travel time; tolls; and schedule delay, i.e., arriving earlier or later than desired. These costs combine to form the trip price; the trip price of arriving at time t on route r for an agent in family i with desired arrival time t^* is

$$p\left(t, r; i, t^*\right) = \alpha_i T_r\left(t\right) + \tau_r\left(t\right) + D_i\left(t^* - t\right); \tag{4}$$

where α is the cost per unit time traveling, i.e., the agent's value of time, and D_i is family i's schedule delay cost function. Schedule delay costs are piecewise linear,

$$D_i\left(t^* - t\right) = \left(t^* - t\right) \begin{cases} \beta_i & t \leq t^* \\ -\gamma_i & t > t^* \end{cases}.$$

where β is the cost per unit time early to work, and γ is the cost per unit time late to work. Each of these parameters is the agent's willing to pay money to reduce travel time or schedule delay by one unit of time. The ratios β/α and γ/α are the agent's willingness to pay in travel time to reduce schedule delay (early and late respectively) by one unit of time.



(A) Free route (B) Priced route

FIGURE 4. Agents' indifference curves by route

I assume that $\beta_i < \alpha_i$ for all i. This means that agents would rather wait for work to start at the office than wait in traffic and is needed to prevent the departure rate from being infinite.[21]

To simplify the problem I assume $\gamma_i = \xi\beta_i$ for all i. This means that those who dislike being early also dislike being late, while those who don't mind being early similarly don't mind being late.

We can define an agent's indifference curve over arrival time, travel time, and tolls using (4). Since on a free route there will be no toll, we can define the indifference curve on a free route as

$$\breve{T}(t; i, t^*, p) = \alpha_i^{-1}\left[p(t, \text{free}; i, t^*) - D_i(t - t^*)\right]. \tag{5}$$

Similarly, since there is to congestion related travel time on a priced route, we can define the indifference curve on a priced route as

$$\breve{\tau}(t; i, t^*, p) = p(t, \text{toll}; i, t^*) - D_i(t - t^*). \tag{6}$$

Figure 4a shows the indifference curve on a free route for three different trip prices. It shows that an agent is indifferent between arriving early or late, say near t_{01} or t_{10}, and having little to no travel time, and arriving exactly on-time at t^* but facing a long commute. The slope of the indifference curve is the agent's willingness to trade travel time for schedule delay, so before t^* the slope is β/α , and after it is $-\gamma/\alpha$.

Figure 4b shows the indifference curve on a priced route, also for three different trip prices. It shows how agents are indifferent between arriving early or late and paying little to no toll, and arriving exactly on-time but paying a high toll. The slope of this indifference curve is the agent's willingness to trade money for schedule delay; before t^* the slope is β and after it is $-\gamma$.

[21]See Small (1982) for empirical evidence that this is true for the morning commute. It holds in most of his specifications. Note that this is not the result of a statistical test but just comparing point estimates. See his Table 3 for the clearest evidence; the first column is essentially β/α.

On both routes the agent's bliss point is at $(t^*, 0)$, arriving right on-time, paying no toll, and dealing with no congestion. The lower the trip price p, the lower the indifference curve.

Agents can differ in their value of time, schedule delay costs, and desired arrival time. I define a *family* of agents as the set of agents with the same value-of-time and schedule delay costs.

The primary source of heterogeneity in drivers' value of time is variation in their income, and so if $\alpha_i > \alpha_j$ then family i is *richer* than family j. While there are other sources of heterogeneity in drivers' value of time,[22] by using α as a proxy for income we can directly discuss the primary concern with congestion pricing, that it helps the rich and hurts everyone else.

The ratio of an agent's schedule delay costs and value of time is a measure of how inflexible his schedule is, and so if $\beta_i/\alpha_i > \beta_j/\alpha_j$ then family i is more *inflexible* than family j. The main source of heterogeneity in drivers' flexibility is differences in occupation, as the opportunity cost of time early or late is different for those with different types of jobs. If a shift worker is late he generally face penalties and when he is early he passes the time talking with co-workers. Since for the shift worker there is not much difference between spending time traveling or being at work early, his β/α will be close to one. Similarly due to the penalty when late, γ/α will be large. In contrast, an academic can just start working whenever he gets to the office and so will have a very low marginal disutility from being early or late and so his β/α and γ/α will be closer to zero. Thus variation in β/α will be driven by in schedule flexibility, where jobs that are more flexible lead to a lower β/α.[23]

Within each family agent's desired arrival times are uniformly distributed over $[t_s, t_e]$.[24] Let n_i denote the density of agents of family i who desired to arrive at any given time in $[t_s, t_e]$ and $N_i = (t_e - t_s)\, n_i$ be the total mass of agents in family i. Demand is perfectly inelastic. Were demand not perfectly inelastic then the distribution of desired arrival times would no longer be uniform once tolls were added to the highway and different agents saw their trip prices change by different amounts. Furthermore, $\sum n_i > s^*$ so that it is impossible for all agents to arrive at their desired arrival time; thus some will need to arrive early or late.

It may seem more natural to assume an agent's desired arrival time falls into some discrete set, such as 7:00, 7:30, 8:00, 8:30, or 9:00 in the morning. However, what we care

[22] A driver's value of time reflects his marginal disutility of travel time and so can be driven by how comfortable his vehicle is or his taste for driving in congestion in addition to the standard labor-leisure trade-off. Other empirically important sources of heterogeneity are trip purpose, distance, and mode, with the last likely driven by selection (Small and Verhoef, 2007; Abrantes and Wardman, 2011).

[23] How flexible a worker's personal life is will also affect the ratio, as leaving early for work means leaving home earlier and going to bed earlier; and similarly leaving late for work likely implies working later to make up for lost time.

[24] In section 8 I will provide evidence that this is a reasonable approximation to the truth.

most about is when agents want to reach the end of the highway. Because the distribution of distances between the end of the highway and work will be continuous, the distribution of desired arrival times at the end of the highway is continuous.

Let $\{r, t\} = \sigma(i, t^*)$ be the strategy of an agent in family i with desired arrival time t^*; $\sigma : \mathcal{G} \times [t_s, t_e] \to \{\text{free},\ \text{toll}\} \times [0, 24]$.

3.3. **Definition of equilibrium.** We are looking for a perfect information, pure strategy Nash equilibrium. No agent will be able to reduce his trip cost by changing his arrival time or route choice.[25]

4. Finding Equilibrium

The fundamental scarcity is that there are times where more agents who want to arrive than are able. Since not everyone can arrive at their desired arrival time, some agents must arrive early or late. For agents to be indifferent between arriving at different times and thus facing different schedule delay costs, they must also face different travel time costs or tolls.

Since on a free route no toll is charged, travel times must vary. The only way to have non-zero travel time is for there to be queuing, and so there will always be congestion on the free route during rush hour, except for at the very start and end, a zero measure set. Note that congestion doesn't necessarily mean long travel times, just that there is additional travel time due to congestion. Because a queue will form, throughput will fall and the arrival rate on the free route is $\lambda_{\text{free}} \cdot s$ for all of rush hour.

On the priced route the toll can vary to keep agents indifferent. One virtue of the PPF implied by the bottleneck model is that it has a unique optimal point that maximizes speed and throughput (see Figure 3) and so the optimal toll is the one that keeps us at this point. Were we to restrict throughput to less than s^*, i.e., move to the right along the PPF, we would have unnecessary schedule delay. Were we to allow more that s^* vehicles to depart, thereby allowing allow queuing, we would move to the dotted portion of the PPF which would waste time and decrease throughput. This means the toll is set to eliminate queuing and maximize throughput. Agents depart at the rate the priced route can handle except possibly on a set of measure zero, and there is no variable travel time. Since a queue never forms, the arrival rate on the tolled route is $\lambda_{\text{toll}} \cdot s^*$ for all of rush hour.

These observations allow me to simplify the notation. Since there is no travel time on the priced route and no toll on the free route I will drop the route specific subscripts for τ and T. Further, define

$$s_r = \begin{cases} s & r = \text{free} \\ s^* & r = \text{toll} \end{cases}.$$

[25]This is similar Wardop's first principle of equilibrium which requires that no agent can unilaterally reduce his travel costs by changing to another route (Wardrop, 1952).

With these results we can simplify think of solving for equilibrium as assigning agents to arrival times. On a free route we can assign $\lambda_{\text{free}} \cdot s$ agents to each arrival time and on a priced route we can assign $\lambda_{\text{toll}} \cdot s^*$. Once we know which agents arrive when, we can back out the travel time profile or toll schedule necessary to implement it.

Framed this way, the bottleneck model when the road is completely free or priced is very similar to the Hotelling (1929) differentiated goods model. We have continuum of differentiated goods (arrival times), and agents have unit demand and bear a cost of purchasing a good different from the one they prefer (schedule delay costs). The key difference is that each good is "provided" by firms in a perfectly competitive market who in aggregate inelastically supply s_r units of the good.

It is also very much like the von Thünen (1930) model of land use. Instead of land use we are modeling the use of arrival times, and we replace transportation costs with schedule delay costs. When all agents have the same desired arrival time and the cost of being late is the same as the cost of being early the models are identical.

4.1. **Free route.** The ability to arrive at the most desirable times is allocated to those who are willing to pay the most for it. For a free route the currency that is used is travel time, while on a priced route it is dollars. Thus on a free route those willing to bear long travel times get to arrive on-time, while on a priced route those willing to pay the most get to do so.

An agents' inflexibility is his willingness to pay in travel time to reduce schedule delay, or in other words, his willingness to pay in travel time to arrive at the peak. Thus on a free route those who are very inflexible will arrive at the peak. This is formalized in the following lemma. The proof, along with all other omitted proofs, is in appendix A.

Lemma 1. *If family i is more inflexible than family j (i.e., $\beta_i/\alpha_i > \beta_j/\alpha_j$) then if an agent from family i with desired arrival time t^* arrives at t on a free route then no agent from family j arrives between t^* and t on a free route.*

Assume for now that the travel time profile has a single local maximum, $t^{\max}$. This is the peak of rush hour and an agent who arrives at $t^{\max}$ faces the longest commute of any agent. We will prove that there is a single peak later in proposition 12.

Define $t_i^{\max}$ as the time such that the agent in family i with desired arrival time $t_i^{\max}$ is indifferent between arriving early or late. When the travel time profile has a single local maximum, any agent from family i who has desired arrival time $t^* < t_i^{\max}$ will strictly prefer to arrive early or on-time, and similarly if $t^* > t_i^{\max}$ then they will strictly prefer to arrive late or on-time. I use the superscript max for two reasons, first, the agent from family i with desired arrival time $t_i^{\max}$ will have the largest trip price of any agent in family i, second, the peak of rush hour, $t^{\max}$, will occur at at least one families $t_i^{\max}$.

We assign agents to arrival times as follows. First, assume we know $t^{\max}$ and $t_i^{\max}$ for all $i \in \mathcal{G}$. Then starting at $t^{\max}$ and working our way backward, we assign to each arrival

time t the most inflexible agents of those who want to arrive early or on-time at t and are not yet assigned an arrival time until we have filled the available capacity. Likewise we start at $t^{\max}$ and work forward, assigning the most inflexible agents who want to arrive late or on-time at t. We break ties by allowing those with an earlier desired arrival time to arrive earlier.

Once we have assigned agents to arrival times we can back out what the travel time profile is from agents' preferences. If an agent arrives early or late on a free route it must be true that his indifference curve is tangent to the "budget line" so the marginal rate of substitution between schedule delay and travel time is equal to the marginal rate of substitution the equilibrium travel time profile offers; i.e., the slope of the travel time profile at the time he arrives must equal his inflexibility if he is early and $-\xi$ times his inflexibility if he is late. If an agent arrives exactly at his desired arrival time all we know is that his schedule delay costs are such that he is unwilling to accept schedule delay given the travel time profile. I formalize these results in the following lemma.[26]

Lemma 2.

$$\{t, \text{free}\} \in \sigma(i, t^*) \Rightarrow \begin{cases} \frac{dT}{dt}(t) = \alpha_i^{-1} \frac{dD_i}{dt}(t) & \text{if } t \neq t^*, \\ -\frac{\gamma_i}{\alpha_i} \leq \frac{dT}{dt}(t^*) \leq \frac{\beta_i}{\alpha_i} & \text{if } t = t^*. \end{cases}$$

To finish defining the travel time profile we add the initial condition that the travel time at the start of rush hour is zero.

Now that we know the travel time profile as a function of $t^{\max}$ and $t_i^{\max}$ for all $i \in \mathcal{G}$ we can pin down what these variables are in equilibrium. We pin down $t^{\max}$ from the requirement that there be no travel time at the end of rush hour, and require that for each family $i \in G$ the agent with desired arrival time $t_i^{\max}$ is indifferent between arriving early or late, as required by its definition.

4.2. **Priced route.** We can find equilibrium on a priced route in much the same way as on a free route. The key difference is that now travel times are allocated by willingness to pay in money, rather than by willingness to pay in travel time. An agent's willingness to pay in money to reduce schedule delay is his β, and so on a priced route those with a high β will arrive at the peak; which I formalize in the following lemma.

Lemma 3. *If $\beta_i > \beta_j$ then if an agent from family i with desired arrival time t^* arrives at t on the priced route then no agent from family j arrives between t^* and t on the priced route.*

[26]This lemma also implies that the presence of a kink in the schedule delay cost function is necessary in order to have inframarginal agents. As long as the travel time profile has continuous first derivatives, which it will almost everywhere, then by the intermediate value theorem the marginal rate of substitution between arrival time and schedule delay cannot go from very positive to very negative without at some point being tangent to the travel time profile. If there is a point of tangency then we have an interior solution and so agents would never be inframarginal. On a priced route the logic will be the same, just replace the travel time profile with the toll schedule.
It is not important that the kink in the schedule delay cost function is at the agent's desired arrival time, just that there is a kink and so the intermediate value theorem doesn't apply.

We assign agents to arrival times in the same manner as on a free route, except the we prioritize agents based on their β instead of their inflexibility. Assuming we know $t^{\max}$ (now the time with the highest toll) and $t_i^{\max}$ for all $i \in \mathcal{G}$ we start at $t^{\max}$ and work our way backwards. We assign to each arrival time t the agents with the highest β of those who want to arrive early or on-time at t and are not yet assigned an arrival time until we have filled the available capacity. Likewise we start at $t^{\max}$ and work forward, assigning the agents with the highest β who want to arrive late or on-time at t. As before, we break ties by allowing those with an earlier desired arrival time to arrive earlier.

In similar fashion to before, we can back out the toll schedule from agents' preferences. If an agent from family i arrives early or late on a priced route it must be true that his indifference curve is tangent to the "budget line" so his marginal rate of substitution between schedule delay and money is equal to the marginal rate of substitution the toll schedule offers. Thus the slope of the toll schedule at the time he arrives must equal β_i if he is early and $-\xi\beta_i$ if he is late. If an agent arrives exactly at his desired arrival time all we know is that his schedule delay costs are such that he is unwilling to accept schedule delay given the toll schedule and so $\gamma_i \leq d\tau/dt\,(t^*) \leq \beta_i$. We formalize these results in the following lemma.

Lemma 4.

$$\{t, \text{toll}\} \in \sigma\,(i, t^*) \Rightarrow \begin{cases} \frac{d\tau}{dt}\,(t) = \alpha_i^{-1} \frac{dD_i}{dt}\,(t) & \text{if } t \neq t^*, \\ -\gamma_i \leq \frac{d\tau}{dt}\,(t^*) \leq \beta_i & \text{if } t = t^*. \end{cases}$$

To finish defining the toll schedule I make the reasonable, but arbitrary, assumption that the toll is zero when the road is uncongested and so is zero at the start of rush hour. Allowing negative tolls is an effective way to "spend" the revenue raised by congestion pricing to improve congestion pricing's distributional impacts, but ruling negative tolls out we make it harder to find a Pareto improvement.

Since we can write the toll schedule as a function of $t^{\max}$ and $t_i^{\max}$ for all $i \in \mathcal{G}$ we can solve for these remaining variables by requiring the toll at the end of rush hour be zero and imposing that $t_i^{\max}$ match its definition.

4.3. **Value pricing.** When we have both a free route and a priced route we need to assign agents to routes, and then we can use the methods above to assign them to travel times on their routes. Agents will travel on the route that gives them to lowest cost. I will save most of the details for how we do this until later, as we will use approach it differently when there are two families than when there are an arbitrary number of families; but want to make one point now.

The start and end of rush hour will be the same on each route. If not, then there would be a time where there was congestion on the free route, but no toll on the priced route, and so an agent arriving at this time on the free route would deviate and arrive at the

same time on the priced route. Similarly there cannot be a positive toll on the priced route while there is no congestion on the free route.

5. Homogeneous Agents

Let's start by considering when every agent is identical; i.e., there is one family and all agents have the same desired arrival time t^*.[27] This is the canonical case worked out by Arnott et al. (1990, 1993). We will start by looking at what implications agents' preferences have for equilibrium. When agents' are homogeneous many of these implications will hold regardless of how we model congestion.

When agents are homogeneous we can use their indifference curves to solve for equilibrium. In equilibrium all agents will be indifferent between arriving at any point during rush hour, and so must be on the same indifference curve. To solve all we need to do is find the indifference curve which is above the x-axis for long enough for all agents to arrive at work. That is, for a given trip price the indifference curve tells us the length of rush hour and hence the supply of trips, and so we are finding the price that equates supply and demand. The indifference curves directly give us the equilibrium travel time profile on a free route and the equilibrium toll schedule on a priced route.[28]

This general point of using an indifference curve to find the length of rush hour and thus the supply of travel times holds regardless of how we model congestion. For a free route the ability to back out travel times from the indifference curve generalizes as well; however for a priced route things are more complicated. This is because it will no longer be true that there is no variable travel time on a priced route. When the PPF is convex you can increase throughput at the cost of slower speeds, and so travel times will not be constant.[29] Fortunately, when agents all have the same value of time we can convert the cost of travel time into dollars and put it on the same axis as the cost of the toll, so the general point about using indifference curves to find equilibrium holds.

In any congestion model with continuous time the first agent to arrive on either route pays no toll and faces no congestion. This must be true since the first agent could shift his arrival time forward by an infinitesimal amount and would then be arriving outside of rush hour. He would then face no travel time and pay no toll.[30] The only cost this first agent to arrive bears are the schedule delay costs from arriving so early.

[27] So $t_s = t_e$ and n_1 is a point mass.

[28] This is the same answer that lemmas 2 and 4 give us.

[29] This is because the marginal social benefit of having an agent arrive near the peak is higher than having him arrive further from the peak, and so in equilibrium the marginal social cost will be higher near the peak than further from the peak. An easy way to understand this is to consider a proposed equilibrium where travel time is constant over rush hour. If you need to add one more agent, at what time do you assign him to travel? If the cost is the same everywhere, then you assign him to arrive right at his desired time. Since you are not indifferent this cannot be an equilibrium.

[30] When there is no one else on the road a driver imposes no externality on others and so the socially optimal toll is zero.

This means we can use changes in the start of rush hour as a sufficient statistic for whether congestion pricing helps all road users when all agents are homogeneous. If congestion pricing leads rush hour being longer, so rush hour starts earlier, then congestion pricing hurts the first agent to arrive because he now has more schedule delay. Since all agents are identical, they must all have the same equilibrium trip price regardless of when they arrive, and so if congestion pricing hurts the first agent to arrive, it must hurt all agents. Likewise, if congestion pricing leads to rush hour being shorter, so rush hour starts later, then congestion pricing helps the first agent to arrive because he now has less schedule delay. Since all agents are identical, if the first agent is better off then all agents are better off.

If we believe that highway traffic is always on the PPF, then we cannot escape the conclusion that congestion pricing, while Kaldor-Hicks efficient, hurts road users before the revenue is redistributed.[31]

Proposition 5 (Prior literature)**.** *If agents are homogeneous then in any congestion model with a strictly negative relationship between flow and speeds, congestion pricing makes all agents worse off before the revenue is redistributed.*

When traffic is on the PPF the goal of congestion pricing is to reduce throughput so that the remaining agents can travel faster.[32] It is this logic that leads the U.S. Department of Transportation to teach that "congestion pricing works by shifting purely discretionary rush hour highway travel to other transportation modes or to off-peak periods" (2006, 1). But reducing throughput means rush hour must be longer, and so when agents are homogeneous this means all are harmed.

The standard bottleneck model $(s = s^*)$ assumes away the traditional trade-off between throughput and speed. Making this assumption is justified because it makes it tractable to model the dynamics of rush hour and doing so is important because congestion is inherently dynamic: what happens on the road at 6 a.m. affects traffic at 7 a.m.[33] As an unanticipated and unappreciated side effect, however, it also changes the welfare effects of congestion pricing.

Proposition 6 (Vickrey, 1969)**.** *If agents are homogeneous then in the bottleneck model with $s = s^*$, congestion pricing does not change consumer welfare prior to revenue redistribution.*

[31]Those familiar with the literature on congestion pricing may wonder how the result is consistent with Henderson (1974) who finds that "per person costs of traveling including the toll may decline with the imposition of tolls" (346). As Chu (1995) shows, in Henderson's proposed equilibrium a agent would be better off by leaving after the end of rush hour and so it is not an equilibrium. Chu reformulates the model to avoid this problem and finds that "the optimal toll increases the equilibrium private trip cost" (336).

[32]More precisely stated, the strategic goal of congestion pricing is to maximize social welfare, but when traffic is on the PPF the tactical goal becomes reducing throughput.

[33]See Arnott and Kraus (1993); Arnott et al. (1993) for additional arguments on the importance of using a dynamic model.

Because reducing the rate at which vehicles pass through the bottleneck doesn't increase speeds, the goal of pricing is no longer to reduce throughput, but rather to prevent a queue from forming. We set prices to reduce the departure rate at the beginning of rush hour and increase it at the end. This helps because it eliminates variable travel time, but because the length of rush hour is unchanged it does not affect consumer welfare.

The literature has not recognized the importance of assuming away the traditional trade-off between throughput and speed for explaining the differences between the welfare effects of congestion pricing in the bottleneck and other models. For example, Arnott et al. (1993) and van den Berg and Verhoef (2011) both use the bottleneck model and find that that the welfare impacts of congestion pricing are much more favorable than previous research reported, but attribute the difference to using a dynamic model rather than the assumption about how pricing affects throughput. Chu (1995) compares the bottleneck model to a different dynamic model and finds results consistent with propositions 5 and 6, but does not attribute it to different assumptions about the relationship between throughput and speed.

If, however, the traffic engineers are right, and queuing creates additional frictions that reduce throughput, then congestion pricing will be a Pareto improvement when agents are identical.

Proposition 7. *If all agents are homogeneous then in the bottleneck model with $s < s^*$, congestion pricing helps all agents before the revenue is redistributed.*

When queues reduce throughput the goal of pricing is to increase throughput by eliminating the queue and its attending frictions. We are able to decrease travel times while increasing throughput. Because rush hour is shorter, all agents are better off.

When agents are heterogeneous we can no longer use changes in the first agent's welfare as a sufficient statistic for changes in all agents welfare, and so changes in the length of rush hour don't map directly into whether or not all agents are better off. With heterogeneous agents congestion pricing can help some agents while hurting others regardless of the relationship between throughput and travel times. However, the results of propositions 5–7 will hold, at a minimum, for the first and last agent to depart when the road is free. Because the results hold for at least one agent, we can conclude that if increasing speeds requires reducing throughput then it is impossible for congestion pricing to be a Pareto improvement before spending the revenue. If, however, congestion pricing can increase throughput while increasing speeds, then it is possible for congestion pricing to be a Pareto improvement regardless of how the revenue is spent.

When agents are homogeneous there is no reason to price only some of the lanes. Leaving some lanes unpriced means leaving them congested and with lower throughput. Homogeneous agents will be indifferent between traveling on either route at any point in rush hour, and so there is no benefit to allowing some to pay with their time instead

of their money. By pricing all of the lanes we maximize total highway throughput which maximizes the reduction in the length of rush hour which maximizes the welfare gains.

However, when agents are heterogeneous it may be necessary to pricing only some of the lanes if we wish to congestion pricing to be a Pareto improvement. As a preview of our future results, consider what would happen if there was a small group of poor agents who also used the road, so small that they don't affect equilibrium. Were we to price all of the road there would be no guarantee they are not worse off; however were we to price just a portion of the lanes we can know that they are better off.

Proposition 8. *If all agents, except for a zero measure family, are homogeneous, then in the bottleneck model with $s < s^*$, pricing a portion, but not all, of the lanes will help all agents before the revenue is redistributed.*

Proof. Since the zero measure family of agents has no impact on equilibrium, we know by proposition 7 that all agents in the family with positive measure are better off. For those agents in the positive measure family who are on the free lanes to be better off, travel times must have fallen at each point in time. Thus if the zero measure agents travel on the free lanes at the same time they traveled before, then they will have shorter travel times and be better off. Since they have an option that gives them a lower trip price than before, whatever they choose must make them better off. Thus all agents are better off. □

The intuition behind this proof also leads to a nice empirical test for whether pricing a portion of the lanes was a Pareto improvement; we can check if travel times on the free lanes fell for every point in time. If so, pricing must have helped every road user.

6. Two Families

Since the thesis of this paper is that value pricing can lead to a Pareto improvement for all users of the road, we need to explicitly allow for heterogeneous preferences. As the primary concern with value pricing is that it hurts the poor and middle class, the main distinction we will make is between high and low income agents.

We will now consider when there are two families of agents, and will define family 1 as rich and family 2 as poor, i.e., $\alpha_1 > \alpha_2$. First we will solve for equilibrium when the road is completely free or priced, and use those results to determine when pricing all of the road is a Pareto improvement. Then we will solve for equilibrium when just a portion of the road is priced, and use those results to determine when value pricing is a Pareto improvement.

6.1. **Road is completely free or priced.**

For simplicity, define family A as the family that will arrive off-peak, and family B as the family that arrives on-peak. This will reduce the number of cases we need to solve and we can map A and B into rich and poor as needed.

Lemma 1 implies that on a free route $\beta_A/\alpha_A < \beta_B/\alpha_B$ and lemma 3 implies that on a priced $\beta_A < \beta_B$.

When the entire road is either free or priced one of two subcases will apply, either $n_B \leq s$ or $n_B > s$. We will consider each case in turn. We will use subscripts to distinguish which case equilibrium trip prices, travel times, and tolls pertain to. We will use the subscript I, free when the agents arriving at the peak are inframarginal and the highway is free, the subscript M, toll when they are the marginal driver when they arrive and the highway is priced, and the permutations of these subscripts.[34]

6.1.1. *Equilibrium when family B is inframarginal.* When $n_B \leq s$ there is enough capacity for the inflexible agents to all arrive exactly at their desired arrival time. This means that only flexible agents arrive early or late. We can use lemma 2 to define the equilibrium travel time profile as the solution to

$$\frac{dT_I}{dt}(t) = \begin{cases} \beta_A/\alpha_A & t_{0A} \leq t < t_A^{\max} \\ -\gamma_A/\alpha_A & t_A^{\max} \leq t < t_{A0} \\ 0 & \text{otherwise} \end{cases}, \text{ and} \tag{7}$$

$$T_I(t_{0A}) = 0. \tag{8}$$

This allows us to write equilibrium travel times as a function of t_{0A}, t_{A0}, and $t_A^{\max}$. The requirements of equilibrium give us three equations that allow us to solve for these three unknowns.

The first equation requires that the demand for early arrivals by agents in family A equals the supply. The supply for early arrivals is the capacity available between start of rush hour and the peak. In this period of time $(t_A^{\max} - t_{0A})\, s$ agents can arrive. However, we need to account for the capacity used by the agents in family B with a desired arrival time in this period of time. Since all agents in family B will arrive on-time, $(t_A^{\max} - t_s)\, n_B$ of the capacity available for early arrivals is used by agents of family B. All agents in family A with a desired arrival time before $t_A^{\max}$ will arrive early, and so demand for early arrivals by agents in family A is $(t_A^{\max} - t_s)\, n_A$. Thus in equilibrium

$$(t_A^{\max} - t_{0A})\, s - (t_A^{\max} - t_s)\, n_B = (t_A^{\max} - t_s)\, n_A. \tag{9}$$

The second equation is similar to the first, and requires that that the demand for late arrivals by agents in family A equals the supply. By similar reasoning as above, in equilibrium we need

$$(t_{A0} - t_A^{\max})\, s - (t_e - t_A^{\max})\, n_B = (t_e - t_A^{\max})\, n_A. \tag{10}$$

[34]To be precise, an agent is the marginal driver at time t if increasing the travel time or toll by a small amount would cause him to choose a different arrival time, he is inframarginal if it would not affect his choice of arrival time.

The final equation comes from requiring that travel time at the end of rush hour be zero,

$$T_I\left(t_{A0}\right) = 0. \tag{11}$$

The way we find equilibrium when the road is priced is essentially the same. As $n_B \leq s$ there is enough capacity for all agents in family B to arrive on-time, and since $\beta_B \geq \beta_A$ they are willing to pay to do so. Using lemma 4 we can define the equilibrium toll schedule as the solution to

$$\frac{d\tau_I}{dt}\left(t\right) = \begin{cases} \beta_A & t_{0A} \leq t < t_A^{\max} \\ -\gamma_A & t_A^{\max} \leq t < t_{A0} \\ 0 & \text{otherwise} \end{cases}, \text{ and} \tag{12}$$

$$\tau_I\left(t_{0A}\right) = 0. \tag{13}$$

Again we have three variables still to determine, and the equations that define them are similar. Because capacity on a priced route increases to s^*, we replace s with s^* in (9) and (10), as well as changing subscripts to denote that we are considering a priced route. Finally, we replace (11) with

$$\tau_I\left(t_{A0}\right) = 0. \tag{14}$$

We now know enough to find equilibrium trip prices. By solving (9), (10), and (11), or the equivalent equations for a priced route, we can determine the equilibrium travel time profile or toll schedule. We can then find each types' equilibrium trip price $\bar{p}\left(i, t^*\right) = \min_{t,r} p\left(t, r; i, t^*\right)$. The equilibrium trip prices for agents in family A for $r \in \{\text{free}, \text{toll}\}$ are

$$\bar{p}_{I,r}\left(A, t_A^{\max}\right) = \beta_A\left(N_A + N_B\right)\frac{1}{s_r}\frac{\xi}{1+\xi}, \text{ and} \tag{15}$$

$$\bar{p}_{I,r}\left(A, t^*\right) = \bar{p}_{I,r}\left(t_A^{\max}\right) - \left(t_A^{\max} - t^*\right)\begin{cases} \beta_A & t^* \leq t_A^{\max} \\ -\xi\beta_A & t^* > t_A^{\max} \end{cases}. \tag{16}$$

For family B agents on a free route equilibrium trip prices are

$$\bar{p}_{I,\text{free}}\left(B, t^*\right) = \frac{\alpha_B}{\alpha_A}\bar{p}_I\left(A, t^*\right), \tag{17}$$

while on a priced route they are

$$\bar{p}_{I,\text{toll}}\left(B, t^*\right) = \bar{p}_I\left(A, t^*\right). \tag{18}$$

While (16)–(18) can be calculated directly, and are, along with (15), in appendix C.1, they are also fairly intuitive. First, note that due to the slope of the travel time profile and toll schedule every agent in family A who arrives early is indifferent between arriving

at their desired arrival time or earlier, and likewise those who are late are indifferent between arriving at their desired arrival time or later.

To see the intuition behind (16) consider two agents in family A, one with desired arrival time of $t_A^{\max}$ and the other of t^*. They are both willing to arrive at t^*, and were they to do so the only difference in their trip price would be the difference in their schedule delay costs at t^*. This means we can write the trip price of the second as the trip price of the first minus the difference in their schedule delay costs at t^*.

The see the intuition for (17) and (18) now consider two agents with desired arrival time t^*, one from each family. Both are all willing to arrive at t^*. When arriving at t^* on a free route neither of them have any schedule delay costs and they face the same travel time, and so the only difference in their trip price is due to the difference in their value of time. By dividing the family A agent's trip price by his value of time we recover the travel time at t^*, which we then multiply by the family B agent's value of time to obtain the family B agent's trip price. Similarly, on a tolled route they face the same toll and have no schedule delay or travel time, and so their trip prices are identical.

When one family is inframarginal their preferences do not affect equilibrium or the marginal family's trip price. Equation (15) is the same as (15) in Arnott et al. (1993), with $N = N_A + N_B$. Further, travel times and tolls are the same. This is our first clue that the intuition of proposition 8 and so value pricing will be a Pareto improvement when will hold when $w_B < s$.

6.1.2. *Equilibrium when family B is marginal.* When $n_B > s$ there is no longer enough capacity for the inflexible agents to all arrive exactly at their desired arrival time, and so they must also arrive early or late. Family B agents will use all of the capacity near the peak, and family A agents will use all of the capacity off-peak. We can use lemma 2 and the requirement that the travel time at the end of rush hour to define the equilibrium travel time profile as the solution to

$$\frac{dT_M}{dt}(t) = \begin{cases} \beta_A/\alpha_A & t_{0A} \leq t < t_{AB} \\ \beta_B/\alpha_B & t_{AB} \leq t < t_B^{\max} \\ -\gamma_B/\alpha_B & t_B^{\max} \leq t < t_{BA} \\ -\gamma_A/\alpha_A & t_{BA} \leq t < t_{A0} \\ 0 & \text{otherwise} \end{cases}, \text{ and} \tag{19}$$

$$T_M(t_{0A}) = T_M(t_{A0}) = 0. \tag{20}$$

Now we have three additional variables to solve for in order to find equilibrium travel times. As before, we will use the requirement that supply equals demand for early and late arrivals, but now we will do so for both types, as well as the requirement that there be

no travel time at the end of rush hour. These requirements give us the following equations.

$$(t_{AB} - t_{0A})\, s = (t_A^{\max} - t_s)\, n_A, \tag{21}$$

$$(t_{A0} - t_{BA})\, s = (t_e - t_A^{\max})\, n_A, \tag{22}$$

$$(t_B^{\max} - t_{AB})\, s = (t_B^{\max} - t_s)\, n_B, \tag{23}$$

$$(t_{BA} - t_B^{\max})\, s = (t_e - t_B^{\max})\, n_B, \text{ and} \tag{24}$$

When the inflexible types were all able to arrive exactly at their desired arrival time $t_B^{\max}$ was not defined and so we didn't need to solve for it, now that they are unable to arrive on-time we do. The requirement that travel time at the end of rush hour be zero now pins down $t_B^{\max}$ rather than $t_A^{\max}$ and so we need to add an equation imposing the definition of $t_A^{\max}$,

$$p\left(t_{AB}, \text{free}; A, t_A^{\max}\right) = p\left(t_{BA}, \text{free}; A, t_A^{\max}\right). \tag{25}$$

What matters is that the cost of arriving early equals the cost of arriving late for the agent in family A with desired arrival time $t_A^{\max}$, so the exact times we evaluate this equation at don't matter as long as one is in $[t_{0A}, t_{AB}]$ and the other is in $[t_{BA}, t_{A0}]$.

As when $n_B \leq s$, the equations which define the equilibrium toll schedule are essentially the same as those that define the equilibrium travel time profile. By lemma 4 and the requirement that the toll at the end of rush hour be zero we know

$$\frac{d\tau_M}{dt}(t) = \begin{cases} \beta_A & t_{0A} \leq t < t_{AB} \\ \beta_B & t_{AB} \leq t < t_B^{\max} \\ -\gamma_B & t_B^{\max} \leq t < t_{BA} \\ -\gamma_A & t_{BA} \leq t < t_{A0} \\ 0 & \text{otherwise} \end{cases}, \text{ and} \tag{26}$$

$$\tau_M(t_{0A}) = \tau_M(t_{A0}) = 0. \tag{27}$$

As before, we replace s with s^* in (21)–(24) and change subscripts. Finally, we update the definition of $t_A^{\max}$ for a priced route by replacing "free" with "toll" throughout (25).

We now know enough to find equilibrium trip prices. By solving (21)–(25), or the equivalent equations for a priced route, we can determine the equilibrium travel time profile or toll schedule. We can then find each types' equilibrium trip price. The equilibrium trip prices for the agents with desired arrival time $t_i^{\max}$ in each family are

$$\bar{p}_{M,\text{free}}\left(A, t_A^{\max}\right) = \bar{p}_{M,\text{toll}}\left(A, t_A^{\max}\right) = \beta_A \left(N_A + N_B\right) \frac{1}{s_r} \frac{\xi}{1+\xi}, \tag{28}$$

and for family B on a free route

$$\bar{p}_{M,\text{free}}\left(B, t_B^{\max}\right) = \alpha_B \left(N_A \frac{\beta_A}{\alpha_A} + N_B \frac{\beta_B}{\alpha_B} \right) \frac{1}{s} \frac{\xi}{1+\xi}, \tag{29}$$

and for family B on a priced route

$$\bar{p}_{M,\text{toll}}\left(B, t_B^{\max}\right) = \left(N_A \beta_A + N_B \beta_B \right) \frac{1}{s^*} \frac{\xi}{1+\xi}. \tag{30}$$

We can then define the trip price for all the other agents in reference to (28)–(30),

$$\bar{p}_{M,r}\left(i, t^*\right) = \bar{p}_{M,r}\left(i, t_i^{\max}\right) - \left(t_i^{\max} - t^*\right) \begin{cases} \beta_i & t^* \leq t_i^{\max} \\ -\xi\beta_i & t^* > t_i^{\max} \end{cases} \text{ for } i \in \{A, B\}. \tag{31}$$

These are derived in appendix C.1.[35]

The intuition behind (31) is that an agent with desired arrival time $t_i^{\max}$ is willing to arrive at the same time as an agent in his family with desired arrival time t^*, let t be the time they arrive. Should they both arrive at t then the only difference in their trip price is the difference in their schedule delay costs at t. Thus, the difference in their schedule delay costs of arriving at t will be the difference in their equilibrium trip prices, regardless of when they actually arrive.

Notice that (25) is the same as (15), and that (31) matches (16). The equilibrium trip price for an agent who is willing to arrive at the start or end of rush hour is pinned down by the length of rush hour. It doesn't matter whether or not the other families agents are all able to arrive at their desired arrival time and it doesn't matter whether the road is priced or free, except indirectly through the effect of pricing on road capacity. Furthermore, the preferences of the family arriving at the peak doesn't effect the equilibrium trip prices of those arriving off-peak.

Now that we know what equilibrium trip prices are when the road is free as well as when it is priced we can find when pricing the entire road helps all agents.

6.1.3. *When is pricing the entire road a Pareto improvement?* While charging time varying tolls can increase throughput by preventing the destructive effects of queuing, it also requires

[35]Note that the equilibrium travel time profile, toll schedule, and trip prices for and trip prices for the agents who desire to arrive at $t_i^{\max}$ are the same as they would be if all agents desired to arrive at $t_i^{\max}$ (cf. Arnott et al., 1994).
Actually, as long as every agent is on the margin of trading schedule delay for travel time (or toll), i.e., every agent is either early or late, and rush hour has a single peak, the equilibrium travel time profile and toll schedule will be the same as the equations defining equilibrium are unchanged. Thus the result of Hendrickson and Kocur (1981) that equilibrium travel times and tolls are invariant to changes in the distribution of desired arrival time as long as agents are homogeneous except for their desired arrival time and rush hour has a single peak generalizes only when all agents are either early or late.
However, this is unlikely to hold unless we strongly restrict the number of families because it requires that for each time t the density of agents from most inflexible family who desire to arrive at that time be greater than highway capacity.

changing the currency used to allocate arrival times from time to money. Although both of these effects are Kaldor-Hicks efficiency enhancing, changing the currency used hurts poorer drivers. Whether pricing is a Pareto improvement will depend on whether the gains in throughput outweigh the harm from changing the currency.

The direct effect of changing the currency is that it makes desirable arrival times relatively cheaper for richer agents. This means a poor agent who had been traveling at the peak will either need to pay more to outbid the rich agent to continue to travel at the peak, or travel further off-peak, thereby increasing his schedule delay.

Changing the currency also can hurt the poor indirectly; because congestion pricing will lower the cost for richer agents it will induce more rich agents to travel. This will counteract some of the benefit to existing drivers of increasing throughput. If demand for trips by the rich is sufficiently elastic it is even possible that rush hour will be longer once congestion pricing is implemented. [36]

To find when pricing the entire road is a Pareto improvement we need to find when equilibrium trip prices when the road is tolled are weakly lower than equilibrium trip prices when the road is free for all agents in both families, with at least one agent's trip price strictly lower. Because equilibrium trip prices take a different form depending on what family is traveling at the peak and whether that family is marginal or inframarginal, we end up with ten different cases to check.[37] The results of doing so are reported in the following proposition, with the proof relegated to appendix A, and shown visually in Figure 5.

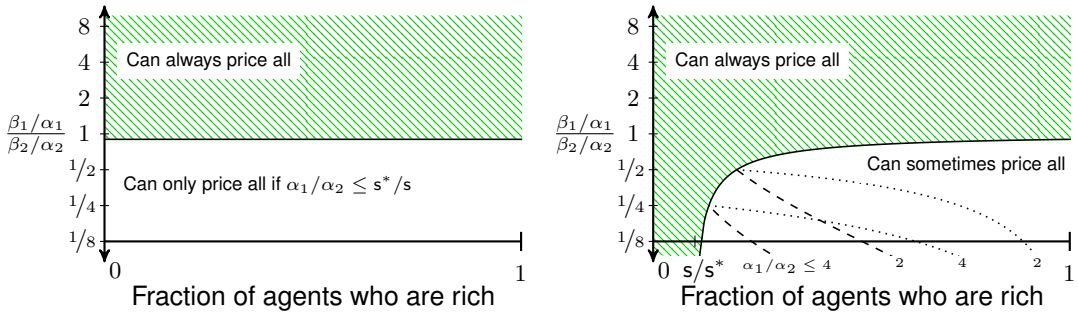
Proposition 9. *If there are two families, with uniformly distributed desired arrival time over a common support, and perfectly inelastic demand, then pricing the entire road never hurts the rich and is a Pareto improvement prior to spending the toll revenue if and only if*

$$\frac{\beta_1/\alpha_1}{\beta_2/\alpha_2} \geq \min\left\{1, \frac{\beta_1}{\beta_2}\right\} \frac{s}{s^*} - \frac{n_2}{n_1} \begin{cases} 0 & \text{if } n_2 \leq s, \\ 1 - \min\left\{1, \frac{\beta_1}{\beta_2}\right\} \frac{s}{s^*} & \text{if } s^* \geq n_2 > s\,. \\ 1 - \frac{s}{s^*} & \text{if } n_2 > s^*; \end{cases} \tag{32}$$

It will be easier to understand the intuition of (32) if we first simplify it by replacing the right hand side with its maximum value, and then later add back in the nuance of the entire expression. Doing so allows us to say that pricing the entire road is a Pareto

[36] For a more detailed discussion of how these three effects can hurt the poor, as well as illustrations of how they manifest themselves in the bottleneck model, see Hall (2012).

[37] A case is (1) the type that arrives at the peak when the road is free, (2) whether they are marginal or inframarginal, (3) the type that travels at the peak when the road is priced, and (4) whether they are marginal or inframarginal. This would suggest there are sixteen possible cases, but a number of the cases are logically impossible. For example, the poor cannot arrive off-peak when the road is free but on-peak when it is priced because this would imply $\alpha_2 > \alpha_1$. Similarly, if the poor always travel at the peak then they cannot be inframarginal when the road is free but marginal when it is priced.



(A) $n_2 < s$ (B) $n_2 \geq s$

FIGURE 5. Parameter values where pricing leads to a Pareto improvement. In the areas we can only sometimes price our ability to achieve a Pareto improvement depends on whether α_1/α_2 is small enough. Several threshold levels of α_1/α_2 are drawn with dotted lines for when $s \leq n_2 \leq s^*$ or dashed lines for when $n_2 > s^*$. Figures drawn with $s/s^* = 0.9$.

improvement if

$$\frac{\beta_1/\alpha_1}{\beta_2/\alpha_2} \geq \frac{s}{s^*}. \tag{33}$$

It is no great surprise to find that pricing the entire road never hurts the rich, while only in some situations does it help the poor. This occurs because the rich find it easier to outbid the poor for desirable arrival times when the currency used is money rather than travel time. The question then becomes whether the increase in capacity is enough to overcome this harm.

It is always possible for the increase in capacity to be large enough for pricing all of the road to be a Pareto improvement, regardless of the other parameters. As the left-hand side of (33) is strictly positive we can always find an s/s^* such that (33) holds.

How much the poor drivers are harmed by the change in currency depends on how their preferences compare to those of the rich. If the poor are less inflexible then the rich, so that the left-hand side of (33) is greater than one and we are in the top half of figures 5a and 5b, then they are not harmed at all and so pricing all of the road is a Pareto improvement. This holds even if there is no throughput drop at bottlenecks. Mathematically this follows because $s \leq s^*$. The intuition behind this is that when the poor arrive off-peak regardless of whether or not the road is priced then pricing cannot displace them nor require them to outbid the rich in a currency that they hold more dear.

When the poor are more inflexible than the rich then we are in the bottom half of figures 5a and 5b, and it is difficult for pricing all of the road to be a Pareto improvement. It becomes easier the more similar are the poor and rich families' level of inflexibility, i.e., the closer the left-hand side of (33) is to one. This is because the more similar agents are in

their willingness to pay with travel time to avoid schedule delay, the more alike the slope of the travel time profile at any point in time during rush hour, and so the less agents prefer their actual arrival time to any other during rush hour. Because agents only slightly prefer their actual arrival time, the harm to an agent of changing their arrival time to be further from the peak is small and so it is easier for the benefit from shrinking rush hour to outweigh this harm.[38]

Returning to (32) we see that the right-hand side is reduced, and so it becomes easier obtain a Pareto improvement, when β_1/β_2 is less than one and small, n_2/n_1 is large, $n_2 > s$, and $s^* \geq n_2 > s$. Note that these effects only matter when the poor are more inflexible than the rich, so that they arrive at the peak when the road is free.

Lowering β_1/β_2 reduces the damage pricing does to the poor. Notice that

$$\frac{\beta_1}{\beta_2} = \frac{\beta_1/\alpha_1}{\beta_2/\alpha_2} \times \frac{\alpha_1}{\alpha_2},$$

and so when we lower β_1/β_2 while holding fixed each families inflexibility we are actually lowering α_1/α_2. In other words, we are reducing the degree of income inequality, and so reducing the difference in the two families' exchange rates between travel time and money. Hence, pricing gives less of an advantage to the rich and so is more likely to lead to a Pareto improvement. The ratio β_1/β_2 is inside of a min operator because if $\beta_1 > \beta_2$ then the rich are willing to pay more in money to reduce schedule delay than the poor and so the poor don't care how much more the rich are willing to pay, just that they are willing to do so. If, however, $\beta_1 < \beta_2$, then the poor are affected by the actual amount the rich are willing to pay since they must pay more than that amount.

When n_2/n_1 is large there are not many rich agents relative to the number of poor agents. Should the rich displace the poor from the peak, the amount of displacement is small. If the rich don't displace the poor then the tolls the poor face are higher due to their presence, but when there are not many rich agents the tolls are not much higher. The harm done to the poor from tolling is less when there are not many rich, and so when n_2/n_1 is large it is easier for pricing to be a Pareto improvement.

It is easier to get a Pareto improvement when $n_2 > s$ because it means that the poor were not getting the advantage of being inframarginal when the road was free. When an agent is inframarginal they are able to arrive at the time they want while paying a price that reflects the preferences of those who don't value being on-time as much as they do. Should the poor be inframarginal when the road is free then pricing does them great harm by taking this advantage away from them and it becomes difficult for pricing to be a Pareto improvement. Even if they are still inframarginal when the road is priced they

[38]Note that the harm from being displaced is an upper bound on how much pricing all of the road hurts the poor because the poor may prefer to respond to pricing by paying higher tolls and arriving at the peak to arriving off-peak.

are hurt because the difference between their own willingness to pay to avoid schedule delay and that of the rich shrunk when the currency changed.

When the poor are more inflexible than the rich the best case for being able to price all of the road is when $s < n_2 < s^*$ and $\beta_2 > \beta_1$. In this case pricing helps the poor in an additional way, it allows them to become the inframarginal driver and thus pay a price based on the preferences of those who don't value being on-time as much as they do. This additional benefit makes it more likely that the gains from pricing will outweigh the harm it does.

If our only option is to price all of the road or none of it, then it is difficult to obtain a Pareto improvement unless the rich are more inflexible than the poor. If, however, we can price just part of the road, then it becomes significantly easier to do so.

6.2. **Value pricing.** Solving for equilibrium with two routes becomes more complicated because now agents choose which route they take as well as their arrival time. There are two results that will make assigning agents to routes simpler. First, the same family will arrive off-peak on both routes, or at least be indifferent about doing so. This is because for all agents the cost of arriving at the very start or end of rush hour is the same on both routes because at the start and end of rush hour there is no toll or travel time, just schedule delay. The second result formalizes the intuition that the rich prefer to be on the priced route and the poor prefer the free route and is given in the following lemma.

Lemma 10. *If there are two families and two routes, one priced and one free, then in equilibrium the rich will never be on the free route unless the poor are too and the poor will never be on the priced route unless the rich are too.*

Equilibrium can fall in one of eight cases depending on the parameters and we will use subscripts to denote which case equilibrium trip prices, travel times, and tolls belong to. The three dimensions the cases differ on are (1) which family is not arriving off-peak, (2) whether they are on one or two routes, and (3) whether some agents in this family are inframarginal or if they are all marginal. We will use the subscript $2R, I,$ poor when the poor family does not arrive off-peak, some poor agents are inframarginal, and agents in the poor family travel on both routes; the subscript $1R, M,$ rich when the rich family does not arrive off-peak, all rich agents are marginal, and agents in the rich family only travel on one route, and the permutations of these subscripts for remaining six cases. Some equilibrium objects will not depend on which family is not arriving off-peak, and for those objects we will omit the third part of the subscript.

As we did when solving for equilibrium when the road is completely free or completely priced, define A as the family arriving off-peak, but let's now define B simply as the other

family. Let C and D denote families A and B preferred routes, that is

$$C(A) = \begin{cases} \text{toll} & \text{if } A = 1 \\ \text{free} & \text{if } A = 2 \end{cases} \text{ and}$$

$$D(B) = \begin{cases} \text{toll} & \text{if } B = 1 \\ \text{free} & \text{if } B = 2 \end{cases}.$$

Using this notation we can write down four sets of equations that define equilibrium for all eight possible cases.

We will leave all but the two simplest and most important cases; $1R, I,$ poor and $1R, I,$ rich; for appendix C.2.[39]

6.2.1. *One family on just one route, and this family is inframarginal* $(1R, I)$. In this case family A travels on both routes and is always marginal while family B travels only on route D and is inframarginal. For this to be possible there must be enough capacity on D for all of the members of family B to arrive on-time, i.e., $n_B \leq \lambda_D s_D$.

Tolls and travel times are the same as when the entire road is free or priced and family B is inframarginal. The travel time profile is defined by (7), (8), and (11), and the toll schedule is defined by (12)–(14).

We then require that for family A the supply for arrival times equals the demand, both for early and late arrivals. This gives us the final two equations we need to define equilibrium.

$$\left(t_A^{\max} - t_{0A}\right)\left(\lambda_{\text{toll}} s^* + (1 - \lambda_{\text{toll}})\, s\right) - \left(t_A^{\max} - t_s\right) n_B = \left(t_A^{\max} - t_s\right) n_A, \text{ and}$$

$$\left(t_{A0} - t_A^{\max}\right)\left(\lambda_{\text{toll}} s^* + (1 - \lambda_{\text{toll}})\, s\right) - \left(t_e - t_A^{\max}\right) n_B = \left(t_e - t_A^{\max}\right) n_A.$$

Solving this system of equations gives us the equilibrium travel time profile and toll schedule. We can then follow the same steps as in section C.1.1 to find the trip prices, which are

[39] In two of the cases left for the appendix; $2R, I,$ poor and $2R, I,$ rich; the toll schedule or travel time profile is not completely defined by lemmas 2 and 4 and so I use another indifference relation to characterize part of the toll schedule or travel time profile. I believe the need to do so goes away if there is a continuum of families.

$$\bar{p}_{1R,I}\left(A, t_A^{\max}\right) = \beta_A \frac{N_A + N_B}{\lambda_{\text{toll}} s^* + (1 - \lambda_{\text{toll}})\, s} \frac{\xi}{1 + \xi}, \tag{34}$$

$$\bar{p}_{1R,I}\left(A, t\right) = \bar{p}_{1R,I}\left(A, t_A^{\max}\right) - \left(t_A^{\max} - t\right) \begin{cases} \beta_A & t \leq t_A^{\max} \\ -\xi\beta_A & t > t_A^{\max} \end{cases}, \tag{35}$$

$$\bar{p}_{1R,I,\text{poor}}\left(B, t\right) = \frac{\alpha_B}{\alpha_A} \bar{p}_{1,RI}\left(A, t\right), \text{ and} \tag{36}$$

$$\bar{p}_{1,RI,\text{rich}}\left(B, t\right) = \bar{p}_{1,RI}\left(A, t\right). \tag{37}$$

The only difference between these prices and those when the entire road is free or priced is that the highway capacity is different. Compared to when the highway is free, capacity is higher, and so every agents trip price is lower. Thus when the one of the families is inframarginal when the road is free, pricing a portion of the road is a Pareto improvement.

The intuition is simple, as long as we keep the poor inframarginal, then their preferences do not effect equilibrium tolls or travel times. By only pricing a portion of the lanes we preserve the ability of the poor to pay with their time instead of money. As a result, when we price some of the lanes and reverse the destructive effects of queuing for those lanes, we increase total highway capacity. This allows rush hour to be shorter on all lanes, reducing all road users trip prices.

If the rich travel at the peak, meaning they are more inflexible than the poor, then as we have already seen, we can price the entire road and so it isn't a surprise that we can also price part of it.

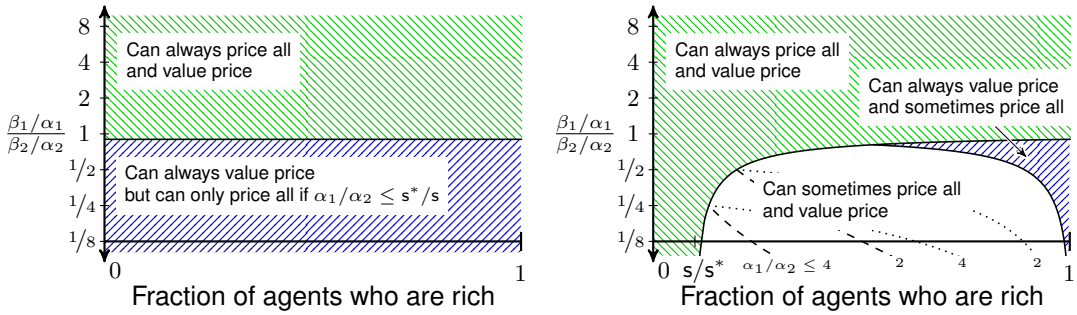
6.2.2. *When else is pricing part of the road a Pareto improvement?* By comparing equilibrium trip prices in the remaining six cases that equilibrium can fall into when value pricing to the four cases that can apply when the road is free we can find the rest of the parameter values for which pricing part of the road is a Pareto improvement; these are summarized in the following proposition and shown in Figure 6.

Proposition 11. *If there are two families, with uniformly distributed demand over a common support, and perfectly inelastic demand, then there exists a $\lambda_{toll} \in (0,1)$ such that pricing λ_{toll} of the road is a Pareto improvement if and only if one of the following is true:*

$$n_2 < s, \tag{38}$$

$$\frac{\beta_1/\alpha_1}{\beta_2/\alpha_2} > \frac{\left(s/s^*\right)\left(n_2/n_1\right)}{\left(1 - s/s^*\right) + n_2/n_1}, \tag{39}$$

or if (32) holds strictly.



(A) $n_2 < s$ (B) $n_2 \geq s$

FIGURE 6. Parameter values where pricing leads to a Pareto improvement. In the areas we can only sometimes price our ability to achieve a Pareto improvement depends on whether α_1/α_2 is small enough. Several threshold levels of α_1/α_2 are drawn with dotted lines for when $s \leq n_2 \leq s^*$ or dashed lines for when $n_2 > s^*$. Figures drawn with $s/s^* = 0.9$.

First, if pricing all of the road helps every single agent, then pricing almost all of the road will too, and so if (32) holds strictly then pricing part of the road will be a Pareto improvement.

By only pricing a portion of the road we are able to further expand when we are able to obtain a Pareto improvement to when (39) holds. As with (32) it is easier to price the greater the ratio of the inflexibility of the rich to the poor and the greater the size of the throughput drop. In (39) it is easier to greater the fraction of drivers who are rich, this is the opposite of (32). The new area that (39) allows us to price is in the bottom right area of Figure 6b.

The intuition for why value pricing can be Pareto improving when most drivers are rich is as follows. When we price a portion of the road the road we increase the measure of arrival times the poor need on the free route. This hurts the poor. But we also reduce the length of rush hour, which directly helps the rich, reducing travel times at any point when the rich are traveling. Because the rich are better off, the poor don't need to pay as much in travel time to outbid them for the desirable arrival times. The first effect is proportional to the number of agents who are poor, while the second is proportional to the total number of agents. The greater the fraction of agents who are rich, the more likely the second effect dominates the first. Should that happen, both the rich and poor are better off and value pricing is a Pareto improvement.

We can approximate the conditions for being able to obtain a Pareto improvement from value pricing as follows, we can price part of the road if there are rich agents arriving at the peak when the road is free. Should there be rich agents arriving at the peak, one of two things must be true: either the rich are more inflexible than the poor or the poor are

inframarginal. In both of these cases pricing part of the road is a Pareto improvement. The only area of the parameter space we miss is the bottom half of Figure 6b; for most of that area we are unable to price. The intuition for this approximation is that when the rich are arriving at the peak we can price part of the highway while still allowing the poor agents who were arriving at the peak to travel on an unpriced route. This condition is likely to hold; casual empiricism says that Lexus and Kias regularly share the road.

7. Continuum of Families

Now let's turn to solving the model with a continuum of families; which we will use to estimate the distribution of agent preferences and evaluate the impact of pricing all or part of the highway.

Working with a continuum of types will be easier if we adjust our notation slightly. We can still index families by i, but it will often be easier to directly refer to a family by their value of time and inflexibility.[40] Let $t^{\max}(\alpha, \delta)$ denote the arrival time of the agent in family $\{\alpha, \delta\}$ who is indifferent between arriving early or late. Let $n(\alpha, \delta, t^*)$ be the density of agents with value of time α, inflexibility $\delta = \beta/\alpha$, and desired arrival time t^*. Let $n_\delta(\delta)$ be the marginal distribution of inflexibility, and $n_\alpha(\alpha)$ be the marginal distribution of value of time.

7.1. **All free.**

We can find equilibrium travel times by first assigning agents to arrival times using the algorithm from section 4 and then using lemma 2 to find travel times. To do so, we first need to confirm rush hour has a single peak, which I do in the following proposition.

Proposition 12 (Rush hour has single peak)**.** *If there is a continuum of families, the support of their distribution is a connected set, and there is a single δ or β that is marginal over the support of desired arrivals then rush hour has a single peak.*

First we assign agents to arrival times. Starting at the, as of yet unknown, peak of rush hour $t^{\max}$ and working backwards, we assign to each arrival time t the s most inflexible agents of those who want to arrive early or on-time at t and are not yet assigned an arrival time. We then work forwards assigning the most inflexible agents of those who want to arrive on-time or late. Problematically, in order to know which agents want to arrive early or on-time at t and are not yet assigned an arrival time we need to know which agents with desired arrival times later than the peak of rush hour $t^{\max}$ but who want to arrive early or on-time will not be able to arrive on-time and so should be considered for

[40] Now the set of families $\mathcal{G}$ will be the set of non-negative real numbers. We can use a real number to index a tuple of real numbers by interleaving the digits of each number in the tuple. We do this separately for the integer and decimal parts; that is, we interleave the decimal parts of the numbers in the tuple to form the decimal part of the index, and we interleave the integer parts of the numbers in the tuple to form the integer part of the index.

assignment to t. That is, we need to solve the assignment problem simultaneously for early and late arrivals.

We are able to make the assignment problem much easier by assuming that desired arrival times are uniformly distributed and independent of value of time and inflexibility. If we only consider those who want to arrive on-time at t, then the problem of finding the lowest inflexibility to arrive at t amounts to finding the $\hat{\delta}$ such that

$$\int_{\hat{\delta}}^{1}\int_{0}^{\infty} n\left(\alpha,\delta,t\right)\, d\alpha d\delta = s. \tag{40}$$

Because of the assumptions we made about the distribution of desired arrival times the solution to this equation is the same for all $t \in [t_s, t_e]$. If we assign all agents with $\delta \geq \hat{\delta}$ to arrive on-time then we will fill the road to capacity during $[t_s, t_e]$ and will satisfy our algorithm and avoid violating lemma 1. At each time $t \in [t_s, t^{\max}]$ the set of agents who want to arrive early at t only contains agents with $\delta < \hat{\delta}$ and so the s most inflexible agents of those who want to arrive early or on-time are those who want to arrive on-time and have $\delta \geq \hat{\delta}$. Similarly, for each time $t \in [t^{\max}, t_s]$ the set of agents who want to arrive late at t only contains agents with $\delta < \hat{\delta}$ and so the s most inflexible agents of those who want to arrive on-time or late are those who want to arrive on-time and have $\delta \geq \hat{\delta}$.

Once we have filled all of the arrival times in $[t_s, t_e]$ we can use the algorithm to assign arrivals before t_s and after t_e. All of the remaining agents to be assigned are either early or late, and for arrivals before t_s we can just work backwards assigning the most inflexible remaining agents who want to be early to each arrival time. Likewise, for arrivals after t_e we can just work forwards assigning the most inflexible remaining agents who want to be late to each arrival time.

With a few definitions we can write down when an agent arrives. The mass of agents who arrive prior to the peak of rush hour is the sum of the mass who arrive early and the mass who arrives on time prior to the peak:

$$m_e = \int_0^{\hat{\delta}}\int_0^{\infty}\int_{t_s}^{t^{\max}(\alpha,\delta)} n\left(\alpha,\delta,t^*\right)\, dt^* d\alpha d\delta + \int_{\hat{\delta}}^{1}\int_0^{\infty}\int_{t_s}^{t^{\max}} n\left(\alpha,\delta,t^*\right)\, dt^* d\alpha d\delta.$$

The mass of agents arriving after the peak $m_l = 1 - m_e$, where the subscripts e and l stand for early and late. The cumulative distribution function of inflexibility for those arriving before the peak is

$$N_{\delta,e}\left(\delta\right) = m_e^{-1}\begin{cases} \int_0^{\delta}\int_0^{\infty}\int_{t_s}^{t^{\max}(\alpha,\delta')} n\left(\alpha,\delta',t^*\right)\, dt^* d\alpha d\delta' & \delta \leq \hat{\delta} \\ N_{\delta,e}\left(\hat{\delta}\right) + \int_{\hat{\delta}}^{\delta}\int_0^{\infty}\int_{t_s}^{t^{\max}} n\left(\alpha,\delta,t^*\right)\, dt^* d\alpha d\delta & \delta > \hat{\delta} \end{cases}.$$

By lemma 1 an agent with inflexibility $\delta < \hat{\delta}$ arrives before everyone who is more inflexible and after those who are less inflexible. This implies an agent whom arrives before the peak and has inflexibility δ and desired arrival time t^* arrives at

$$A_e(\delta, t^*) = \begin{cases} t^{\max} - [1 - N_{\delta,e}(\delta)] m_e/s & \delta < \hat{\delta} \\ t^* & \delta \geq \hat{\delta} \end{cases} \tag{41}$$

The top line (41) is the peak of rush hour minus the amount of time it will take everyone who is more inflexible to arrive. This implies the start of rush hour is at

$$t_{01} = t^{\max} - m_e/s. \tag{42}$$

We can use lemma 2 to derive travel times, but first we need to know the marginal type arriving at each time. The marginal type is the type whose preferences determine the slope of the travel time profile. We can then integrate over the slope at each arrival time before t to find the travel time at t.

We can find the marginal type at each arrival time by finding the least inflexible agent to arrive at each arrival time. For $t \in [t_s, t_e]$ the marginal type has inflexibility $\hat{\delta}$. For $t \in [t_{01}, t_s)$ we can find the marginal type by inverting $A_e(\delta, t^*)$ for $\delta < \hat{\delta}$. Doing so and using (42) to simplify, we find the marginal type is at t is

$$B_e(t) = \begin{cases} N_{\delta,e}^{-1}\left(\frac{t-t_{01}}{t^{\max}-t_{01}}\right) & t_{01} \leq t < t_s \\ \hat{\delta} & t_s \leq t \leq t^{\max} \end{cases}.$$

By lemma (2) the travel time at $t < t_s$ is

$$\begin{aligned} T(t) &= \int_{t_{01}}^{t} B_e(t')\, dt' \\ &= \begin{cases} 0 & t < t_{01} \\ \int_{t_{01}}^{t} N_{\delta,e}^{-1}\left(\frac{t'-t_{01}}{t^{\max}-t_{01}}\right) dt & t_{01} \leq t < t_s \\ T(t_s) + (t - t_s)\hat{\delta} & t_s \leq t \leq t^{\max} \end{cases}. \end{aligned}$$

We can rewrite the middle line in a way that will make it easier to interpret later. Let's do a change of variables with $\delta' = N_{\delta,e}^{-1}([t' - t_{01}] / [t^{\max} - t_{01}])$, so that $(t' - t_{01}) / (t^{\max} - t_{01}) = N_{\delta,e}(\delta)$ and $dt = (t^{\max} - t_{01}) n_{\delta,e}(\delta)\, d\delta$, which gives us

$$T(t) = (t^{\max} - t_{01}) \int_0^{N_{\delta,e}^{-1}\left(\frac{t-t_{01}}{t^{\max}-t_{01}}\right)} \delta' n_{\delta,e}(\delta')\, d\delta'.$$

The amount of schedule delay an agent has is the difference between his desired arrival time and when he actually arrives,

$$C_e(\delta, t^*) = t^* - A_e(\delta, t^*).$$

Combining the travel time costs and schedule delay costs gives us an agent's trip price:

$$\bar{p}_e(\alpha, \delta, t^*) = \alpha T \circ A_e(\delta, t^*) + \alpha\delta C_e(\delta, t^*)$$

for $\delta < \hat{\delta}$

$$= \alpha \left(t^{\max} - t_{01}\right) \int_0^{\delta} \delta' n_{\delta,e}\left(\delta'\right)\, d\delta' + \alpha\delta \left\{ t^* - \left[t^{\max} - \frac{m_e}{s}\left(1 - N_{\delta,e}\left(\delta\right)\right)\right]\right\}$$

$$= \alpha \frac{m_e}{s} \left[\int_0^{\delta} \delta' n_{\delta,e}\left(\delta'\right)\, d\delta' + \delta\left(1 - N_{\delta,e}\left(\delta\right)\right)\right] + \alpha\delta\left(t^* - t^{\max}\right)$$

$$= \alpha \frac{m_e}{s} \left[\int_0^{1} \min\left\{\delta', \delta\right\} n_{\delta,e}\left(\delta'\right)\, d\delta'\right] + \alpha\delta\left(t^* - t^{\max}\right) \tag{43}$$

for $\delta \geq \hat{\delta}$

$$= \alpha \left[\left(t^{\max} - t_{01}\right) \int_0^{\hat{\delta}} \delta' n_{\delta,e}\left(\delta'\right)\, d\delta' + \left(t^* - t_s\right)\hat{\delta}\right]$$

substituting $t_s = t^{\max} - \left[1 - N_{\delta,e}\left(\hat{\delta}\right)\right] m_e/s$yields

$$= \alpha \left[\frac{m_e}{s} \int_0^{\hat{\delta}} \delta' n_{\delta,e}\left(\delta'\right)\, d\delta' + \left(t^* - t^{\max}\right)\hat{\delta} + \frac{m_e}{s}\left[1 - N_{\delta,e}\left(\hat{\delta}\right)\right]\hat{\delta}\right]$$

$$= \alpha \frac{m_e}{s} \left[\int_0^{1} \min\left\{\delta', \hat{\delta}\right\} n_{\delta,e}\left(\delta'\right)\, d\delta'\right] + \alpha\hat{\delta}\left(t^* - t^{\max}\right). \tag{44}$$

We can summarize (43) and (44) as

$$\bar{p}_e\left(\alpha, \delta, t^*\right) = \alpha \frac{m_e}{s} \left[\int_0^{1} \min\left\{\delta', \delta, \hat{\delta}\right\} n_{\delta,e}\left(\delta'\right)\, d\delta'\right] + \alpha \min\left\{\delta, \hat{\delta}\right\}\left(t^* - t^{\max}\right).$$

Repeating all of these steps for late arrivals give us

$$\bar{p}_l\left(\alpha, \delta, t^*\right) = \alpha \frac{m_l}{s} \xi \left[\int_0^{1} \min\left\{\delta', \delta, \hat{\delta}\right\} n_{\delta,l}\left(\delta'\right)\, d\delta'\right] - \alpha\xi \min\left\{\delta, \hat{\delta}\right\}\left(t^* - t^{\max}\right).$$

Everything we have done so far in this section has taken $t^{\max}\left(\alpha, \delta\right)$ as given; now we need to solve for it by finding the desired arrival time for each family $\{\alpha, \delta\}$ that is indifferent between arriving early or late. Recall that $t^{\max}\left(\alpha, \delta\right)$ is only defined for $\delta \leq \hat{\delta}$ since those with $\delta > \hat{\delta}$ will arrive on time.[41] This gives us the following functional equation:

$$\bar{p}_e\left(\alpha, \delta, t^{\max}\left(\alpha, \delta\right)\right) = \bar{p}_l\left(\alpha, \delta, t^{\max}\left(\alpha, \delta\right)\right)$$

$$\Rightarrow t^{\max}\left(\alpha, \delta\right) = t^{\max} + \left(\delta + \xi\delta\right)^{-1} \int_0^{1} \min\left\{\delta', \delta\right\} \left[\xi \frac{m_l}{s} n_{\delta,l}\left(\delta'\right) - \frac{m_e}{s} n_{\delta,e}\left(\delta'\right)\right] d\delta' \tag{45}$$

[41]The families with inflexibility $\hat{\delta}$ will arrive on-time and be indifferent between arriving a little earlier or later, depending on whether they are arriving before or after the peak. The agent with inflexibility $\hat{\delta}$ who arrives exactly at the peak will be indifferent between arriving early or late.

Problematically, $n_{\delta,l}$ and $n_{\delta,e}$ are functions of $t^{\max}(\alpha,\delta)$; however, we can solve (45) with a lucky guess. When there are just two families $t_i^{\max} = (t_s + \xi t_e) / (1+\xi)$ for $i \in \{1,2\}$.[42] Let's guess that this result holds even with a continuum of types.

If $t^{\max}(\alpha,\delta)$ is a constant then $N_{\delta,e}(\delta) = N_{\delta,l}(\delta) = N_\delta(\delta)$. Furthermore, if $t^{\max}(\alpha,\delta) = (t_s + \xi t_e) / (1+\xi)$ then $m_e = \xi/(1+\xi)$ and $m_l = 1/(1+\xi)$. Plugging these results into (45) we find that the final factor in the integral is zero and so the entire second term is zero. Thus we find $t^{\max}(\alpha,\beta) = t^{\max} = (t_s + \xi t_e) / (1+\xi)$ and so this is indeed a solution.

Plugging these results into our formulas for equilibrium trip price we find

$$\bar{p}_{\text{free}}(\alpha,\delta,t^*) = \alpha \frac{1}{s}\frac{\xi}{1+\xi}\left[\int_0^1 \min\left\{\delta',\delta,\hat{\delta}\right\} n_\delta\left(\delta'\right)\, d\delta'\right] - \left(t^{\max} - t^*\right)\alpha \min\left\{\delta,\hat{\delta}\right\}\begin{cases}1 & t^* \le t^{\max}\\ -\xi & t^* > t^{\max}\end{cases}. \tag{46}$$

7.2. **All toll.** To find the equilibrium trip prices when the entire road is priced we follow the same steps as when it is free. The only difference is that we replace δ with β. So instead of finding $\hat{\delta}$ we find $\hat{\beta}$, and instead of finding the marginal distribution of δ, we find the marginal distribution of $\beta = \alpha\delta$, etc. Doing so yields

$$\bar{p}_{\text{toll}}(\alpha,\delta,t^*) = \frac{1}{s^*}\frac{\xi}{1+\xi}\left[\int_0^\infty \min\left\{\beta',\alpha\delta,\hat{\beta}\right\} n_\beta\left(\beta'\right)\, d\beta'\right] - \left(t^{\max} - t^*\right)\min\left\{\alpha\delta,\hat{\beta}\right\}\begin{cases}1 & t^* \le t^{\max}\\ -\xi & t^* > t^{\max}\end{cases}. \tag{47}$$

We have a closed form solution for each agents trip price on a completely free or priced route, up to the possible need to solve the integrals numerically.

There is some nice intuition behind (46) and (47). We can write the equilibrium trip price generically as

$$\text{trip price} = \frac{\xi}{1+\xi} \times \text{length rush hour} \times \begin{array}{c}\text{censored mean of}\\ \text{willingness to pay}\end{array} - \begin{array}{c}\text{adjustment for}\\ \text{desired arrival time}\end{array}. \tag{48}$$

Let's work through each term of (48).

The ratio $\xi/(1+\xi)$ is a measure of how the cost of being late compares to the cost of being early and, as we saw above, is equivalent to the fraction of agents who arrive before the peak of rush hour. If ξ is zero then it is costless to be late, as a result agents can wait to travel until there is no traffic or toll; everyone will be late and have a trip price of zero. As ξ increases the costs of being late increases and so a larger share of agents arrive before the peak. Because drivers care more about arriving on-time, travel times (or tolls) are higher and everyone's trip price increases.

[42] This is shown in appendix C.

The next factor in the first term is the length of rush hour, which on a free route is s^{-1} and on a priced route is s^{*-1} because we normalized the mass of agents to one. A longer rush hour means more schedule delay and higher travel times or tolls, and so increases trip prices.

The final factor of the first term is the most interesting; the integral in (46) and (47) is the censored mean of an agent's willingness to pay in whatever currency the route requires. So on a free route it is the censored mean of the agent's inflexibility, or willingness to pay in travel time to reduce schedule delay, while on the priced route it is the censored mean of the agent's willingness to pay in dollars to reduce schedule delay. On the free route we then multiply this by the agent's value of time to convert from travel time to dollars.

The censoring occurs at the willingness to pay of the marginal agent who arrives at the same time as the agent whose trip price we are considering. For an agent with $\delta < \hat{\delta}$ on a free route or $\beta < \hat{\beta}$ on a priced route this is his own willingness to pay. This means he doesn't care about the actual preferences of those with a higher willingness to pay; whether they are willing to pay a cent more or a thousand dollars more for the most desirable arrival times doesn't matter, either way they will outbid him for the most desirable arrival times and so all that matters is how much of the desirable arrival time they will use. In contrast, he cares very much about the preferences of those whom he must outbid, since he must actually outbid them.

If an agent is inframarginal, so $\delta > \hat{\delta}$ on a free route or $\beta > \hat{\beta}$ on a priced route, then the censoring occurs at the marginal willingness to pay of the marginal agent at the time they arrive. This has similar to what we see in models of perfectly competitive markets, where prices are based on the preferences of the marginal agent.

The final term is an adjustment for differences in desired arrival times. Those who want to arrive at the peak of rush hour will pay the highest prices, while those who prefer to arrive further from the peak will pay lower prices.

7.3. **Value pricing.** When we price part of the road we give agents an additional choice to make: which route to take. While when there were just two families we were able to solve for the value pricing equilibrium analytically, we will no longer be able to do so, and so will need to solve equilibrium numerically. Fundamentally, we first assign agents to routes and then solve for equilibrium on each route. Solving numerically will require me to use several approximations and I will choose them such that we can used the closed form solutions above for a completely free or priced highway to find equilibrium on a route given the agents who are on it.

The assignment of agents to routes is made simpler by the following lemma, which allows us to divide the space of agents' preference parameters into those on the free route and those on the priced route using a continuous function.

Lemma 13. *For a given flexibility and desired arrival time there is a value of time, $\hat{\alpha}(\delta, t^*)$ such that all agents with a higher value of time travel on the priced route and all agents with a lower value of time travel on the free route. Furthermore, $\hat{\alpha}$ is a continuous function if the travel time profile and toll schedule are continuous.*

This is not as strong as lemma 10 but still nicely reflects our intuition that the rich will be on the priced route and the poor will be on the free route.

It is unlikely that after conditioning on route choice the distribution of desired arrival times will be uniform and independent of α and δ. This means that the marginal type will not be constant over $[t_s, t_e]$, however, I will approximate it with a constant. This will allow us to apply (46) and (47) to each route individually, adjusting for route capacity and the distribution of agents on the route, to find the trip prices:

$$\bar{p}_{\text{free}}(\alpha, \delta, t^*) = \alpha \frac{1}{(1-\lambda)s} \frac{\xi}{1+\xi} \left[\int_{t_s}^{t_e} \int_0^1 \int_0^{\hat{\alpha}(\delta', t')} \min\left\{\delta', \delta, \hat{\delta}\right\} n\left(\alpha', \delta', t'\right) \, d\alpha' d\delta' dt' \right] - \alpha \min\left\{\delta, \hat{\delta}\right\} (t^{\max} - t^*) \begin{cases} 1 & t^* \leq t^{\max} \\ -\xi & t^* > t^{\max} \end{cases}, \tag{49}$$

$$\bar{p}_{\text{toll}}(\alpha, \delta, t^*) = \frac{1}{\lambda s^*} \frac{\xi}{1+\xi} \left[\int_{t_s}^{t_e} \int_0^1 \int_{\hat{\alpha}(\delta', t')}^{\infty} \min\left\{\beta', \alpha\delta, \hat{\beta}\right\} n\left(\alpha', \delta', t'\right) \, d\alpha' d\delta' dt' \right] - \min\left\{\alpha\delta, \hat{\beta}\right\} (t^{\max} - t^*) \begin{cases} 1 & t^* \leq t^{\max} \\ -\xi & t^* > t^{\max} \end{cases}. \tag{50}$$

Without making this assumption we would not have a closed form solution for trip prices. In my main specification the largest error in the marginal type less than 0.2%. Given the small size of the approximation error and how much it helps in solving for equilibrium, this approximation seems a reasonable approach.

The intuition for (46) and (47) also apply to (49) and (50); we just need to multiply and divide the first term by the mass of agents on that route. The mass of agents divided by the route's capacity gives the length of rush hour on that route, and dividing the integral by the mass of agents on that route turns it back into a censored mean, this time using the distribution of agents on the route.

I will use projection methods to solve for $\hat{\alpha}(\delta, t^*)$, specifically Chebyshev collocation. Given our approximation of the marginal type over $[t_s, t_e]$ we can further simplify $\hat{\alpha}$ using the next lemma, which shows that $\hat{\alpha}$ is often flat in one dimension.

Lemma 14. *All agents in a family that is not inframarginal regardless of which route they are on will travel on the same route or be indifferent between both routes. Similarly, all agents who are inframarginal regardless of which route they are on and who have the same value of time and desired arrival time will travel on the same route or be indifferent between both routes.*

The intuition for the first claim is that when an agent is not inframarginal his desired arrival time does not determine his actual arrival time, except for whether he is early or late, and so his trip price differs from the other agents in his family only by the adjustment for desired arrival time. This adjustment is the same on both routes and so cancels out looking at the difference between trip prices on either route, and so if one route is preferred by one agent in a family, it must be preferred by all agents.

The proof for the second claim is that if an agent is inframarginal regardless of which route he chooses, then he arrives on-time regardless of the route he chooses. This means his cost on the free route is $\alpha T\left(t^*\right)$ and his cost on the priced route is $\tau\left(t^*\right)$, and he will chose whichever route has the lowest cost. This holds for any agent who is inframarginal regardless of which route he chooses and who has the same value of time and desired arrival time, and so all of these agents will make the same choice.

Given the approximation of a constant marginal type over the support of the distribution of desired arrivals, lemma 14 implies

$$\delta < \hat{\delta} \text{ and } \alpha\delta < \hat{\beta} \Rightarrow \frac{\partial \hat{\alpha}\left(\delta, t^*\right)}{\partial t^*} = 0, \quad \text{and}$$
$$\delta > \hat{\delta} \text{ and } \alpha\delta > \hat{\beta} \Rightarrow \frac{\partial \hat{\alpha}\left(\delta, t^*\right)}{\partial \delta} = 0.$$

Because of this I approximate $\hat{\alpha}\left(\delta, t^*\right)$ as

$$\hat{\alpha}\left(\delta, t^*\right) = \begin{cases} \hat{\alpha}_M\left(\delta\right) & \delta < \hat{\delta} \\ \hat{\alpha}_I\left(t^*\right) & \delta \geq \hat{\delta} \end{cases}, \tag{51}$$

where $\hat{\alpha}_M\left(\delta\right)$ and $\hat{\alpha}_I\left(t^*\right)$ are solved for using Chebyshev collocation. This approximation is wrong over the small area where $\left(\delta - \hat{\delta}\right)\left(\hat{\alpha}\left(\delta, t^*\right)\delta - \hat{\beta}\right) < 0$, and in contrast to the two dimensional Chebyshev approximation of $\hat{\alpha}\left(\delta, t^*\right)$, it will not converge to the true $\hat{\alpha}\left(\delta, t^*\right)$ regardless of the degree of the Chebyshev polynomial.

However, using (51) has significantly better small sample performance. In my baseline model the approximation error, measured as the largest welfare loss from traveling on the route assigned by $\hat{\alpha}\left(\delta, t^*\right)$ instead of the route that actually minimizes trip price, is less than a tenth of a cent using (51) with tenth degree Chebyshev polynomials, for twenty nodes total, while the approximation error is nearly a dollar using the tensor product of two tenth degree Chebyshev polynomials, for one hundred nodes in total.

Formally, solving for equilibrium is finding the $\hat{\beta}, \hat{\delta}$, and $\hat{\alpha}$ such that

$$\int_{t_s}^{t_e} \int_0^{\hat{\delta}} \int_0^{\hat{\alpha}(\delta,t^*)} n\left(\alpha', \delta', t'\right)\, d\alpha' d\delta' dt' = (1-\lambda)\, s\left(t_e - t_s\right),$$

$$\int_{t_s}^{t_e} \int_0^{1} \int_{\max\left\{\hat{\alpha}(\delta,t^*), \delta^{-1}\hat{\beta}\right\}}^{\infty} n\left(\alpha', \delta', t'\right)\, d\alpha' d\delta' dt' = \lambda s^*\left(t_e - t_s\right), \quad \text{and}$$

$$\bar{p}_{\text{free}}\left(\hat{\alpha}\left(\delta, t^*\right), \delta, t^*\right) = \bar{p}_{\text{toll}}\left(\hat{\alpha}\left(\delta, t^*\right), \delta, t^*\right) \quad \text{for all } \{\delta, t^*\} \in \mathcal{C}.$$

where $\mathcal{C}$ is the set of Chebyshev collocation nodes.

8. Estimating Distribution of Consumer Preferences

As we saw when there were just two families, whether value pricing can make all road users better off depends on agents' preferences. We now turn to estimating the distribution of agents' preferences, along with a few other parameters, so we can determine the distribution and size of the welfare gains from congestion pricing.

The main structural object we need to estimate is the joint distribution of agents' inflexibility, value of time, and desired arrival time. My general approach will be to estimate the joint distribution of agents' preferences separately for two broad categories of agents, those who are able to choose when they arrive at their destination and those who can't, and I will call them the inflexible and flexible categories. I will assume that within each category, inflexibility, value of time, and desired arrival time are independent, but that the marginal distributions can differ across categories.[43]

8.1. Data.

We will estimate this joint distribution for drivers on a segment of California State Route 91 (SR-91). The segment we will focus on is thirty-three miles long and runs from the center of Corona to the junction of SR-91 and I-605. The primary reason to focus on SR-91 is that good data is available on those who use SR-91 because SR-91 was the first highway to have a portion of its lanes priced and so it has been extensively studied. I choose this specific segment because is roughly represents a median commute for those living in Corona who use SR-91.

I use data from three sources. The first, California Polytechnic State University's State Route 91 Impact Study (SR-91 IS) (Sullivan, 1999), is a series of surveys conducted between 1995 and 1999 of drivers who use SR-91 The second data set, the 2009 National Household Travel Survey (NHTS) (U.S. Department of Transportation, 2009), is a national survey. The SR-91 IS allows us to measure variables specifically for those who use SR-91, while the NHTS allows us to confirm that these results are representative of other large metropolitan statical areas (MSA) and how they compare to rest of the nation.[44]

[43] This is necessary in part because I cannot observe the inflexibility of those in the inflexible category or the desired arrival time of those in the flexible category. For evidence that the desired arrival times of those in the inflexible category are not strongly correlated with income see appendix D.1.

[44] I define a large MSA as one with a population above three million.

TABLE 2. Fraction of drivers and trips that are flexible

		NHTS	
Fraction of ...	SR-91 IS	Large MSAs	All
Drivers who leave early or late to avoid traffic	.57 [.55, .60]		
Workers who commute via interstate who can choose work arrival time	.50 [.47, .53]	.47 [.45, .49]	.44 [.43, .45]
Trips on interstate during morning that are flexible	.43 [.40, .47]	.35–.60 [.32, .62]	.30–.59 [.29, .60]

Notes: 95% confidence intervals in brackets. For second and third column confidence intervals calculated using Jackknife-2 replicate weights. A trip is flexible if the driver and all passengers can choose when to arrive at their destination, unless the destination is the driver's home, in which case they must be able to choose their departure time. A trip is a series of trip segments which ends when the driver stays at one destination for more than thirty minutes.

The final data set is the California Department of Transportation's Performance Measurement System (PeMS) (California Department of Transportation, 1999). PeMS includes road detector data from almost all of the highways in California. From this database I can calculate travel times for each arrival time for every non-holiday weekday in 2004.[45] I do this for the start of every five minute interval between 4:00 a.m. and 10:00 a.m.

8.2. **Fraction of drivers who are flexible.** The first task is to estimate the relative sizes of the two categories of agents using surveys of road users from the SR-91 IS and NHTS. Table 2 reports three different measures how many on the highway are flexible for road users from three samples.

The main result is that roughly half of all road users are flexible and can adjust when they travel. The first row is a revealed preference measure of flexibility; fully 57% of road users on SR-91 reported that in their typical peak period travel they left early or late to avoid traffic congestion. Furthermore, about half of the workers who commute to work using the interstate are able to choose what time they arrive at work.[46]

However, the interstate gets used for more than just traveling to and from work, and including these other trips during the morning reduces my estimate the relative sizes of

[45] I define holidays as the ten United States Federal Holidays.

[46] The NHTS asks whether a worker can set or change when he starts work. Problematically for my purposes this means that a worker who can start work whenever he wants but who must drop his children off at school at a specific time would look flexible to me, even though he is not. The SR-91 IS does not have this problem as it asks directly whether the respondent could set their arrival time for the trip.

the two categories of agents.[47] I report a range of values for the fraction of trips that are flexible for the NHTS because it only reports if arrival times can be chosen for work trips, and so I need to make assumptions about what kind of non-work trips are flexible. Assuming that only work trips can be flexible leads to estimates that thirty percent of trips are flexible, while assuming that other trips where the driver probably has control over when it begins, such as shopping, doctors appointments, and visiting friends, are flexible almost doubles the percentage of trips that are flexible.

I will use the fraction of trips in the morning that are flexible from the SR-91 IS as the primary measure of the fraction of agents who are flexible.

8.3. **Distribution of value of time.** To estimate the distribution of the value of time I first map household income into an individual value of time and then fit a log-normal distribution to the data using maximum likelihood. I do this separately for the flexible and inflexible drivers. To map household income to value of time I use the U.S. Department of Transportation's formula for local personal travel: an individual's value of time is half their hourly household income, which is their annual household income divided by 2,080 hours per year (Belenky, 2011, p. 12).[48]

The results of doing so are in Table 3. In every case but the first the median value of time of the flexible drivers is higher than that of the inflexible drivers, as shown by the positive and statistically significant Goodman and Kruskal's rank correlation. The first definition of flexibility is a measure of whether the agent is flexible while the third is a measure of whether the trip is flexible. Thus we find that higher income people have more flexible trips, but due to other factors in their lives are not actually any more flexible than the poor. These results are consistent with the intuition that better paid jobs tend to be more flexible, but that better paid workers tend to be older and have more constraints in their personal lives, such as needing to take care of their children. By using the second column as our main specification we make it harder to find a Pareto improvement.

Using the definition of flexibility based on whether the trip was flexible, the SR-91 IS results are similar to those for large MSAs in the NHTS, except that the interquartile range is much larger in the NHTS data. Likely this is due to selection into living in Riverside County and commuting on SR-91. This smaller group is likely to be more similar that the larger group of those who live in a large MSA.

[47]The morning is defined in the SR-91 IS as 4–10 a.m. and so for consistency I maintain that definition with the NHTS.

[48]The U.S. Department of Transportation uses this formula to estimate a median value of time based on median household income, I am going further in using it by applying it to individuals.
There is a large literature estimating the mean or median value of time, which generally finds it is half the mean or median wage, though it is higher when roads are congested. There is a much smaller and much more recent literature that looks at the distribution of the value of time. See Small and Verhoef (2007, p. 53) for a literature review.

TABLE 3. Distribution of value of time for morning highway users

	SR-91 IS		NHTS	
	All	All	Large MSAs	All
Definition of flexibility†	1	3	3	3
Flexible				
Median	22.12	25.95	26.05	20.41
	(0.64)	(0.88)	(0.34)	(0.13)
Interquartile range	16.4	20.0	32.21	25.41
	(1.0)	(1.8)	(0.89)	(0.35)
N	413	303	7,059	21,342
Inflexible				
Median	22.71	22.16	22.52	19.02
	(0.73)	(0.56)	(0.27)	(0.11)
Interquartile range	16.4	15.19	24.55	18.95
	(1.3)	(0.88)	(0.59)	(0.19)
N	292	433	4,270	12,995
Rank correlation‡	-0.053	0.20***	0.157***	0.108***
	(0.059)	(0.057)	(0.037)	(0.028)

Notes: Standard errors in parentheses. Standard errors in columns one and two are calculated by bootstrapping. The data for columns three and four are weighted using individual weights and their standard errors calculated using Jackknife-2 replicate weights.

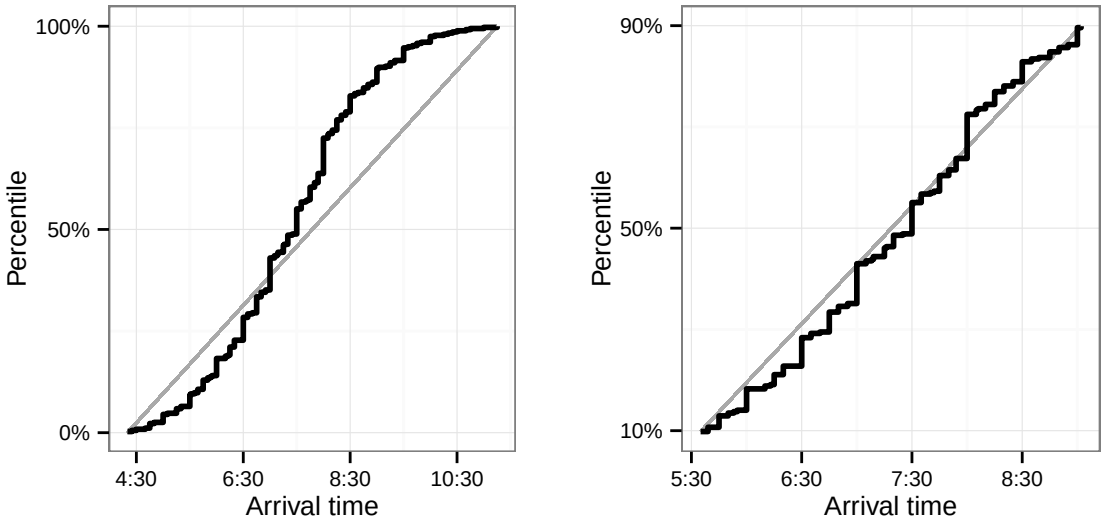
† Refers to the row of table 2 that is used to define who is flexible. I use the most generous numbers for definition 3 and the NHTS data.

‡ Goodman and Kruskal's γ between income and flexibility.

*** $p < .001$

The SR-91 IS results are fairly similar to those of Small et al. (2005), which uses revealed and stated preference data to measure the distribution of value of time for road users on SR-91. While they do not estimate the distribution separately for flexible and inflexible agents, I can compare how my more crude method compares when applied to the entire population. Adjusting for inflation, they find that the median value of time is \$29.54 and the interquartile range is \$10.47, while I find a median of 23.58 and an interquartile range of 17.06. The lower median means I will undervalue the time savings while the larger interquartile range means I will have increased inequality and make it harder to find a Pareto improvement.

8.4. **Distribution of desired arrival time.** We now want to measure the distribution of desired arrival times. What we care about is when agents want to arrive at the highway exit, rather than when they want to arrive at their destinations. Unfortunately, the data reports the second instead of the first and so I have actual arrival times rather than desired,



(A) All data

(B) 10th–90th percentiles

FIGURE 7. Cumulative distribution function of arrival time for agents who cannot choose when they arrive at their destination and arrive before noon. Data from SR-91 IS.

and the arrival time at the destination rather than at the highway exit. I can avoid the first problem by focusing on those who are unable to choose their arrival time, and so are in the inflexible category. For them their actual arrival time is their desired arrival time. While I only observe the distribution of desired arrival times for those who are inflexible, I will assume the distribution is the same for both categories. I will fix the second problem by estimating the first desired arrival time at the highway exit as part of estimating the distribution of inflexibility.

Using the SR-91 IS I can estimate the distribution of desired arrival times for these inflexible agents; the CDF of this distribution is shown in Figure 7a. If the distribution was uniform then the CDF would lie along the 45 degree line; it is clear that the distribution of desired arrival times is not uniform, contrary to what we have assumed throughout this paper. However, when we remove the first and last ten percent of drivers to arrive, as in Figure 7b, then the distribution is close to being uniform. This is a robust finding and holds within the NHTS data as well; in appendix D I redo Figure 7 using the NHTS for several MSAs as well as the entire sample.

The distribution of when agents want to arrive at the end of the highway will be a smoothed version of the distribution of when agents want to arrive at their destination. This is because the distance from the end of the highway to their destinations will vary, and so among those who want to reach their destination at 7:00 a.m. there will be some

who want to reach the end of the highway at 6:40 a.m. and others who want to reach it at 6:55 a.m. Thus, while Figure 7 has a few times when a significant number of all drivers want to arrive, as seen by when the CDF is vertical, such as at 7:00 and 8:00 a.m., it is unlikely the distribution of desired arrival times at the highway exit has the same mass points.

The distribution of desired arrival times of the inflexible agents is not uniform, however, assuming it is uniform is a reasonable approximation and is relatively innocuous. By truncating the extreme deciles I am ignoring those who want to arrive extremely early or late. Some of these workers are arriving outside of rush hour, and so they are not relevant for my analysis, and the others are among those who are least harmed by congestion pricing. They are already traveling off-peak, and so cannot be displaced by rich drivers who decide to travel at the peak once the road is tolled, and because congestion pricing can reduce the length of rush hour they may find that after pricing they are traveling outside of rush hour and so pay no toll. Should congestion pricing help those who want to arrive at the peak of rush hour it almost certainly also helps those who want to arrive at the tails.

The assumption that the distribution of desired arrival times is the same for both categories of agents is somewhat harmless. If an agent is never inframarginal then his desired arrival time does not affect when he travels, other than whether he is early or late; he arrives off-peak and when he arrives is determined by his δ or β depending on whether the route is free or priced. Ascribing the wrong desired arrival time to these agents will not affect equilibrium or the change in the agents' trip prices.[49] However, this will not be true for agents who would be inframarginal on one or both routes. Then by assuming an agent has a different desired arrival time then he actually does I miscalculate which agents are inframarginal, and so miscalculate the slope of the travel time profile or toll schedule.

I estimate the endpoints of the distribution of desired arrivals times by matching the largest and smallest remaining observation to the expected value of their order statistics. Doing so gives me an unbiased estimate of the endpoints and length of the uniform distribution which would generate the truncated distribution in Figure 7b.[50] Following this procedure gives me an estimate of 4.41 hours for the length of desired arrivals, as is reported in Table 4.

8.5. **Distribution of inflexibility.** The bottleneck model provides a mapping between model parameters and the travel time profile. By inverting this mapping we can estimate our remaining parameters: the distribution of inflexibility; the ratio of the cost of being early to late, ξ; and the length of rush hour on a free route, $1/s$.

[49]It will affect the level of their trip prices, but in a consistent way so that it differences out.
[50]See appendix D.2 for a proof.

I will only be able to estimate the distribution of inflexibility for those drivers who do not arrive on-time. For all other drivers I will only have a lower bound. This follows from lemma 2 and is because the travel time profile does not reflect the preferences of the inframarginal drivers.

I estimate the distribution of inflexibility in two ways; my primary approach is to use GMM to chose the model parameters to best fit the empirical travel time profile. Specifically, my moment conditions are that the model predicted travel times matches the sample average travel times. My second approach is to non-parametrically fit the model, this second approach is only partially identified and its value is in allowing us to check the restrictiveness of our functional form assumptions.

For the GMM estimator I assume that the distribution of inflexibility for those agents who are in the flexible category is uniform on $\left[0, \tilde{\delta}\right]$, where $\tilde{\delta}$ is unknown and must be estimated. In order to map the model to the data I also estimate the earliest time that agents want to reach the intersection of SR-91 and I-605, t_s, as well as free flow travel times, T^f. I use the estimates of the fraction of agents who are flexible and the length of desired arrivals from above in fitting the model; I estimate them separately because I have natural measures of them and so want to match those particular "moments" exactly.

When non-parametrically fitting the model I choose the travel time at the start of each interval as well as when the peak of rush hour occurs to minimize the GMM criterion subject to three sets of constraints imposed by the theory:

(1) Travel times are positive
(2) Travel times are increasing before the peak and decreasing after
(3) Travel times are convex before the peak and convex after the peak

The first constraint is never binding and the third constraint makes the second constraint redundant for all but the first and last arrival times.

The empirical travel time profile along with the predicted travel time profiles from both methods are shown in Figure 8. The two predicted travel time profiles are very similar, and the root GMM criterion is only 7.7% higher when assuming inflexibility is uniformly distributed. The two predicted travel time profiles differ the most at 5:00 and 10:00 a.m. The difference at 10:00 largely results from not imposing the assumption that $\gamma_i = \xi\beta_i$ for all families i in the non-parametric estimation. The small difference in the root GMM criterion of the non-parametric and GMM estimates suggests that it is innocuous to assume the distribution of inflexibility for those agents who are flexible is uniform and $\gamma_i = \xi\beta_i$ for all families i.

The results from the GMM estimation are reported in Table 4. I estimate that the inflexibility of those in the flexible category is uniformly distributed on $[0, 0.228]$ and then assume that the inflexibility of those in the inflexible category has a beta$(5, 0.5)$ distribution transformed to have support $[0.228, 1]$. I choose the parameters of the beta distribution to put most of the weight near one and have the mode be at one. The mixture of these

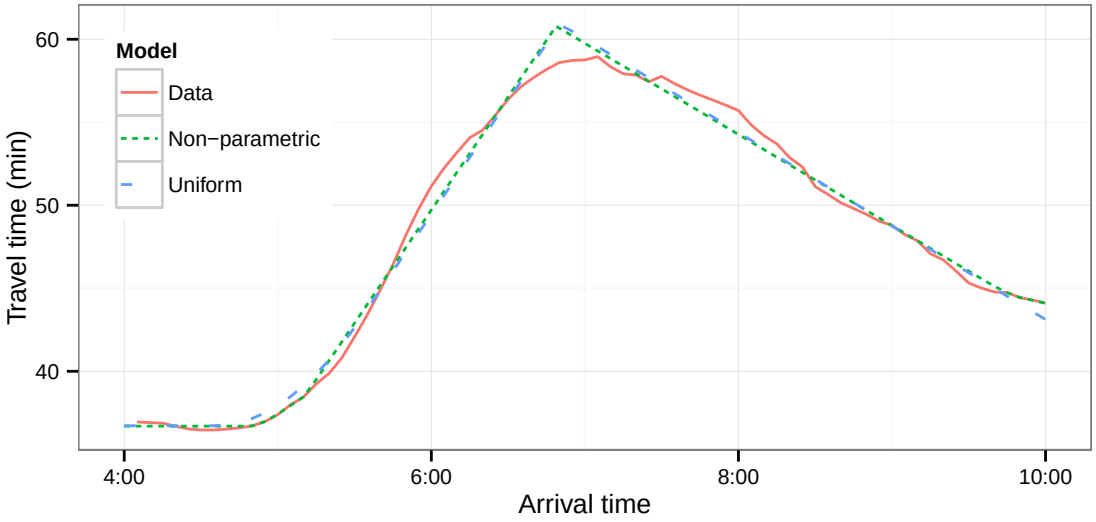


FIGURE 8. Actual vs predicted travel times

TABLE 4. Remaining base case parameter values

Parameter	Point estimate	Standard error
Maximum inflexibility of flexible agents	0.228	(0.045)
Ratio of schedule delay costs late to early	0.411	(0.031)
Length of rush hour on free route (hours)	7.73	(0.39)
Length of desired arrivals (hours)	4.40	(0.23)

Note: Standard errors calculated by bootstrapping. The first three rows report the GMM estimates (N = 250), the final row comes from fitting the largest and smallest observations of the trimmed sample of desired arrival times to the expected value of their order statistics (N = 489).

distributions, weighted by the fraction of agents in each category, gives the marginal distribution of inflexibility and is plotted in Figure 9. While we would probably expect a smoother distribution than Figure 9 shows, by pushing most of the inflexible agents towards one I make it harder to find a Pareto improvement.

I find that the cost of being late is less than the cost of being early. This contrasts with our intuition and with previous research, however, it is largely a result of how I estimate the ratio of the cost of being late to early.[51] I only observe β/α for the marginal drivers arriving before the peak of rush hour and γ/α for the marginal drivers arriving after the peak. I assume a functional form for the distribution of β/α and assume that for each agent the cost of being late is just some constant multiple of the cost of being early ($\gamma_i = \xi\beta_i$). My finding that the cost of being late is less than the cost of being early

[51] cf. Small (1982); Hendrickson and Plank (1984); Parthasarathi et al. (2010).

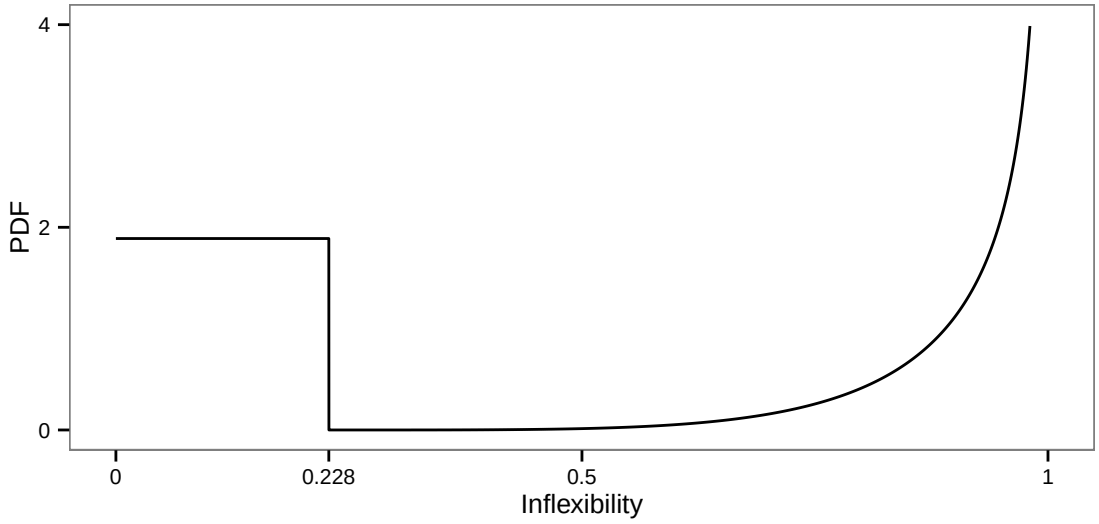


FIGURE 9. Marginal distribution of inflexibility

is best interpreted as being that the marginal driver who is late pays less of a cost then the marginal driver who is early; there is nothing unreasonable about this. Furthermore, being late doesn't necessarily mean literally arriving late to an appointment, but can be that you would prefer to go to the doctor at 9 a.m. but instead schedule the appointment for 11 a.m. to avoid traffic. You arrive exactly on-time to your 11 a.m. appointment, but still have schedule delay costs.

I estimate that rush hour is more than seven and a half hours long; starting before five in the morning and not ending until a little after noon. While this sounds extreme, it follows from my definition of rush hour as the period when travel times are higher than they would be in free flow conditions, rather just when travel times are terrible.

9. COUNTERFACTUALS

Given the estimates of the distribution of driver preferences we can use the results of section 7 to solve for equilibrium under counterfactual congestion pricing regimes. This allows us to predict what would happen if we added time varying tolls to the highway, either all or some of the lanes.

The final parameter we need to construct counterfactuals is the amount that throughput falls once a queue forms. Based on the transportation engineering literature summarized in Table 1 I assume throughput falls ten percent once the queue forms.

Table 5 reports measures of the aggregate effects of pricing all or a fourth of the highway. The headline result is that pricing a fourth of the lanes helps all road users, while pricing all of it only helps 75% of them. Pricing a portion of the lanes is a Pareto improvement, while pricing all of the road is not.

TABLE 5. Aggregate effects of congestion pricing

	All free	All tolled	Pricing 1/4th of lanes
Fraction agents better off		0.73	1
Toll revenue ($ per capita)		3.78	1.45
Largest increase in cost ($)		6.81	0
Annual welfare gains ($ per capita)[†]			
Social		2384	1037
Private		491	309
Excess travel times (min)			
Average	9.73	0	9.85
Peak	24.18	0	23.92
Toll ($)			
Average	0	3.79	5.39
Peak	0	9.49	13.10

[†] I assume two trips per working day and 250 working days per year.

Pricing all of the road raises about two and a half times the revenue that pricing just a fourth does. The decrease in revenue is not proportional to the decrease in the fraction of the road priced because the tolls are higher when just a fourth of the road is priced. Tolls are higher because when fewer lanes are priced there are fewer agents in the priced lanes, and so the marginal agent has a higher value of time. The tolls reflect the marginal agents preferences and so are higher.

While pricing all of the road raises significant revenue, it is not enough for a lump sum rebate to make pricing the entire road a Pareto improvement. The worst off agent is almost seven dollars worse off each trip, approaching twice the revenue raised per agent per trip. This means that if we want to use the revenue to make pricing the entire road a Pareto improvement we will need to spend it in a way that targets those who are harmed.

By only pricing a fourth of the lanes we are able to get a Pareto improvement even before using the revenue, however, this Pareto improvement comes at a cost. To obtain a Pareto improvement we need to give up more than half of the social welfare gains available from congestion pricing. That said, if by making congestion pricing a Pareto improvement we are able to actually implement congestion pricing then we are trading $2384 per person per year of potential, unrealized, welfare gains for $1037 per person per year of actual welfare gains. This is a good trade.

The magnitude of the welfare gains available from congestion pricing are large, over a thousand dollars per road user per year. Pricing a fourth of the lanes would be equivalent to increasing the median income of these agents by almost 1.5%, and pricing all of the road would increase median income by over 3%. Most of the welfare gain comes from changing the currency used to pay for desired arrival times from time to money. The time spent in traffic is a social loss while the money spent on tolls is just a transfer. This

portion of the welfare gains accrues to whomever gets to keep the toll revenue. However, a significant amount of the welfare gains go to the road users themselves. Even if the toll revenue is burned the average road user will be $309 better off each year due to value pricing.

Value pricing captures a large portion of the welfare gains available, even though we are pricing only a fourth of the lanes. This contrasts with Verhoef et al. (1996) as well as Liu and McDonald (1998, 1999) who find that pricing part of the highway only yields a small portion of the welfare gains available from congestion pricing. They find this result because if congestion pricing requires reducing throughput, then when we price only part of the road we need to take into account that by pricing drivers out of the priced lanes we make traffic worse in the free lanes. This leads to lower tolls and more congestion on the priced route, and so a smaller share of the welfare gains. In contrast, when pricing increases throughput we do not have this additional concern.

Pricing a fourth of the lanes reduces travel times at every point in time, on average by 1.5%, and so we pass the simple test for a Pareto improvement we constructed at the end of section 5. However, because agents arrive over a smaller period of time, pricing a fourth of the lanes increases the average travel time agents experience.[52]

Figure 10 shows how congestion pricing affects different families of agents. The left panel of Figure 10 shows the change in trip price due to pricing the entire road for each inflexibility and all values of time below fifty dollars an hour, averaged over all desired arrival times. [53] The agents harmed the most by pricing all of the road are the inflexible poor, those who need to arrive to work exactly on-time and who would strongly prefer to pay with their time to do so instead of their money. The curve of darkest red in the lower right of Figure 10 lies along the curve $\alpha = \hat{\beta} \cdot \delta$; these are the agents who were able to arrive exactly on time when the road is free, but when the road is priced they are displaced from the peak of rush hour by flexible rich drivers who start arriving during the peak. The inflexible rich are the best off, when the road is free they arrive on-time but bear large travel time costs, and they are thrilled to be able to pay with money to avoid the travel time costs. The flexible are not much affected by adding tolls to the highway, they avoided paying with travel time by arriving off-peak and they will avoid paying with money by continuing to arrive on peak. They are better off since they will have less schedule delay, but as they have a low cost of schedule delay the improvement in their welfare is slight.

[52]Put differently, there are now times on the free route when travel times are zero, but as one one travels at these times they are not included in the average.

[53]Because the right panel of Figure 10 shows the change in trip price averaged over desired arrival times it shows that the worse off family is hurt by three dollars while Table 5 reports the agent most harmed by pricing the entire road is almost seven dollars worse off.

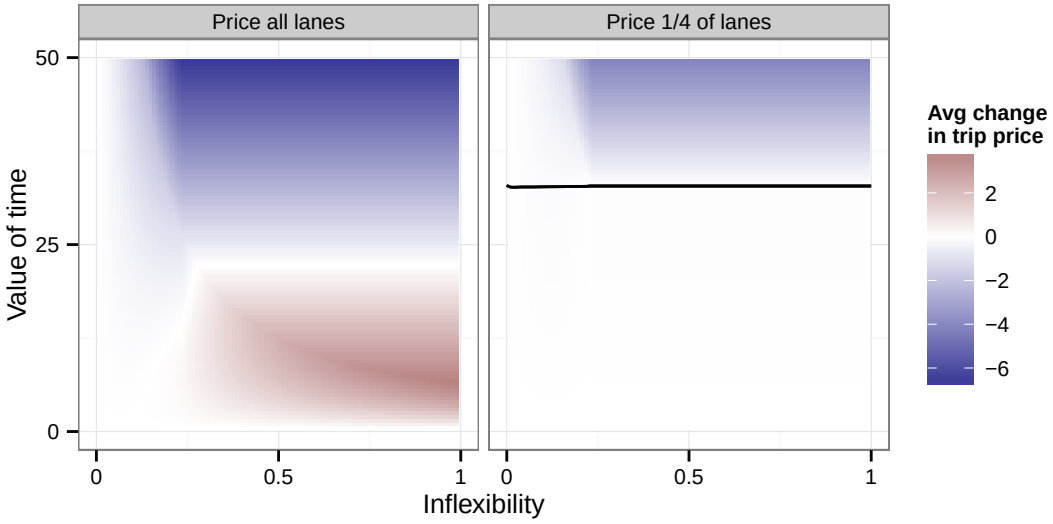


FIGURE 10. Average change in trip price by family. The black lines in the right panel are the maximum and minimum values of $\hat{\alpha}(\delta, t^*)$ for each δ.

By pricing just a fourth of the lanes we preserve the ability of the poor to pay with their time, and so as the right panel of Figure 10 shows, avoid hurting the inflexible poor. We reduce the benefits to the inflexible rich, but we have a Pareto improvement.

The right panel of Figure 10 also shows which agents are on which road. The black lines are the maximum and minimum values of $\hat{\alpha}(\delta, t^*)$ for each δ, and so separate the space of families into those on the priced route and those on the free route. Those above both lines are on the priced route, those below both are on the free route, and for those families whose parameters are between the two lines the route an agent is on depends on his desired arrival time.

In both panels of 10 the change in trip price is constant for a given value of time across a range of high levels of inflexibility. This occurs for the same reason $\hat{\alpha}(\delta, t^*)$ is flat for $\delta > \hat{\delta}$ and $\hat{\alpha}(\delta, t^*) > \hat{\beta}/\hat{\delta}$: if an agent is inframarginal regardless of whether the road is free or priced then he arrives exactly on time and so his actually inflexibility doesn't affect his trip price or the change in his trip price.

If we are willing to relax the requirement that pricing be a Pareto improvement and instead put some bound on the maximum harm done then we can enjoy a greater portion of the potential welfare gains. Figure 11 shows this trade off. The largest drop in the maximum harm comes from leaving at least some of the lanes unpriced, because the inflexible poor would prefer to have a more congested free option where they can pay with their time to needing to pay with their money. By pricing 75% of the road we can reap 80% of the social welfare gains while inflicting only 50% of the maximum harm.

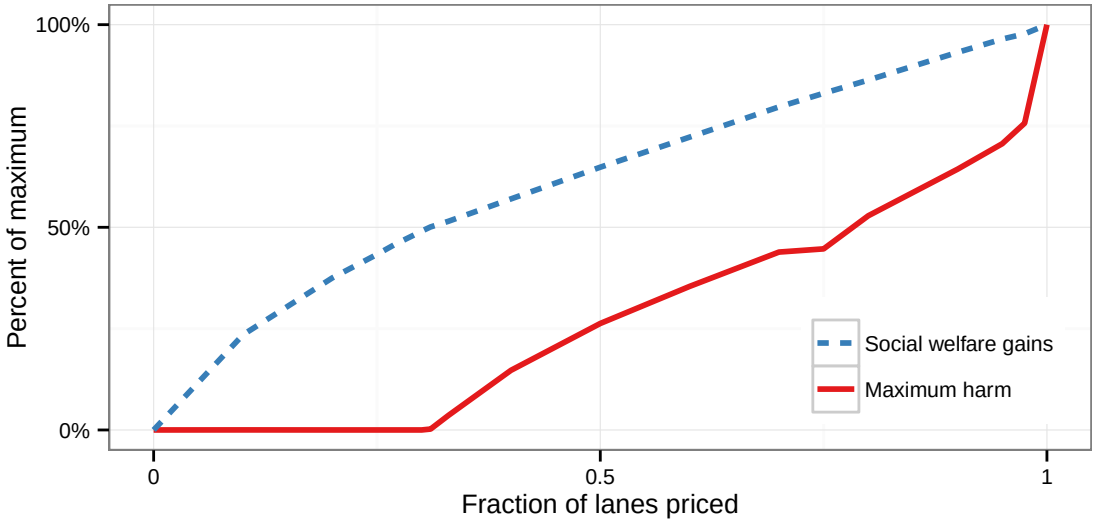


FIGURE 11. Trade-off between maximum harm and social efficiency gains

10. CONCLUSION

Our current understanding is that the goal of congestion pricing is to reduce the rate at which drivers arrive at their destination so that they can travel there faster, and that doing so, while efficient, creates winners and losers. However, recent research by transportation engineers shows that when too many vehicles are on the highway, fewer get through; this implies that time varying tolls can actually increase the rate at which drivers arrive. By increasing the arrival rate congestion pricing can reduce the length of rush hour, potentially making all road users better off. Unfortunately, because doing so requires changing the currency used to allocate desirable arrival times from time to money, congestion pricing will likely harm the poor. By leaving some of the lanes free we can preserve the ability of poor to pay with their time instead of their money. As long some rich drivers are using the highway at the peak, we can price some of the lanes while still allowing all of the poor drivers who had been traveling at the peak to continue to do so without paying a toll. We will have priced some of the lanes without hurting and road users, we will have a Pareto improvement.

There are at least two ways to strengthen the case I have made for value pricing being a Pareto improvement. First, we can strengthen the case for value pricing by using the revenue to help those whom congestion pricing harms; so far I explicitly ignored what was done with the revenue. Options include cutting regressive taxes or directly rebating the revenue by having negative tolls off-peak on both routes. On the priced route this would lower tolls at every point in time and on the free route it would eliminate travel time due to congestion off-peak and reduce travel times at the peak.

Second, we can include in our analysis alternate ways for the poor to pay with time instead of money for use of the highway during rush hour. Car pooling takes extra time but allows those doing so to split the cost of the tolls; a discount can even be offered to reduce the cost further. Likewise, buses that use the priced lanes will offer better travel times than were available before there were priced lanes, be cheaper than driving alone, but take addition time in getting to the stop, waiting for the bus, etc.

Because of the simplicity of the production possibility frontier implied by the bottleneck model we miss out learning more about how to set optimal tolls. In the standard models where we have a convex PPF we need to know agents' preferences in order to find the right point on the PPF, and this informational burden has rightly been considered an impediment to pricing. However, the speed that maximizes throughput is high, perhaps 50–60 mph (see Figure 2 for some evidence), and so finding the optimal point on the upper portion of the PPF is of second order concern compared to getting back on the PPF. This means we can forget about needing detailed measurements of consumer preferences when setting tolls; just find the toll schedule that maximizes throughput.

References

AASHTO (2005). *A policy on design standards–interstate system.* (5th ed.). Washington D.C.: American Association of State Highway and Transportation Officials.

Abrantes, P. A. and M. R. Wardman (2011, January). Meta-analysis of UK values of travel time: An update. *Transportation Research Part A: Policy and Practice 45*(1), 1–17.

Arnott, R., A. de Palma, and R. Lindsey (1990, January). Economics of a bottleneck. *Journal of Urban Economics 27*(1), 111–130.

Arnott, R., A. de Palma, and R. Lindsey (1999). Recent developments in the bottleneck model. In *Road pricing, traffic congestion and the environment: Issues of efficiency and social feasibility*, pp. 79–112. Edward Elgar.

Arnott, R. and E. Inci (2010, November). The stability of downtown parking and traffic congestion. *Journal of Urban Economics 68*(3), 260–276.

Arnott, R. and M. Kraus (1993, March). The Ramsey problem for congestible facilities. *Journal of Public Economics 50*(3), 371–396.

Arnott, R., A. d. Palma, and R. Lindsey (1993, March). A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *The American Economic Review 83*(1), 161–179.

Arnott, R., A. d. Palma, and R. Lindsey (1994, May). The welfare effects of congestion tolls with heterogeneous commuters. *Journal of Transport Economics and Policy 28*(2), 139–161.

Banks, J. (1991). Two-capacity phenomenon at freeway bottlenecks: a basis for ramp metering? *Transportation Research Record 1320*, 83–90.

Banks, J. H. (1990). Flow processes at a freeway bottleneck. *Transportation Research Record* (1287), 20–28.

Belenky, P. (2011, September). Revised departmental guidance on valuation of travel time in economic analysis. Technical report, U.S. Department of Transportation, Washington, D.C.

Bertini, R. and M. Leal (2005). Empirical study of traffic features at a freeway lane drop. *ASCE Journal of Transportation Engineering 131*(6), 397–407.

Bertini, R. and S. Malik (2004, January). Observed dynamic traffic features on freeway section with merges and diverges. *Transportation Research Record: Journal of the Transportation Research Board 1867*(-1), 25–35.

California Department of Transportation (1999). Performance measurement system. Technical report, Sacramento, California.

Cassidy, M. J. and R. L. Bertini (1999, February). Some traffic features at freeway bottlenecks. *Transportation Research Part B: Methodological 33*(1), 25–42.

Cassidy, M. J. and J. Rudjanakanoknad (2005, December). Increasing the capacity of an isolated merge by metering its on-ramp. *Transportation Research Part B: Methodological 39*(10), 896–913.

Chu, X. (1995, May). Endogenous trip scheduling: The Henderson approach reformulated and compared with the Vickrey approach. *Journal of Urban Economics 37*(3), 324–343.

Chung, K., J. Rudjanakanoknad, and M. J. Cassidy (2007, January). Relation between traffic density and capacity drop at three freeway bottlenecks. *Transportation Research Part B: Methodological 41*(1), 82–95.

Cohen, Y. (1987). Commuter welfare under peak-period congestion tolls: Who gains and who loses. *International Journal of Transport Economics 14*(3), 239–266.

Council, N. R. (2000). *Highway capacity manual.* Washington, DC :: Transportation Research Board, National Research Council,.

Currie, J. and R. Walker (2011, January). Traffic congestion and infant health: Evidence from e-ZPass. *American Economic Journal: Applied Economics 3*(1), 65–90.

Daganzo, C. (1996). The nature of freeway gridlock and how to prevent it. In *Traffic and Transportation Theory*, Lyon, France, pp. 629–646. Pergamon.

Daganzo, C. F. (1994, August). The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research Part B: Methodological 28*(4), 269–287.

de Palma, A. and R. Lindsey (2002, January). Comparison of morning and evening commutes in the vickrey bottleneck model. *Transportation Research Record: Journal of the Transportation Research Board 1807*(-1), 26–33.

Elefteriadou, L., R. P. Roess, and W. R. McShane (1995). Probabilistic nature of breakdown at freeway merge junctions. *Transportation Research Record 1484*, 80–89.

Environmental Protection Agency (2011). Greenhouse gas emissions from a typical passenger vehicle. Technical Report EPA-420-F-11-041, Washington, D.C.

Fosgerau, M. and K. A. Small (2013, July). Hypercongestion in downtown metropolis. *Journal of Urban Economics 76*, 122–134.

Giuliano, G. (1992). An assessment of the political acceptability of congestion pricing. *Transportation 19*(4), 335–358.

Guan, Y., J. Zhu, N. Zhang, and X. Yang (2009). Traffic flow characteristics of bottleneck segment with ramps on urban expressway. In *International Conference on Transportation Engineering 2009*, pp. 1679–1684. American Society of Civil Engineers.

Hall, F. and K. Agyemang-Duah (1991). Freeway capacity drop and the definition of capacity. *Transportation Research Record 1320*, 1–98.

Hall, J. D. (2012, August). *Pareto Improvements from Lexus Lanes: The case for value pricing on heavily congested highways*. Dissertation, University of Chicago, Chicago, Illinois.

Halvorson, R. and K. R. Buckeye (2006, January). High-occupancy toll lane innovations: I-394 MnPASS. *Public Works Management & Policy 10*(3), 242 –255.

Harrington, W., A. J. Krupnick, and A. Alberini (2001, February). Overcoming public aversion to congestion pricing. *Transportation Research Part A: Policy and Practice 35*(2), 87–105.

Hårsman, B. and J. M. Quigley (2010, September). Political and public acceptability of congestion pricing: Ideology and self-interest. *Journal of Policy Analysis and Management 29*(4), 854–874.

Helbing, D. and M. Treiber (1998, October). Gas-kinetic-based traffic model explaining observed hysteretic phase transition. *Physical Review Letters 81*(14), 3042.

Henderson, J. V. (1974, July). Road congestion: A reconsideration of pricing theory. *Journal of Urban Economics 1*(3), 346–365.

Hendrickson, C. and G. Kocur (1981, February). Schedule delay and departure time decisions in a deterministic model. *Transportation Science 15*(1), 62–77.

Hendrickson, C. and E. Plank (1984, January). The flexibility of departure times for work trips. *Transportation Research Part A: General 18*(1), 25–36.

Hotelling, H. (1929). Stability in competition. *Economic Journal 39*, 41—57.

Hurdle, V. F. and P. K. Datta (1983). Speeds and flows on an urban freeway: some measurements and a hypothesis. *Transportation Research Record 905*, 127–137.

Hymel, K. (2009, March). Does traffic congestion reduce employment growth? *Journal of Urban Economics 65*(2), 127–135.

Ison, S. (2000, October). Local authority and academic attitudes to urban road pricing: a UK perspective. *Transport Policy 7*(4), 269–277.

Johnson, M. B. (1964, April). On the economics of road congestion. *Econometrica 32*(1/2), 137–150.

Jones, P. M. (1991). UK public attitudes to urban traffic problems and possible countermeasures: a poll of polls. *Environment and Planning C: Government and Policy 9*, 245–256.

Knight, F. H. (1924). Some fallacies in the interpretation of social cost. *The Quarterly Journal of Economics 38*(4), 582 –606.

Lave, C. (1994, March). The demand curve under road pricing and the problem of political feasibility. *Transportation Research Part A: Policy and Practice 28*(2), 83–91.

Leclercq, L., J. A. Laval, and N. Chiabaut (2011). Capacity drops at merges: an endogenous model. *Procedia - Social and Behavioral Sciences 17*, 12–26.

Lighthill, M. J. and G. B. Whitham (1955, May). On kinematic waves. II. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences 229*(1178), 317–345.

Lindsey, R. and E. Verhoef (2008). Congestion modeling. In D. Hensher and K. Button (Eds.), *Handbook of Transportation Modelling* (2 ed.)., Number 1 in Handbooks in Transport, pp. 417–441. New York: Elsevier.

Liu, L. N. and J. F. McDonald (1998, November). Efficient congestion tolls in the presence of unpriced congestion: A peak and off-peak simulation model. *Journal of Urban Economics 44*(3), 352–366.

Liu, L. N. and J. F. McDonald (1999, April). Economic efficiency of second-best congestion pricing schemes in urban highway systems. *Transportation Research Part B: Methodological 33*(3), 157–188.

Muñoz, J. C. and C. F. Daganzo (2002, July). The bottleneck mechanism of a freeway diverge. *Transportation Research Part A: Policy and Practice 36*(6), 483–505.

Newell, G. F. (1988, February). Traffic flow for the morning commute. *Transportation Science 22*(1), 47.

Oh, S. and H. Yeo (2012, December). Estimation of capacity drop in highway merging sections. *Transportation Research Record: Journal of the Transportation Research Board 2286*(-1), 111–121.

Parthasarathi, P., A. Srivastava, N. Geroliminis, and D. Levinson (2010, September). The importance of being early. *Transportation 38*(2), 227–247.

Perez, B. G. and G.-C. Sciara (2003, March). A guide for HOT lane development. Technical Report FHWA-OP-03-009, Federal Highway Administration, Washington, D.C.

Persaud, B., S. Yagar, and R. Brownlee (1998, January). Exploration of the breakdown phenomenon in freeway traffic. *Transportation Research Record: Journal of the Transportation Research Board 1634*(-1), 64–69.

Pigou, A. C. (1912). *Wealth and welfare*. London: Macmillan and co., limited.

Podgorski, K. V. and K. M. Kockelman (2006). Public perceptions of toll roads: A survey of the texas perspective. *Transportation Research Part A: Policy and Practice 40*(10), 888–902.

Richards, P. I. (1956). Shock waves on the highway. *Operations research 4*(1), 42—51.

Rudjanakanoknad, J. (2005, May). *Increasing Freeway Merge Capacity Through On-Ramp Metering*. Dissertation, University of California at Berkeley, Berkeley, CA.

Samaras, C. and K. Meisterling (2008, May). Life cycle assessment of greenhouse gas emissions from plug-in hybrid vehicles: Implications for policy. *Environmental Science & Technology 42*(9), 3170–3176.

Schrank, D., B. Eisele, and T. Lomax (2012, December). 2012 urban mobility report. Technical report, Texas A&M Transportation Institute, College Station, Texas.

Small, K. and E. Verhoef (2007). *The economics of urban transportation*. New York: Routledge.

Small, K. A. (1982, June). The scheduling of consumer activities: Work trips. *The American Economic Review 72*(3), 467–479.

Small, K. A. (1983). The incidence of congestion tolls on urban highways. *Journal of urban economics 13*(1), 90–111.

Small, K. A. (1992, December). Using the revenues from congestion pricing. *Transportation 19*(4), 359–381.

Small, K. A. and X. Chu (2003, September). Hypercongestion. *Journal of Transport Economics and Policy 37*(3), 319–352.

Small, K. A., C. Winston, and J. Yan (2005, July). Uncovering the distribution of motorists' preferences for travel time and reliability. *Econometrica 73*(4), 1367–1382.

Srivastava, A. and N. Geroliminis (2013, May). Empirical observations of capacity drop in freeway merges with ramp control and integration in a first-order model. *Transportation Research Part C: Emerging Technologies 30*, 161–177.

Starkie, D. (1986, March). Efficient and politic congestion tolls. *Transportation Research Part A: General 20*(2), 169–173.

Stiglitz, J. (1998, April). Distinguished lecture on economics in government: The private uses of public interests: Incentives and institutions. *The Journal of Economic Perspectives 12*(2), 3–22.

Sullivan, E. (1999). State route 91 impact study datasets. Technical report, California Polytechnic State University, San Luis Obispo, California.

Sullivan, E. (2002, January). State route 91 value-priced express lanes: Updated observations. *Transportation Research Record: Journal of the Transportation Research Board 1812*(-1), 37–42.

Sullivan, E. and J. Harake (1998, January). California route 91 toll lanes impacts and other observations. *Transportation Research Record: Journal of the Transportation Research Board 1649*(-1), 55–62.

Treiber, M., A. Hennecke, and D. Helbing (2000). Congested traffic states in empirical observations and microscopic simulations. *Physical Review E 62*(2), 1805.

U.S. Department of Transportation (2006, December). Congestion pricing: A primer. Technical Report FHWA-HOP-07-074, Washington, D.C.

U.S. Department of Transportation (2009). 2009 national household travel survey. Technical report, Washington, D.C.

U.S. Energy Information Administration (2013). Annual energy outlook 2013: Supplemental tables. Technical report.

van den Berg, V. and E. T. Verhoef (2011, August). Winning or losing from dynamic bottleneck congestion pricing?: The distributional effects of road pricing with heterogeneity in values of time and schedule delay. *Journal of Public Economics 95*(7–8), 983–992.

Verhoef, E., P. Nijkamp, and P. Rietveld (1996, November). Second-best congestion pricing: The case of an untolled alternative. *Journal of Urban Economics 40*(3), 279–302.

Verhoef, E. T. (1999, May). Time, speeds, flows and densities in static models of road traffic congestion and congestion pricing. *Regional Science and Urban Economics 29*(3), 341–369.

Verhoef, E. T. (2001, May). An integrated dynamic model of road traffic congestion based on simple car-following theory: Exploring hypercongestion,. *Journal of Urban Economics 49*(3), 505–542.

Vickrey, W. (1987). Marginal and average cost pricing. In S. N. Durlauf and L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics* (1 ed.). Basingstoke: Palgrave Macmillan.

Vickrey, W. S. (1963, May). Pricing in urban and suburban transport. *The American Economic Review 53*(2), 452–465.

Vickrey, W. S. (1969, May). Congestion theory and transport investment. *The American Economic Review 59*(2), 251–260.

von Thünen, J. H. (1930). *Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*. Jena: Gustav Fischer.

Walters, A. A. (1961, October). The theory and measurement of private and social cost of highway congestion. *Econometrica 29*(4), 676–699.

Wardrop, J. (1952, January). Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers, Part II 1*(3), 325–378.

Zhang, L. and D. Levinson (2004, January). Some properties of flows at freeway bottlenecks. *Transportation Research Record: Journal of the Transportation Research Board 1883*(-1), 122–131.