

Pareto Improvements from Lexus Lanes: The case for value pricing on heavily congested highways

Jonathan Hall*

Working Draft: Preliminary and Incomplete

September 20, 2011

Abstract

In the standard economic model of road pricing adding tolls to just one of multiple routes leads to high value of time individuals traveling faster while those with a low value of time travel slower. While it is a Kaldor-Hicks improvement, most individuals are worse off before toll revenues are distributed. In this paper I use insights from the traffic engineering literature to show that in heavy traffic restricting the number of vehicles on a highway can increase throughput and speed and so a carefully designed toll on one route is Pareto improving. Both high value of time and low value of time individuals travel faster and everyone is better off regardless of what is done with toll revenue.

JEL Classification: R41, R48, D62.

Contents

1	Introduction	3
2	Understanding traffic congestion	6
2.1	Evidence traffic is off the production possibilities frontier	9
3	Basic model	11
3.1	Congestion technology	11
3.2	Agents and preferences	13
3.3	Equilibrium concept and information	14

*Email: jonathan.hall@uchicago.edu. I am grateful for helpful feedback from Gary Becker, Eric Budish, William Hubbard, Ethan Lieber, Devin Pope, Mark Phillips, Allen Sanderson, Chad Syverson, and George Tolley, as well as participants in the IO Working Group and Micro Lunch.

4	Homogeneous preferences	16
5	Heterogeneous preferences	23
5.1	Finding equilibrium	24
5.2	Homogeneous value of time	26
5.3	Potential barriers to a Pareto improvement	27
5.3.1	Poor must pay more costly tolls	29
5.3.2	Rich displace poor from peak	30
5.3.3	Rich displace poor from road	30
5.4	When can pricing be a Pareto improvement?	31
5.4.1	Two types	32
5.4.2	Arbitrary number of types	35
6	Conclusion	40
A	Solving for equilibrium in the basic model	41
A.1	Proofs	41
A.2	Key constants	41
A.3	Toll lanes	42
A.4	Free lanes	43
A.5	Total costs	44
A.6	Changes in social welfare	45
A.7	Calculating percentage savings from tolling when we know ratio of fixed travel time to variable travel time	45
B	Proof of proposition 4 (Existence)	46
B.1	Overview of proof	46
B.2	Individual cost minimization	47
B.3	Supply equals demand	48
B.3.1	Construct a nonempty, compact, and convex set \mathbb{P}	48
B.3.2	Show that $ \mathcal{T}_{g,r} $ is convex and compact.	48
B.3.3	Show that $ \mathcal{T}_{g,r} $ upper hemicontinuous.	50
B.3.4	Define an excess demand function, \mathbf{E} , show that it is upper hemicontinuous and the set $\mathbf{E}(\mathbf{p})$ is compact and convex for all $\mathbf{p} \in \mathbb{P}$	55
B.3.5	Define a correspondence $\zeta : \mathbb{P} \rightarrow \mathbb{P}$ that is upper hemicontinuous from \mathbb{P} into itself with the property that the set $\zeta(\mathbf{p})$ is nonempty and convex for all $\mathbf{p} \in \mathbb{P}$. Finally ζ has a fixed point only if excess demands are zero.	55
B.3.6	Use the Kakutani fixed point theorem to show there exists a $\mathbf{p} \in \mathbb{P}$ such that $\mathbf{p} \in \zeta(\mathbf{p})$	57

B.3.7 Confirm the departure rate is finite	58
C Proof of uniqueness	58
D Other proofs	58

1 Introduction

[I]n some circumstances, it would be possible, by shifting a few carts from route B to route C, greatly to lessen the trouble of driving those still left on B, while only slightly increasing the trouble of driving along C. In these circumstances a rightly chosen measure of differential taxation against road B would create an “artificial” situation superior to the “natural” one. (Pigou, 1920, p. 194)

In the ninety years since Arthur C. Pigou introduced the idea that tolls could be used to alleviate congestion, carts have given way to automobiles and congestion has grown to consume more than 4 billion hours each year (Schrank et al., 2010). Despite his insights, traffic congestion is a large and growing problem facing major cities worldwide. Concerns that pricing roads hurts many road users, particularly the poor, has led to significant political opposition to congestion pricing and limited its adoption. This paper will use discoveries from the traffic engineering literature to show that in the case of heavily congested highways, judicious design of a toll applied to a portion of the lanes can make all road users better off, even before toll revenue is spent. Congestion pricing can lead to a Pareto improvement.

Our intuition on congestion pricing starts with the observation that each additional car on a road creates an externality by slowing down all of the other vehicles using the road. Pigou’s insight was that we can apply a tax to correct this externality. By raising the cost of driving at rush hour, fewer people use the road, leading to less congestion and faster travel times for those remaining. If drivers shift to other roads, then congestion will get worse on those roads, but pricing the first route will still be efficiency enhancing since we will have sorted high value of time individuals onto fast routes, and those with a low value of time onto slow routes. In addition, some trips whose value was below social marginal cost will no longer take place. This is a Kaldor-Hicks improvement; the winners gain more than the losers lose.¹

¹It is quite possible that all road users are worse off, with only the toll collector being better off. In order to have fewer people using the road we must have increased the total cost of traveling during rush hour for the marginal drivers. We reduced the travel time costs but increased the financial costs. If the inframarginal drivers have the same value of time as the marginal driver, then their total cost will have increased as well. It is only for those drivers whose value of time is sufficiently larger than that of the marginal driver who are actually better off.

Consider how this applies specifically to value pricing, the practice of taking a freeway and pricing only some of its lanes. These lanes are sometimes called HOT (high occupancy/toll) lanes when those carpooling can use them for free or a reduced price, and also go by the epithet of “Lexus Lanes” due to the idea that only those who can afford a Lexus can afford to pay the tolls. Consider, for example, a freeway where traffic goes 30 mph during the morning rush hour. The standard intuition would say that if we price some of the lanes, say half for this example, so that traffic flows are split 40/60 between the priced and free lanes, then those in the priced lanes will be able to go 42.5 mph while the free lanes are slowed down to 19.5 mph.² It is this intuition that has lead AAA Mid-Atlantic to write an op-ed for *The Washington Post* on July 1, 2005 to criticize value pricing because it “would create a two-tier system of public roads where the rich can roll while the poor will poke”. The welfare effects seem clear, CEOs get to work quickly while janitors face even worse traffic.

However, this intuition presumes the roads are already being used to produce a given number of trips at the lowest cost. In other words the standard intuition is about moving along the production possibilities frontier, trading off the number of trips made with the speed those trips happen at. However, in heavily congested freeway traffic, defined as congestion such that speeds are below 50 mph, we are often far from the production possibilities frontier. This is because there is another externality to driving beyond slowing down other travelers; in heavily congested traffic adding another vehicle to the road reduces throughput. As a result, one additional driver attempting to make a trip can reduce the rate at which drivers complete their trip, all while travel times are climbing.

As an example of this perverse and counter-intuitive result is found in the experience of the eleven mile morning commute from Bellevue to Seattle in Washington state. Between 2005 and 2007 the average peak travel time increased by six percent and the length of rush hour³ climbed forty minutes, all while peak period vehicle miles traveled fell by three percent. Twenty-three of the thirty-eight commutes the Washington State Department of Transportation tracks had weakly increasing travel times and length of rush hours with a decrease in traffic flow from 2005 to 2007 (2008, pp. 20-21). These costs add up and are severe; the Washington State Department of Transportation estimates that if in 2007 they had been on the production possibilities frontier drivers would have saved at least 102,000 hours every weekday (2008, p. 40). A similar estimate comes from Kwon et al. (2006), who estimate that if the recurrent bottlenecks on a forty-five mile section of I-880 in the San Francisco Bay area operated at ninety percent of capacity and 60 mph then it would eliminate a third of the congestion delay on that road. This would amount to 5,500

²Quite literally this is the standard intuition. These are precisely the results that Liu and McDonald (1998, 1999); Small and Yan (2001); Small et al. (2006); Light (2009) would predict for inelastic demand. All of these papers use the travel time formula of U.S. Bureau of Public Roads (1964).

³Defined as the amount of time that average speed fell below seventy percent of the posted speed limit.

hours a day.

There are at least three reasons high density can cause lower throughput: queue spillovers, driver error propagation, and throughput drops at bottlenecks. We will discuss these in more detail along with the empirical evidence for them in section 2.

The solution to these problems is to restrict flow onto the highway to prevent traffic density from reaching the point where highway throughput begins to fall. Fortunately, Pigou already taught us how to do so, by pricing the roads.

Tolls can be used to move road usage back to the production possibilities frontier. As a result, priced lanes can move more traffic per lane than free lanes. Returning to our example, rather than traffic flows being split 40/60, they could be split 60/40. The tolled lanes could travel 50 mph while increasing the speed of the free lanes to 35 mph. Everyone has a faster travel time, and since we know those in the priced lanes are better off despite paying a toll by revealed preference, we have achieved a Pareto improvement.

This raises the question: why not price the entire road? By pricing just a portion of the lanes we are leaving efficiency gains on the table, but we are also preserving the option to pay with time and so protect the poor. The goal is to have a pricing mechanism that naturally generates a Pareto improvement, instead of relying on transfers. If we set out to maximize social welfare subject to the constraint of generating a Pareto improvement, then in the majority of cases we will be unable to price the entire road.

The intuition above is inherently static and doesn't consider the driver's choice of when to travel. By using a dynamic model I will reveal some potential problems but also illustrate cases where pricing the entire road can yield a Pareto improvement. The intuition is also deterministic and so overlooks the large welfare gains that come from providing a reliable route to those who need it. Adding this will increase the range of cases where value pricing will lead to a Pareto improvement.

In building the case that pricing highways can be a Pareto improvement I will, however, overlook the sizable potential public health improvements that can come from reducing congestion. These have been estimated by Levy et al. (2010) and Currie and Walker (2011).

In this paper I will modify the basic bottleneck model of Arnott et al. (1990), which is itself a formalization of Vickrey (1969), to consider value pricing and hypercongestion. As in their work, there will be a single road connecting where people live and where they work. The only source of congestion will be a bottleneck along this road. We will consider three policy regimes, no road pricing, tolling the entire road, and value pricing, where we price only a portion of the lanes. In addition, we will vary the model along three dimensions: no throughput drop versus throughput drop, homogeneous versus heterogeneous preferences, and complete information versus uncertainty.

The main result is that the careful application of value pricing can lead to a Pareto

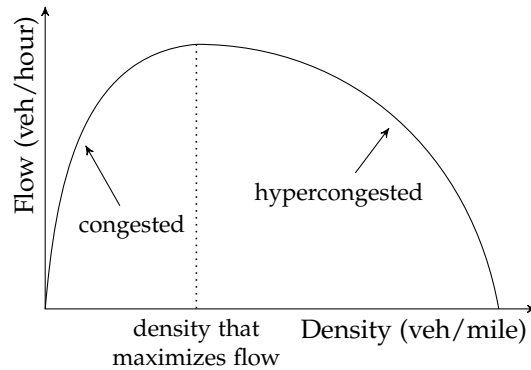


Figure 1: Fundamental diagram of traffic flow

improvement. In contrast to the current literature I show that congestion pricing is welfare enhancing regardless of what is done with the toll revenue and that while it may be difficult to achieve a Pareto improvement without knowledge of consumer preferences, most of the welfare gains can be realized without this information.

After showing that value pricing can lead to a Pareto improvement in theory, I will estimate the potential gains available from value pricing in southern California.

2 Understanding traffic congestion

My paper depends critically on the claim that tolls can be used to increase traffic flow. To understand this we must first understand a peculiar fact about traffic, which I referred to earlier. When there are too many cars on the road, fewer get through. This relationship is shown by the *fundamental diagram of traffic flow*, shown in figure 1.⁴ This figure shows the relationship between flow, the number of vehicles passing a given location in an hour, and density, the number of vehicles physically on a mile of the road.

One easy way to think about this relationship is to picture a race track. If your goal is to maximize the number of laps completed in an hour (i.e. flow), how many cars do you want on the track (i.e. density)? Consider what happens when you move from one vehicle to two; their speed will hardly change, and so the number of laps completed will nearly double. So when density is low adding vehicles increases the flow. But if you take this too far and fill every available inch of track with vehicles, then they will not be able to move at all. No motion, no laps completed; flow will be zero. Somewhere in the middle you will find the right balance between having a lot of cars on the track and still

⁴The fundamental diagram of traffic flow was first estimated in Greenshields et al. (1935) and has been estimated countless times since. For a book that reviews much of the empirical evidence see May (1990).

allowing them to move quickly which will maximize the number of laps completed.⁵ The nomenclature of urban economics is to call traffic congested if density is less than this point, and hypercongested if it is greater.⁶

The fundamental diagram of traffic flow is no secret to economists and can even be found in economics textbooks such as Small and Verhoef (2007). Walters (1961) and Johnson (1964) even use it to argue that tolls can increase speed and flow.

It has, however, largely been rejected as a causal relationship. Verhoef (1999, 2001) shows that in a static model a hypercongested equilibrium is unstable. Newell (1988) shows that in the popular LWR hydrodynamic model of congestion an active bottleneck will always operate a full throughput, hypercongestion will not reduce it.⁷ Small and Chu (2003) argue that hypercongestion is unimportant for studying freeways.

This view comes from the fact that we can observe hypercongestion when low traffic flows, due to downstream constraints, lead to high densities. As an illustration, consider a ten lane highway that approaches a one lane bridge. When more cars arrive than the bridge can handle there will be queuing. If we are observing the highway a mile upstream from the bridge, then we will observe moderate densities and high flows as the queue builds up downstream. Once the queue reaches us, density will be very high and flow will fall to a tenth of its maximum. We observe high density and low flows but the causal relationship is that low flows, due to a downstream bottleneck, are causing high density; rather than high density being the cause of low flows. It is exactly this possibility that Lindsey and Verhoef (2008) have in mind when they conclude that “[t]he emerging view seems to be that hypercongestion is a transient phenomenon, occurring in queues immediately upstream of bottlenecks, that is best studied with dynamic models” (p. 421).

I offered two examples, in the race track example high density can cause low flows, while in the bridge example the causal relationship is reversed. The fundamental difference between them is that in the race track example every point is a potential bottleneck while in the bridge example there is just one bottleneck. On the race track the queue for each bottleneck interferes with the other bottlenecks, resulting in a none of them operating at their full capacity, but in the bridge example the bottleneck always operates at its full capacity. Both examples are illustrative of real world phenomenon. The race track example has a nearly precise analog in the beltways or ring roads that go around major cities, such as I-495, which encircles Washington D.C., and Boulevard Périphérique, which encircles Paris. It has been long recognized that these roads are especially prone

⁵Flow is the product of speed times density, that is $F = S \times D$. This means that as long as the elasticity of speed with respect to flow, $\epsilon_{S,D} = -\frac{\partial S}{\partial D} \frac{D}{S} < 1$ flow will be increasing and when $\epsilon_{S,D} > 1$, flow will fall.

⁶Traffic engineers use different terminology, they call traffic uncongested or unsaturated when on the left side of the figure, and congested or saturated when on the right.

⁷The LWR model is named after those who first proposed it, Lighthill and Whitham (1955) and Richards (1956).

to crippling congestion.⁸ Furthermore, the race track example serves well as an analogy for congestion on urban streets.⁹ And while we rarely see ten lanes go down to just one, most freeway on-ramps involve fitting $n + 1$ lanes of traffic into n lanes.

As mentioned in the introduction, there are at least three reasons high density can cause lower throughput. Let us now look at each of these in turn.

The first reason high density can cause lower throughput is that as queues grow they can block upstream off-ramps. Then those wanting to use the upstream off-ramp must now wait in a line for a congested resource they themselves do not want to use. This is the causal mechanism at work in the race track/beltway example and is what Vickrey (1969) called a triggerneck. Similarly, a queue can grow at a busy off-ramp, spilling onto the mainline of the freeway and blocking traffic. This has been modeled by Daganzo (1996) and Daganzo (1998).

The second reason high density can lead to lower throughput is that small mistakes propagate when density is high and cause long lasting disruptions. An example is one driver taps on his breaks, causing the driver behind to slam on his breaks, and the effect ripples backwards. This means that when density is high traffic can spontaneously breakdown and experience lower speeds and flows. This has been studied by Kerner and Rehborn (1997); Helbing and Huberman (1998); Kerner and Klenov (2002); Orosz et al. (2009), and Flynn et al. (2009).

The final reason is that bottlenecks have lower throughput once a queue forms behind them. Traffic bottlenecks usually occur at places where drivers need to merge, such as a lane drop or on-ramp. When traffic is very heavy it is difficult for the vehicles that need to merge to change lanes, at some point some of them will slow down, or even stop, before merging. After changing lanes they are still moving at a very low speed and this reduces the flow through the bottleneck. The bottleneck spreads laterally to the other lanes as oncoming vehicles start changing lanes to avoid the slowdown. 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.

These last two reasons are similar in that both drivers take actions or make mistakes that reduce flow. The difference is that in the second reason this happens whenever density gets to high while in the third it happens at specific places, bottlenecks. Daganzo et al. (1999) show that the empirical evidence provided by Kerner and Rehborn (1997) in support of driver error propagation actually is better explained by bottlenecks.

It is the last causal relationship that we will incorporate into a standard model.

⁸For example, see Vickrey (1969, p. 252) and Daganzo (1996). There is an animation on Daganzo's website that illustrates this nicely at www.its.berkeley.edu/volvocenter/gridlock/.

⁹It should be noted that Arnott (1990) and Small and Chu (2003) discuss hypercongestion in central business districts as a causal relationship.

2.1 Evidence traffic is off the production possibilities frontier

There is a large literature by traffic engineers documenting the throughput drop at bottlenecks, which they generally call the two-capacity hypothesis. 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 effected the merging lanes, not the entire road. The estimates for the size of the drop range from 2 to 15 percent; their estimates are presented in table 1.¹⁰ This phenomenon has also been modeled in the physics literature in Helbing and Treiber (1998), and Treiber et al. (2000).

Table 1: Findings of traffic 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)	3.2	I-8, San Diego, site 1
	−.1	I-8, San Diego, site 2
	−.7	I-805, San Diego
	−1.2	SR-163, San Diego
Persaud et al. (1998)	11.6	Highway 401, Toronto, site 1 a.m.
	15.3	Highway 401, Toronto, site 1 p.m.
	10.6	Highway 401, Toronto, site 2
Cassidy and Bertini (1999)	8.7	Queen Elizabeth Way, Toronto
	7.4	Gardiner Expressway, 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

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. The data comes either from video cameras or loop detectors.

Figure 2, reprinted from Bertini and Leal (2005, p. 402), shows the detrended cumulative flows (N) and cumulative speeds (V) at a bottleneck on M4 near London on November 18, 1998. The flows in the 17 minutes before a queue forms average 3,690

¹⁰There are papers not expressly testing the two-capacity hypothesis who 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.

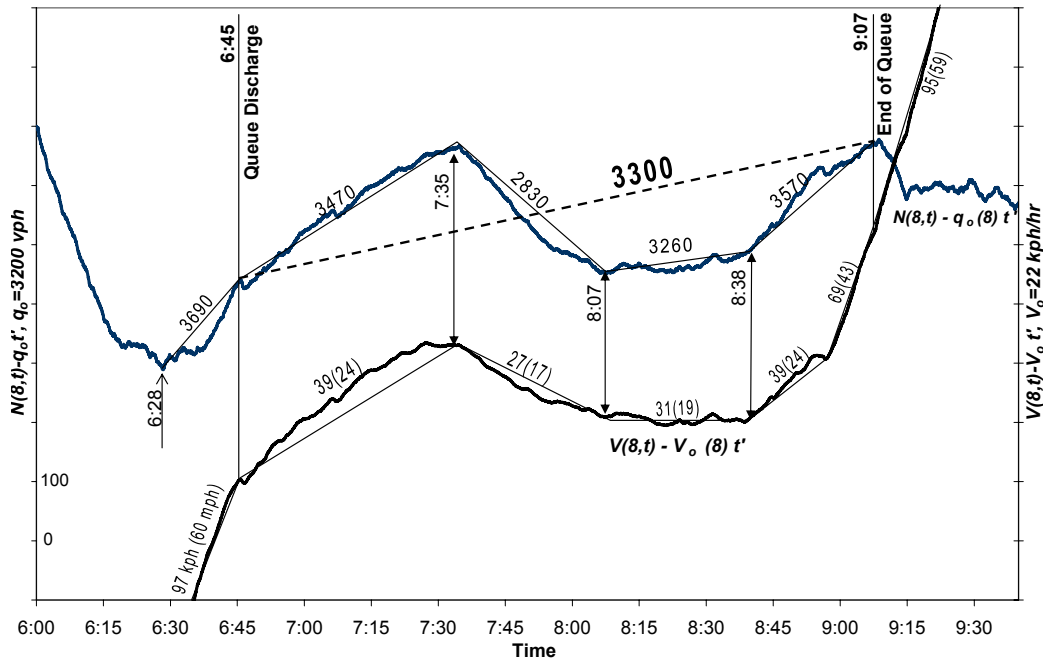


Figure 2: Detrended cumulative flow and velocity on M4 for November 18, 1998. Reprinted from Bertini and Leal (2005, p. 402).

vehicles per hour and vehicles are traveling 60 mph. For the following 142 minutes for which the queue remains flows average 3,300, a drop of 10.6 percent, and speeds fall below 25 mph for most of time. In this paper the start and end of the queue were identified by comparing detrended cumulative flows at detectors before and after the bottleneck. When they diverged there was a larger than usual number of vehicles on the freeway segment between the two detectors, a queue.¹¹ Other researchers have used sharp increases in speed or sharp declines in occupancy, the fraction of time a loop detector has a vehicle over it, between detectors to identify when a bottleneck is active and has a queue.

There is a much smaller literature, to my knowledge only one paper, documenting how queue spillovers can reduce throughput. Muñoz and Daganzo (2002) find that flows on I-880N fall by as much as 25 percent when a queue from traffic exiting onto I-238 backs up onto I-880N and blocks through traffic. Even though I-880N is five lanes wide through the stretch they were studying in the San Francisco Bay area the queue frequently spread across all of the lanes.

¹¹They plot this in figure 4 of their paper.

3 Basic model

The standard model for studying dynamic congestion is the bottleneck model of Vickrey (1969), which was formalized by Arnott et al. (1990, 1993). I will make two modifications to the model. The first will incorporate the finding by traffic engineers that bottleneck throughput falls once a queue forms as well as serve as shorthand for the effects of queue spillovers, and the second will allow us to consider value pricing, where some lanes are tolled and others remain free.

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 $\lambda_{\text{toll}} + \lambda_{\text{free}} = 1$.¹² 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 arrival rate on a route, $r_r(t)$, exceeds its capacity, $\lambda_r \cdot s^*$, a queue develops. Once a queue forms the throughput of the bottleneck for that route falls to $\lambda_r \cdot s = \lambda_r(1 - \theta) \cdot s^*$, where $\theta \in [0, 1)$ is the size of the throughput drop. Therefore, queue length, measured as the number of vehicles in the queue, evolves according to

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

Definition 1. In the standard bottleneck model $s = s^*$ and $\lambda_{\text{toll}} \in \{0, 1\}$.

Definition 2. In the modified bottleneck model $s < s^*$ and $\lambda_{\text{toll}} \in [0, 1]$.

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

Definition 3. Value pricing is when $\lambda_{\text{toll}} \in (0, 1)$.

Travel time along each route is

$$T_r(t) = T^f + T_r^v(t) \quad r \in \{\text{free, toll}\}, \quad (2)$$

¹²Implicit in this is the assumption it is costless to split the road into two routes. The Federal Highway Administration recommends a three to four foot buffer between the priced and unpriced lanes when a pylon barrier is used (Perez and Sciara, 2003, p. 39-40). On I-394 in Minnesota there is a two foot buffer without any barrier between the priced and unpriced lanes (Halvorson and Buckeye, 2006, p. 246). Federal standards call for twelve foot lanes on interstates (AASHTO, 2005, p. 3), though there are exceptions.

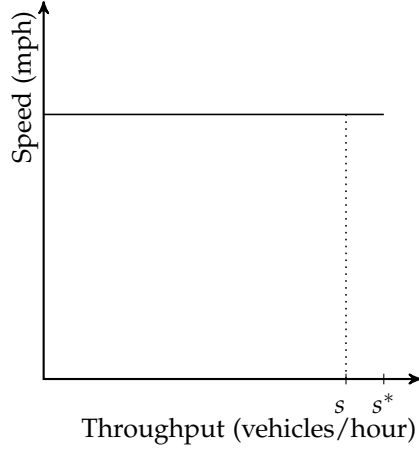


Figure 3: Production possibility frontier for bottleneck model

where T^f is fixed travel time, the amount of time it takes to travel the road absent any congestion, and $T_r^v(t)$ 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

$$T_r^v(t) = \frac{Q_r(t)}{\lambda_r \cdot s}. \quad (3)$$

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.

For simplicity, and without loss of generality, we will assume that $T^f = 0$.¹³ Throughout the rest of this paper when we discuss travel time we are only referring to the variable, congestion related, travel time.

¹³With a few redefinitions we could leave the model general. Redefine t as the time agents choose to arrive at the bottleneck, so if we let t_{depart} be the time they leave home then $t = t_{\text{depart}} + T_{\text{before bottleneck}}^f$. Similarly, redefine t^* as the time one needs to leave the bottleneck to arrive at work on-time, so t_{true}^* is the actual ideal arrival time then $t^* = t_{\text{true}}^* - T_{\text{after bottleneck}}^f$. The primary reason for not doing so is because it would require reversing the standard interpretation of arrive and depart. Instead of discussing departures from home and arrivals at work, we would discuss arrivals at the bottleneck and departures from the bottleneck.

3.2 Agents and preferences

There are G types of atomistic agents; let $i \in 1, \dots, G$ denote an arbitrary type. Agents choose when to leave and which route to take in order to minimize their trip cost. The price of a trip on route r that starts at time t for an agent in type i , $p_{i,r}(t)$, is the sum of the cost of travel time, schedule delay costs $D_i(\cdot)$, and the toll $\tau_r(t)$. Schedule delay costs are a function of the difference between the desired arrival time t_i^* and the actual arrival time $t_a = t + T_r(t)$;

$$p_{i,r}(t) = \alpha_i T_r(t) + D(t_a - t_i^*) + \tau_r(t). \quad (4)$$

I assume schedule delay costs are piecewise linear, so that

$$D_i(t_a - t_i^*) = \begin{cases} -\beta_i(t_a - t_i^*), & \text{for } t_a \leq t_i^*, \\ \gamma_i(t_a - t_i^*), & \text{for } t_a > t_i^*. \end{cases}$$

With piecewise linear schedule delay costs we can rewrite the trip price as

$$p_{i,r}(t) = \alpha_i T_r(t) + \beta_i(\text{time early}) + \gamma_i(\text{time late}) + \tau_r(t), \quad (5)$$

where α_i is the shadow price of travel time, β_i is the shadow price of time spent at work early, and γ_i is the shadow price of time late to work. Each of these parameters is the agents' willingness to pay money to reduce travel time or delay by one unit of time. I assume that $\gamma_i > \alpha_i > \beta_i > 0$. This means that agents will prefer an extra minute of travel time if it means they are one minute less late to work, as well as that they would rather wait for work to start at the office than wait in traffic.¹⁴

The mass of agents in type i who are willing to pay trip price p_i to travel is $N_i(p_i)$. I assume N_i is differentiable with $\partial N_i / \partial p_i \leq 0$, and $N_i(p_i) \geq 0$ for all $p_i \in \mathbb{R}_+$. I also assume $N_i(0) > 0$ and that there exists a p_i^{\max} such that $N(p_i^{\max}) = 0$.¹⁵

Let $\tilde{\mathcal{T}}$ be the *supply-feasible departure times*, the set of times which agents are allowed to depart, and assume this is a non-binding constraint and $|\tilde{\mathcal{T}}| < \infty$. This amounts to the observation that in the case of commuting we can restrict attention to a single day without artificially constraining agents choice set. Due to assuming no fixed travel time, we could just as well call this the supply-feasible arrival times.

To state the assumption that set of supply-feasible departure times is a non-binding

¹⁴See 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 statistical test but just comparing point estimates. See his table 3 for the clearest evidence; the first two columns are essentially β/α and γ/α .

¹⁵Note that the assumption that $N_i(0) < \infty$ would not need to be made if I did not assume the fixed travel time was zero. A strictly positive travel time along with a strictly positive shadow price of travel time yields a strictly positive minimum trip price.

In addition, while the assumption that there is a price such that demand is zero rules out perfectly inelastic demand, there can still be an arbitrarily large region where demand is perfectly inelastic.

constraint formally, first define p'_i as type i 's minimum trip price of leaving at an endpoint of $\bar{\mathcal{T}}$,

$$p'_i = \min \{D_i(t - t_i^*) \mid t = \sup \bar{\mathcal{T}} \text{ or } t = \inf \bar{\mathcal{T}}\}.$$

Then assume that $p_i^{\max} < p'_i$. This means that the set of supply-feasible departure times is so large that no one wants to travel outside of it.

While agents choose their departure time t it will often be simpler to discuss their choice of arrival time t_a . I will prove a little later in Proposition 1 that when D_i is continuous then in equilibrium there is a one-to-one mapping between departure and arrival times, so this is a harmless simplification.

3.3 Equilibrium concept and information

Our equilibrium concept is perfect information, pure strategy Nash equilibrium. No agent will be able to reduce his trip cost by changing his time of departure or route choice.¹⁶ Let \mathcal{L} denote the set of routes.

Definition 4. *Given a bottleneck congestion game specified by*

$$\left(\{\alpha_i, D_i(\cdot), t_i^*, N_i(\cdot)\}_{i=1}^G, \{s, s^*, \bar{\mathcal{T}}\}, \{\lambda_r\}_{r \in \mathcal{L}} \right)$$

a trip price vector $\mathbf{p} = (\bar{p}_1, \dots, \bar{p}_G)$, a set of arrival times $\{\{\mathcal{T}_{i,r}\}_{i=1}^G\}_{r \in \mathcal{L}}$, a set of departure rates $\{\{r_{i,r}(t)\}_{i=1}^G\}_{r \in \mathcal{L}}$, and toll schedules $\{\tau_r(t)\}_{r \in \mathcal{L}}$ constitute a deterministic arrival time equilibrium if

1. For every $i \in \{1, 2, \dots, G\}$ all agents minimize their trip price; that is,

$$\begin{aligned} p_{i,r}(t_a) &= \bar{p}_i & \text{for } t_a \in \mathcal{T}_{i,r}, r \in \mathcal{L} \\ p_{i,r}(t) &\geq \bar{p}_i & \text{for } t \in \bar{\mathcal{T}}, r \in \mathcal{L} \end{aligned}$$

2. For every $i \in \{1, 2, \dots, G\}$ there is enough time for all $N_i(\bar{p}_i)$ agents to travel, and no more; supply of travel time equals demand for travel time. That is,

$$\begin{aligned} N_i(\bar{p}_i) &= s \sum_{r \in \mathcal{L}} |\{t_a \mid t_a \in \mathcal{T}_{i,r}, \text{ and } Q_r(t) > 0 \text{ or } r_{i,r}(t) > s^*\}| \\ &\quad + s^* \sum_{r \in \mathcal{L}} |\{t_a \mid t_a \in \mathcal{T}_{i,r}, Q_r(t) = 0 \text{ and } r_{i,r}(t) \leq s^*\}|. \end{aligned}$$

Notice that there is nothing determining how the tolls are set and so, presuming equilibrium exists, there will likely be a continuum of equilibria. I am focusing on a particular

¹⁶This is similar to Wardrop's first principle of equilibrium which says that no agent can unilaterally reduce his travel costs by changing to another route (Wardrop, 1952).

kind of equilibrium, one where one route is priced, the other is free, and tolls are set to maximize consumer welfare.¹⁷ This allows me to give a narrower definition of equilibrium.

Definition 5. *Given a bottleneck congestion game specified by*

$$\left(\{\alpha_i, D_i(\cdot), t_i^*, N_i(\cdot)\}_{i=1}^G, \{s, s^*, \bar{\mathcal{T}}\}, \{\lambda_{\text{free}}, \lambda_{\text{toll}}\} \right)$$

a trip price vector $\mathbf{p} = (\bar{p}_1, \dots, \bar{p}_G)$, a set of arrival times $\{\mathcal{T}_{i,\text{free}}, \mathcal{T}_{i,\text{toll}}\}_{i=1}^G$, a set of departure rates $\{r_{i,\text{free}}(t), r_{i,\text{toll}}(t)\}_{i=1}^G$, and toll schedules $\{\tau_{\text{free}}(t), \tau_{\text{toll}}(t)\}$ constitute a value pricing equilibrium if

1. For every $i \in \{1, 2, \dots, G\}$ all agents minimize their trip price; that is,

$$p_{i,r}(t_a) = \bar{p}_i \quad \text{for } t_a \in \mathcal{T}_{i,r}, r \in \{\text{free}, \text{toll}\}, \text{ and} \quad (6)$$

$$p_{i,r}(t) \geq \bar{p}_i \quad \text{for } t \in \bar{\mathcal{T}}, r \in \{\text{free}, \text{toll}\}. \quad (7)$$

2. For every $i \in \{1, 2, \dots, G\}$ there is enough time for all $N_i(\bar{p}_i)$ agents to travel, and no more; supply of travel time equals demand for travel time. That is,

$$\begin{aligned} N_i(p_i) = s & \sum_{r \in \{\text{free}, \text{toll}\}} |\{t_a | t_a \in \mathcal{T}_{i,r}, \text{ and } Q_r(t) > 0 \text{ or } r_{i,r}(t) > s^*\}| \\ & + s^* \sum_{r \in \{\text{free}, \text{toll}\}} |\{t_a | t_a \in \mathcal{T}_{i,r}, Q_r(t) = 0 \text{ and } r_{i,r}(t) \leq s^*\}|. \end{aligned} \quad (8)$$

3. Tolls are set on the tolled route to maximize social welfare,

$$\tau_{\text{toll}}(t) = \arg \max_{\tau_{\text{toll}}(t)} \left(\sum_{i=1}^G \int_{\bar{p}_i}^{\infty} N_i(p) dp + \int_{t \in \bar{\mathcal{T}}} \tau_{\text{toll}}(t) \sum_{i=1}^G r_{i,\text{toll}}(t) dt \right).$$

4. No toll is charged on the free route,

$$\tau_{\text{free}}(t) \equiv 0 \quad \forall t \in \bar{\mathcal{T}}.$$

Proposition 1. *In a value pricing equilibrium there is a one-to-one mapping between departure time and arrival time.*

Proposition 2. *In a value pricing equilibrium the tolled route is used to its maximum capacity*

¹⁷In section X we will look at what happens when tolls are set to maximize profits, as well as with other objectives.

for all of rush hour,

$$\sum_{i=1}^G r_{i,\text{toll}}(t) = \lambda_{\text{toll}} \cdot s^* \quad \forall t \in \bigcup_{i=1}^G \mathcal{T}_{i,\text{toll}},$$

except possibly on a set of measure zero, which means

$$T_{\text{toll}}^v(t) = 0 \quad \forall t \in \bar{\mathcal{T}},$$

and the arrival rate on the tolled route is $\lambda_{\text{toll}} \cdot s^*$ for all of rush hour.

The intuition is that if the aggregate arrival rate is higher than route capacity then there is needless queuing, and if it is lower then there is needless schedule delay. All proofs not in the text, such as this one, can be found in appendix D.

Proposition 3. *In a value pricing equilibrium there is always congestion on the free route during rush hour, except for at the very start and end of rush hour (a zero measure set).*

$$T_{\text{free}}^v(t) > 0 \quad \forall t \in \bigcup_{i=1}^G \text{Int}(\mathcal{T}_{i,\text{free}}),$$

which means the arrival rate on the free route is $\lambda_{\text{free}} \cdot s$ for all of rush hour.

Note that congestion doesn't mean long travel times, just that variable travel delay is positive. The intuition for this result is that since not everyone can arrive at their desired arrival time, some agents must arrive early or late. For agents to be indifferent while facing different schedule delay costs, they must also face different travel time costs.

Because of propositions 2 and 3 we can rewrite (8) as

$$\lambda_{\text{free}} \cdot s |T_{i,\text{free}}| + \lambda_{\text{toll}} \cdot s^* |T_{i,\text{toll}}| = N_i(\bar{p}_i). \quad (9)$$

Proposition 4 (Existence). *Under assumptions X a value pricing equilibrium exists.*

Proposition 5 (Uniqueness). *Under assumptions X a value pricing equilibrium exists where the trip price of each type, aggregate departure rates, and toll schedules are unique.*

4 Homogeneous preferences

Let us start by considering when all agents are identical, except for the value they place on completing the trip; that is, we have a downward sloping demand curve. Define t_0 as the endogenous time that the first agent arrives, t_e as the endogenous time the last agent arrives.

Just knowing preferences tells us much about the form of equilibrium, independent of the congestion technology. Equations (6) and (7) hold for any congestion model, and while the functional form of (8) would be different with another model, we still have the requirement that rush hour is long enough for everyone to travel.

We can use (4) to define an isocost curve which reflects their willingness to trade-off travel time for schedule delay; it is implicitly defined by

$$T^v(t_a|\bar{p}) = \alpha^{-1}(\bar{p} - D(t_a - t^*)). \quad (10)$$

Similarly we can use (4) define an isocost curve that reflects agents' willingness to trade money for schedule delay,

$$\tau^v(t_a|\bar{p}) = \bar{p} - D(t_a - t^*). \quad (11)$$

Figure 4 shows the isocost curve defined by (10) for three different \bar{p} . Note that it is plotted in terms of departure time, which has a one-to-one mapping to schedule delay. It shows the agents' indifference between leaving at the start, t_0 , or end, t_e , of rush hour and having no variable travel time but high schedule delay; and arriving exactly on-time at t^* but having the longest commute. The slope of the isocost curve for early arrivals is the ratio of the shadow prices of time early and travel time, β/α . Similarly, the slope of the isocost curve for late arrivals is the ratio of the shadow prices of time late and travel time, $-\gamma/\alpha$.

As agents have the lowest trip price when they arrive at t^* having dealt with no congestion the bliss point is at $(t^*, 0)$. So a higher isocost curve, with a longer rush hour and worse peak travel time, means all agents are worse off, and similarly, a lower isocost curve, with a shorter rush hour and lower peak travel time, means all agents are better off. These alternate isocost curves are shown in figure 4 by the dashed lines.

Since in equilibrium all agents will be indifferent between arriving at any point during rush hour, this curve gives equilibrium travel times on a route where there is no toll charged.¹⁸ To solve, we just need to find the isocost curve such that the length of rush hour is long enough for all of the agents to travel at the trip price implied by the isocost curve. It is at this point that we need to know the congestion technology.

This also gives us an easy way to see some of the properties of equilibrium in the bottleneck model. A queue will form on free lanes immediately and last for the entire rush hour. This is necessary to generate the strictly positive travel times during the interior of rush hour; only the first and last vehicle escape queuing. Also, agents will depart over an interval, since were there any gaps in departures for the free lane the last person to leave before the gap could delay their departure without changing the time they exit the queue.

¹⁸Actually, it give us equilibrium travel times on any route where the toll is constant.

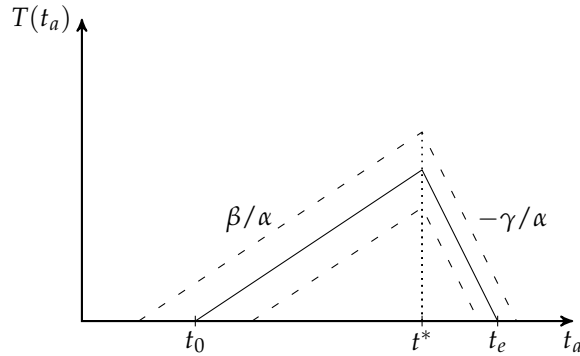


Figure 4: Isocost curve for homogeneous agents on unpriced lanes

Thus they could reduce travel time without changing arrival time. This would lower their trip price and so could not hold in equilibrium.

Just as the isocost curve helps us understand equilibrium on unpriced lanes, we can use the isocost curve from (11), reflecting agents willingness to trade-off money for schedule delay to understand equilibrium on priced lanes and shown in figure 5, to understand equilibrium on priced lanes. This isocost curve shows the agents' indifference between leaving at the start or end of rush hour and paying no toll but having high schedule delay, and leaving to arrive exactly on-time at t^* but paying a large toll.

For this isocost curve to be used in finding equilibrium for priced lanes we need to hold travel time constant over the rush period.¹⁹ For the bottleneck model this is true by Proposition 2. However, for a more general model with flow congestion travel times will not be constant on priced lanes.²⁰ Regardless of the congestion model used agents will be indifferent between traveling at any point during rush hour and the first and last agents to depart will face no congestion,²¹ so we can compare how the start of rush hour changes under different policy regimes and models and use that as a sufficient statistic for the trip price.

Proposition 6 (Identical agents worse off under pricing if on PPF). *In any congestion model, dynamic or static, with a strictly negative relationship between flows and speeds, when agents are*

¹⁹There are three things the agent cares about, travel time, schedule delay, and toll; so the true isocost curve is three dimensional. I am just looking at a slice of it.

²⁰This is because the marginal social benefit of having a agent arrive near the peak is higher than having him arrive further from the peak, and so 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. Thus it cannot be an equilibrium to have travel times be the same over the entire rush hour.

²¹If the first agent faced congestion then he could depart an insignificant amount of time sooner, with almost no change to his schedule delay, and face no congestion. Since he reduced his travel time without increasing schedule delay, he is better off. Thus the first agent to depart cannot have any congestion. Similarly for the last agent to depart.

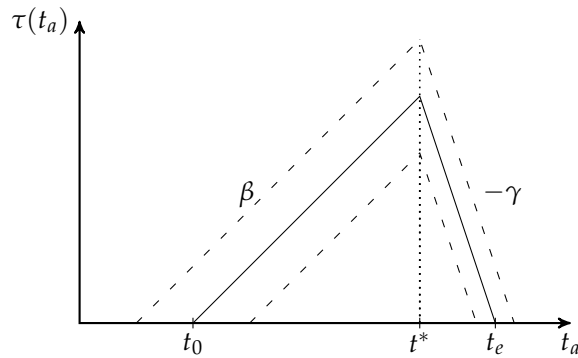


Figure 5: Isocost curve for homogeneous agents on priced lanes

homogeneous congestion pricing makes all agents worse off prior to revenue redistribution.

Proof. Since there is a congestion externality, it is optimal to price the road to reduce flows and increase speeds.

In a dynamic model this means the length of rush hour must be longer. As a result the first agent must leave earlier and so faces greater schedule delay costs and is worse off. Since all agents are identical and one agent is worse off, all agents are worse off.

In a static model this means some agents must be priced off the road. In order to have fewer people using the road we must have increased the total cost of traveling during rush hour for the marginal agents. We reduced the travel time costs but increased the financial costs. Since the inframarginal agents have the same value of time as the marginal agent, then their total cost will have increased as well. Everyone is worse off. \square

To be clear, if road usage is already on the production possibilities frontier than congestion pricing is still efficiency enhancing; it is just not a Pareto improvement, though it can become one depending on how the revenue is spent.

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” (p. 346). As Chu (1995) shows, in Henderson’s proposed equilibrium a agent would be better off by leaving after the end of rush hour; it does not satisfy equation (7). Chu reformulates the model to avoid this problem and finds that “[t]he optimal toll increases the equilibrium private trip cost” (p. 336).

This result does not hold in the bottleneck model because without congestion pricing we will not be on the production possibilities frontier. If there is no throughput drop at bottlenecks the arrival rate is same regardless of whether the road is priced. This observation gives us our next result:

Proposition 7 (Vickrey, 1969). *In the standard bottleneck model with identical agents, optimal road pricing does not change consumer welfare prior to revenue redistribution.*

Proof. Since pricing the road does not change the arrival rate, it does not change the length of rush hour. As a result the first and last agent to depart are no better off. Since all agents are indifferent between traveling at the start or end of rush hour and when they do travel, no one is better off. \square

Pricing will raise revenue and reduce travel time, but from the agents' point of view they have just swapped travel time for an equally costly toll. It is more efficient because the toll revenue is not lost, while the time spent waiting in the queue was.²²

However, when we modify the model to incorporate throughput drops at bottlenecks then tolling can increase throughput and the arrival rate at work.

Proposition 8 (Pareto improvement if throughput drop). *In the modified bottleneck model with identical agents, value pricing improves consumer welfare for all agents. Furthermore, consumer welfare is strictly increasing in the portion of the road tolled.*

Proof. Since road pricing increases the arrival rate, it reduces the length of rush hour. As a result the first and last agents have less schedule delay and so are better off. Since all agents are indifferent between traveling at the start or end of rush hour and when they do travel, they are all better off. \square

Both Proposition 7 and 8 can be derived mathematically as well. To solve for \bar{p} we use the equal price condition, (6), evaluated at t_0 and t_e , and the requirement that everyone is able to travel during rush hour, (8).

In order to satisfy the no profitable deviation requirement, (7), we need $t_0 \leq t^*$ and $t_e \geq t^*$. If $t_0 > t^*$ then $p(t^*) = 0 < p(t_0)$ since there would be no congestion at t^* , and no schedule delay by definition, while the agent leaving at t_0 does have schedule delay. Similarly if $t_e < t^*$ then $p(t^*) = 0 < p(t_e)$.

Using this and the fact that the first and last agent to depart face no congestion, we can write

$$\begin{aligned}\bar{p} &= \beta(t^* - t_0), \text{ and} \\ \bar{p} &= \gamma(t_e - t^*).\end{aligned}$$

Substituting in for λ_{free} and s turns (9) into

$$(t_e - t_0) \cdot s^*(1 - \theta + \theta\lambda_{\text{toll}}) = N(\bar{p}).$$

²²As Johnson (1964, p. 142) put it "[g]iven the existing social fabric of the system as we know it, time cannot be collected for later use."

Which gives us three simple linear equations in three unknowns. Solving these yields

$$\bar{p} = \left(\frac{\beta\gamma}{\beta + \gamma} \right) \frac{N(\bar{p})}{s^*(1 - \theta + \theta\lambda_{\text{toll}})}. \quad (12)$$

Differentiating this with respect to λ_{toll} yields

$$\frac{\partial \bar{p}}{\partial \lambda_{\text{toll}}} = - \frac{\theta}{(1 - \theta + \theta\lambda_{\text{toll}})} \frac{\bar{p}}{1 - \epsilon} \quad (13)$$

where ϵ is the price elasticity and is always weakly negative. Notice that as demand becomes more elastic the decrease in trip price becomes smaller as the increase in capacity due to pricing is used by new agents. If $\theta = 0$, i.e. $s = s^*$, then $\frac{\partial \bar{p}}{\partial \lambda_{\text{toll}}} = 0$, and so there is no change in the trip price as we toll more of the lanes. This is just Proposition 7. However, if $\theta > 0$ then $\frac{\partial \bar{p}}{\partial \lambda_{\text{toll}}} > 0$ and so as we toll more of the lanes we are decreasing trip price and so increasing consumer welfare. This is Proposition 8.

Including the throughput drop at bottlenecks also increases the social welfare gains that come from congestion pricing.

Proposition 9. *Social cost falls to*

$$\frac{(1 - \theta)(1 - \theta(1 - \lambda_{\text{toll}}) - \lambda_{\text{toll}}/2)}{(1 - \theta(1 - \lambda_{\text{toll}}))^2}$$

of its previous level.

If $\lambda_{\text{toll}} = 1$ this simplifies to

$$\frac{1 - \theta}{2}$$

of its previous level.

Proof. See appendix A.6 □

In the standard bottleneck model, where $\theta = 0$, tolling cuts the total social costs of travel in half, this comes from eliminating variable travel time (Arnott et al., 1993, p. 166). When there is a throughput drop the savings are even larger since schedule delay is also reduced.

The size of these savings are so large because fixed travel time is zero. Schuman (2011) finds that average peak period variable travel time was 95–348 percent of fixed travel time on the fifty most congested corridors in 2010 (p. 23). This would imply that total social costs would fall by 29–39 percent when $\theta = 0$, and 33–44 percent when $\theta = 0.1$.²³

Since consumer welfare is strictly increasing as we increase the amount of the lanes that are priced, it is optimal to price all of them. There is no need to keep some of the lanes

²³See appendix A.7 for details on how this was calculated. The percentage decrease in social costs due to tolling is given by (65).

free when everyone is the same. However, to capture the intuition from the introduction, let's consider if there was a small group of poor who also used the road, so small that they don't affect equilibrium.

Proposition 10. *If everyone but a non-measurable group is the same then pricing a portion, but not all, of the lanes, will lead to a Pareto improvement.*

Proof. Since the non-measurable group has no impact on equilibrium, we know by Proposition 8 that the measurable agents are better off. For the measurable agents on the free lanes to be better off, travel time must have fallen at each point in time. Thus if the non-measurable agents travel on the free lanes at the same time they traveled before then they will have shorter travel times. Since they have an option that gives them a lower trip price than before, whatever they choose gives them a lower trip price, so the non-measurable agents are better off.

Since all agents are better off, pricing a portion of the lanes leads to a Pareto improvement. \square

Proposition 10 does not mean it will not be a Pareto improvement to price the entire road, but rather that we can be sure it is a Pareto improvement if we preserve the option to pay with time by taking the unpriced lanes.

The intuition used in this proof also leads to a nice empirical test for whether pricing a portion of the lanes caused a Pareto improvement; we can check if average travel times on the free lanes fell at every point in time.

One last point I would like to make before we move on to explicitly considering multiple types is that the cost curves are not policy invariant. That is, the price of allowing N agents to travel during rush hour changes when the road is priced. This point was first made by Arnott and Kraus (1993), but seems to have been ignored. As they point out, the implication is that we should not just assume the form of the social or private cost functions, but derive them from the consumers' time-of-use decisions and the congestion technology.

Proposition 11 (Arnott and Kraus (1993)). *In the standard bottleneck model the social cost curve is not policy invariant.*

Proof. In appendix A I solve for total social cost (TSC). If we substitute $\theta = 0$ (i.e. $s = s^*$) into (64) we find that

$$\text{TSC}(\lambda_{\text{toll}}) = \frac{\beta\gamma}{\beta + \gamma} \frac{N^2}{\hat{s}} \left(\frac{1}{2} + \frac{1 - \lambda_{\text{toll}}}{2} \right),$$

and so $\partial\text{TSC}/\partial\lambda_{\text{toll}} < 0$. Thus the social cost curve is not policy invariant. \square

In the bottleneck model, and any other dynamic model, agents will change when they depart in response to a toll. So rather than just moving along a cost curve we actually shift it. In addition, in any model where we are off of the production possibilities frontier, we shift the cost curve by increasing throughput.

Proposition 12. *In the bottleneck model with throughput drops the social cost curve and private cost curve are not policy invariant.*

Proof. When $\epsilon = 0$, i.e. we are holding the number of agents fixed, and $\theta > 0$, so there is a throughput drop at bottlenecks, then $(13) < 0$. Thus the private cost curve is not policy invariant.

As in Proposition 11, using (64) we find that and so $\partial TSC / \partial \lambda_{\text{toll}} < 0$. Thus the social cost curve is not policy invariant. \square

5 Heterogeneous preferences

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 its effect on the poor, the main distinction we will make is between high and low income agents.

Recall that the preference parameters, α, β , and γ are respectively the shadow prices of travel time, time spent at work early, and time late to work. More explicitly, α is the absolute value of the ratio of the marginal utility of travel time over the marginal utility of wealth, and similarly for β and γ . We can write this out as:

$$\alpha = -\frac{\frac{\partial \text{utility}}{\partial \text{travel time}}}{\frac{\partial \text{utility}}{\partial \text{wealth}}}, \quad \beta = -\frac{\frac{\partial \text{utility}}{\partial \text{time early}}}{\frac{\partial \text{utility}}{\partial \text{wealth}}}, \quad \text{and} \quad \gamma = -\frac{\frac{\partial \text{utility}}{\partial \text{time late}}}{\frac{\partial \text{utility}}{\partial \text{wealth}}}.$$

Notice that for ratios of these parameters the marginal utility of wealth cancels out, so that we just have the marginal rate of substitution. Specifically, consider that

$$\frac{\beta}{\alpha} = -\frac{\frac{\partial \text{utility}}{\partial \text{time early}}}{\frac{\partial \text{utility}}{\partial \text{wealth}}} \times \left(-\frac{\frac{\partial \text{utility}}{\partial \text{travel time}}}{\frac{\partial \text{utility}}{\partial \text{wealth}}} \right)^{-1} = \frac{\frac{\partial \text{utility}}{\partial \text{time early}}}{\frac{\partial \text{utility}}{\partial \text{travel time}}}.$$

Thus while β is an individual's willingness to pay in money to reduce the time he spends at work prior to t^* , β/α is his willingness to pay for the same thing with his time. Similarly for γ and γ/α . Since α is the willingness to pay in money to avoid travel time it is the value of time measured by many transportation studies.

Table 2: Sources of variation in preference parameters

Parameters	Source of variation	Correlation
α, β, γ	income	+
$\frac{\beta}{\alpha}, \frac{\gamma}{\alpha}$	job flexibility	-

The implication of this is that an individual's level of wealth will have strong effects on his preference parameters through the marginal utility of wealth, however, it does not have as direct of an effect on the ratios of these parameters. This is because time early and time late mean fundamentally different things for workers with different types of jobs, and correspondingly the marginal disutility of being early or late will vary 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 there is not much difference between spending time traveling or being at work for the shift worker, his β/α will be close to one. 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 will be closer to zero. Similar logic holds for γ/α . Thus variation in the ratio of β/α and γ/α will come from differences in job flexibility, where jobs that are more flexible lead to lower ratios. These results are summarized in Table 2.²⁴

While what a worker is able to do if he is early and the consequences he faces for being late will vary with his job, the fundamental opportunity cost of travel time is the same; he has lost that time. This means the affect of wealth is strongest on α and it will be what we use to discuss differences in wealth.

Definition 6. *If $\alpha_i > \alpha_j$ then type i is richer than type j . If there are are just two values of α then any type with the larger α is rich and any type with the smaller α value is poor.*

I will show that heterogeneity in α has the effect that we typically think heterogeneity in wealth as having, making it a good proxy. Specifically, in Proposition 14 I will show that when otherwise heterogeneous agents all share the same α then congestion pricing is a weak Pareto improvement, even when there is no throughput drop at bottlenecks.

5.1 Finding equilibrium

First let's consider a completely unpriced highway. In equilibrium each agent minimizes his trip price; he cannot prefer to travel at a different time. For an agent to have no profitable deviations on an unpriced route we must be able to draw the indifference curve

²⁴How flexible a workers 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.

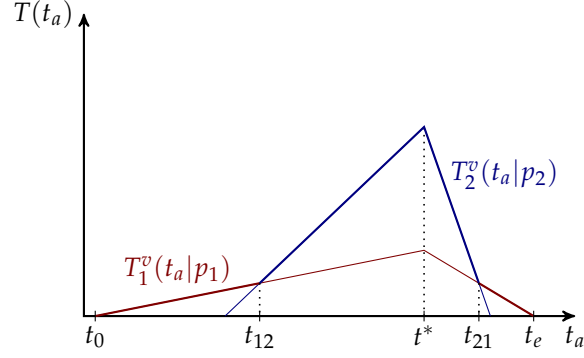


Figure 6: Isocost curves for two types of agents on an unpriced route

consistent with his trip price and find that equilibrium travel times are always at or above his indifference curve. Since this must be true for all agents of all types, equilibrium travel times must be at least as large as the pointwise maximum of the agents' isocost curves. Note that at any time where equilibrium travel time was higher than anyone's isocost curve, no one would be willing to depart for work. So given a vector of trip prices, $\mathbf{p} = (p_1, \dots, p_i)$, equilibrium travel times experienced by agents are pinned down by the non-negative upper envelope of agents' isocost curves. Given that all agents of the same type must be on the same isocost curve we can take this maximum over types,

$$T^v(t|\mathbf{p}) = \max \{T_1^v(t|p_1), \dots, T_G^v(t|p_G), 0\},$$

where $T_i^v(t|p_i)$ is the isocost curve defined in (10) for type i .

Agents of type i are willing to travel when equilibrium travel times lie along their indifference curve and so the set of demand feasible departure times on an unpriced route are

$$\bar{T}_{i,\text{free}}(\mathbf{p}) = \{t | t \in \bar{T}, T^v(t|\mathbf{p}) = T_i^v(t|p_i)\}.$$

It is possible that agents of multiple types are willing to depart at the same time, so denote the set of times agents of type i actually travel as $\mathcal{T}_{i,\text{free}} \subset \bar{T}_{i,\text{free}}$.

We can see a proposed equilibrium graphically in figure 6. The isocost curves for two groups are plotted showing the trade off each type faces between schedule delay and travel time. The start and end of rush hour are marked by t_0 and t_e . In bold is the upper envelope, and t_{ij} marks when type i stops leaving for work and type j begins. In this example $\mathcal{T}_{1,\text{free}} = [t_0, t_{12}] \cup [t_{21}, t_e]$ and $\mathcal{T}_{2,\text{free}} = (t_{12}, t_{21})$.

The final requirement for equilibrium on a free route is that the supply of travel time equals demand from each type at the price they face. Graphically this amounts to changing the height of the isocost curves until the amount of time each types isocost curve is in

the upper contour set equals the amount of time needed for all $N_i(p_i)$ of them to travel.

The same reasoning holds on a priced route, equilibrium tolls are given by the non-negative upper envelope of agents' isocost curves,

$$\tau(t|\mathbf{p}) = \max\{\tau_1(t|p_1), \dots, \tau_G(t|p_G), 0\},$$

where $\tau_i(t|p_i)$ is the isocost curve defined in (11) for type i . Here the constraint that tolls be non-negative is arbitrary, though realistic. The demand feasible departure times for type i are

$$\bar{T}_{i,\text{toll}}(\mathbf{p}) = \{t|t \in \bar{T}, \tau(t|\mathbf{p}) = \tau_i(t|p_i)\}$$

and the set of times agents of type i actually travel is denoted $\mathcal{T}_{i,\text{toll}} \subset \bar{T}_{i,\text{toll}}$.

To find equilibrium when pricing a portion of the freeway and leaving the rest unpriced we add the additional restriction that each type faces the same trip price on both routes and evaluate the supply of travel time for the highway in total, not each route individually.

5.2 Homogeneous value of time

While the bottleneck model is dynamic and adds several additional dimensions of heterogeneity, such as preferred arrival time and schedule delay costs, the distributional affects of congestion pricing are still primarily driven by differences in agents' value of time. When we price highways we change the currency used to buy access from time to money. If everyone has the same exchange rate between these two currencies then congestion pricing is able to internalize the congestion externality without tilting the playing field in favor any type of agent over another.

For the bottleneck model adding heterogeneity in preferred arrival time and schedule delay costs, but maintaining a homogeneous value of time, does not substantively change any of the results of the previous section. Just as in Proposition 7, if pricing doesn't change highway throughput then it doesn't change consumer welfare.

Proposition 13. *In the standard bottleneck model with otherwise heterogeneous agents who have the same value of time, α , value pricing does not change consumer welfare prior to revenue redistribution.*

Proof. Outline:

1. Let X be equilibrium when highway free
2. Show that X is still an equilibrium when highway priced

(a) Show $\bar{T}_{i,\text{toll}}(\mathbf{p}) = \bar{T}_{i,\text{free}}(\mathbf{p}) \forall i$.

(b) So we can choose $\mathcal{T}_{i,\text{toll}} = \mathcal{T}_{i,\text{free}}$.

(c) Since $N_i(p_i) = s |\mathcal{T}_{i,\text{free}}|$, $N_i(p_i) = s |\mathcal{T}_{i,\text{toll}}|$

3. So X is still an equilibrium with the exact same trip prices.

□

This result conforms with Arnott et al. (1994) who consider the welfare effects of congestion pricing in the standard bottleneck model with linear schedule delay costs and two types. They add heterogeneity in one dimension at a time and find that congestion pricing is welfare neutral when otherwise identical agents have different desired time of arrival or cost of being late (p. 157).

Similarly to Proposition 8, if pricing increases throughput then it will be a Pareto improvement.

Proposition 14 (Pareto improvement if homogeneous value of time). *In the modified bottleneck model with otherwise heterogeneous agents who have the same value of time, α , value pricing improves consumer welfare prior to revenue redistribution. Furthermore, consumer welfare is strictly increasing in the portion of the road tolled.*

Proof. Forthcoming.

Outline:

1. Assume false, at least one type with a higher trip price.
2. Then either they have a neighbor who also has a higher trip price or they have more capacity.
3. If they have more capacity then this is not an equilibrium, so they must have a neighbor who also has a higher trip price.
4. Then the first group plus the neighbor either have a neighbor who has a higher trip price, or jointly have more capacity.
5. Continue this, run out of types who can have a higher trip price.
6. So you cannot have any type with a higher trip price.

□

5.3 Potential barriers to a Pareto improvement

Heterogeneity in agent's value of time, however, can prevent pricing from being a Pareto improvement for at least two reasons. The standard problem is that changing the currency

from time to money reduces the cost for the rich while raising the cost for the poor, potentially pricing them off the road. By considering choice of when to travel explicitly we give the poor another way to react to pricing. This creates an additional problem since now the rich may displace the poor from peak travel times but also gives a way for pricing the entire road can be a Pareto improvement.

With both of these problems the question is whether the increase in total capacity due to pricing will be enough to offset the harm.

Proposition 15. *For any set of parameters there exists a large enough capacity drop such that value pricing leads to a Pareto improvement.*

Proof. Forthcoming.

For θ arbitrarily large I can have the length of rush hour shrink arbitrary small. For inelastic demand if we let N_1 be the mass of the type with the largest β/α then for $\theta > 1 - N_1$ the length of rush hour shrink enough so that everyone is traveling when the type with the highest β/α was traveling. This means everyone is better off. \square

In some ways this is a ridiculous proposition, since if $\theta = 1$ then without pricing no one can travel at all, and so pricing will be a Pareto improvement regardless of what the other parameters are. However, the inverse is not true, that for any set of parameters there is a small enough capacity drop such that value pricing cannot lead to a Pareto improvement.²⁵

To illustrate these potential problems we will compare a completely free highway to a completely priced one when there is no throughput drop at bottlenecks, $s = s^*$. I will simplify to two types who share the same desired arrival time, $t^* = t_1^* = t_2^*$. They will have piecewise linear schedule delay costs, and to further reduce the number of dimensions of heterogeneity I assume $\gamma_i = \zeta\beta_i$ for $i = 1, 2$; where $\zeta > 1$. For the first two problems I will assume demand is inelastic. With these assumptions this is the model of Arnott et al. (1994) Section 3. Without loss of generality I assume $\alpha_1 > \alpha_2$, so type 1 agents are rich and type 2 agents are poor.

Looking back at figure 6 observe the slope of the isocost curves are increasing prior to t^* and decreasing after. This means that types with the largest values β/α arrive closest to t^* on the free route and those with the largest β arrive closest to t^* on the priced route. Those who are willing to pay the most, in time on a free route or money on a paid one, get the most desirable travel times.

Proposition 16. *When all agents share the same desired time of arrival (t^*) and have piecewise linear schedule delay costs, agents on unpriced routes arrive in increasing order of β/α before t^* and in decreasing order of γ/α after. Similarly on priced routes they arrive in increasing order of*

²⁵Of course, were we to allow a negative "capacity drop" then it could be true.

β before t^* and decreasing order of γ after. Note that not all types will necessarily arrive before and after t^* .²⁶

Formally, this is:

- If there exists a $t_i \in \mathcal{T}_{i,\text{free}}$ and $t_j \in \mathcal{T}_{j,\text{free}}$ such that $t_i < t_j \leq t^*$, then $\beta_i/\alpha_i \leq \beta_j/\alpha_j$.
- If there exists a $t_i \in \mathcal{T}_{i,\text{free}}$ and $t_j \in \mathcal{T}_{j,\text{free}}$ such that $t_i > t_j \geq t^*$, then $\gamma_i/\alpha_i \leq \gamma_j/\alpha_j$.
- If there exists a $t_i \in \mathcal{T}_{i,\text{toll}}$ and $t_j \in \mathcal{T}_{j,\text{toll}}$ such that $t_i < t_j \leq t^*$, then $\beta_i \leq \beta_j$.
- If there exists a $t_i \in \mathcal{T}_{i,\text{toll}}$ and $t_j \in \mathcal{T}_{j,\text{toll}}$ such that $t_i > t_j \geq t^*$, then $\gamma_i \leq \gamma_j$.

Proof. Forthcoming. □

This result will make finding equilibrium much simpler once we know who travels on which route. Solving the remaining problem of knowing who travels on which route is significant, and is made much easier by only having two types. This problem goes away completely when we only consider all free and all priced.

5.3.1 Poor must pay more costly tolls

The standard problem with pricing a good that was previously allocated by queuing is that it raises the price for the poor relative to the rich.²⁷ This will be a problem when the poor travel at the peak on free and priced routes. By Proposition 16 this requires $\beta_2/\alpha_2 > \beta_1/\alpha_1$ and $\beta_2 > \beta_1$, so the poor are willing to pay more in travel time and money to travel at the peak than the rich. Since $\alpha_1 > \alpha_2$,

$$\frac{\beta_2/\alpha_2}{\beta_1/\alpha_1} > \frac{\beta_2}{\beta_1},$$

meaning the difference between the poor and rich agents' willingness to pay with time is greater than the difference between their willingness to pay with money.

We can see this graphically in figure 7, which shows equilibrium isocost curves on a completely free highway and a completely priced highway. When the road is priced the difference between the two types slopes is smaller and so the poor end up on a higher isocost curve. Note that the isocost curve for the rich doesn't change. If there is no throughput drop at bottlenecks then pricing doesn't change the length of rush hour. Since in both situations a rich agent must travel at beginning and end of rush hour and those times are unchanged, their welfare is unchanged. In this case prior to redistributing the toll revenue pricing hurts the poor without helping the rich.

²⁶Can generalize to single peaked instead of same t^* .

²⁷Perhaps the word problem should be in quotes, it is a problem for generating a Pareto improvement, but often the purpose of pricing is to change the allocation of the good to those who value it more.

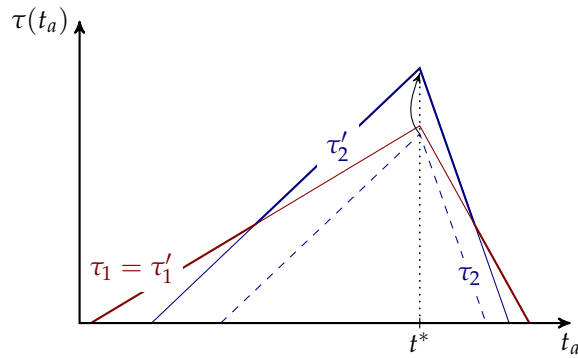


Figure 7: Equilibrium isocost curves when road is priced

5.3.2 Rich displace poor from peak

The wealthy with flexible schedules present another barrier in the way of achieving a Pareto improvement from congestion pricing. Because they have flexible schedules, when the road is unpriced and so heavily congested they can just arrive at work early or late so as to avoid wasting their valuable time in traffic. However, they are rich enough that when there is a priced option they will choose to pay and arrive at his desired time, avoiding the inconvenience of getting to work early or late.

The problem is that the wealthy and flexible are using capacity that had been available for those with less flexible schedules. Now those who had been traveling at the peak must outbid each other for the now more scarce peak travel times or travel off peak. Either way they are worse off.

Notice that the rich are better off. Before they had been traveling at the start or end of rush hour and facing no congestion; they still has that option and so by revealed preference prefers to travel at the peak and pay the toll. When the rich displace the poor from the peak we are improving efficiency by reallocating the most desirable arrival times to those who have the highest willingness to pay.

By Proposition 16 we know that for this to be a problem we need $\beta_2/\alpha_2 > \beta_1/\alpha_1$, so that on a free route the poor travel at the peak, and $\beta_1 > \beta_2$, so that on a priced route the rich travel at the peak.

Figure 8 shows how pricing affects welfare graphically.

5.3.3 Rich displace poor from road

When pricing raises the price the poor must pay they can either just pay it or travel at a different time, as above, but they also can react by not traveling at all or choosing some option outside of the model, such as taking another highway or public transportation. This

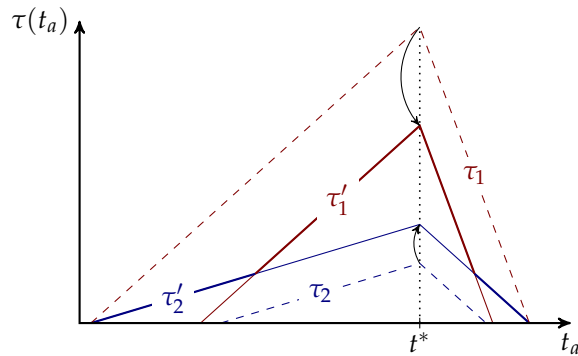


Figure 8: Equilibrium isocost curves when rich displace poor from peak

FIGURE: Show isocost curves on priced and free lanes next to each other. Assume $\alpha_2 = 1$ so that slope of poor type's isocost curve is same in both figures (and for consistency between figures). Show old iso-cost curves for both types on paid lanes, with arrow showing change.

Figure 9: Equilibrium isocost curves when rich displace poor from road

was ruled out above by the assumption of inelastic demand. The assumption of inelastic demand, however, masks another potential barrier to achieving a Pareto improvement. Even when pricing doesn't directly raise the cost of the poor, by virtue of reducing the cost for the rich it can induce more rich to travel and displace poor from the road.

To shut down the previous two effects, I now assume $\beta_1/\alpha_1 > \beta_2/\alpha_1$ and $\beta_1 > \beta_2$, so that the rich always travel at the peak and the poor always travel off-peak.²⁸

Figure 9 shows this graphically.

In many ways this potential problem is similar to the previous problem, when the rich displace the poor from the peak. That is when the rich switch from off-peak travel times to peak travel times, pushing the poor to the off-peak. In this case the rich are switching from out of model routes or modes and pushing the poor further off-peak.

5.4 When can pricing be a Pareto improvement?

With our understanding of the potential barriers that can prevent pricing highways from being a Pareto improvement we can turn to tightening the bounds on income inequality needed to assure we can obtain a Pareto improvement from pricing. Proposition 14 gives us the extreme bound, if there is no inequality and pricing can increase throughput then pricing is always a Pareto improvement.

The bound on inequality will depend on how much variation in flexibility there is

²⁸Note that we have now considered every possible combination ordering of these parameters.

between agents. The more similar the flexibility across agents the greater the allowed variation in wealth. With inelastic demand the worst possible outcome for the poor is that they start traveling at the start and end of rush hour. The more similar they are to the rich in terms of flexibility, the less costly of a change that will be. We can see this graphically as the distance between where their original isocost curve crosses the x-axis and t_0 in figure 6.

Taking this to the extreme we get the following proposition.

Proposition 17. *If $D_i(x)/\alpha_i = D_j(x)/\alpha_j$ for all types i and j and all x and demand is perfectly inelastic then pricing is always a Pareto improvement.*

Proof. On a completely free route everyone is indifferent between traveling at any time during rush hour, so even if pricing changes when types travel it only puts them somewhere they were indifferent between. Basically, the first two problems above cannot happen. \square

5.4.1 Two types

Let's continue with the assumptions of two types, one rich and one poor, piecewise schedule delay costs, proportional cost of being late to early, and homogeneous desired time of arrival. I will also maintain the assumption of inelastic demand.

Let's start by finding equilibrium trip price on an upriced highway. To avoid needing to solve for all of the different cases, let type A travel at the peak and type B travel off peak. Then as needed we will match types 1 and 2 to types A and B .

In equilibrium there must be enough time for all agents of both types to travel,

$$N_A = (t_{BA} - t_{AB}) s, \text{ and} \quad (14)$$

$$N_B = [(t_{AB} - t_0) + (t_e - t_{BA})] s. \quad (15)$$

These two equations give us four critical times, when rush hour starts and stops and when the type of agent that is traveling changes. In equilibrium type B agents must be indifferent between traveling at t_0 and t_e with no congestion, and similarly type A agents must be indifferent between traveling at t_{BA} and t_{AB} ;

$$\bar{p}_B = p_B(t_0) = p_B(t_e), \text{ and} \quad (16)$$

$$\bar{p}_A = p_A(t_{BA}) = p_A(t_{AB}). \quad (17)$$

There will be congestion at t_{BA} and t_{AB} and we can determine how much using type B 's isocost curves;

$$T_{\text{free}}^v(t_{BA}) = T^v(t_{BA}|\bar{p}_b), \text{ and} \quad (18)$$

$$T_{\text{free}}^v(t_{AB}) = T^v(t_{AB}|\bar{p}_b). \quad (19)$$

This gives us eight equations in eight unknowns, and because we assumed piecewise linear schedule delay costs these equations are all linear; solving this system of linear equations yields

$$\bar{p}_{B,\text{free}} = \beta_B \frac{N_A + N_B}{s} \frac{\xi}{1 + \xi}, \text{ and} \quad (20)$$

$$\bar{p}_{A,\text{free}} = \left(\beta_A f_A + \frac{\alpha_A}{\alpha_B} \cdot \beta_B (1 - f_A) \right) \frac{N_A + N_B}{s} \frac{\xi}{1 + \xi}. \quad (21)$$

where $f_B = N_B / (N_A + N_B)$, the fraction of agents of type B .

The system of equations for a priced highway is very similar. The only changes we need to make are to replace s with s^* in (14) and (15), and replace (18) and (19) with

$$\tau_{\text{toll}}(t_{BA}) = \tau(t_{BA}|\bar{p}_b), \text{ and} \quad (22)$$

$$\tau_{\text{toll}}(t_{AB}) = \tau(t_{AB}|\bar{p}_b). \quad (23)$$

Solving this system of equations yields

$$\bar{p}_{B,\text{toll}} = \beta_B \frac{N_A + N_B}{s^*} \frac{\xi}{1 + \xi}, \text{ and} \quad (24)$$

$$\bar{p}_{A,\text{toll}} = (\beta_A f_A + \beta_B (1 - f_A)) \frac{N_A + N_B}{s^*} \frac{\xi}{1 + \xi}. \quad (25)$$

Proposition 16 tells us how to tell whether the rich or poor agents are type A ; on a free highway the type A agent is the one with the larger β/α and on a priced highway the type A agent is the one with the larger β .

Now that we have solved for equilibrium trip price we can find when pricing the entire highway is a Pareto improvement by finding when $\bar{p}_{1,\text{toll}} \leq \bar{p}_{1,\text{free}}$ and $\bar{p}_{2,\text{toll}} \leq \bar{p}_{2,\text{free}}$, with one type strictly better off.

Proposition 18. *With two types, piecewise linear schedule delay costs, homogeneous desired time of arrival, and inelastic demand, pricing the entire road is a Pareto improvement prior to revenue redistribution only if*

$$\frac{\alpha_2}{\alpha_1} \geq \frac{\beta_2}{\beta_1} \left(\min \left\{ 1, \frac{\beta_1}{\beta_2} \right\} (1 - \theta) - \frac{1 - f_1}{f_1} \theta \right). \quad (26)$$

Proof. Forthcoming. □

The rich are always weakly better off under pricing, and for a positive throughput drop they are always strictly better off, so it comes down to checking if the poor are weakly better off.

The minimum term comes from the fact that if $\beta_1 > \beta_2$ then the rich travel at the peak on the priced route and then it doesn't matter how much more the rich are willing to pay than the poor, just that they are willing to pay more. But when the poor are willing to pay more than the rich for the peak travel times it does matter since they have to outbid the rich.

Equation 26 gives us a lower bound on the ratio of the poor agents' value of time to the rich agents' value of time. Notice that the right hand side is always less than one, so that if both types have the same value of time then pricing the whole road is a Pareto improvement. This matches Proposition 14

We can however, increase the range of parameter values where pricing can lead to a Pareto improvement by considering pricing only $\lambda_{\text{toll}} \in (0, 1)$ of the lanes.

Before we had three cases, where different ordering of parameters lead to different systems of equations. Now for two of the cases we have two subcases. In some subcases both types are on both routes while in others only one type is on one of the routes. Since working through five different cases would be tedious, it is done as part of the proof of Proposition 19, which is in the appendix.

In each case we write down equations requiring that the supply of travel time equals the demand for travel time, and then evaluate each types trip price when they start and stop traveling. To determine the toll or travel time when types switch we use the isocost curve of the type that is traveling off-peak on that route.

We only get the additional subcases when $\beta_2/\alpha_2 > \beta_1/\alpha_1$, so that the poor are traveling at the peak on the free route.

If $\beta_1 > \beta_2$ then on an entirely priced highway the poor will travel off peak. This means the rich will fill up the priced route until there are no rich agents left so we change subcases when $\lambda_{\text{toll}} > f_1$.

If $\beta_2 > \beta_1$ then on an entirely priced highway the poor will travel at the peak, so we need to check for when the poor start traveling on the paid route. The poor will switch when the cost of paying the toll at t^* is the same as their trip price on the free route. Since the trip price at t^* on the paid route is entirely due to the toll and so is the same for both types, we change subcases when $\bar{p}_1(\lambda_{\text{toll}}) = \bar{p}_2(\lambda_{\text{toll}})$.

Proposition 19. *With two types, piecewise linear schedule delay costs, homogeneous desired time of arrival, and inelastic demand, pricing a portion of the road is a Pareto improvement prior to revenue redistribution if*

Figure:

Figure 10: Parameter values where pricing leads to a Pareto improvement

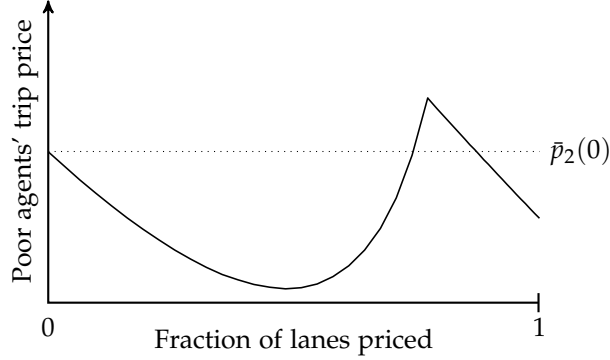


Figure 11: Example of effect of pricing λ_{toll} of the highway on the poor type's trip price with perfectly inelastic demand

$$\frac{\alpha_2}{\alpha_1} > \frac{\beta_2}{\beta_1} \left(\min \left\{ 1, \frac{\beta_1}{\beta_2} \right\} (1 - \theta) - \frac{1 - f_1}{f_1} \theta \right), \text{ or} \quad (27)$$

$$\frac{\alpha_2}{\alpha_1} > \frac{\beta_2 (1 - f_1)(1 - \theta)}{\beta_1 (1 - f_1(1 - \theta))}. \quad (28)$$

If pricing all of the road is a Pareto improvement, it doesn't follow that pricing half of the road is as well. Figure 11 shows the trip price for the poor type as a function of the fraction of the road priced. The first portion of the curve is essentially quadratic because there are two effects on the poor types trip price, both of which are basically linear.²⁹ The poor are better off since rush hour is shrinking, but worse off because the rich are displacing them from the peak. Notice that their trip price peaks at the point when all of the rich are on the paid lanes and then falls almost linearly after that. This is when the poor start traveling on the priced lane, at this point they are as bad off as they can be since they are traveling at the start and end of rush hour and they benefit from the shrinking of rush hour.

5.4.2 Arbitrary number of types

Now we turn to generalizing the results to allow for more than just two types. To do so we will use a geometric shortcut for determining an agent's trip price. First let's look at an unpriced route, shown in figure 12. An agent who arrives at his desired arrival time

²⁹Meaning that when you look at a plot of it, it looks linear.

Figure: Lots of isocost curves, single peak but different tStar. All but the one in question should be faded somehow (dashed line perhaps). Mark pBar on y-axis, t0, tStar, tE, perhaps when group in question starts and stops traveling. Draw right triangle. Line segments leading to peak of agents triangle should be in bold.

Figure 12: Finding trip price on a free route

pays no schedule delay costs, and so the entirety of his costs are travel time costs. The height of his type's isocost curve at t^* gives us his travel time, and his travel time maps directly into the trip price that all agents of his type face: we just multiply by his value of time, α .

The height of the isocost curve at t^* is the same as the height of the right triangle with vertices at $(t_0, 0)$, $(t_i^*, 0)$, and $(t_i^*, \bar{p}/\alpha)$. The slope of the hypotenuse, m , is the height divided by the length of the base of the triangle (rise over run), and a little rearranging gives us

$$\bar{p}_i = \alpha_i m (t_i^* - t_0). \quad (29)$$

This raises two questions: what is the slope of the hypotenuse and what economic meaning does it have? In figure 12 m is the average slope of the line segments in bold. That is, it is the average of the willingness to pay in travel time to avoid time early of those agents who arrive before the agent in question, plus his willingness to pay in travel time times the mass of agents who arrive between him and his preferred arrival time. To put it another way, it is the average of the willingness to pay in travel time to avoid time early (β/α) of those who arrive before the agent's desired arrival time (t_i^*), where we replace the preferences of those arriving after the agent with his own.

The logic is the same on a priced route as well as for agents who arrive late, with the replacement of the willingness to pay in travel time with the willingness to pay in money in the case of a priced route and the replacement of time early with time late for those who are late. These results are summarized in the following proposition.

Proposition 20. *In any bottleneck model with finite types who have linear schedule delay costs and where there is a single peak, then if agent i arrive early on an unpriced route his trip price only depends on the average willingness to pay in travel time to avoid time early of those who arrive before him, the length of rush hour prior to his desired arrival time, and his preferences and arrival time.*

Similarly, if agent i arrives early on a priced route then his trip price only depends on the average willingness to pay in money to avoid time early of those who arrive before him, the length of rush hour prior to his desired arrival time, and his willingness to pay in money to avoid time early and arrival time.

If he arrives late than we just replace time early with time late in the above statements.

Specifically, agent i 's trip price is

$$\bar{p}_{i,\text{free}} = \begin{cases} \alpha_i \left(\text{Ave} \left(\frac{\beta_j}{\alpha_j} \mid t_j < t_i \right) \cdot (t_i - t_0) + \frac{\beta_i}{\alpha_i} (t_i^* - t_i) \right) (t_i^* - t_0) & \text{if } t_i \leq t^*, \\ \alpha_i \left(\text{Ave} \left(\frac{\gamma_j}{\alpha_j} \mid t_j < t_i \right) \cdot (t_i - t_0) + \frac{\gamma_i}{\alpha_i} (t_i^* - t_i) \right) (t_e - t_i^*) & \text{if } t_i > t^*; \end{cases} \text{ or} \quad (30)$$

$$\bar{p}_{i,\text{toll}} = \begin{cases} (\text{Ave}(\beta_j \mid t_j < t_i) \cdot (t_i - t_0) + \beta_i (t_i^* - t_i)) (t_i^* - t_0) & \text{if } t_i \leq t^*, \\ (\text{Ave}(\gamma_j \mid t_j < t_i) \cdot (t_i - t_0) + \gamma_i (t_i^* - t_i)) (t_e - t_i^*) & \text{if } t_i > t^*. \end{cases} \quad (31)$$

Please note that while everywhere else I index types and discuss types, here we are indexing and discussing agents.

Proof. Forthcoming. □

Note that Proposition 20 holds independent of how congestion is modeled on unpriced routes. On a paid route it only holds when the level of congestion is level over rush hour, which will not be true under more general congestion technologies. It can also be generalized to non-linear schedule delay costs, but then costs will no longer be in terms of the censored mean but just of a weighted average and so the interpretation is less clear. For types that travel on both sides of the peak we could also find their trip price as a function of both the censored mean of β/α and γ/α and drop the dependence on t^* .

Unfortunately Proposition 20 does not give us trip prices in terms of the primitives of the model, to get this I will need to make some additional assumptions.

Proposition 21. *In any bottleneck model with finite types who have linear schedule delay costs, homogeneous desired time of arrival, and $\gamma_i = \zeta \beta_i$ for all i , then type i 's trip price is*

$$\bar{p}_{i,\text{free}} = \alpha_i \left(\frac{\bar{\beta}}{\alpha} \right)_i^c \frac{N}{s} \frac{\zeta}{1 + \zeta}, \text{ or} \quad (32)$$

$$\bar{p}_{i,\text{toll}} = \bar{\beta}_i^c \frac{N}{s^*} \frac{\zeta}{1 + \zeta}. \quad (33)$$

where

$$\left(\frac{\bar{\beta}}{\alpha} \right)_i^c = \int_0^{\beta_i/\alpha_i} x f_{\beta/\alpha}(x) dx + \beta_i/\alpha_i \left(1 - F_{\beta/\alpha}(\beta_i/\alpha_i) \right), \text{ and} \quad (34)$$

$$\bar{\beta}_i^c = \int_0^{\beta_i} x f_{\beta}(x) dx + \beta_i/\alpha_i \left(1 - F_{\beta}(\beta_i) \right) \quad (35)$$

are the right censored means, censored at type i 's level of the parameter in question, and f_x and F_x are the probability and cumulative distribution functions of x .

Proof. Forthcoming. Can use recursive method here. □

This proposition says three things matter for an agent's trip price, the total length of rush hour, the mass of agents with the same or higher willingness to pay, and the average preferences of those with a lower willingness to pay.

The actual preferences of those with a higher willingness to pay don't matter, whether they are willing to pay a cent more or a thousand dollars more for the prime arrival times doesn't matter; either way they will outbid the agent for those spots and so all that matters is how much of the desirable space they will use.

The preferences of who that the agent must outbid do matter however, since he must actually outbid them. He only directly cares about those with a willingness to pay right below him, but since they in turn care about those with a willingness to pay directly below them, he cares as well. You must outbid me, and I must outbid someone else. If the person I must outbid suddenly has a higher willingness to pay then I must raise my bid, and for you to continue outbidding me, you too must raise your bid.

Using Proposition 21 we can set some limits on the amount of income inequality a region can have and still expect to get a Pareto improvement from pricing the entire highway.

Proposition 22 (Ideal, not yet achieved). *If some measure of the covariance of β and α is less than s^*/s then pricing the entire road is a Pareto improvement. Ideally it would have a direct measure of income inequality in it.*

Proposition 23. *In the modified bottleneck model with finite types who have linear schedule delay costs, inelastic demand, homogeneous desired time of arrival, and $\gamma_i = \xi\beta_i$ for all i , and $\beta_i \geq \beta_j \Leftrightarrow \alpha_i \geq \alpha_j$ for all i, j , then pricing the entire road is a Pareto improvement.*

Proof. Forthcoming. □

In many situations pricing the entire road will not be Pareto improving, even though it a Kaldor-Hicks improvement. In many of these situations pricing just a portion of the road will be a Pareto improvement. By giving up some of the efficiency gains (due to not pricing the entire road) we make pricing politically feasible.

The formulas for trip price are very similar to those in Proposition 21. The only changes are that we now take the censored mean conditional on those on each route and that the term for the length of rush hour is different.

Proposition 24. *In any bottleneck model with finite types who have linear schedule delay costs, homogeneous desired time of arrival, and $\gamma_i = \xi\beta_i$ for all i , then type i 's trip price is*

$$\bar{p}_{i,\text{free}}(\lambda_{\text{toll}}) = \alpha_i \left(\frac{\bar{\beta}}{\alpha}\right)_i^c (\lambda_{\text{toll}}) \frac{N}{s\lambda_{\text{free}} + s^*\lambda_{\text{toll}}} \frac{\zeta}{1 + \zeta}, \text{ or} \quad (36)$$

$$\bar{p}_{i,\text{toll}}(\lambda_{\text{toll}}) = \bar{\beta}_i^c (\lambda_{\text{toll}}) \frac{N}{s\lambda_{\text{free}} + s^*\lambda_{\text{toll}}} \frac{\zeta}{1 + \zeta}. \quad (37)$$

where

$$\left(\frac{\bar{\beta}}{\alpha}\right)_i^c (\lambda_{\text{toll}}) = \int_0^{\beta_i/\alpha_i} x f_{\beta/\alpha}(x; \lambda_{\text{toll}}) dx + \beta_i/\alpha_i \left(1 - F_{\beta/\alpha}(\beta_i/\alpha_i; \lambda_{\text{toll}})\right), \text{ and} \quad (38)$$

$$\bar{\beta}_i^c (\lambda_{\text{toll}}) = \int_0^{\beta_i} x f_{\beta}(x; \lambda_{\text{toll}}) dx + \beta_i/\alpha_i \left(1 - F_{\beta}(\beta_i; \lambda_{\text{toll}})\right) \quad (39)$$

are the right censored means, censored at type i 's level of the parameter in question, and f_x and F_x are the probability and cumulative distribution functions of x for the population of agents on the free route when $x = \beta/\alpha$ and the unpriced route when $x = \beta$.

Proof. Forthcoming. □

Determining which types travel on which route is still a problem, and prevents me from being able to state Proposition 24 strictly in terms of the primitives of the model. It is however, easy to test if we have the assignment of types to routes wrong. An agent will travel on the route where he faces the lowest cost, for example if agents of type i are traveling on the priced route then $\bar{p}_{i,\text{toll}}(\lambda_{\text{toll}}) \leq \bar{p}_{i,\text{free}}(\lambda_{\text{toll}})$.

However, building on the intuition that the rich will prefer the priced lanes we get the following proposition.

Proposition 25. *If there are two types with equally flexible schedules, then as we increase the portion of the lanes that are priced the richer type will start traveling on the priced lanes before the poorer type.*

If $\beta_i/\alpha_i = \beta_j/\alpha_j$ for $i \neq j$ and $\alpha_i > \alpha_j$ then if for a given λ_{toll} type j is on the priced route, so is type i ; and if type i is on the free route, so is type j .

Proof. Forthcoming. □

The examples I used to illustrate the potential pitfalls of pricing had a silly feature; the rich and poor never traveled at the same time. Casual empiricism says that Lexus' and Kia's regularly share the road. If at any given moment half of agents are rich, then we can price half the road and have the rich take the priced lanes leaving the poor in the free lanes. We would have a Pareto improvement.

Proposition 26. *In the modified bottleneck model with homogeneous desired time of arrival and $\gamma_i = \zeta\beta_i$ for all i , if there is an arbitrary number of levels of flexibility, and at each level of*

flexibility there was a rich and a poor type. Then if ι is the minimum fraction of agents of a given level of flexibility that are rich, pricing ι of the lanes will be a Pareto improvement prior to revenue redistribution.

Proof. Forthcoming. □

The basic idea here is that all that changes is the length of rush hour, the distribution of β/α on the free lane remains unchanged. Since those on the free lane are better off, those who switched could have stayed and been better off. Thus, by revealed preference the switchers are better off too.

6 Conclusion

Our current understanding is that congestion pricing, while efficient, creates winners and losers (e.g. TODO: Add citations). A lot of work has gone into how to use the revenue raised to compensate losers so that it will be a Pareto improvement (e.g. TODO: add citations). Using insights from Vickrey and the traffic engineering literature I have shown that congestion pricing can naturally generate a Pareto improvement before the revenue is spent. In many cases this will require sacrificing some of the potential efficiency gains by only pricing a portion of the capacity in order to leave a free option for those who prefer to pay with their time. Even if policy makers choose to price the entire road, this research shows that the cost to the poor will be much smaller than previously believed.

The harm to the poor comes from taking a resource that had been allocated by time and allocating it by money. This is a good thing since allocating by time is inherently wasteful but benefits the rich at the expense of the poor. There are, however, several ways of allowing the poor to pay with their time for access to priced lanes. 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 the existence of the priced lanes, have cheaper financial costs than driving solo, but take additional time in getting to the stop, waiting for the bus, etc.

I explicitly abstracted away from what was to be done with the revenue, and it could be used in a way to strengthen the case that pricing is a Pareto improvement. If we included good uses of the revenue the optimal toll may even be higher than throughput maximizing because of the benefits. As Vickrey (TODO: Add citation) pointed out, when pricing externalities you are imposing a tax with a negative dead-weight loss and can use that revenue to reduce other taxes that impose a dead-weight loss, increasing the benefits even more.

By recognizing that pricing can increase the throughput of a highway we are able to turn pricing from a Kaldor-Hicks improvement to a Pareto improvement. It further changes other results in the field; for example, it has been suggested that the presence of agglomeration externalities weaken the case for congestion pricing (e.g. TODO: add citations (arnott, others)), but if pricing can increase throughput then agglomeration externalities increase the benefits from pricing and add another reason to price to maximize throughput.

A Solving for equilibrium in the basic model

Before we get started, let us write out the equation (6), the equal trip price condition, in detail. It is

$$\bar{p} = \alpha T_{\text{free}}(t) + \beta(t^* - t - T_{\text{free}}(t)) \quad \text{for } t \in [t_0, \tilde{t}_{\text{free}}], \quad (40)$$

$$= \alpha T_{\text{free}}(t) + \gamma(t + T_{\text{free}}(t) - t^*) \quad \text{for } t \in (\tilde{t}_{\text{free}}, t_e], \quad (41)$$

$$= \alpha T^f + \beta(t^* - t) + \tau_{\text{toll}}(t) \quad \text{for } t \in [t_0, \tilde{t}_{\text{toll}}], \quad (42)$$

$$= \alpha T^f + \gamma(t - t^*) + \tau_{\text{toll}}(t) \quad \text{for } t \in (\tilde{t}_{\text{toll}}, t_e]. \quad (43)$$

We solve for equilibrium in four parts. First we will prove some results about the nature of the equilibrium. The following three steps then derive the exact form, and the order of them is somewhat irrelevant. We need to find some key constants, the start of the rush hour, t_0 , the end of rush hour, t_e , the times that commuters need to depart in order to arrive ontime on each route, \tilde{t}_{free} and \tilde{t}_{toll} , as well as the cost of making a trip. We also need to find $r_{\text{free}}(t)$ for the free lanes. Finally we need to find the optimal tolls, $\tau_{\text{toll}}(t)$.

A.1 Proofs

Proposition 27. *Arrivals occur over a connected interval.*

Proposition 28. *Toll road will never have a queue.*

Proposition 29. *If there is a queue in the free lane then the toll road will be being used to capacity.*

Proposition 30. *A queue forms in the free lane immediately. Whenever it is used, there is a queue.*

A.2 Key constants

Now we want to solve for t_0 , t_e , \tilde{t}_{free} , \tilde{t}_{toll} , and \bar{p} . Note that the capacity of the road is

$$\hat{s} = \lambda_{\text{free}} \cdot s + \lambda_{\text{toll}} \cdot s^* = s^*(1 - \theta \lambda_{\text{free}}). \quad (44)$$

There are five linear equations which define these four times and the trip price.

$$\begin{aligned}\hat{s}(t_e - t_0) &= N, \\ \bar{p} &= \alpha T^f + \beta(t^* - t_0 - T^f), \\ \bar{p} &= \alpha T^f + \gamma(t_e + T^f - t^*), \\ \bar{p} &= \alpha(t^* - \tilde{t}_{\text{free}}), \text{ and} \\ \tilde{t}_{\text{toll}} &= t^* - T^f.\end{aligned}$$

The first equation says that rush hour is long enough for all N commuters to make the trip. The next three equations are that the trip price is the same for the agent who leaves first, last, and who receives an on-time arrival. The last equation uses the result that there is no congestion on the tolled route to define \tilde{t}_{toll} , the toll that is paid by those agents who arrive at t^* . Using linear algebra to solve this yields

$$t_0 = t^* - T^f - \left(\frac{\gamma}{\beta + \gamma}\right) \frac{N}{\hat{s}}, \quad (45)$$

$$t_e = t^* - T^f + \left(\frac{\beta}{\beta + \gamma}\right) \frac{N}{\hat{s}}, \quad (46)$$

$$\tilde{t}_{\text{free}} = t^* - T^f - \frac{1}{\alpha} \left(\frac{\beta\gamma}{\beta + \gamma}\right) \frac{N}{\hat{s}}, \quad (47)$$

$$\tilde{t}_{\text{toll}} = t^* - T^f, \text{ and} \quad (48)$$

$$\bar{p} = \alpha T^f + \left(\frac{\beta\gamma}{\beta + \gamma}\right) \frac{N}{\hat{s}}. \quad (49)$$

A.3 Toll lanes

Next let us consider the toll lanes. By propositions 28 and 29 we know that

$$r_{\text{toll}}(t) = \begin{cases} \lambda_{\text{toll}} \cdot s^* & \text{for } t \in [t_0, t_e] \\ 0 & \text{otherwise} \end{cases} \quad (50)$$

We can simply read to optimal tolls right off of the equal trip price condition. We find the optimal toll

$$(42) \Rightarrow \tau_{\text{toll}} = \bar{p} - \alpha T^f + \beta(t - t^*) = \left(\frac{\beta\gamma}{\beta + \gamma}\right) \frac{N}{\hat{s}} + \beta(t - t^* + T^f) \quad \text{for } t \in [t_0, \tilde{t}_{\text{toll}}].$$

$$(43) \Rightarrow \tau_{\text{toll}} = \bar{p} - \alpha T^f + \gamma(t^* - t) = \left(\frac{\beta\gamma}{\beta + \gamma}\right) \frac{N}{\hat{s}} - \gamma(t - t^* + T^f) \quad \text{for } t \in (\tilde{t}_{\text{toll}}, t_e]$$

Therefore, letting $a = \left(\frac{\beta\gamma}{\beta + \gamma}\right) \frac{N}{\hat{s}}$, we get

$$\tau_{\text{toll}}(t) = \begin{cases} a + \beta (t - t^* + T^f) & \text{for } t \in [t_0, \tilde{t}_{\text{toll}}], \\ a - \gamma (t - t^* + T^f) & \text{for } t \in (\tilde{t}_{\text{toll}}, t_e], \\ 0 & \text{otherwise.} \end{cases} \quad (51)$$

A.4 Free lanes

Now we turn the free lanes. As with the tolls, can simply read T_{free}^v off of the equal trip price condition. This yields

$$\begin{aligned} (40) \Rightarrow T_{\text{free}}^v(t) &= \frac{\bar{p}}{\alpha - \beta} - T^f + \frac{\beta}{\alpha - \beta} (t - t^*) \\ &= \frac{1}{\alpha - \beta} \left(\frac{\beta\gamma}{\beta + \gamma} \frac{N}{\hat{s}} + \beta (t - t^* + T^f) \right) \quad \text{for } t \in [t_0, \tilde{t}_{\text{free}}]. \end{aligned} \quad (52)$$

$$\begin{aligned} (41) \Rightarrow T_{\text{free}}^v(t) &= \frac{\bar{p}}{\alpha + \gamma} - T^f + \frac{\gamma}{\alpha + \gamma} (t - t^*) \\ &= \frac{1}{\alpha + \gamma} \left(\frac{\beta\gamma}{\beta + \gamma} \frac{N}{\hat{s}} - \gamma (t - t^* + T^f) \right) \quad \text{for } t \in (\tilde{t}_{\text{free}}, t_e]. \end{aligned} \quad (53)$$

We want to use $T^v(t)$ to find $r(t)$ and $Q(t)$. To do so, differentiate (3) to find

$$\frac{\partial T_i^v(t)}{\partial t} = \frac{\partial Q_i(t)}{\partial t} (\lambda_i s)^{-1}$$

and substituting in (1) for $\frac{\partial Q_i(t)}{\partial t}$ gives us

$$\frac{\partial T_i^v(t)}{\partial t} = (r_i(t) - \lambda_i s) (\lambda_i s)^{-1}. \quad (54)$$

Differentiating (52) yields

$$\frac{\partial T_{\text{free}}^v(t)}{\partial t} = \frac{\beta}{\alpha - \beta}. \quad (55)$$

Setting (54) and (55) equal yields

$$(r_{\text{free}}(t) - \lambda_{\text{free}} s) (\lambda_{\text{free}} s)^{-1} = \frac{\beta}{\alpha - \beta} \Rightarrow r_{\text{free}}(t) = \frac{\alpha}{\alpha - \beta} \lambda_{\text{free}} s \quad \text{for } t \in [t_0, \tilde{t}_{\text{free}}]. \quad (56)$$

Similarly, using (54) and (53) we can show that

$$r_{\text{free}}(t) = \frac{\alpha}{\alpha + \gamma} \lambda_{\text{free}} s \quad \text{for } t \in (\tilde{t}_{\text{free}}, t_e]. \quad (57)$$

This gives us

$$r_{\text{free}}(t) = \begin{cases} \frac{\alpha}{\alpha - \beta} \lambda_{\text{free}} s & \text{for } t \in [t_0, \tilde{t}_{\text{free}}], \\ \frac{\alpha}{\alpha + \gamma} \lambda_{\text{free}} s & \text{for } t \in (\tilde{t}_{\text{free}}, t_e], \\ 0 & \text{otherwise.} \end{cases} \quad (58)$$

The length of the queue can be found either by integrating (58) or by substituting (52) and (53) into (3). Both methods yield

$$Q_{\text{free}}(t) = \begin{cases} \frac{1}{\alpha - \beta} \left(\frac{\beta\gamma}{\beta + \gamma} \frac{N}{\hat{s}} + \beta (t - t^* + T^f) \right) \lambda_{\text{free}} s & \text{for } t \in [t_0, \tilde{t}_{\text{free}}], \\ \frac{1}{\alpha + \gamma} \left(\frac{\beta\gamma}{\beta + \gamma} \frac{N}{\hat{s}} - \gamma (t - t^* + T^f) \right) \lambda_{\text{free}} s & \text{for } t \in (\tilde{t}_{\text{free}}, t_e] \\ 0 & \text{otherwise.} \end{cases} \quad (59)$$

A.5 Total costs

Arrivals at work are simply \hat{s} for $t \in [t_0 + T^f, t_e + T^f]$, and so we can integrate the schedule delay function against this to find schedule delay costs:

$$\text{SDC}(\lambda_{\text{toll}}) = \beta \int_{t_0 + T^f}^{t^*} (t^* - t) \hat{s} dt + \gamma \int_{t^*}^{t_e + T^f} (t - t^*) \hat{s} dt = \frac{1}{2} \frac{\beta\gamma}{\beta + \gamma} \frac{N^2}{\hat{s}}. \quad (60)$$

Variable travel time costs are given by

$$\text{VTC}(\lambda_{\text{toll}}) = \alpha \int_{t_0}^{t_e} T_{\text{free}}^v(t) \cdot r(t) dt = \frac{1}{2} \frac{\beta\gamma}{\beta + \gamma} \frac{N^2}{\hat{s}} \frac{\lambda_{\text{free}} \cdot s}{\hat{s}}. \quad (61)$$

Fixed travel time costs are

$$\text{FTC} = \alpha \cdot N \cdot T^f. \quad (62)$$

Toll revenues are

$$\text{TR}(\lambda_{\text{toll}}) = \int_{t_0}^{t_e} \tau_{\text{toll}}(t) \cdot r_{\text{toll}}(t) dt = \frac{1}{2} \frac{\beta\gamma}{\beta + \gamma} \frac{N^2}{\hat{s}} \frac{\lambda_{\text{toll}} \cdot s^*}{\hat{s}}. \quad (63)$$

Notice that the last term in toll revenues and variable travel time is the fraction of total throughput available on the priced or free route, which is the same as the fraction of agents that use that route.

A.6 Changes in social welfare

Total social cost is simply the sum of fixed travel time costs, (62), variable travel time costs, (61), and schedule delay costs, (60);

$$\text{TSC}(\lambda_{\text{toll}}) = \frac{\beta\gamma}{\beta + \gamma} \frac{N^2}{s} \left(\frac{1}{2} + \frac{(1 - \lambda_{\text{toll}})s}{2s} \right) + \alpha NT^f. \quad (64)$$

If we assume $T^f = 0$, then we can solve for the main result of proposition 9.

$$\frac{\text{TSC}(\lambda_{\text{toll}})}{\text{TSC}(0)} = \frac{(1 - \theta)(1 - \theta(1 - \lambda_{\text{toll}}) - \lambda_{\text{toll}}/2)}{(1 - \theta(1 - \lambda_{\text{toll}}))^2}.$$

A.7 Calculating percentage savings from tolling when we know ratio of fixed travel time to variable travel time

When $\lambda_{\text{toll}} = 0$, variable travel time costs, (61), equal schedule delay costs, (60). If x is average peak period variable travel time as a fraction of fixed travel time, then

$$x\text{VTC}(0) = \text{FTC}.$$

Also notice that

$$\frac{\text{SDC}(1)}{\text{SDC}(0)} = \frac{s}{s^*} = 1 - \theta.$$

Finally, recall that tolling ($\lambda_{\text{toll}} = 1$) eliminates variable travel time.

Using these results we can solve for total social costs as a function of variable travel costs.

$$\begin{aligned} \text{TSC}(0) &= \text{FTC} + \text{VTC}(0) + \text{SDC}(0) \\ &= \left(2 + \frac{1}{x}\right) \text{VTC}(0) \\ \text{TSC}(1) &= \text{FTC} + \text{VTC}(1) + \text{SDC}(1) \\ &= \left(1 - \theta + \frac{1}{x}\right) \text{VTC}(0) \end{aligned}$$

We can use these to find the ratio of total social costs when the road is completely priced to when it is not.

$$\begin{aligned} \frac{\text{TSC}(1)}{\text{TSC}(0)} &= \frac{1 - \theta + \frac{1}{x}}{2 + \frac{1}{x}} \\ &= \frac{1 + x + x\theta}{1 + 2x} \end{aligned}$$

Finally, the percentage decrease in social costs as a result of tolling the entire road is:

$$100 \left(1 - \frac{\text{TSC}(1)}{\text{TSC}(0)} \right) = 100 \left(\frac{x(1+\theta)}{1+2x} \right) \quad (65)$$

B Proof of proposition 4 (Existence)

B.1 Overview of proof

Proposition 31. *Under assumptions X a value pricing equilibrium exists.*

This proof is based on Lindsey (2004) proof of existence. It is generalized to deal with multiple routes and tolling, and several mistakes are fixed.

To prove existence I will show there is a vector of trip prices, \mathbf{p} , associated arrival time sets, $\left\{ \mathcal{T}_{g,\text{free}}, \mathcal{T}_{g,\text{toll}} \right\}_{g=1}^G$, and toll schedules $\{\tau_{\text{free}}, \tau_{\text{toll}}\}$ that satisfy individual cost minimization (2.5,2.6), supply of travel times equal demand for travel times (2.8), no toll is charged on free route and toll is charged optimally on tolled route.

I will show that individual cost minimizations holds by showing how to find travel time or tolls so that no agent wants to change his arrival time, and then showing there is a one-to-one mapping between departure times and arrival times. As a result, no agent will want to change his departure time.

To show supply of arrival times equals demand of arrival times (2.8), we will use Kakutani's fixed point theorem. The steps are as follows:

1. Construct a nonempty, compact, and convex set \mathbb{P} .
2. Show that $|\mathcal{T}_{g,r}|$ is convex, compact, and upper hemicontinuous
3. Define an excess demand function, show that it is compact, convex, and upper hemicontinuous
4. Define a correspondence ζ that is nonempty, convex, maps from $\mathbb{P} \rightarrow \mathbb{P}$, upper hemicontinuous, and gives a fixed point when excess demands are zero.
5. Use the Kakutani fixed point theorem to show there exists a $\mathbf{p} \in \mathbb{P}$ such that $\mathbf{p} \in \zeta(\mathbf{p})$.
6. Confirm the departure rate is finite

Define group 0 as a fictitious group that occupies those arrival times in \bar{T} not chosen by any other group, and so has an isocost curve always equal to zero, $\rho_{0,r}(t|p_0) = 0 \forall r$. For this to happen we need to fix $p_0 = 0$, $D_0(t) = 0 \forall t$, and $\alpha_0 > 0$. Let $N_0(p_0) = 0$. Define $\bar{\mathcal{G}} = \mathcal{G} \cup \{0\}$ as the set of groups that includes group 0.

B.2 Individual cost minimization

Using trip price as defined in (4) we can define indifference curves as

$$T_g^v(t|p_g) = \alpha_g^{-1} \left(p_g - D_g \left(t - t_g^* \right) \right), \text{ and}$$

$$\tau_g(t|p_g) = p_g - D_g \left(t - t_g^* \right).$$

By assumption there is no toll on the free route and by Proposition 2 there is no congestion on the tolled route. Then define the upper, nonnegative, envelope of all the groups' isocost curves as

$$T^v(t|\mathbf{p}) = \max \{ T_1^v(t|p_1), \dots, T_G^v(t|p_G), 0 \}, \text{ and}$$

$$\tau(t|\mathbf{p}) = \max \{ \tau_1(t|p_1), \dots, \tau_G(t|p_G), 0 \}.$$

Lemma 1. *If travel times are given by the non-negative upper envelope of isocost curves then there are no profitable deviations in arrival times on an unpriced route. Similarly, if tolls are given by the non-negative upper envelope of isocost curves then there are no profitable deviations in arrival times on a priced route.*

Proof. Consider an arbitrary $t \in \bar{\mathcal{T}}$. On a free route

$$T^v(t|\mathbf{p}) = \max \{ T_1^v(t|p_1), \dots, T_G^v(t|p_G), 0 \} \geq T_g^v(t|p_g)$$

$$\Rightarrow \alpha_g T^v(t|\mathbf{p}) + D_g \left(t - t_g^* \right) \geq \alpha_g T_g^v(t|p_g) + D_g \left(t - t_g^* \right)$$

$$\Rightarrow p_{g,\text{free}}(t) \geq p_g.$$

Similarly on a priced route

$$\tau(t|\mathbf{p}) = \max \{ \tau_1(t|p_1), \dots, \tau_G(t|p_G), 0 \} \geq \tau_g(t|p_g)$$

$$\Rightarrow D_g \left(t - t_g^* \right) + \tau(t|\mathbf{p}) \geq D_g \left(t - t_g^* \right) + \tau_g(t|p_g)$$

$$\Rightarrow p_{g,\text{toll}}(t) \geq p_g.$$

□

Agents will only travel when the travel time or toll lies along the indifference curve implied by their equilibrium trip price; I define the set of demand-feasible arrival times

for group g on route r as

$$\begin{aligned}\bar{\mathcal{T}}_{g,\text{free}}(\mathbf{p}) &= \left\{ t \mid t \in \bar{\mathcal{T}}, T^v(t|\mathbf{p}) = T_g^v(t|p_g) \right\}, \text{ and} \\ \bar{\mathcal{T}}_{g,\text{toll}}(\mathbf{p}) &= \left\{ t \mid t \in \bar{\mathcal{T}}, \tau(t|\mathbf{p}) = \tau_g(t|p_g) \right\}.\end{aligned}$$

By the same algebra as in the lemma above, $t \in \bar{\mathcal{T}}_{g,r} \Rightarrow p_{g,r}(t) = \bar{p}_g$. It is possible that more than one type of agent is willing to travel at a given time, so further define $\mathcal{T}_{g,r} \subseteq \bar{\mathcal{T}}_{g,r}$ as the set of times where agents of type g actually arrive at work using route r .

Lemma 2. *If XXX and travel times are given by the non-negative upper envelope of isocost curves then there are no profitable deviations in departure times.*

Proof. By Lemma 1 no agent wants to change his arrival time, and since by Lemma ?? no agent can change his departure time without changing his arrival time. Therefore, no agent wants to change his departure time. \square

B.3 Supply equals demand

B.3.1 Construct a nonempty, compact, and convex set \mathbb{P} .

Define $\mathbb{P} \equiv \{0\} \times [0, p_1^{\max}], \dots, [0, p_G^{\max}] \subset \mathbb{R}^{G+1}$. By construction \mathbb{P} is nonempty, compact, and convex.

B.3.2 Show that $|\mathcal{T}_{g,r}|$ is convex and compact.

Define $Z \equiv 2^{G+1} \setminus \emptyset$, the power set of $\bar{\mathbf{G}}$ without the empty set. Let $z \in Z$ be a typical element.

Define $\hat{T}_{z,r}(\mathbf{p}) \equiv \bigcap_{g \in z} \bar{\mathcal{T}}_{g,r}(\mathbf{p}) \setminus \left(\bigcup_{g \notin z} \bar{\mathcal{T}}_{g,r}(\mathbf{p}) \right)$, $z \in Z$ as the set valued function that maps from the demand feasible arrival times to the times that groups $g \in z$ are the only groups willing to travel. This gives places where indifference curve overlap or intersect and are the highest indifference curves.

When multiple groups overlap the order which groups arrive and the mass which arrives is indeterminate. Let $f_{g,z,r}$ denote the fraction of $\hat{T}_{z,r}$ occupied by group g , where

$$f_{g,z,r} \geq 0; \quad f_{g,z,r} = 0 \quad \text{if } g \notin z, \quad g \in \bar{\mathbf{G}} \text{ and } r \in \{\text{free, toll}\}; \quad \sum_{g \in z} f_{g,z,r} = 1. \quad (66)$$

Thus the time group g travels on route r , $T_{g,r}$, is when they are the only group with the highest indifference curve, $\hat{T}_{g,r}$, and for $f_{g,z,r}$ of $\hat{T}_{z,r}$, and so the amount of time a group g has on route r is

$$|T_{g,r}| = \sum_{z \in Z} f_{g,z,r} |\hat{T}_{z,r}(\mathbf{p})|.$$

Define the vector of these sets as

$$\mathbf{B}_r(\mathbf{p}) \equiv \{(|T_{0,r}|, \dots, |T_{G,r}|) \mid f_{g,z,r} \text{ satisfies (66)}\}.$$

To avoid needing to write proofs separately for priced and free lanes define

$$\begin{aligned} \rho_r(t|\mathbf{p}) &= \begin{cases} T^v(t|\mathbf{p}) & \text{if } r = \text{free, and} \\ \tau(t|\mathbf{p}) & \text{if } r = \text{toll;} \end{cases} \\ \rho_{g,r}(t|p_g) &= \begin{cases} T_g^v(t|p_g) & \text{if } r = \text{free, and} \\ \tau_g(t|p_g) & \text{if } r = \text{toll;} \end{cases} \quad \text{and} \\ s_r &= \begin{cases} s & \text{if } r = \text{free, and} \\ s^* & \text{if } r = \text{toll.} \end{cases} \end{aligned}$$

Lemma 3. *If D_g is measurable for all $g = 1, \dots, G$ then the set $\tilde{\mathcal{T}}_{g,r}(\mathbf{p})$ is measurable for all \mathbf{p} and for all $g = 1, \dots, G$.*

Proof. Fix an arbitrary \mathbf{p} . Recall that the linear combination of measurable functions is measurable (Royden and Fitzpatrick, 2010, p. 56). By assumption D_g is measurable, so $p_{g,r}$ and $\rho_{g,r}$ are measurable. Since p_r is the maximum of measurable functions, it is measurable (Royden and Fitzpatrick, 2010, p. 58); and so ρ_g is measurable.

Note that $\rho_{g,r} - \rho_r$ is measurable and so for each real number c , the sets $\{t \in \tilde{\mathcal{T}} \mid \rho_{g,r} - \rho_r > c\}$ and $\{t \in \tilde{\mathcal{T}} \mid \rho_{g,r} - \rho_r \geq c\}$ are measurable. Note that

$$\tilde{\mathcal{T}}_{g,r}(\mathbf{p}) = \{t \in \tilde{\mathcal{T}} \mid \rho_r(t|\mathbf{p}) = \rho_{g,r}(t|p_g)\} = \{t \in \tilde{\mathcal{T}} \mid \rho_{g,r} - \rho_r \geq 0\} \setminus \{t \in \tilde{\mathcal{T}} \mid \rho_{g,r} - \rho_r > 0\}$$

and so the set $\tilde{\mathcal{T}}_{g,r}(\mathbf{p})$ is measurable. \square

Lemma 4. *If D_g is measurable for all $g = 1, \dots, G$ then the set $\hat{\mathcal{T}}_{z,r}(\mathbf{p})$ is measurable for all \mathbf{p} and all $z \in \mathbb{Z}$.*

Proof. By lemma 3 $\tilde{\mathcal{T}}_{g,r}(\mathbf{p})$ is measurable for all \mathbf{p} , g , and r ; and since unions, intersections, and complements of measurable sets are measurable, $\hat{\mathcal{T}}_{z,r}(\mathbf{p})$ is measurable. \square

Lemma 5. $\mathbf{B}_r(\mathbf{p})$ is a nonempty convex set. (Alternatively, the image of \mathcal{B}_r is nonempty and convex.)

Proof. Note that the set of $f_{g,z,r}$ defined by 66 is a simplex and so is nonempty and convex. Since $\mathbf{B}_r(\mathbf{p})$ is linear in $f_{g,z,r}$, it is convex.

Since the set of $f_{g,z,r}$ is nonempty, $\mathbf{B}_r(\mathbf{p})$ is nonempty. \square

B.3.3 Show that $|\mathcal{T}_{g,r}|$ upper hemicontinuous.

I will suppress the dependence of $\bar{T}_{g,r}$, on \mathbf{p} .

Let $C_{g,r} = \lim_{k \rightarrow \infty} \bar{T}_{g,r}^k$ under the discrete metric. This means that $\limsup_{k \rightarrow \infty} \bar{T}_{g,r}^k = \liminf_{k \rightarrow \infty} \bar{T}_{g,r}^k = C_{g,r}$. A point is in the \liminf if it occurs for all but finitely many k , and a point is in the \limsup if it occurs for infinitely many k . Formally,

$$\liminf_{k \rightarrow \infty} \bar{T}_{g,r}^k = \bigcup_{i=1}^{\infty} \bigcap_{j=1}^{\infty} \bar{T}_{g,r}^k = \left\{ t \mid \exists k^* \in \mathbb{Z} \text{ such that for } k > k^*, t \in \bar{T}_{g,r}^k \right\}, \text{ and}$$

$$\limsup_{k \rightarrow \infty} \bar{T}_{g,r}^k = \bigcap_{i=1}^{\infty} \bigcup_{j=1}^{\infty} \bar{T}_{g,r}^k = \left\{ t \mid \exists \text{ subsequence for which } t \in \bar{T}_{g,r}^k \text{ for all } k \right\}.$$

Lemma 6. *There is a subsequence of \mathbf{p}^k such all $t \in \bar{T}$ belongs to at least one $C_{g,r}$ for all routes.*

Proof. The proof proceeds in three steps, first we construct our subsequence, second we show that for all $t \in \bar{T}$, $t \in \liminf_{k \rightarrow \infty} \bar{T}_{g,r}^k$ for some $g \in \bar{\mathbb{G}}$, and finally we show that $t \in \limsup_{k \rightarrow \infty} \bar{T}_{g,r}^k$ implies $t \in \liminf_{k \rightarrow \infty} \bar{T}_{g,r}^k$. Then we can conclude for all $t \in \bar{T}$, $t \in \lim_{k \rightarrow \infty} \bar{T}_{g,r}^k$ for some $g \in \bar{\mathbb{G}}$,

Step 1: Construct a subsequence of \mathbf{p}^k such that $p_i^k - p_j^k$ and $\alpha_i^{-1} p_i^k - \alpha_j^{-1} p_j^k$ are monotone for each pair of groups. We can iteratively construct this sequence because every sequence has a monotone subsequence and there are a finite number of pairs of groups and routes.

Step 2: Let $i \succeq_t j$ mean that there exists a $k^* \in \mathbb{Z}$ such that $k > k^* \Rightarrow \rho_{i,r}^k(t) \geq \rho_{j,r}^k(t)$. Consider an arbitrary $t \in \bar{T}$ and two arbitrary groups, $i, j \in \bar{\mathbb{G}}$. Without loss of generality let $p_i^k - p_j^k$ be monotone increasing if we are considering the tolled route, or let $\alpha_i^{-1} p_i^k - \alpha_j^{-1} p_j^k$ be monotone increasing if we are considering the free route. This implies that if there exists a k^* such that $\rho_{i,r}^{k^*}(t) > \rho_{j,r}^{k^*}(t)$ then $\rho_{i,r}^k(t) > \rho_{j,r}^k(t)$ for all $k > k^*$, and so $i \succeq_t j$. If such a k^* does not exist then $\rho_{i,r}^k(t) \leq \rho_{j,r}^k(t)$ for all k , i.e. $k^* = 1$, and so $j \succeq_t i$. Thus every pair of groups is comparable.

Notice that $i \succeq_t i$ and so the comparison is reflexive.

Also note that these comparisons are transitive, meaning $h \succeq_t i$ and $i \succeq_t j$ implies $h \succeq_t j$, since if there exists a k_1^* such that $\rho_{h,r}^k(t) \geq \rho_{i,r}^k(t)$ for all $k > k_1^*$ and a k_2^* such that $\rho_{i,r}^k(t) \geq \rho_{j,r}^k(t)$ for all $k > k_2^*$, then for $k > k^* = \max\{k_1^*, k_2^*\}$, $\rho_{h,r}^k(t) \geq \rho_{i,r}^k(t) \geq \rho_{j,r}^k(t)$.

Since we have a finite number of groups and a binary relation \succeq_t on that set of groups that satisfies reflexivity, comparability, and transitivity, there exists a group g for which there exists a k^* such that for all $k > k^*$, $\rho_{g,r}^k(t) \geq \rho_{i,r}^k(t) \forall i \in \bar{\mathbb{G}}$ and so $t \in \liminf_{k \rightarrow \infty} \bar{T}_{g,r}(\mathbf{p}^k)$. Since this holds for an arbitrary t , it holds for all $t \in \bar{T}$.

Step 3: If $t \in \limsup_{k \rightarrow \infty} \bar{T}_{g,r}(\mathbf{p}^k)$ under the discrete metric then there is a subsequence such that $\rho_{g,r}(t | p_g^k) = \rho_r(t | \mathbf{p}^k)$ for infinitely many values of k . Suppose by way

of contradiction that there is a $t \in \limsup_{k \rightarrow \infty} \bar{T}_{g,r}(\mathbf{p}^k)$ such that $t \notin \liminf_{k \rightarrow \infty} \bar{T}_{g,r}(\mathbf{p}^k)$. Let $t \in \liminf_{k \rightarrow \infty} \bar{T}_{h,r}(\mathbf{p}^k)$, and so for all but finitely many k , $\rho_{h,r}(t|p_h^k) = \rho_r(t|\mathbf{p}^k)$. This implies there are infinitely many k such that $\rho_{g,r}(t|p_g^k) = \rho_{h,r}(t|p_h^k)$. However, monotonicity of $p_g^k - p_h^k$ and $\alpha_g^{-1}p_g^k - \alpha_h^{-1}p_h^k$ means that $\rho_{g,r}(t|p_g^k) = \rho_{h,r}(t|p_h^k)$ for $k = k_1$ and $k = k_2$ implies that $\rho_{g,r}(t|p_g^k) = \rho_{h,r}(t|p_h^k)$ for $k \in [k_1, k_2]$, as a result $\rho_{g,r}(t|p_g^k) = \rho_{h,r}(t|p_h^k) = \rho_r(t|\mathbf{p}^k)$ for all but finitely many k and so $t \in \liminf_{k \rightarrow \infty} \bar{T}_{g,r}(\mathbf{p}^k)$.

Thus $\limsup_{k \rightarrow \infty} \bar{T}_{g,r}(\mathbf{p}^k) = \liminf_{k \rightarrow \infty} \bar{T}_{g,r}(\mathbf{p}^k)$, and so $C_{g,r} = \lim_{k \rightarrow \infty} \bar{T}_{g,r}(\mathbf{p}^k)$. \square

Lemma 7. $t \in C_{g,r} \Rightarrow t \in \bar{T}_{g,r}(\mathbf{p}^\infty)$.

Proof. Note that since $\rho_r(t|\mathbf{p})$ is continuous in \mathbf{p} , $\rho_r(t|\mathbf{p}^k) \rightarrow \rho_r(t|\mathbf{p}^\infty)$.

$$\begin{aligned}
t \in C_{g,r} &\Rightarrow \exists k^* \in \mathbb{Z} \text{ such that for } k > k^*, \rho_{g,r}(t|p_g^k) = \rho_r(t|\mathbf{p}^k) \\
&\Rightarrow \text{for } k > k^*, \alpha_g^{-1}(p_g^k - D_g(t - t_g^*)) = \rho_r(t|\mathbf{p}^k) \\
&\Rightarrow \text{for } k > k^*, P_g^k = \alpha_g \cdot \rho_r(t|\mathbf{p}^k) + D_g(t - t_g^*) \\
&\Rightarrow P_g^k \rightarrow \alpha_g \cdot \rho_r(t|\mathbf{p}^\infty) + D_g(t - t_g^*) \\
&\Rightarrow P_g^\infty = \alpha_g \cdot \rho_r(t|\mathbf{p}^\infty) + D_g(t - t_g^*) \\
&\Rightarrow \alpha_g^{-1}(P_g^\infty - D_g(t - t_g^*)) = \rho_r(t|\mathbf{p}^\infty) \\
&\Rightarrow \rho_{g,r}(t|P_g^\infty) = \rho_r(t|\mathbf{p}^\infty) \\
&\Rightarrow t \in \bar{T}_{g,r}(\mathbf{p}^\infty)
\end{aligned}$$

\square

Definition 1. $\hat{C}_{z,r} \equiv \bigcap_{g \in z} C_{g,r} \setminus \left(\bigcup_{g \notin z} C_{g,r} \right) = \{t | t \in C_{g,r} \forall g \in z \text{ and } t \notin C_{g,r} \forall g \notin z\}$.

Lemma 8. $\lim_{k \rightarrow \infty} \hat{T}_{z,r} = \hat{C}_{z,r}$.

Proof. I find the lim inf and the lim sup and show that they both equal $\hat{C}_{z,r}$.

$$\begin{aligned}
\liminf_{k \rightarrow \infty} \hat{T}_{z,r}^k &= \bigcup_{i=1}^{\infty} \bigcap_{j=i}^{\infty} \left(\bigcap_{g \in z} \bar{T}_g^k \setminus \bigcup_{g \notin z} \bar{T}_g^k \right) \\
&= \bigcup_{i=1}^{\infty} \bigcap_{j=i}^{\infty} \left(\bigcap_{g \in z} \bar{T}_g^k \cap \bigcap_{g \notin z} (\bar{T}_g^k)^c \right) \\
&= \bigcap_{g \in z} \left(\bigcup_{i=1}^{\infty} \bigcap_{j=i}^{\infty} \bar{T}_g^k \right) \cap \bigcap_{g \notin z} \left(\bigcup_{i=1}^{\infty} \bigcap_{j=i}^{\infty} (\bar{T}_g^k)^c \right) \\
&= \bigcap_{g \in z} C_{g,r} \cap \bigcap_{g \notin z} \left(\bigcap_{i=1}^{\infty} \bigcup_{j=i}^{\infty} \bar{T}_g^k \right)^c \\
&= \bigcap_{g \in z} C_{g,r} \cap \bigcap_{g \notin z} C_{g,r}^c \\
&= \bigcap_{g \in z} C_{g,r} \setminus \bigcup_{g \notin z} C_{g,r} \\
&= \hat{C}_{g,r}
\end{aligned}$$

$$\begin{aligned}
\limsup_{k \rightarrow \infty} \hat{T}_{z,r}^k &= \bigcap_{i=1}^{\infty} \bigcup_{j=i}^{\infty} \left(\bigcap_{g \in z} \bar{T}_g^k \setminus \bigcup_{g \notin z} \bar{T}_g^k \right) \\
&= \bigcap_{i=1}^{\infty} \bigcup_{j=i}^{\infty} \left(\bigcap_{g \in z} \bar{T}_g^k \cap \bigcap_{g \notin z} (\bar{T}_g^k)^c \right) \\
&= \bigcap_{g \in z} \left(\bigcap_{i=1}^{\infty} \bigcup_{j=i}^{\infty} \bar{T}_g^k \right) \cap \bigcap_{g \notin z} \left(\bigcap_{i=1}^{\infty} \bigcup_{j=i}^{\infty} (\bar{T}_g^k)^c \right) \\
&= \bigcap_{g \in z} C_{g,r} \cap \bigcap_{g \notin z} \left(\bigcup_{i=1}^{\infty} \bigcap_{j=i}^{\infty} \bar{T}_g^k \right)^c \\
&= \bigcap_{g \in z} C_{g,r} \cap \bigcap_{g \notin z} C_{g,r}^c \\
&= \bigcap_{g \in z} C_{g,r} \setminus \bigcup_{g \notin z} C_{g,r} \\
&= \hat{C}_{g,r}
\end{aligned}$$

Since $\limsup_{k \rightarrow \infty} \hat{T}_{z,r}^k = \liminf_{k \rightarrow \infty} \hat{T}_{z,r}^k = \hat{C}_{g,r}$, $\lim_{k \rightarrow \infty} \hat{T}_{z,r} = \hat{C}_{z,r}$. \square

Lemma 9. $\hat{T}_{z,r}(\mathbf{p}^\infty) \cap \hat{C}_{z,r} \neq \emptyset \Rightarrow \hat{T}_{z,r}(\mathbf{p}^\infty) \subset \hat{C}_{z,r}$. Alternative statement, if $t_1 \in \hat{T}_{z,r}(\mathbf{p}^\infty) \cap$

$\hat{C}_{\tilde{z},r}$ then $t_2 \in \hat{T}_{z,r}(\mathbf{p}^\infty) \Rightarrow t_2 \in \hat{C}_{\tilde{z},r}$.

Proof. Let $t_1 \in \hat{T}_{z,r}(\mathbf{p}^\infty) \cap \hat{C}_{\tilde{z},r}$, which exists by assumption, and let t_2 be an another arbitrary point such that $t_2 \in \hat{T}_{z,r}(\mathbf{p}^\infty)$. First note that $t_1 \in \hat{C}_{\tilde{z},r} \Rightarrow t_1 \in C_{g,r}$ for all $g \in \tilde{z}$, which by lemma 7 implies $t_1 \in \bar{T}_g(\mathbf{p}^\infty)$ for all $g \in \tilde{z}$, and so $\tilde{z} \subset z$.

By lemma 6 $t_2 \in C_{h,r}$ for at least one $h \in \bar{G}$. By lemma 7 $t_2 \in \bar{T}_{h,r}(\mathbf{p}^\infty)$ and so $t_2 \in z$, which means $t_1 \in \bar{T}_{h,r}(\mathbf{p}^\infty)$. Observe that

$$\begin{aligned} t \in \bar{T}_{h,r}(\mathbf{p}^\infty) &\Rightarrow \rho_{h,r}(t|p_h^\infty) = \rho_{g,r}(t|p_g^\infty) \forall g \in z \\ &\Rightarrow p_h^\infty - D_h(t - t_h^*) = p_g^\infty - D_g(t - t_g^*) \\ &\Rightarrow p_h^\infty - p_g^\infty = D_h(t - t_h^*) - D_g(t - t_g^*) \end{aligned}$$

and so $D_h(t_1 - t_h^*) - D_g(t_1 - t_h^*) = D_h(t_2 - t_h^*) - D_g(t_2 - t_h^*) \forall h, g \in z$.

Now observe that $t_2 \in C_{h,r}$ means that $\rho_{h,r}(t_2|p_h^k) \geq \rho_{g,r}(t_2|p_g^k) \forall g \in \bar{G}$ for all but finite k , which means $p_h^k - p_g^k \geq D_h(t_2 - t_h^*) - D_g(t_2 - t_g^*) \forall g \in \bar{G}$ for all but finite k . Likewise for all $i \in \tilde{z}$, $t_1 \in C_{i,r}$ implies $p_i^k - p_g^k \geq D_i(t_1 - t_i^*) - D_g(t_1 - t_g^*) \forall g \in \bar{G}$ for all but finite k . Since $i \in z$, this implies

$$\begin{aligned} D_h(t_1 - t_h^*) - D_i(t_1 - t_g^*) &\geq p_h^k - p_i^k \geq D_h(t_1 - t_h^*) - D_i(t_1 - t_g^*) \text{ for all but finite } k, \\ \Rightarrow p_h^k - p_i^k &= D_h(t_1 - t_h^*) - D_i(t_1 - t_g^*) \text{ for all but finite } k, \\ \Rightarrow \rho_{h,r}(t|p_h^k) &= \rho_{i,r}(t|p_i^k) = \rho_r(t|\mathbf{p}^k) \text{ for all but finite } k, \\ \Rightarrow t_1 &\in C_{h,r} \\ \Rightarrow h &\in \tilde{z}. \end{aligned}$$

Similarly, it can be shown $t_2 \in C_{i,r}$. Since for $t_1, t_2 \in \hat{T}_{z,r}(\mathbf{p}^\infty)$, $t_1 \in C_{g,r} \Leftrightarrow t_2 \in C_{g,r} \forall g$, then $t_1 \in C_{g,r} \not\Leftrightarrow t_2 \in C_{g,r} \forall g$, and so $t_1 \in \hat{C}_{z,r} \Leftrightarrow t_2 \in \hat{C}_{z,r}$. \square

Lemma 10. *If D_g is measurable for all $g = 1, \dots, G$ then $\lim_{k \rightarrow \infty} |\hat{T}_{z,r}| = |\hat{C}_{z,r}|$ for all $z \in \mathcal{Z}$.*

Proof. By lemma 4 the set $\tilde{T}_{z,r}(\mathbf{p})$ is measurable. Define

$$\mathbf{1}_x(t) \equiv \begin{cases} 1 & t \in x \\ 0 & t \notin x \end{cases}.$$

Note that $|\hat{T}_{z,r}^k| \leq |\bar{T}| \forall k$ and that $\mathbf{1}_{\lim_{k \rightarrow \infty} \hat{T}_{z,r}^k} = \lim_{k \rightarrow \infty} \mathbf{1}_{\hat{T}_{z,r}^k}$. (See a proof in solutions to Schilling, René L. 2006). Also note that $\mathbf{1}_x$ is measurable if x is a measurable set. Then by

Lebesgue's dominated convergence theorem

$$\lim_{k \rightarrow \infty} |\hat{T}_{z,r}| = \lim_{k \rightarrow \infty} \int \mathbf{1}_{\hat{T}_{z,r}^k} du = \int \lim_{k \rightarrow \infty} \mathbf{1}_{\hat{T}_{z,r}^k} du = \int \mathbf{1}_{\lim_{k \rightarrow \infty} \hat{T}_{z,r}^k} du = \left| \lim_{k \rightarrow \infty} \hat{T}_{z,r} \right| = |\hat{C}_{z,r}|.$$

□

Lemma 11. *If D_g is measurable for all $g = 1, \dots, G$ then $\mathbf{B}_r(\mathbf{p}) = (|T_{0,r}(\mathbf{p})|, \dots, |T_{G,r}(\mathbf{p})|)$ is upper hemicontinuous for $r \in \{\text{free, toll}\}$.*

Proof. To show $\mathbf{B}_r(\mathbf{p})$ is upper hemicontinuous it is necessary and sufficient to show that for any convergent sequence $\mathbf{p}^k \in \bar{p}$, $k = 1, 2, \dots$ such that $\mathbf{p}^k \rightarrow \mathbf{p}^\infty$, there is a convergent subsequence of $\{\mathbf{y}^k\}$, $\mathbf{y}^k = \mathbf{B}_r(\mathbf{p}^k) \in [0, \lambda_r \cdot s_r |\bar{T}|]^{G+1}$, that converges to a point $\mathbf{y}^\infty \in \mathbf{B}_r(\mathbf{p}^\infty)$. Since $[0, \lambda_r \cdot s_r |\bar{T}|]^{G+1}$ is compact, there exists a convergent subsequence of $\{\mathbf{y}^k\}$. By lemma 6 there is a subsequence of this subsequence such that all $t \in \bar{T}$ are in at least one $C_{g,r}$, $g \in \bar{G}$. Let us consider this subsequence.

Recall that $|T_{g,r}^k| = \sum_{z \in Z} f_{g,z,r}^k |\hat{T}_{z,r}^k|$. Since $f_{g,z,r}^k$ is in $[0, 1]$ a compact space, it has a convergent subsequence. Choose the subsequence such that $\{f_{g,z,r}^k\}$ converges for all $g \in \bar{G}$, $z \in Z$, and $r \in \{\text{free, toll}\}$. We can do so since the set of groups, routes, and Z are finite. Let $f_{g,z,r}^\infty = \lim_{k \rightarrow \infty} f_{g,z,r}^k$. Because of the set of $f_{g,z,r}$ is closed, $f_{g,z,r}^\infty$ belongs to that set. Now by lemma 10

$$y_g^\infty = \lim_{k \rightarrow \infty} |T_{g,r}(\mathbf{p}^k)| = \sum_{z \in Z} \lim_{k \rightarrow \infty} f_{g,z,r}^k \cdot \lim_{k \rightarrow \infty} |\hat{T}_{z,r}^k| = \sum_{z \in Z} f_{g,z,r}^\infty |\hat{C}_{z,r}|.$$

Now we will show that there exists weights $f'_{g,z,r}$ such that

$$y_g^\infty = \sum_{z \in Z} f'_{g,z,r} |\hat{T}_{z,r}(\mathbf{p}^\infty)| \in |T_{g,r}(\mathbf{p}^\infty)| \quad \forall g \in \bar{G}.$$

Choose $f'_{g,z,r}$ as follows:

1. For each $z \in Z$ and $r \in \{\text{free, toll}\}$ pick an arbitrary $t \in \hat{T}_{z,r}(\mathbf{p}^\infty)$.
2. Find the $\tilde{z} \in Z$ such that $t \in \hat{C}_{\tilde{z},r}$.
3. Set $f'_{g,z,r} = f_{g,\tilde{z},r}^\infty$ for all $g \in \bar{G}$.

Since the $f_{g,\tilde{z},r}^\infty$ meet the requirements of (66), so do the $f'_{g,z,r}$. Furthermore, since $\bigcup_{z \in Z} (\hat{T}_{z,r}(\mathbf{p}^\infty) \cap \hat{C}_{\tilde{z},r}) = \bar{T} \cap \hat{C}_{\tilde{z},r} = \hat{C}_{\tilde{z},r}$, we know that if we define $Z_{\tilde{z}} = \{z | \hat{T}_{z,r}(\mathbf{p}^\infty) \cap \hat{C}_{\tilde{z},r} \neq \emptyset\}$ then since the

$\hat{T}_{z,r}$ are disjoint, $\sum_{z \in Z_z} |\hat{T}_{z,r}(\mathbf{p}^\infty)| = |\hat{C}_{z,r}|$. Then for all $g \in \mathbf{G}$,

$$\begin{aligned} \sum_{z \in Z} f'_{g,z,r} |\hat{T}_{z,r}(\mathbf{p}^\infty)| &= \sum_{z \in Z} f_{g,\bar{z},r}^\infty |\hat{T}_{z,r}(\mathbf{p}^\infty)| \\ &= \sum_{\bar{z} \in Z} \sum_{z \in Z_z} f_{g,\bar{z},r}^\infty |\hat{T}_{z,r}(\mathbf{p}^\infty)| \\ &= \sum_{\bar{z} \in Z} f_{g,\bar{z},r}^\infty |\hat{C}_{\bar{z},r}| \\ &= y_g^\infty. \end{aligned}$$

Since this holds for all $g \in \mathbf{G}$, $\mathbf{y}^\infty \in \mathbf{B}_r(\mathbf{p}^\infty)$ and so $\mathbf{B}_r(\mathbf{p})$ is upper hemicontinuous. \square

B.3.4 Define an excess demand function, \mathbf{E} , show that it is upper hemicontinuous and the set $\mathbf{E}(\mathbf{p})$ is compact and convex for all $\mathbf{p} \in \mathbb{P}$.

The excess demand of group g is

$$E_g(\mathbf{p}) \equiv N_g(p_g) - \left((1 - \lambda_{\text{toll}}) s \left| T_{g,\text{free}}(p_g) \right| + \lambda_{\text{toll}} \cdot s^* \left| T_{g,\text{toll}}(p_g) \right| \right). \quad (67)$$

This is the mass of agents that would like to travel at price p_g but are unable to. Let $\mathbf{E} = (E_0, \dots, E_g) = \mathbf{N} - ((1 - \lambda_{\text{toll}}) s \mathbf{B}_{\text{free}} + \lambda_{\text{toll}} \cdot s^* \mathbf{B}_{\text{toll}})$, $\mathbf{E} : \mathbb{P} \rightarrow \mathbb{R}$.

Lemma 12. *If demand, \mathbf{N} , is continuous and the measure of travel time each group receives on each routes, \mathbf{B}_{free} and \mathbf{B}_{toll} , is upper hemicontinuous, then the excess demand function \mathbf{E} is upper hemicontinuous.*

Proof. Since \mathbf{N} is a continuous function, it is also an upper hemicontinuous correspondence. Since \mathbf{E} is the linear combination of upper hemicontinuous functions it is upper hemicontinuous. \square

Lemma 13. *If demand, \mathbf{N} , is continuous and defined over \mathbb{P} , and $\mathbf{B}_r(\mathbf{p})$ is a nonempty, convex set for all $\mathbf{p} \in \mathbb{P}$, then the set $\mathbf{E}(\mathbf{p})$ is nonempty and convex for all $\mathbf{p} \in \mathbb{P}$.*

Proof. Since \mathbf{N} is single valued the set $\mathbf{N}(\mathbf{p})$ is nonempty and convex for all $\mathbf{p} \in \mathbb{P}$.

Since $\mathbf{N}(\mathbf{p})$, $\mathbf{B}_{\text{free}}(\mathbf{p})$, and $\mathbf{B}_{\text{toll}}(\mathbf{p})$ are nonempty, $\mathbf{E}(\mathbf{p})$ is nonempty.

The set $\mathbf{E}(\mathbf{p})$ is convex because \mathbf{E} is linear in \mathbf{N} , \mathbf{B}_{free} , and \mathbf{B}_{toll} , and $\mathbf{N}(\mathbf{p})$, $\mathbf{B}_{\text{free}}(\mathbf{p})$, and $\mathbf{B}_{\text{toll}}(\mathbf{p})$ are convex sets. \square

B.3.5 Define a correspondence $\zeta : \mathbb{P} \rightarrow \mathbb{P}$ that is upper hemicontinuous from \mathbb{P} into itself with the property that the set $\zeta(\mathbf{p})$ is nonempty and convex for all $\mathbf{p} \in \mathbb{P}$. Finally ζ has a fixed point only if excess demands are zero.

We need to define a correspondence ζ with the following properties,

1. Maps into itself; $\zeta : \mathbb{P} \rightarrow \mathbb{P}$.
2. Upper hemicontinuous.
3. Has a nonempty, convex image; $\zeta(\mathbf{p})$ is a non,empty convex set.
4. Has a fixed point when excess demands are zero; $\mathbf{0} \in \mathbf{E}(\mathbf{p}) \Leftrightarrow \mathbf{p} \in \zeta(\mathbf{p})$.

The basic idea is that this correspondence is an adjustment mechasism that could be used to find equilibrium. When excess demand is positive it will raise the price a group is willing to pay in order to obtain more time for the group, and similarly when excess demand is negative it will lower the price a group is willing to pay. Define

$$f_g(x) = p_g + \begin{cases} \frac{x}{N_g(0)} (p_g^{\max} - p_g) & \text{if } x > 0 \\ \frac{x}{(\lambda_{\text{free}S} + \lambda_{\text{toll}S^*}) |\bar{T}|} p_g & \text{if } x \leq 0 \end{cases}$$

I define $\zeta_0(\mathbf{p}) \equiv 0$, and

$$\sim_g(\mathbf{p}) \equiv f_g \circ \mathcal{E}_g = \{f(x) \mid x \in \mathcal{E}_g(\mathbf{p})\}, \quad 1, \dots, G.$$

Lemma 14. ζ maps into itself, that is, the image of ζ is a subset of its domain.

Proof. Consider an arbitrary $x \in \mathcal{E}_g(\mathbf{p})$. If $x \geq 0$ then $x \in [0, N_g(0)]$ and so $x/N_g(0) \in [0, 1]$, which implies $\sim_g \subset [p_g, p_g^{\max}]$. If $x \leq 0$ then $x \in [-(\lambda_{\text{free}S} + \lambda_{\text{toll}S^*}) |\bar{T}|, 0]$ and so $x/((\lambda_{\text{free}S} + \lambda_{\text{toll}S^*}) |\bar{T}|) \in [-1, 0]$, which implies $\sim_g \subset [0, p_g]$.

So if $p_g \in [0, p_g^{\max}]$ then $\zeta_g \subset [0, p_g^{\max}]$ for all $g = 1, \dots, G$, and since $\zeta_0(\mathbf{p}) \equiv 0$, $\zeta(\mathbf{p}) \subset \mathbb{P}$ for all $\mathbf{p} \in \mathbb{P}$. \square

Lemma 15. If \mathbf{E} is upper hemicontinuous then ζ is upper hemicontinuous.

Proof. Since f is piecewise linear I just need to show $\lim_{x \rightarrow 0^+} f(x) = \lim_{x \rightarrow 0^-} f(x)$ to prove f is continuous. Evaluating these limits I find

$$\begin{aligned} \lim_{x \rightarrow 0^+} f(x) &= \lim_{x \rightarrow 0^+} p_g + \frac{x}{N_g(0)} (p_g^{\max} - p_g) && = p_g, \text{ and} \\ \lim_{x \rightarrow 0^-} f(x) &= \lim_{x \rightarrow 0^-} p_g + \frac{x}{(\lambda_{\text{free}S} + \lambda_{\text{toll}S^*}) |\bar{T}|} p_g && = p_g. \end{aligned}$$

Since f is continuous it is upper hemicontinuous. Since \mathbf{E} is upper hemicontinuous, \mathcal{E}_g is as well. The composition of two upper hemicontinuous functions is upper hemicontinuous (? , p. 48), and so $\zeta_g = f_g \circ \mathcal{E}_g$ is upper hemicontinuous. Thus $\zeta = \{0\} \times f_g \circ \mathcal{E}_1 \times \dots \times f_G \circ \mathcal{E}_G$ is the Cartesian product of upper hemicontinuous functions and so is upper hemicontinuous ?, p. 47. \square

Lemma 16. *If the set $\mathbf{E}(\mathbf{p})$ is nonempty, compact, and convex for all $\mathbf{p} \in \mathbb{P}$ then the set $\zeta(\mathbf{p})$ is nonempty, compact, and convex for all $\mathbf{p} \in \mathbb{P}$.*

Proof. Let e be an element of $\mathbf{E}(\mathbf{p})$. Such an element exists since $\mathbf{E}(\mathbf{p})$ is nonempty. Then $f(e) \in \zeta(\mathbf{p})$ and so $\zeta(\mathbf{p})$ is nonempty.

Consider an arbitrary group g . Because $\mathcal{E}_g(\mathbf{p})$ is convex and compact it is a closed interval in \mathbb{R} . Since $\mathcal{E}_g(\mathbf{p})$ a closed interval in \mathbb{R} and f_g is continuous, by the intermediate value theorem $f_g(\mathcal{E}_g(\mathbf{p}))$ is a closed interval and so is convex and compact. Since this holds for an arbitrary group, it holds for all.

Therefore $\zeta(\mathbf{E}(\mathbf{p})) = \{0\} \times f_1(\mathcal{E}_1(\mathbf{p})) \times \dots \times f_G(\mathcal{E}_G(\mathbf{p}))$ is the Cartesian product of convex and compact sets and so is convex and compact. \square

Lemma 17. *If $N_g(0) > 0$ and $N_g(p_g^{\max}) = 0$ for all $g = 1, \dots, G$; then $\zeta(\mathbf{p})$ has a fixed point if and only if $\mathbf{0} \in \mathbf{E}_{-0}(\mathbf{p})$.*

Proof. If $\mathbf{p} \in \zeta(\mathbf{p})$ then for all $g = 1, \dots, G$

$$p_g = p_g + \begin{cases} \frac{x}{N_g(0)} (p_g^{\max} - p_g) & \text{if } x > 0 \\ \frac{x}{(\lambda_{\text{free}}^s + \lambda_{\text{toll}}^{s^*}) |\bar{\mathcal{T}}|} p_g & \text{if } x \leq 0 \end{cases} \quad \text{for some } x \in \mathcal{E}_g(\mathbf{p})$$

and so either $\frac{x}{N_g(0)} (p_g^{\max} - p_g) = 0$ and $x > 0$, or $\frac{x}{(\lambda_{\text{free}}^s + \lambda_{\text{toll}}^{s^*}) |\bar{\mathcal{T}}|} p_g = 0$ and $x \leq 0$. This is true only if $x = 0$, $p_g = 0$ and $x < 0$, or $p_g = p_g^{\max}$ and $x > 0$. If $p_g = 0$ then $|T_{g,\text{free}}(p_g)| = |T_{g,\text{free}}(p_g)| = 0$ and by assumption $N_g(p_g) > 0$. This implies $\mathcal{E}_g > 0$ and contradicts $x < 0$. If $p_g = p_g^{\max}$ then by assumption $N_g(p_g) = 0$. Since measures are weakly positive, this implies $\mathcal{E}_g \leq 0$ and contradicts $x > 0$. Therefore $x = 0$, which means $0 \in \mathcal{E}_g(\mathbf{p})$ for all g and so $\mathbf{0} \in \mathbf{E}_{-0}(\mathbf{p})$.

If $\mathbf{0} \in \mathbf{E}_{-0}(\mathbf{p})$ then

$$\zeta(\mathbf{p}) \ni p_g + \frac{0}{(\lambda_{\text{free}}^s + \lambda_{\text{toll}}^{s^*}) |\bar{\mathcal{T}}|} p_g = p_g \quad \forall g = 1, \dots, G$$

and so $\mathbf{p} \in \zeta(\mathbf{p})$. \square

B.3.6 Use the Kakutani fixed point theorem to show there exists a $\mathbf{p} \in \mathbb{P}$ such that $\mathbf{p} \in \zeta(\mathbf{p})$.

I have shown that \mathbb{P} is a nonempty, compact, convex set, and that $\zeta : \mathbb{P} \rightarrow \mathbb{P}$ is an upper hemicontinuous correspondence from \mathbb{P} into itself with the property that the set $\zeta(\mathbf{p}) \subset \mathbb{P}$ is nonempty and convex for all $\mathbf{p} \in \mathbb{P}$. Thus we can invoke Kakutani's fixed point theorem, there is a $\mathbf{p} \in \mathbb{P}$ such that $\mathbf{p} \in \zeta(\mathbf{p})$.

B.3.7 Confirm the departure rate is finite

Lemma 18. *The departure rate is finite.*

Proof. If there is no queuing then the departure rate is finite (i.e. an infinite departure rate implies queuing). Can just cite Lindsey 2004. \square

C Proof of uniqueness

Proposition 32. *Under assumptions X a value pricing equilibrium exists where the trip price for each type, aggregate departure rates, and toll schedules are unique.*

Assume by way of contradiction that there are two distinct equilibrium trip price vectors, \mathbf{p} and \mathbf{p}' , $\mathbf{p} \neq \mathbf{p}'$. Then there must exist some type i for whom $\bar{p}_i \neq \bar{p}'_i$, assume without loss of generality that $\bar{p}_i < \bar{p}'_i$.

Then there are two possibilities for the effect of this increase in trip price on the supply of travel time to type i , either it increases or there is another type j who also has a higher trip price $\bar{p}_j < \bar{p}'_j$.

If there is another equilibrium where I have a higher trip price, then either
I have more arrival time
or someone I border also has a higher trip price
Rinse and repeat.

D Other proofs

Proofs that are in progress are stored in proofs.lyx.

References

- AASHTO (2005). *A policy on design standards—interstate system*. (5th ed.). Washington D.C.: American Association of State Highway and Transportation Officials.
- Arnott, R. (1990, June). Signalized intersection queuing theory and central business district auto congestion. *Economics Letters* 33(2), 197–201.
- Arnott, R., A. de Palma, and R. Lindsey (1990, January). Economics of a bottleneck. *Journal of Urban Economics* 27(1), 111–130.
- 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.
- 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.
- 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.
- 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. (1998, February). Queue spillovers in transportation networks with a route choice. *Transportation science* 32(1), 3–11.
- Daganzo, C. F., M. J. Cassidy, and R. L. Bertini (1999, June). Possible explanations of phase transitions in highway traffic. *Transportation Research Part A: Policy and Practice* 33(5), 365–379.

- Elefteriadou, L., R. P. Roess, and W. R. McShane (1995). Probabilistic nature of breakdown at freeway merge junctions. *Transportation Research Record* 1484, 80–89.
- Flynn, M. R., A. R. Kasimov, J. Nave, R. R. Rosales, and B. Seibold (2009, May). Self-sustained nonlinear waves in traffic flow. *Physical Review E* 79(5), 056113.
- Greenshields, B. D., J. R. Bibbins, W. S. Channing, and H. H. Miller (1935). A study of traffic capacity. *Highway Research Board Proceedings* 14, 448–477.
- Hall, F. and K. Agyemang-Duah (1991). Freeway capacity drop and the definition of capacity. *Transportation Research Record* 1320, 1–98.
- Halvorson, R. and K. R. Buckeye (2006, January). High-Occupancy toll lane innovations: I-394 MnPASS. *Public Works Management & Policy* 10(3), 242–255.
- Helbing, D. and B. A. Huberman (1998, December). Coherent moving states in highway traffic. *Nature* 396(6713), 738–740.
- 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.
- 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.
- Johnson, M. B. (1964, April). On the economics of road congestion. *Econometrica* 32(1/2), 137–150. ArticleType: primary_article / Full publication date: Jan. - Apr., 1964 / Copyright © 1964 The Econometric Society.
- Kerner, B. S. and S. L. Klenov (2002). A microscopic model for phase transitions in traffic flow. *Journal of Physics A: Mathematical and Theoretical* 35(3).
- Kerner, B. S. and H. Rehborn (1997, November). Experimental properties of phase transitions in traffic flow. *Physical Review Letters* 79(20), 4030.
- Kwon, J., M. Mauch, and P. Varaiya (2006, January). Components of congestion: Delay from incidents, special events, lane closures, weather, potential ramp metering gain, and excess demand. *Transportation Research Record* 1959(1), 84–91.
- Leclercq, L., J. A. Laval, and N. Chiabaut (2011). Capacity drops at merges: an endogenous model. *Procedia - Social and Behavioral Sciences* 17, 12–26.
- Levy, J. I., J. J. Buonocore, and K. von Stackelberg (2010). Evaluation of the public health impacts of traffic congestion: a health risk assessment. *Environmental Health* 9(1), 65.

- Light, T. (2009, September). Optimal highway design and user welfare under value pricing. *Journal of Urban Economics* 66(2), 116–124.
- 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. (2004, August). Existence, uniqueness, and trip cost function properties of user equilibrium in the bottleneck model with multiple user classes. *Transportation Science* 38(3), 293–314.
- 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.
- May, A. (1990). *Traffic flow fundamentals*. Englewood Cliffs N.J.: Prentice Hall.
- 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.
- Orosz, G., R. E. Wilson, R. Szalai, and G. Stépan (2009, October). Exciting traffic jams: Nonlinear phenomena behind traffic jam formation on highways. *Physical Review E* 80(4), 046205.
- Perez, B. G. and G. 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. (1920). *The economics of welfare*, (1 ed.). London,: Macmillan and co., ltd.
- Richards, P. I. (1956). Shock waves on the highway. *Operations research* 4(1), 42–51.

- Royden, H. and P. Fitzpatrick (2010, January). *Real Analysis* (4 ed.). Prentice Hall.
- Rudjanakanoknad, J. (2005, May). *Increasing Freeway Merge Capacity Through On-Ramp Metering*. Dissertation, University of California at Berkeley, Berkeley, CA.
- Schrank, D., T. Lomax, and S. Turner (2010). 2010 urban mobility report. Technical report, Texas Transportation Institute, College Station, Texas.
- Schuman, R. (2011, March). INRIX national traffic scorecard: 2010 annual report. Technical report, INRIX, Kirkland, WA.
- Small, K. and E. Verhoef (2007). *The economics of urban transportation*. New York: Routledge.
- Small, K., C. Winston, and J. Yan (2006). Differentiated road pricing, express lanes, and carpools: Exploiting heterogeneous preferences in policy design. *Brookings-Wharton Papers on Urban Affairs*, 53–96.
- Small, K. A. (1982, June). The scheduling of consumer activities: Work trips. *The American Economic Review* 72(3), 467–479.
- Small, K. A. and X. Chu (2003, September). Hypercongestion. *Journal of Transport Economics and Policy* 37(3), 319–352. ArticleType: primary_article / Full publication date: Sep., 2003 / Copyright © 2003 The London School of Economics and Political Science and University of Bath.
- Small, K. A. and J. Yan (2001, March). The value of "Value pricing" of roads: Second-Best pricing and product differentiation. *Journal of Urban Economics* 49(2), 310–336.
- 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. Bureau of Public Roads (1964). *Traffic assignment manual for application with a large, high speed computer*. Washington, D.C.: GPO.
- 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. S. (1969, May). Congestion theory and transport investment. *The American Economic Review* 59(2), 251–260.

- 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.
- Washington State Department of Transportation (2008, November). The gray notebook. Technical Report 31, Washington State Department of Transportation.
- 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.