University of Toronto
Department of Economics

Working Paper 548

Pareto Improvements from Lexus Lanes: The effects of pricing a portion of the lanes on congested highways

By Jonathan D. Hall

October 15, 2015

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

JONATHAN D. HALL
UNIVERSITY OF TORONTO

October 15, 2015

Abstract. This paper shows that a judiciously designed toll applied to a portion of the lanes of a highway can generate a Pareto improvement even before the resulting revenue is spent. I obtain this new result by extending a standard dynamic congestion model to reflect an important additional traffic externality recently identified by transportation engineers: additional traffic does not simply increase travel times, but also introduces frictions that reduce throughput. In particular, I show that as long as some rich drivers use the highway at the peak of rush hour, then adding tolls to a portion of the lanes yields a Pareto improvement. To confirm the relevance of this theoretical possibility in practice, I use survey and travel time data to estimate the joint distribution of driver preferences over arrival time, travel time, and tolls, and use these results to estimate the effects of adding optimal time-varying tolls. I find that adding tolls on up to half of the lanes yields a Pareto improvement, and that the social welfare gains of doing so are substantial—up to $1,740 per road user per year.

Keywords: congestion pricing, value pricing, Pareto improvement

*I am especially grateful for the guidance and support that Gary Becker and Eric Budish have given me. I am also grateful for helpful feedback from Richard Arnott, Ian Fillmore, Mogens Fosgerau, Edward Glaeser, Brent Hickman, William Hubbard, Kory Kroft, Ethan Lieber, Robin Lindsey, Robert McMillan, John Panzar, Devin Pope, Mark Phillips, Allen Sanderson, Ken Small, Chad Syverson, George Tolley, Vincent Van den Berg, Jos Van Ommeren, Clifford Winston, and Glen Weyl, as well as seminar audiences at the University of Chicago, Northwestern University, University of Toronto, Brigham Young University, Clemson University, Technical University of Denmark, Tinbergen Institute, RSAI, Kumho-Nectar, World Bank Conference on Transport and ICT, and the NBER Summer Institute. All remaining errors are my own. *Email:* jonathan.hall@utoronto.ca.

1. Introduction

In the ninety years since Arthur Pigou introduced the idea that tolls could be used to alleviate traffic congestion and increase social welfare, carts have given way to automobiles and congestion has grown to consume 42 hours per commuter annually in the United States—nearly an entire work week (Schrank, Eisele, Lomax, and Bak, 2015). In addition to the 6.9 billion hours drivers lost to additional travel time in 2014, congestion wasted 3.1 billion gallons of fuel (Schrank, Eisele, Lomax, and Bak, 2015), releasing an additional 28 million metric tons of carbon dioxide into the atmosphere, as well as a host of other pollutants. This additional pollution amounts to more than six times the annual emissions saved by the current fleet of hybrid and electric vehicles,[1] and is responsible for up to 8,600 preterm births a year (Currie and Walker, 2011). Congestion also retards economic growth; cutting congestion delay in half would raise employment growth by an estimated 1 percent per year (Hymel, 2009).

Despite the significance of these costs, the vast majority of roads remain unpriced. A major barrier to implementing congestion pricing is the received wisdom among economists, policy makers, and the public that it would make many, if not most, road users worse off.[2] That is, under the standard view, congestion pricing generates a Kaldor-Hicks improvement, meaning the winners gain more than the losers lose, but not a Pareto improvement that helps all road users.

Since, under this view, congestion pricing generates a Kaldor-Hicks improvement, there exists a set of transfers we could implement that would turn it into a Pareto improvement; however, there are at least two problems with doing so. First, it is difficult to target the transfers precisely enough to actually make all road users better off.[3] Second, even when we can design transfers that make a policy Pareto-improving, they can still be difficult to implement. As Stiglitz (1998) points out, it may not be enough to identify such transfers because the transfers are transparent and thus harder to defend than the implicit transfers the status quo entails; further, the government cannot commit to maintaining the transfers. This makes policies that naturally generate a Pareto improvement all the more valuable.

The main result of this paper is that, contrary to the received wisdom, a carefully designed toll on a portion of the lanes of a highway can generate a Pareto improvement, even before the toll revenue is spent. To price a portion of the lanes we split a highway into two routes using a barrier or painted lines, and add tolls to one of the routes. This

[1] See Appendix B.1 for the sources and calculations for this claim.

[2] See Appendix B.2 for a brief explanation of this standard result, evidence that this result is the received wisdom, and a discussion of other barriers to congestion pricing.

[3] For example, Small (1983, 1992) makes practical proposals regarding how to use the revenue to improve the distributional effects of congestion pricing but is careful to state that it is very unlikely that following his proposals would generate a Pareto improvement.

practice is generally called 'value pricing,' and the priced lanes are called 'HOT lanes,' 'express lanes,' or more derisively, 'Lexus lanes.'[4]

I obtain this new result by extending the bottleneck congestion model of Vickrey (1969) and Arnott, de Palma, and Lindsey (1990, 1993) to reflect an important additional traffic externality that has been identified by the transportation engineering literature but that has largely been ignored in the economics literature.[5] Not only does each additional vehicle slow others down, but, in heavy enough traffic, additional vehicles can create additional frictions which reduce throughput.

To understand these two externalities better, consider a two-lane highway that merges down to one lane at some point, creating a bottleneck. When the arrival rate at the bottleneck exceeds its capacity a queue forms. Each additional vehicle that travels during rush hour *lengthens the queue*, increasing the travel time of all those behind it by a few seconds. This lengthening of the queue is the standard externality. However, what this simple externality fails to capture is the fact that a queue creates additional frictions that *reduce throughput*, reducing the rate at which vehicles can pass through the bottleneck and further increasing travel times. This contrasts with most queues; while a long line at the grocery store means you have to wait a while, it does not affect the rate at which customers are served.[6]

It is by reducing this second externality that tolling can lead to a Pareto improvement. Time-varying tolls can smooth the rate that people depart for work, increasing speeds and throughput,[7] and when agents are homogeneous, this is enough to conclude everyone is better off.

In practice, road users are not homogeneous, and allowing for heterogeneity makes it likely that pricing all of the lanes will hurt some road users, even when pricing increases throughput. Adding tolls reduces the time costs of traveling while increasing the financial costs. As not all road users value their time equally, this can hurt some road users.

---

[4]The name 'value pricing' refers to drivers' option of paying more for something of greater value, and the acronym HOT stands for High-Occupancy/Toll. The epithet 'Lexus lanes' is intended to convey the accusation that only those who can afford a Lexus can afford to drive in them. The empirical evidence indicates that drivers of all income levels use the priced lanes, although the rich use them more frequently (Sullivan and Harake, 1998, Sullivan, 2002, Patterson and Levinson, 2008).

[5]See Small (2015) for a recent review of the literature using the bottleneck model.

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

[7]While it seems counterintuitive that adding tolls can increase both speeds and throughput, I show how this is possible in Section 2.3.

However, we can still generate a Pareto improvement by pricing a portion of the lanes. Doing so increases speeds and throughput in the priced lanes allowing them to carry a more-than-proportional share of traffic. This means the free lanes are carrying a less-than-proportional share of traffic, and so travel times in the free lanes are better than they were before. Since travel times in the free lanes are better, those who continue to use the free lanes are better off. Those in the priced lanes could have stayed in the free lanes, and so by revealed preference are better off. We have generated a Pareto improvement.

My main theoretical contribution involves characterizing the set of parameter values for which pricing some or all of the lanes generates a Pareto improvement when there are two groups of agents.[8] Furthermore, I identify potential exceptions to the intuition above and provide an intuitive sufficient condition for value pricing to yield a Pareto improvement: we simply need some rich drivers to be using the highway at the peak of rush hour.[9]

These theoretical results build on Walters (1961), who first conjectured that too many cars on the road could reduce throughput, and Vickrey (1987) who gave this second externality a name: hypercongestion.[10] My results extend those of Johnson (1964) and de Meza and Gould (1987), who showed that Walters' conjecture implied congestion pricing would generate a Pareto improvement when agents are homogeneous, by modeling the mechanism by which throughput falls and showing that we can still obtain a Pareto improvement when agents are heterogeneous.[11]

I make two additional theoretical contributions. First, I show how the bottleneck model's implicit assumption that throughput is unaffected by pricing explains the differences between the welfare effects of congestion pricing in the bottleneck model relative to other models. Second, I extend the bottleneck model to allow for a continuum of desired arrival times. This feature, with otherwise homogeneous agents, appeared in the initial papers using the bottleneck model (Vickrey, 1969, Hendrickson and Kocur, 1981), but was subsequently dropped as it did not affect equilibrium outcomes.[12] However, once agents

[8]A group is a set of agents with the same value of time and schedule inflexibility but with heterogeneous desired arrival times.

[9]When there are more than two groups we need some of the richest group of drivers to be using the highway at the peak of rush hour.

[10]Rotemberg (1985) shows an additional way equilibrium throughput can be lower than socially optimal: if drivers internalize all the costs of an accident, then driving marginally closer to the vehicle ahead of them (holding speed constant) increases highway throughput without changing the social cost of accidents, and so increases social welfare. This externality cannot be internalized with a toll.

[11]This work also builds on a literature studying the distributional effects of pricing a portion of the lanes (eg., Small, Winston, and Yan, 2006, Light, 2009, van den Berg and Verhoef, 2011), and is related to a more recent literature on hypercongestion in urban centers, which is the context in which Vickrey defined it (e.g., Small and Chu, 2003, Arnott and Inci, 2010, Arnott, 2013, Fosgerau and Small, 2013).

[12]The two other papers to consider agents with a continuum of desired arrival times who are heterogeneous in other dimensions are Newell (1987), who shows analytically that equilibrium travel times and tolls only depend on the preferences of some drivers, and de Palma and Lindsey (2002), who numerically solve for the equilibrium when there are no tolls. I build on this work by finding closed form solutions for the equilibrium when either none or all of the lanes are priced, and solving numerically for the equilibrium when some of the lanes are priced.

are heterogeneous along other dimensions then allowing for agents' desired arrival time to be continuously distributed has significant effects on equilibrium outcomes and is vital for matching the model to the data.

My main empirical contributions are twofold: confirming the practical relevance of the theoretical possibility of generating a Pareto improvement, and measuring the size of the social welfare gains available from congestion pricing. I generalize my model to allow agent preferences to vary continuously along three dimensions: value of time, desired arrival time, and schedule inflexibility. I then estimate the joint distribution of agent preferences over these three dimensions for road users on California State Route 91. This is the first time the distribution of inflexibility has been estimated, despite its importance in dynamic congestion models, as well as the first time this joint distribution has been measured. I then use these estimates to evaluate the effects of congestion pricing. I find that the welfare gains from congestion pricing are large. Pricing all of the lanes increases social welfare by $2,400 per road user per year, but at the cost of hurting some road users by $2,390 per year. However, by pricing just half of the lanes we obtain a Pareto improvement while still capturing 73 percent of the social welfare gains.

I extrapolate my results to the rest of the United States and find that applying a throughput-maximizing toll to half the lanes on all urban highways would increase social welfare by over $30 billion per year, or $850 per year for the typical urban highway commuter.[13]

2. How tolls can increase throughput

The argument in this paper depends critically on the claim that tolls can be used to increase highway throughput. In this section I first explain the two causal mechanisms identified by transportation engineers which reduce throughput, both of which occur when a queue forms at a bottleneck. I then show how a carefully designed toll can prevent these frictions from arising, thereby increasing throughput.

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

2.1. **Queue spillovers.**

The first throughput-reducing friction occurs when the queue behind a bottleneck grows long enough that it blocks other traffic. For example, a queue can grow at a busy off-ramp, spilling onto the mainline of the freeway and blocking through traffic; similarly, a queue on the highway can block upstream exits. Vickrey

[13]The social welfare gains are smaller for the typical urban road user than for those on California State Route 91 because Route 91 is among the most congested highways in America and those who use it have longer-than-average commutes.

(1969) labeled the second situation a trigger neck and transportation engineers call both situations a queue spillover.

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

2.2. **Throughput drop at bottlenecks.** In addition, throughput at the bottleneck itself can fall once a queue forms. On our two-lane highway the vehicles in the right lane need to change lanes before getting to the bottleneck. When traffic is heavy doing so is difficult, and there will be a vehicle that comes to a stop before merging and, rather than waiting for a gap, will force its way over. Transportation engineers call this a destructive lane change, and we can see the damage in two ways. First, the vehicle that forced its way over will be moving very slowly and so go through the bottleneck at a slow speed. Equivalently, it will open up a gap in front of itself; this will be a period of time that the bottleneck, the scarce resource on the highway, is not being used.

There is a large transportation engineering literature documenting that throughput at bottlenecks drops once a queue forms, which they refer to as the two-capacity hypothesis. The name "two-capacity hypothesis" refers to the idea that a road has one capacity, or throughput, when there is no queue and a different capacity when there is a queue. The median estimate for the size of the drop is 10 percent; estimates range as high as 25 percent, and are presented in Appendix Table C.1.

2.3. **Tolls can increase throughput.** We can keep a queue short, thus preventing the two mechanisms above from reducing throughput, by using a time-varying toll to smooth the rate at which vehicles get on the highway. In doing so, it becomes possible for tolls to make all road users better off.

Figure 1 gives a stylized example of how this can work. Consider a two-lane highway that merges down to one lane, creating a bottleneck. When the road is unpriced, drivers depart from home at rate $\rho(t)$. At 7:00 a.m., rush hour begins and 48 vehicles per minute depart from home, but if the highway's maximum throughput is only 40 vehicles per minute, then a queue forms and travel times start climbing. As the queue gets longer, the second externality takes effect and highway throughput falls to just 32 vehicles per minute. As we approach 8:30, the number of vehicles on the highway as well as travel times climb to their peak. At 8:30, the departure rate falls to 8 vehicles per minute, allowing the length of the queue, and thus travel times, to start falling, until eventually everyone has reached work and rush hour ends at 9:20. In equilibrium, homogeneous drivers are indifferent between departing anytime during rush hour; they can either leave early (or late) to avoid

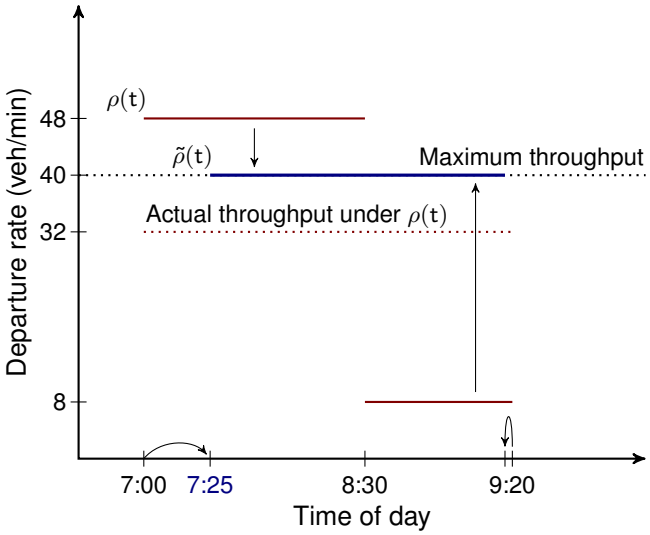


FIGURE 1. Tolls can smooth the departure rate, preventing queuing and increasing throughput.

traffic but get to work earlier (or later) than desired, or leave so as to arrive right on-time but endure a long commute in bad traffic.

Using time-varying tolls we can induce drivers to depart at rate $\tilde{\rho}(t)$, reducing the departure rate before 8:30 and increasing the rate thereafter. By preventing the queue from forming, we eliminate both externalities; there is no queue and throughput remains high at 40 vehicles per minute. Since throughput is higher, rush hour is shorter. In our stylized example, rush hour can start 25 minutes later and end 3 minutes earlier.

By considering the effect of pricing on the first driver to depart in the morning, we can take our first look at the welfare impacts of congestion pricing. When the road is free, this driver does not face any congestion, but leaves for work very early. Adding tolls shortens rush hour, so he does not need to leave as early; and he still faces no congestion and pays no toll (for reasons we will see later), and therefore is better off. If all drivers are identical, then the fact that the first driver to depart is better off means all drivers must be better off; we have obtained a Pareto improvement before spending the revenue.

3. MODEL

I build on the bottleneck model of Vickrey (1969), which was formalized by Arnott, de Palma, and Lindsey (1990, 1993). I make three important modifications to the model. The first is to add the second externality by allowing throughput to fall when a queue forms. This is a natural way to model the throughput drop at bottlenecks and serves as shorthand for the effects of queue spillovers.[14] The second modification is to allow the

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

social planner to choose the fraction of the lanes that are priced; this goes beyond existing work, such as Van den Berg and Verhoef (2011), which considers the welfare implications of pricing a fixed portion of the lanes. The third change is to allow agents' desired arrival time at work to be continuously distributed, as in Vickrey (1969), Hendrickson and Kocur (1981), and de Palma and Lindsey (2002). Allowing for a continuum of desired arrival times is important because it allows drivers to be inframarginal, meaning that if the cost of their chosen arrival time increases by a small amount, they do not change when they arrive. This matters because it is necessary to match the data. The evolution of travel times across the day suggests that the marginal driver at any point in time is quite willing to change when he arrives to save just a little travel time, which means the marginal driver cannot be a shift worker. A model that does not allow for inframarginal drivers must either predict travel times which climb (and fall) much quicker than observed or does not contain agents with very inflexible schedules.

3.1. **Congestion technology.** There is a single road connecting where people live to where they work; this road can be split into two routes, one tolled, the other free. Let λ_{toll} and λ_{free} denote the fraction of capacity devoted to each route, where $\lambda_{\text{toll}} + \lambda_{\text{free}} = 1$.[15] Travel along this road is uncongested, except for a single bottleneck through which at most s^* vehicles can pass per unit time. Letting r denote the route and t the time of departure from home, when the departure rate on a route, $\rho_r(t)$, exceeds its capacity, $\lambda_r \cdot s^*$, a queue develops. Once the queue is more than ϵ vehicles long the throughput of the bottleneck for that route falls to $\lambda_r \cdot s$, where $s \leq s^*$. Therefore, queue length, Q_r, measured as the number of vehicles in the queue, evolves according to

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

I then simplify by taking the limit as $\epsilon \to 0$, so throughput on a congested route is constant.[16]

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

[16]This allows me to keep the model simple while avoiding existence of equilibrium problems which can occur when, if the route is congested, the equilibrium departure rate is too low to create congestion, but when the route is uncongested the equilibrium departure rate is high enough to create congestion.

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

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

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

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

$$T^v_{d,r}(t) = \frac{Q_r(t)}{\lambda_r \cdot s}. \tag{2}$$

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

It will be simpler to focus on arrival times instead of departure times, so define $T_r(t)$ as the travel time on route r for an agent *arriving* at t. Because this model is deterministic, there is a one-to-one mapping between departure times and arrival times, and thus doing so is innocuous.[17]

3.2. **Agent preferences.** Agents choose when to arrive at work and which route to take to minimize the cost of traveling. Agents dislike three aspects of traveling: travel time, tolls, and schedule delay—that is, arriving earlier or later than desired. These costs combine to form the trip cost; the trip cost of arriving at time t on route r for an agent in group i with desired arrival time t^* is

$$p(t, r; i, t^*) = \alpha_i T_r(t) + \tau_r(t) + D_i(t^* - t) \tag{3}$$

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

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

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

[17] See Appendix D.1 for more details.

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

To simplify the problem let $\gamma_i = \xi\beta_i$ for all i. This means that those who dislike being early also dislike being late, while those who do not mind being early similarly do not mind being late.[18]

Agents can differ in their value of time, schedule delay costs, and desired arrival time. A *group* of agents is the set of agents with the same value-of-time and schedule delay costs. Let $\mathcal{G}$ denote the set of groups. We will consider $\mathcal{G} \in \{\{1\}, \{1,2\}, \mathbb{R}^+\}$.

The primary source of heterogeneity in agents' value of time is variation in their income, and so if $\alpha_i > \alpha_j$ then group i is *richer* than group j. While there are other sources of heterogeneity in agents' value of time,[19] by using α as a proxy for income we can directly discuss the primary concern with congestion pricing: that it helps the rich and hurts everyone else.

The ratio of an agent's schedule delay costs to value of time provides a measure of how inflexible his schedule is, so if $\beta_i/\alpha_i > \beta_j/\alpha_j$ then group i is more *inflexible* than group j. The main source of heterogeneity in agents' flexibility arises from differences in occupation, as the opportunity cost of time early or late is different for those with different types of jobs. If a shift worker is late he generally face penalties and when he is early he passes the time talking with co-workers. Since there is not much difference for the shift worker between spending time traveling or being at work early, his β/α is close to one (the largest possible β/α). Similarly, due to the penalty when late, γ/α is large. In contrast, an academic can start working whenever he gets to the office and so has a very low marginal disutility from being early or late and so his β/α and γ/α are closer to zero. Thus variation in β/α is driven by variation in schedule flexibility, where jobs that are more flexible lead to a lower β/α.[20]

[18]Relaxing this assumption would only affect my theoretical results if there are agents who switch from arriving early to arriving late, or vice-versa, when tolls are added to the road. Because of this assumption, my estimator for marginal distribution of β/α (and the distribution of γ/α as it is a transformation of the distribution of β/α) uses information about both early and late arrivals. In Section 8 I also fit a version of the model which relaxes this assumption, among others, and find that these assumptions have a fairly trivial effect on how well I can match the data and on those parameter estimates the relaxed version of the model can recover.

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

[20]How flexible a worker's personal life is also affects 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.

Within each group, agents' desired arrival times are uniformly distributed over $[t_s, t_e]$.[21] Having a continuous distribution of desired arrival times allows a positive measure set of agents to arrive on-time, and thus allows for inframarginal agents; assuming this continuous distribution is uniform keeps the model analytically tractable despite having a continuum of types. While it may seem more natural to assume an agent's desired arrival time falls into some discrete set, such as $\{7{:}00, 7{:}30, \ldots, 9{:}00\}$, what matters is when agents want to arrive at the end of the highway, not when they want to arrive at work. Because the distribution of distances between the end of the highway and work is continuous, the distribution of desired arrival times at the end of the highway is also continuous.

Let n_i denote the density of agents of group i who desire to arrive at a given time in $[t_s, t_e]$ and let $N_i = (t_e - t_s)\, n_i$ be the total mass of agents in group i. Furthermore, $\sum n_i$ is assumed to exceed the road's capacity (s^*), so it is impossible for all agents to arrive at their desired arrival time; thus some will need to arrive early or late.

The mass of agents of each type who use the road is independent of the trip cost, that is, demand for travel along this road is perfectly inelastic. Were demand not perfectly inelastic then the distribution of desired arrival times would no longer be uniform once tolls were added to the highway and different types saw their trip costs change by different amounts. By assuming perfectly inelastic demand, I maintain the benefits of having uniformly-distributed desired arrival times. This assumption fits my empirical context: California State Route 91, a highway through a mountain pass between Riverside County and Orange County. Commuters do not have a reasonable alternative to taking SR-91 and public transit accounts for less than 1 percent of the trips through the pass (Sullivan and Burris, 2006, 192); thus I do not need to worry about agents switching to different roads or modes and the only choice I am missing is the choice of whether to travel.[22]

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

3.3. **Definition of equilibrium.** The relevant equilibrium concept is that of a perfect information, pure strategy Nash equilibrium, in which no agent can reduce his trip cost by changing his arrival time or route choice.

[21] In Section 8 I provide evidence that this is a reasonable approximation to the truth. I am also assuming that the distribution of desired arrival times is independent of value of time and inflexibility; in Appendix H.2 I provide empirical evidence in support of this assumption.

[22] By having perfectly inelastic demand I rule out one way pricing can hurt the poor: because congestion pricing lowers the cost for richer agents it induces more rich agents to travel. This counteracts some of the benefit to existing agents of increasing throughput. If demand for trips by the rich is sufficiently elastic it is even possible rush hour is longer once congestion pricing is implemented. In a previous version of this paper I had elastic demand, and homogeneous desired arrival times, and the elasticity of demand only had minor effects on the results. That said, there is evidence that the long run demand for travel is perfectly elastic (Duranton and Turner, 2011). If demand is perfectly elastic for all types then it is impossible to increase or reduce the cost of travel, and pricing all of the lanes, regardless of the effect it has on throughput, never hurts any road users even before the revenue is used.

I show that an equilibrium exists by construction, and show that equilibrium trip prices, travel times, and tolls are unique in Appendix G.

4. Finding the equilibrium

The fundamental scarcity is that there are times where more agents want to arrive than are able. Since not everyone can arrive at their desired arrival time, some agents must arrive early or late. For some agents to be willing to arrive early or late, they must receive a compensating differential in the form of lower travel times or cheaper tolls.

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

In the bottleneck model the optimal toll eliminates congestion. One virtue of the bottleneck model is that its production possibilities frontier has a unique optimal point that maximizes speed and throughput and so the optimal toll is the one that keeps us at this point. Restricting the departure rate on the priced route to less than $\lambda_{\text{toll}} \cdot s^*$ leaves capacity unused and creates unnecessary schedule delay. Allowing more than $\lambda_{\text{toll}} \cdot s^*$ vehicles to depart generates queuing, which wastes time and decreases throughput. This means the socially optimal toll is set to eliminate queuing and maximize throughput. The toll varies to induce some agents to arrive early or late, so they depart at the rate the priced route can handle; thus a queue never forms, and the departure and arrival rate on the tolled route is $\lambda_{\text{toll}} \cdot s^*$ for all of rush hour.

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

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

Given these results, the bottleneck model when the road is completely free or priced is similar to the Hotelling (1929) differentiated goods model. We have continuum of differentiated goods (arrival times), and agents have unit demand and bear a cost of purchasing a good different from the one they prefer (schedule delay costs). The key difference is that each good is "provided" by firms in a perfectly competitive market who in aggregate inelastically supply s_r units of the good.[23]

[23]It is also analogous to the von Thünen (1930) model of land use. Instead of land use we are modeling the use of arrival times, and we replace transportation costs with schedule delay costs. When all agents have

4.1. **Free route.** The most desirable arrival times are allocated to those who are willing to pay the most for them. For a free route the currency used is travel time. This means those who are very inflexible arrive closer to their desired arrival time because an agent's inflexibility is his willingness to pay in travel time to reduce schedule delay, that is, his willingness to pay in travel time to arrive closer to his desired arrival time. This is formalized in the following lemma. The proof, along with all other omitted proofs, is given in Appendix A.

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

Once we have assigned agents to arrival times, we can use their preferences to back out the travel time profile (i.e., the function T). If an agent arrives early or late on a free route it must be true that his marginal rate of substitution between schedule delay and travel time equals the marginal rate of substitution the equilibrium travel time profile offers; that is, the slope of the travel time profile at the time he arrives must equal his inflexibility if he is early and $-\xi$ times his inflexibility if he is late. If an agent arrives exactly at his desired arrival time all we know is that his schedule delay costs are such that he is unwilling to accept schedule delay given the travel time profile. I formalize these results in the following lemma.[24]

Lemma 2.

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

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

4.2. **Priced route.** For a priced route the currency used to allocate arrival times is money. This means those with a high β arrive closer to their desired arrival time because an agent's β is his willingness to pay in money to reduce schedule delay. This is formalized in the following lemma.

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

In similar fashion to before, once we have assigned agents to arrival times, we can use their preferences to back out the toll schedule (i.e., the function τ). If an agent from

---

the same desired arrival time and the cost of being late is the same as the cost of being early the models are identical.

[24]This lemma also implies that to have inframarginal agents there must be a kink in the schedule delay cost function.

group i arrives early or late on a priced route it must be true that his indifference curve is tangent to the "budget line" so his marginal rate of substitution between schedule delay and money equals the marginal rate of substitution the toll schedule offers. Thus the slope of the toll schedule at the time he arrives must equal β_i if he is early and $-\xi\beta_i$ if he is late. If an agent arrives exactly at his desired arrival time all we know is that his schedule delay costs are such that he is unwilling to accept schedule delay given the toll schedule. We formalize these results in the following lemma.

Lemma 4.

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

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

4.3. **Value pricing.** When there are two routes, one free and the other priced, we need to assign agents to routes, and then we can use the methods above to assign them to arrival times on their routes. Agents travel on the route that gives them the lowest cost. I save most of the details concerning how I do this until later, as I approach it differently when there are two groups than when there are an arbitrary number of groups, though want to make one point now: the start and end of rush hour are the same on each route. If not, then there would be a time where there was congestion on the free route, but no toll on the priced route, and so an agent arriving at this time on the free route would deviate and arrive at the same time on the priced route. Similarly, there cannot be a positive toll on the priced route while there is no congestion on the free route.

5. Homogeneous agents

Let us start by considering the case where every agent is identical.[25] Starting with the simplest version of the model allows me to highlight how the welfare effects of congestion pricing depend crucially on whether tolling increases or decreases throughput.

In any congestion model with continuous time the first agent to arrive on either route pays no toll and faces no congestion. This must be true since the first agent could shift his arrival time forward by an infinitesimal amount and would then be arriving outside of rush hour. He would then face no travel time and pay no toll.[26] The only cost this first agent bears is the schedule delay costs from arriving so early.

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

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

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

This logic implies that if we believe highway traffic is always on the PPF, then we cannot escape the conclusion that congestion pricing, while Kaldor-Hicks efficient, hurts road users before the revenue is spent.

Proposition 1 (Prior literature)**.** *If agents are homogeneous in any congestion model with a strictly negative relationship between flow and speeds, then congestion pricing makes all agents worse off before the toll revenue is spent.*

When traffic is on the PPF the goal of congestion pricing is to reduce throughput so that the remaining agents can travel faster.[27] It is this logic that leads the U.S. Department of Transportation to state that "congestion pricing works by shifting purely discretionary rush hour highway travel to other transportation modes or to off-peak periods" (2006, 1). But reducing throughput means rush hour must be longer, and so when agents are homogeneous this means all are harmed before the revenue is used.

The standard bottleneck model $(s = s^*)$ assumes away the traditional trade-off between throughput and speed, where increasing one requires decreasing the other (cf. Pigou, 1920, Knight, 1924, Walters, 1961). While this makes modeling the dynamics of rush hour tractable, as an unappreciated side effect it also changes the welfare effects of congestion pricing.

Proposition 2 (Vickrey, 1969)**.** *If agents are homogeneous in the bottleneck model with no throughput drop (i.e., $s = s^*$), then congestion pricing neither helps nor hurts any agents before the toll revenue is spent.*

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

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

The literature has not recognized the importance of assuming away the traditional trade-off between throughput and speed for explaining the differences between the welfare effects of congestion pricing in the bottleneck and other models.[28] For example, Arnott, de Palma, and Lindsey (1993) and Van den Berg and Verhoef (2011) both use the bottleneck model and find that the welfare impacts of congestion pricing are much more favorable than previous research reported, but attribute the difference to using a dynamic model rather than the implicit assumption about how pricing affects throughput.

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

Proposition 3. *If all agents are homogeneous in the bottleneck model with a throughput drop (i.e., $s < s^*$), then congestion pricing generates a Pareto improvement and helps all agents before the toll revenue is spent.*

When queues reduce throughput the goal of pricing is to increase throughput by eliminating the queue and its attendant frictions. We are able to increase both speeds and throughput. Because rush hour is shorter, all agents are better off.

Including heterogeneity in agents' preference in our modeling will make the analysis more complicated; however, whether it is possible for congestion pricing to generate a Pareto improvement prior to using the revenue still depends on whether tolling increases throughput. While we can no longer use changes in the start of rush hour as a sufficient statistic for how *everyone's* welfare changes, it still tells us about how *someone's* welfare changes. Because these results hold for at least one agent, we can conclude that if increasing speeds requires reducing throughput then it is *impossible* for congestion pricing to generate a Pareto improvement before spending the revenue, but since congestion pricing can increase throughput while increasing speeds, then it is *possible* for congestion pricing to generate a Pareto improvement regardless of how the revenue is spent.

6. Two groups

While it is possible for pricing all of the lanes to generate a Pareto improvement when agents are heterogeneous, we can expand the set of parameters for which congestion pricing generates a Pareto improvement by pricing only a portion of the lanes. To identify the potential barriers to obtaining a Pareto improvement from congestion pricing and to understand how value pricing can help overcome these barriers, let us now allow for two groups of agents. I end the section with a simple sufficient condition for when congestion pricing leaves all road users better off: as long as some rich agents use the highway at the peak of rush hour then value pricing generates a Pareto improvement. This result

[28]Chu (1995) implicitly makes this point when he shows "the behavior of the [bottleneck] approach is the limit of that of the reformulated Henderson approach as the elasticity of travel delay goes to infinity" (p. 324).

holds even when there are an arbitrary number of groups. This section contains the main theoretical contributions of the paper.

As the primary concern with congestion pricing is that it only helps the rich, the main distinction I make is between high- and low-income agents, and so I define group 1 as rich and group 2 as poor (i.e., $\alpha_1 > \alpha_2$).

As our first look at the benefit of value pricing, consider what happens when the only heterogeneity is due to a small group of poor agents, so small that they do not affect the equilibrium. If we price all of the lanes there is no guarantee they are not worse off;[29] however when we price just a portion of the lanes we can know that they are better off.

Proposition 4. *If all agents except for a zero measure set are homogeneous, then in the bottleneck model with a throughput drop (i.e., $s < s^*$), pricing a portion of the lanes generates a Pareto improvement and helps all agents before the toll revenue is spent.*

Proof. Since the zero measure group of agents has no impact on equilibrium, we know by Proposition 3 that all agents in the group with positive measure are better off. For those agents in the positive measure group who are on the free lanes to be better off, travel times must have fallen at each point in time. Thus if the zero measure agents travel on the free lanes at the same time they traveled before, then they will have shorter travel times and be better off. Since they have an option that gives them a lower trip cost than before, whatever they choose must make them better off. Thus all agents are better off. □

The logic behind this proof leads to a straightforward empirical test for whether value pricing gives rise to a Pareto improvement, even with arbitrary heterogeneity: check if travel times on the free lanes fell for every point in time. If so, pricing must have helped every road user.

Proposition 4 isolates the mechanism by which value pricing makes it easier to generate a Pareto improvement: value pricing increases highway throughput while preserving the ability of agents to pay with their time instead of their money to travel at the peak. While the increase in total throughput is smaller than when pricing all of the lanes, and so the social welfare gains are smaller too, doing so makes it easier to obtain a Pareto improvement.

Once we allow both groups to be large enough to affect the equilibrium there are additional barriers to obtaining a Pareto improvement; however, the intuition of Proposition 4 will still hold. In the rest of this section I solve for the equilibrium with two groups of positive mass and then use these results to determine when pricing all or part of the lanes generates a Pareto improvement.

[29] Congestion pricing reduces travel time costs while increasing monetary costs, and so whether the small group is better off depends on how their value of time compares to that of the other agents.

6.1. **Equilibrium when the road is completely free or priced.** For simplicity, define group A as the group that arrives off-peak, and group B as the group that arrives on-peak. This reduces the number of cases we need to solve and we can map A and B into rich and poor as needed. Lemma 1 implies that on a free route $\beta_A/\alpha_A < \beta_B/\alpha_B$ and Lemma 3 implies that on a priced route $\beta_A < \beta_B$. When the entire road is either free or priced, one of two subcases apply: either $n_B \leq s$ or $n_B > s$.

6.1.1. *Equilibrium when group B is inframarginal.* When $n_B \leq s$ on a free road there is enough capacity for the inflexible agents to all arrive exactly at their desired arrival time. This means that only flexible agents arrive early or late.

Define $t_i^{\max}$ as the time such that the agent in group i with this desired arrival time is indifferent between arriving early or late. Any agent from group i who has desired arrival time $t^* < t_i^{\max}$ strictly prefers to arrive early or on-time, and similarly if $t^* > t_i^{\max}$ then they strictly prefer to arrive late or on-time.[30] I use the superscript "max" for two reasons: first, the agent from group i with desired arrival time $t_i^{\max}$ will have the largest trip cost of any agent in group i; second, the peak of rush hour, $t^{\max}$, occurs at one or more groups $t_i^{\max}$.

Defining t_{ij} as the time when agents from group i stop arriving and agents from group j start arriving, and, for the sake of notation, defining a fictional group 0 who travels when no one else is on the road, we can use Lemma 2 to define the equilibrium travel time profile as the solution to

$$\frac{dT_I}{dt}(t) = \begin{cases} \beta_A/\alpha_A & t_{0A} \leq t < t_A^{\max} \\ -\gamma_A/\alpha_A & t_A^{\max} \leq t < t_{A0} \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

$$T_I(t_{0A}) = 0. \tag{5}$$

The subscript I denotes that these objects belong to the case where some agents are inframarginal.[31]

This allows us to write equilibrium travel times as a function of the start of rush hour, t_{0A}, the end of rush hour, t_{A0}, and the peak of rush hour, $t_A^{\max}$. The requirements of equilibrium give us three equations that can be solved for these three unknowns.

The first equation requires that the demand for early arrivals by agents in group A equals the supply. The supply for early arrivals is the capacity available between the start of rush hour and the peak. In this period of time $(t_A^{\max} - t_{0A})\, s$ agents can arrive. However, we need to account for the capacity used by agents in group B. Since they arrive on-time,

[30]This relies on the travel time profile having a single local maximum, which I prove in the appendix in Proposition G.1.

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

$(t_A^{\max} - t_s)\, n_B$ of the capacity available for early arrivals is used by agents of group B. All agents in group A with a desired arrival time before $t_A^{\max}$ arrive early, and so demand for early arrivals by agents in group A is $(t_A^{\max} - t_s)\, n_A$. Thus in equilibrium

$$(t_A^{\max} - t_{0A})\, s - (t_A^{\max} - t_s)\, n_B = (t_A^{\max} - t_s)\, n_A. \tag{6}$$

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

$$(t_{A0} - t_A^{\max})\, s - (t_e - t_A^{\max})\, n_B = (t_e - t_A^{\max})\, n_A. \tag{7}$$

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

$$T_I\left(t_{A0}\right) = 0. \tag{8}$$

The way we find the equilibrium when the road is priced is essentially the same. As $n_B \leq s$ there is enough capacity for all agents in group B to arrive on-time. Using Lemma 4 we can define the equilibrium toll schedule as the solution to

$$\frac{d\tau_I}{dt}\left(t\right) = \begin{cases} \beta_A & t_{0A} \leq t < t_A^{\max} \\ -\gamma_A & t_A^{\max} \leq t < t_{A0} \\ 0 & \text{otherwise} \end{cases}, \tag{9}$$

$$\tau_I\left(t_{0A}\right) = 0. \tag{10}$$

Again we have three variables still to determine, and the equations that define them are similar to the equations for a free road. Because capacity on a priced route increases to s^*, we replace s with s^* in (6) and (7), as well as changing subscripts to denote that we are considering a priced route. Finally, we replace (8) with

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

We now know enough to find equilibrium trip costs. By solving (6), (7), and (8), or the equivalent equations for a priced route, we can determine the equilibrium travel time profile or toll schedule. We can then find each type's equilibrium trip cost $\bar{p}\left(i, t^*\right) = \min_{t,r} p\left(t, r; i, t^*\right)$, as is done in Appendix F.1. The equilibrium trip costs for agents in group A for $r \in \{\text{free}, \text{toll}\}$ are

$$\bar{p}_{I,r}\left(A, t_A^{\max}\right) = \beta_A \left(N_A + N_B\right) \frac{1}{s_r} \frac{\xi}{1+\xi}, \tag{12}$$

$$\bar{p}_{I,r}\left(A, t^*\right) = \bar{p}_{I,r}\left(A, t_A^{\max}\right) - \left(t_A^{\max} - t^*\right) \begin{cases} \beta_A & t^* \leq t_A^{\max} \\ -\xi\beta_A & t^* > t_A^{\max} \end{cases}. \tag{13}$$

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

$$\bar{p}_{I,\text{free}}(B, t^*) = \frac{\alpha_B}{\alpha_A} \bar{p}_I(A, t^*), \tag{14}$$

while on a priced route they are

$$\bar{p}_{I,\text{toll}}(B, t^*) = \bar{p}_I(A, t^*). \tag{15}$$

While (13)–(15) can be calculated directly, they are also fairly intuitive. First, note that due to the slope of the travel time profile and toll schedule every agent in group A who arrives early is indifferent between arriving at their desired arrival time or earlier, and likewise those who are late are indifferent between arriving at their desired arrival time or later.

To see the intuition behind (13) consider two agents in group A, one with desired arrival time of $t_A^{\max}$ and the other of t^*. They are both willing to arrive at t^*, and were they to do so the only difference in their trip cost would be the difference in their schedule delay costs at t^*. This means we can write the trip cost of the second as the trip cost of the first minus the difference in their schedule delay costs at t^*.

To see the intuition for (14) and (15) consider two agents with desired arrival time t^*, one from each group. Both are willing to arrive at t^*. When arriving at t^* on a free route neither of them have any schedule delay costs and they face the same travel time, so the only difference in their trip cost is due to the difference in their value of time. By dividing the group A agent's trip cost by his value of time we recover the travel time at t^*, which we then multiply by the group B agent's value of time to obtain the group B agent's trip cost. Similarly, on a tolled route they face the same toll and have no schedule delay or travel time, and so their trip costs are identical.

When one group is inframarginal their preferences do not affect the equilibrium or the marginal group's trip cost. Equation (12) is the same as (15) in Arnott, de Palma, and Lindsey (1993), with $N = N_A + N_B$. Further, travel times and tolls are the same. Because the inframarginal group's preferences do not affect equilibrium, the logic behind Proposition 4 carries over to this case, and, as I show later, value pricing generates a Pareto improvement.

6.1.2. *Equilibrium when group B is marginal.* When $n_B > s$ there is no longer enough capacity for the inflexible agents to all arrive at their desired arrival time, and so they must arrive early or late.[32] Group B agents will use all of the capacity near the peak, and group A agents will use all of the capacity off-peak. We can use Lemma 2 and the

[32]Agents arriving early (late) are indifferent between arriving anytime between their desired arrival time and the earliest (latest) someone from their group arrives. Within a group of indifferent agents I choose their arrival times such that if an agent desires to arrive later than another agent, then he actually arrives later than the other agent. This choice means in this case only a set of measure zero agents arrive at their desired arrival time; it would be possible to re-arrange arrival times so that a greater share of group B agents arrived at their

requirement that the travel time at the end of rush hour is zero to define the equilibrium travel time profile as the solution to

$$\frac{dT_M}{dt}(t) = \begin{cases} \beta_A/\alpha_A & t_{0A} \leq t < t_{AB} \\ \beta_B/\alpha_B & t_{AB} \leq t < t_B^{\max} \\ -\gamma_B/\alpha_B & t_B^{\max} \leq t < t_{BA} \\ -\gamma_A/\alpha_A & t_{BA} \leq t < t_{A0} \\ 0 & \text{otherwise} \end{cases}, \tag{16}$$

$$T_M(t_{0A}) = T_M(t_{A0}) = 0. \tag{17}$$

The subscript M denotes that these objects belong to the case where all agents are marginal.

Now we have three additional variables to solve for to find equilibrium travel times. As before, we use the requirement that supply equals demand for early and late arrivals, but now we do so for both groups. These requirements give us the following equations.

$$(t_{AB} - t_{0A})\, s = (t_A^{\max} - t_s)\, n_A, \tag{18}$$

$$(t_{A0} - t_{BA})\, s = (t_e - t_A^{\max})\, n_A, \tag{19}$$

$$(t_B^{\max} - t_{AB})\, s = (t_B^{\max} - t_s)\, n_B, \tag{20}$$

$$(t_{BA} - t_B^{\max})\, s = (t_e - t_B^{\max})\, n_B. \tag{21}$$

For the final equation I impose the definition of $t_A^{\max}$,

$$p(t_{AB}, \text{free}; A, t_A^{\max}) = p(t_{BA}, \text{free}; A, t_A^{\max}). \tag{22}$$

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

$$\frac{d\tau_M}{dt}(t) = \begin{cases} \beta_A & t_{0A} \leq t < t_{AB} \\ \beta_B & t_{AB} \leq t < t_B^{\max} \\ -\gamma_B & t_B^{\max} \leq t < t_{BA} \\ -\gamma_A & t_{BA} \leq t < t_{A0} \\ 0 & \text{otherwise} \end{cases}, \tag{23}$$

$$\tau_M(t_{0A}) = \tau_M(t_{A0}) = 0. \tag{24}$$

As before, we replace s with s^* in (18)–(21) and change subscripts. Finally, we update the definition of $t_A^{\max}$ for a priced route by replacing "free" with "toll" in (22).

---

desired arrival times, but they would still be indifferent between being early or late, and it would have no effect on travel times, tolls, or trip costs.

Once again, we solve the applicable equations to determine the equilibrium travel time profile or toll schedule, and use those to find equilibrium trip costs. The equilibrium trip costs for the agents with desired arrival time $t_i^{\max}$ in each group $i \in \{A, B\}$ on route $r \in \{\text{free}, \text{toll}\}$ are

$$\bar{p}_{M,r}\left(A, t_A^{\max}\right) = \beta_A \left(N_A + N_B\right) \frac{1}{s_r} \frac{\xi}{1+\xi}, \tag{25}$$

$$\bar{p}_{M,\text{free}}\left(B, t_B^{\max}\right) = \alpha_B \left(N_A \frac{\beta_A}{\alpha_A} + N_B \frac{\beta_B}{\alpha_B}\right) \frac{1}{s} \frac{\xi}{1+\xi}, \tag{26}$$

$$\bar{p}_{M,\text{toll}}\left(B, t_B^{\max}\right) = \left(N_A \beta_A + N_B \beta_B\right) \frac{1}{s^*} \frac{\xi}{1+\xi}. \tag{27}$$

We can then define the trip costs for all the other agents in reference to (25)–(27):

$$\bar{p}_{M,r}\left(i, t^*\right) = \bar{p}_{M,r}\left(i, t_i^{\max}\right) - \left(t_i^{\max} - t^*\right) \begin{cases} \beta_i & t^* \leq t_i^{\max} \\ -\xi\beta_i & t^* > t_i^{\max} \end{cases} \text{ for } i \in \{A, B\}. \tag{28}$$

These are derived in Appendix F.1, where I also show that $t_A^{\max} = t_B^{\max}$, a result which also holds in every subsequent case.

The intuition behind (28) is that an agent with desired arrival time $t_i^{\max}$ is willing to arrive at the same time as an agent in his group with desired arrival time t^* and so the only difference in their trip cost is the difference in their schedule delay costs at that time.

Notice that (25) is the same as (12), and that (28) matches (13). The equilibrium trip cost for an agent who is willing to arrive at the start or end of rush hour is pinned down by the length of rush hour. It does not matter whether the other group's agents are all able to arrive at their desired arrival time and it does not matter whether the road is priced or free, except indirectly through the effect of pricing on road capacity. Furthermore, the preferences of the group arriving at the peak do not affect the equilibrium trip costs of the group arriving off-peak.

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

Lemma 5. *If there are two groups and two routes, one priced and one free, then the rich are never on the free route unless the poor are too, and the poor are never on the priced route unless the rich are too.*

Given these results, we can write down modified versions of the linear systems of equations above and solve for equilibrium trip prices for each of the eight value pricing cases.[33,34] I do so in Appendix F.2.

6.3. **When does congestion pricing generate a Pareto improvement?** While charging time-varying tolls can increase throughput by preventing the destructive effects of queuing, it also requires changing the currency used to allocate arrival times from time to money. Although both of these effects are Kaldor-Hicks efficiency-enhancing, changing the currency used hurts poor agents, and in particular it hurts poor inflexible agents. Whether pricing generates a Pareto improvement depends on whether the gains in throughput outweigh the harm from changing the currency for all agents.

Changing the currency hurts agents who are both inflexible and poor because the direct effect of changing the currency is that it makes desirable arrival times relatively cheaper for richer agents. This means a poor agent who had been traveling at the peak—that is, a poor agent who is also inflexible—either needs to pay more to outbid the rich agent to continue to travel at the peak, or travel further off-peak, thereby increasing his schedule delay. As a result, pricing all of the lanes will not always yield a Pareto improvement, as the next proposition shows.

Proposition 5 (Pricing all lanes is not always Pareto improving)**.** *If there are two groups and the mass of poor agents is not too large, then there exists a small enough ratio of the inflexibility of the rich to the inflexibility of the poor,* $(\beta_1/\alpha_1)\ /\ (\beta_2/\alpha_2)$*, such that pricing all of the lanes does not generate a Pareto improvement before the revenue is spent.*

The mass of poor agents is too large for this result to hold when

$$n_2 > s \quad \textit{and} \quad \frac{n_2}{n_1+n_2} > 1 - \frac{1-\frac{s}{s^*}}{1-\left(1-\min\left\{1,\frac{\beta_1}{\beta_2}\right\}\right)\frac{s}{s^*}}.$$

Consider an illustrative example where there are rich and flexible finance professors, and poor and inflexible retail store cashiers. When there are no tolls on the road the finance professors take advantage of their flexibility to avoid rush hour traffic by traveling before or after the peak. After all, they can start working once they get to their office, or work from home for a while and leave late. In contrast, the cashiers travel so as to arrive at work close to their desired arrival time; while they waste time sitting in traffic, that is not much different from getting to work early and wasting time waiting for their shift to

[33]Equilibrium can fall in one of eight cases depending on the parameters. The three dimensions in which the cases differ are (1) which group is not arriving off-peak, (2) whether some agents in this group are inframarginal or if they are all marginal, and (3) whether they are on one or two routes.

[34]In two of the cases the toll schedule or travel time profile is not completely defined by Lemmas 2 and 4 and so I use another indifference relation to characterize part of the toll schedule or travel time profile. The need to use this other indifference relation goes away when there is a continuum of groups.

start. Thus, when the road is unpriced, the cashiers travel at the peak of rush hour and the finance professors travel off-peak.

If the finance professors are sufficiently richer than the cashiers, then when we add tolls to all of the lanes of the highway the order of arrival reverses. The finance professor did not like waking up early to avoid traffic, but was willing to do so because he could start working as soon as he arrived at his office. Now by paying a toll to travel at the peak he can avoid both waking up early and sitting in traffic. Unfortunately, in switching from traveling off-peak to on-peak, the finance professor displaces the cashier, who must now travel off-peak.[35] Unless the increase in capacity due to pricing is large enough, the cashiers are worse off.

That said, if the rich are more inflexible than the poor, so that instead we have relatively poor yet flexible humanities professors, and rich yet inflexible lawyers, then pricing the entire road helps all road users. When the road is free, the flexible humanities professors wake up early to avoid traffic while the inflexible lawyers travel at the peak, putting up with traffic as the price of being on-time to their many meetings. Now when we add tolls the order of arrival does not change: the humanities professors still get to work early (they would rather show up early than pay a hefty toll) and the lawyers still travel at the peak, but are thrilled to pay a toll rather than sit in traffic. The increased capacity of the highway due to pricing means the humanities professors do not need to get to work quite as early and reduces the equilibrium tolls both groups pay. Everyone is better off.

Furthermore, even if we have rich and flexible finance professors and poor and inflexible cashiers, if the highway capacity was large enough when the road was free for all of the cashiers to arrive exactly on-time, and even some of the finance professors were able to arrive on-time (i.e., the cashiers were inframarginal), then by pricing just some of the lanes we can still obtain a Pareto improvement. We need to leave enough of the lanes unpriced so that all the cashiers can continue to travel on an unpriced route and arrive on-time. The finance professors who already had been traveling at the peak will travel on the priced lanes, and so none of the cashiers are displaced by finance professors shifting from off-peak to on-peak. Because we have priced some of the lanes, throughput is higher, rush hour shorter, and all agents are better off.

Combining these last two heuristic arguments suggests we can avoid the harm from congestion pricing if there are already some rich agents traveling at the peak of rush hour. This intuition is formalized in the following proposition.

Proposition 6 (Sufficient condition for pricing to generate a Pareto improvement)**.** *If there are two groups of agents, pricing can increase throughput* $(s^* > s)$*, and there are some rich agents*

[35]Alternatively, if the finance professors are not sufficiently richer than the cashiers, the cashiers will choose to outbid the finance professors for the right to travel at the peak of rush hour. However, doing so still leaves them worse off (unless the throughput drop is large enough).

traveling at the peak of rush hour when the road is free, then there exists a $\lambda_{toll} \in (0, 1]$ such that pricing λ_{toll} of the lanes generates a Pareto improvement even before the revenue is spent.

Casual empiricism finds that at the peak of rush hour the road is filled with both Lexus and Kias, and so the requirements of this proposition seem likely to hold. This suggests it is likely that value pricing will generate a Pareto improvement even before the revenue is spent. However, while Proposition 6 can be generalized beyond two groups, it also has a weakness: being able to price some fraction of the lanes does not mean we can price an economically significant fraction.[36] To show that this result is empirically and economically relevant, I will generalize the model to allow for a continuum of groups, then estimate the distribution of agent preferences, and use that distribution to estimate the welfare effects of pricing part or all of the lanes.

I fully characterize the set of parameters for which pricing part or all of the lanes generates a Pareto improvement in Appendix F.3. Because equilibrium trip costs take a different form depending on how A and B map into rich and poor, whether group B is marginal or inframarginal, and whether group B is on one or two routes, doing so requires solving 19 different cases. Proposition 7 highlights the most important additional results from doing so, where the phrase "more likely" is formally defined as follows.

Definition. Let $\mathcal{V}_{x=x_0}$ be the set of parameters for which outcome Z occurs given that parameter x has value x_0. If $x_0 < x_1 \Leftrightarrow \mathcal{V}_{x=x_0} \subset \mathcal{V}_{x=x_1}$, then outcome Z is *more likely as x increases*. Similarly, if $x_0 > x_1 \Leftrightarrow \mathcal{V}_{x=x_0} \subset \mathcal{V}_{x=x_1}$, then outcome Z is *more likely as x decreases.*

Proposition 7. *If there are two groups then pricing all or part of the lanes is more likely to generate a Pareto improvement prior to spending the toll revenue as*

- *the ratio of inflexibility of rich to poor $[(\beta_1/\alpha_1) \,/\, (\beta_2/\alpha_2)]$ increases,*
- *the throughput drop $(1 - s/s^*)$ increases, and*
- *income inequality (α_1/α_2) decreases.*

In addition

- *for any set of parameters there exists a throughput drop large enough such that pricing the entire road generates a Pareto improvement, and*
- *the set of parameter values such that pricing all of the lanes generates a Pareto improvement is a subset of the closure of the set of parameter values such that pricing a portion of the lanes generates a Pareto improvement.*

[36] To generalize, define the rich as the richest agents. The logic continues to hold that by leaving enough capacity for all those who are not rich who had been traveling at the peak to continue to do so we will generate a Pareto improvement.

7. Continuum of groups

In this section I solve the model with a continuum of groups. The main value of this section is that it provides the tools needed to show the theoretical possibility results from Section 6 are empirically relevant. I use the results of this section to estimate the distribution of agent preferences in Section 8 and then combine the results of this section with those empirical results to evaluate the impact of pricing all or part of the highway in Section 9.

I show that when the rich are more inflexible than the poor, pricing all of the lanes gives rise to a Pareto improvement, and so the intuition for the importance of the correlation between income and inflexibility continues to hold.

Working with a continuum of types will be easier if we adjust our notation slightly. We can still index groups by i, but it will often be easier to directly refer to a group by their value of time and inflexibility.[37] Define $\delta = \beta/\alpha$ as inflexibility, $n(\alpha, \delta, t^*)$ as the density of agents (which for simplicity I normalize to integrate to one), $n_\delta(\delta)$ as the marginal distribution of inflexibility, and $n_\beta(\beta)$ as the marginal distribution of β.

7.1. **Equilibrium when the road is completely free or priced.** We can find equilibrium trip costs for every agent by first assigning agents to arrival times using the algorithm from Section 4, then using Lemma 2 to find travel times, and then combining agents' travel times and schedule delays to find their trip costs. Along the way I confirm that rush hour has a single peak and that equilibrium trip prices, travel times, and tolls are unique. The details are in Appendix G.

I find the following closed form solutions (up to the possible need to solve the integrals numerically) for equilibrium trip costs on a completely free or priced route:

$$\bar{p}_{\text{free}}(\alpha, \delta, t^*) = \frac{\xi}{1+\xi}\frac{1}{s}\left[\alpha \int_0^1 \min\left\{\delta', \delta, \hat{\delta}\right\} n_\delta(\delta')\, d\delta'\right] - (t^{\max} - t^*)\,\alpha \min\left\{\delta, \hat{\delta}\right\} \begin{cases} 1 & t^* \leq t^{\max} \\ -\xi & t^* > t^{\max} \end{cases}, \tag{29}$$

$$\bar{p}_{\text{toll}}(\alpha, \delta, t^*) = \frac{\xi}{1+\xi}\frac{1}{s^*}\left[\int_0^\infty \min\left\{\beta', \alpha\delta, \hat{\beta}\right\} n_\beta(\beta')\, d\beta'\right] \tag{30}$$

[37] Now the set of groups $\mathcal{G}$ is the set of non-negative real numbers. We can use a real number to index a tuple of real numbers by interleaving the digits of each number in the tuple. Since $\beta/\alpha = \delta \in [0, 1]$ and $1 = 0.\bar{9}$, we can write a groups δ only using decimals, and so we interleave the decimal parts of a group's value of time and inflexibility to create the decimal part of their index, and let the integer part of a group's value of time be the integer part of their index.

$$- \left(t^{\max} - t^*\right) \min\left\{\alpha\delta, \hat{\beta}\right\} \begin{cases} 1 & t^* \leq t^{\max} \\ -\xi & t^* > t^{\max} \end{cases},$$

where $\hat{\delta}$ and $\hat{\beta}$ are the marginal type to arrive during $[t_s, t_e]$.

To see the intuition behind these expressions for trip cost, we can rewrite them as

(31)
$$\text{trip cost} = \frac{\xi}{1+\xi} \times \text{length rush hour} \times \begin{array}{c}\text{censored mean of} \\ \text{willingness to pay}\end{array} - \begin{array}{c}\text{adjustment for} \\ \text{desired arrival time}\end{array}.$$

Let us work through each term of (31). The ratio $\xi / (1+\xi)$ is a measure of how the cost of being late compares to the cost of being early and is the fraction of agents who arrive before the peak of rush hour. If ξ is zero then it is costless to be late, as a result agents can wait to travel until there is no traffic or toll; everyone will be late and have a trip cost of zero. As ξ increases the costs of being late increases and so a larger share of agents arrive before the peak. Because agents care more about arriving on-time, travel times (or tolls) are higher and everyone's trip cost increases.

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

The final factor of the first term is the most interesting; the integrals in (29) and (30) are the censored mean of an agent's willingness to pay in the currency the route requires. On a free route it is the censored mean of the distribution of inflexibility, or willingness to pay in travel time to reduce schedule delay, while on the priced route it is the censored mean of the distribution of willingness to pay in dollars to reduce schedule delay. On the free route we then multiply this by the agent's value of time to convert from travel time to dollars.

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

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

We can use the expressions for trip costs, (29) and (30), to show that when the rich are more inflexible than the poor, then pricing all of the lanes generates a Pareto improvement. This helps confirms that our intuition that the correlation between income and inflexibility is a crucial parameter in determining whether pricing generates a Pareto improvement carries over from two groups to a continuum of groups.

Proposition 8. *If the inflexibility and value of time are perfectly rank correlated, so that if one agent is richer than another then he is also more inflexible than the other, then pricing all of the lanes of a highway generates a Pareto improvement before the revenue is spent.*

7.2. **Equilibrium when value pricing.** In contrast to the case where there were just two groups, I must solve for the value pricing equilibrium numerically instead of analytically. To do so, I first assign agents to routes and then solve for the equilibrium on each route. Solving numerically requires me to use several approximations, which I choose so I can use the closed-form solutions for trip prices on a completely free or priced highway, (29) and (30), to find equilibrium trip prices on a route given the agents who are on it.

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

Lemma 6. *For a given flexibility and desired arrival time there is a value of time, $\hat{\alpha}(\delta, t^*)$ such that all agents with a higher value of time travel on the priced route and all agents with a lower value of time travel on the free route. Furthermore, $\hat{\alpha}$ is a continuous function if the travel time profile and toll schedule are continuous.*

It is unlikely that after conditioning on route choice the distribution of desired arrival times will be uniform and independent of α and δ. This means that $\hat{\delta}$ and $\hat{\beta}$ need not be constant over $[t_s, t_e]$; however, in practice they are nearly constant and so I approximate them with a constant. The largest approximation error in $\hat{\delta}$ and $\hat{\beta}$ ranges from 0.2 to 2.9 percent across my main three specifications.[38] Making this approximation allows me to apply (29) and (30) to each route individually, adjusting for route capacity and the distribution of agents on the route.[39] Given the small size of the approximation error and how much it helps in solving for the equilibrium, this approximation seems reasonable.

Given the approximation of $\hat{\delta}$ and $\hat{\beta}$ over $[t_s, t_e]$ I can further simplify $\hat{\alpha}$ using the next lemma, which shows that $\hat{\alpha}$ is often flat in one dimension.

Lemma 7. *All agents in a group that is not inframarginal regardless of which route they are on, travel on the same route or are indifferent between both routes, that is,*

[38]Specifically, given the equilibrium I have found, I find the marginal δ and β for each $t \in [t_s, t_e]$ and compare it to $\hat{\delta}$ and $\hat{\beta}$.

[39]See Appendix G.3 for details.

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

Similarly, all agents who are inframarginal regardless of which route they are on and who have the same value of time and desired arrival time, travel on the same route or are indifferent between both routes, that is,

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

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

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

Based on Lemma 7, I approximate $\hat{\alpha}\left(\delta, t^*\right)$ as

$$\hat{\alpha}\left(\delta, t^*\right) = \begin{cases} \hat{\alpha}_M\left(\delta\right) & \delta < \hat{\delta} \\ \hat{\alpha}_I\left(t^*\right) & \delta \geq \hat{\delta} \end{cases}, \tag{32}$$

where $\hat{\alpha}_M\left(\delta\right)$ and $\hat{\alpha}_I\left(t^*\right)$ are solved for using Chebyshev collocation. This approximation performs significantly better than the two dimensional Chebyshev approximation of $\hat{\alpha}\left(\delta, t^*\right)$: in my main specifications the approximation error[40] is less than a tenth of a cent using (32) with tenth degree Chebyshev polynomials, for twenty nodes total, while the approximation error is nearly a dollar using the tensor product of two tenth degree Chebyshev polynomials, for one hundred nodes in total.[41]

To solve for the equilibrium I find the type that is the marginal agent during $[t_s, t_e]$ on each route ($\hat{\delta}$ on the free route and $\hat{\beta}$ on the priced route), as well as the function $\hat{\alpha}\left(\delta, t^*\right)$ that separates the space of agent preferences by which route they are on, by solving the following set of equations numerically:

[40]Measured as the largest welfare loss from traveling on the route assigned by $\hat{\alpha}\left(\delta, t^*\right)$ instead of the route that actually minimizes trip cost.

[41]Using (32) has worse asymptotic properties than the two dimensional Chebyshev approximation of $\hat{\alpha}\left(\delta, t^*\right)$, as it will not converge to the true $\hat{\alpha}\left(\delta, t^*\right)$ over the small area where $\left(\delta - \hat{\delta}\right)\left(\hat{\alpha}\left(\delta, t^*\right)\delta - \hat{\beta}\right) < 0$, regardless of the degree of the Chebyshev polynomial.

$$\int_{t_s}^{t_e}\int_{\hat{\delta}}^{1}\int_{0}^{\hat{\alpha}(\delta,t^*)} n(\alpha,\delta,t)\, d\alpha d\delta dt = (1-\lambda)\, s(t_e - t_s),$$

$$\int_{t_s}^{t_e}\int_{0}^{1}\int_{\max\left\{\hat{\alpha}(\delta,t^*),\delta^{-1}\hat{\beta}\right\}}^{\infty} n(\alpha,\delta,t)\, d\alpha d\delta dt = \lambda s^*(t_e - t_s),$$

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

where $\mathcal{C}$ is the set of Chebyshev collocation nodes. In Appendix G.4 I show there is a unique solution to this set of equations.

8. Estimating the distribution of consumer preferences

The theoretical analysis above makes clear that whether value pricing can make all road users better off depends on agents' preferences. We now turn to estimating the distribution of agents' preferences, along with other relevant parameters. We will then combine these results with those in Section 7 to evaluate the distribution and size of the welfare gains from congestion pricing in Section 9.

The main structural object I estimate is the joint distribution of agents' inflexibility, value of time, and desired arrival time. My general approach is to split the population into two categories using a measure of whether an agent is, broadly speaking, flexible or inflexible; then, within each category I estimate the marginal distributions of the three preference parameters. Having done so, I combine these marginal distributions into a joint distribution by assuming each preference parameter is independent of the others.[42] This means the correlations between preference parameters manifest themselves through differences in the marginal distributions *between* categories, rather than through the correlations *within* a category.

8.1. Data.

I estimate this joint distribution for road users on a segment of California State Route 91 (SR-91). The segment I focus on is thirty-three miles long and runs from the center of Corona to the junction of SR-91 and I-605. I choose this specific segment because it roughly represents a median commute for those living in Corona who use SR-91.

I use data from three sources. The first, California Polytechnic State University's State Route 91 Impact Study (Sullivan, 1999), is a series of surveys of road users who use SR-91 conducted between 1995 and 1999. I use this data set to estimate the fraction of agents who are flexible, the distribution of the value of time for each category, and the distribution of desired arrival times for the inflexible category.

The second data set is the 2009 National Household Travel Survey ("NHTS", U.S. Department of Transportation, 2009), which I use to confirm that my estimates from

[42] In Appendix H I conduct two tests of the assumption of independence of marginal distributions within a category. In both cases I fail to reject the hypothesis that they are independent.

TABLE 1. Fraction of drivers and trips that are flexible

Fraction of ...	SR-91	Large MSAs
Drivers who typically leave early or late to avoid traffic	.57 [.55, .60]	
Trips on interstate during morning where drivers can choose arrive time	.43 [.40, .47]	.35–.60 [.32, .62]
Trips on interstate to work where drivers can choose arrival time	.50 [.47, .53]	.47 [.45, .49]

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

the SR-91 Impact Study are similar to what I would estimate for other large metropolitan statistical areas (MSAs).[43]

The final data set is the California Department of Transportation's Performance Measurement System ("PeMS", 2014). PeMS includes road detector data from almost all of the highways in California. From this data set, I calculate travel times for every business day in 2004.[44] I use these travel times to estimate the distribution of inflexibility.

8.2. **Fraction of road users who are flexible.** The first task is to estimate the relative sizes of the two categories of agents.

The SR-91 Impact Study contains two measures of flexibility: one focuses on the driver's *typical* trip, the other on a *specific* recent trip. The first row of Table 1 shows that 57 percent of road users on SR-91 report that they typically leave early or late to avoid traffic congestion, and the second row shows that 43 percent of road users could choose what time they arrived at their destination for a specific peak period trip. The strength of the first measure is that it asks whether the road user takes an action which reveals their flexibility, while the strength of the second is that it is about a specific recent trip and so better reveals the fraction of road users on a given morning who are flexible.

While the NHTS does not ask the same questions as the SR-91 Impact Study, it does allow me to make some comparisons between drivers on SR-91 and those in other large MSAs. The NHTS only asks individuals if they can choose their arrival time for work trips, rather than for all trips. For the sake of making a clean comparison between drivers on SR-91 and in the rest of the United States, the third row of Table 1 reports estimates of

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

[44]I define a business day as a weekday which is not one of the ten United States federal holidays.

how many road users can choose when to arrive at work from both data sets.[45] I find that the fraction of workers who are flexible on SR-91 is fairly similar to the fraction in other large MSAs.

I can estimate the fraction of all trips during the morning that are flexible using the NHTS if I make assumptions about what kinds of non-work trips are flexible.[46] Doing so leads to the range of estimates reported in the second column of the second row. The bottom of the range comes from assuming no non-work trips are flexible, while the top comes from assuming that other trips where the driver probably has control over when it begins (such as shopping, doctor's appointments, and visiting friends) are flexible. My estimate of the fraction of trips on SR-91 that are flexible falls roughly in the middle of the range of estimates for large MSAs.

I use the specific-trip measure as my definition of which agents are in the flexible category. Doing so gives more conservative estimates for the maximum fraction of the lanes we can price and still obtain a Pareto improvement, as well as for the private welfare gains from pricing a given fraction of the lanes. All other results are largely unaffected by which definition of flexibility I use.[47]

8.3. **Distribution of the value of time.** To estimate the distribution of the value of time I first map household income into value of time and then fit a log-normal distribution to the data using maximum likelihood. I do this separately for the flexible and inflexible road users.

To map household income to value of time, I use the following U.S. Department of Transportation formula: an individual's value of time is half their hourly household income, which is their annual household income divided by 2,080 hours per year (Belenky, 2011, p. 12).[48,49] While it would be preferable to use annual individual income, or better yet, individual wages, the SR-91 Impact Study and NHTS do not contain this information.

Using this formula means I underestimate the welfare gains from congestion pricing and overstate the difficulty in obtaining a Pareto improvement. This is due to two standard results in the literature on the value of travel time. The first is that drivers particularly dislike traveling in congested traffic, valuing a reduction in it at above half their wage (Small and Verhoef, 2007, p. 53), which means I underestimate the welfare gains. The

[45] I limit the NHTS sample to those who travel on the interstate, so that they are similar to those traveling on SR-91. While SR-91 is not an interstate, it is a limited access highway and so indistinguishable in all but signage from an interstate.

[46] The morning is defined in the SR-91 Impact Study as 4–10 a.m. and so for consistency I maintain that definition with the NHTS. I also continue to limit the NHTS sample to those who travel on the interstate.

[47] The results using the typical-trip measure of flexibility are reported in Appendices H and I.

[48] The U.S. Department of Transportation uses this formula to estimate a median value of time based on median household income, I am going further in using it by applying it to individuals.

[49] There is a large literature estimating the mean or median value of time, which generally finds it is half the mean or median wage, though it is higher when roads are congested. See Small and Verhoef (2007, p. 53) for a literature review.

TABLE 2. Distribution of value of time for morning highway users

	SR-91	Large MSAs
Flexible		
Median	25.95	26.05
	(0.90)	(0.34)
Interquartile range	20.0	32.21
	(1.8)	(0.89)
N	303	7,059
Inflexible		
Median	22.16	22.52
	(0.56)	(0.27)
Interquartile range	15.19	24.55
	(0.87)	(0.59)
N	433	4,270
Rank correlation†	0.200***	0.157***
	(0.055)	(0.037)

Notes: Standard errors in parentheses. Standard errors for SR-91 estimates are calculated by bootstrapping. The data for large MSAs are weighted using individual weights and their standard errors calculated using jackknife-2 replicate weights. I convert household income to value of time using a formula from the USDOT (Belenky, 2011), adjust dollar amounts to 2012 dollars using the CPI, and then fit the categorical data to a log-normal distribution using maximum likelihood. For large MSAs I use the most generous definition of flexibility, which assumes certain non-work trips are flexible.

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

*** $p < .001$

second is that value of time does not increase proportionally with income (ibid.), and so by assuming it does increase proportionally with income I overestimate the variance in value of time. This makes it appear harder to find a Pareto improvement.

I fit a log-normal distribution to the data using maximum likelihood. To write the likelihood function, define y_i as the income category for observation i, and $h(y_i)$ and $l(y_i)$ as the highest and lowest incomes within category y_i. Further define c as the function that converts household income into an estimate of value of time using the formula from Belenky (2011) and adjusts for inflation using the consumer price index, using 2012 as the base year, and F as the cumulative distribution function for a log-normal distribution which depends on the parameter vector $\vec{\theta}$. Using this notation I write the likelihood of observing the data as

$$\mathcal{L}\left(\vec{\theta}|\vec{y}\right) = \prod_{i=1}^{N}\left[F\left(c\left(h\left(y_i\right)\right)|\vec{\theta}\right) - F\left(c\left(l\left(y_i\right)\right)|\vec{\theta}\right)\right]. \tag{33}$$

I estimate the parameters of the distribution of the value of time by finding the parameters $\vec{\theta}$ which maximize (33).

I estimate the distribution of the value of time separately for the flexible and inflexible categories, as mentioned earlier. While the levels of these estimates matter for valuing the time saved by congestion pricing, as Propositions 7 and 8 show, it is the correlation between value of time and flexibility that affects our ability to obtain a Pareto improvement.[50]

The first column of Table 2 reports the results from estimating the distribution of the value of time using the SR-91 Impact Study. I find that flexible agents on average have higher values of time than the inflexible. The Goodman and Kruskal's rank correlation between flexibility and income (reported in the last row) is 0.20, which can be roughly interpreted as meaning that three times out of five a randomly selected flexible agent will have a higher value of time than a randomly selected inflexible agent.[51]

The last column in Table 2 present the results from estimating the distribution of the value of time in large MSAs using the NHTS. In this column I use the most generous form of the NHTS definition of flexibility, which assumes certain non-work trips are flexible.

As with the estimates of the fraction of agents in the flexible category, the results for SR-91 are similar to other large MSAs. Comparing both columns of Table 2 shows that we obtain similar estimates for the median value of time for each category, and a relatively similar rank correlation. While my estimates of the interquartile range are much larger for large MSAs than for SR-91, this is probably because the SR-91 Impact Study only contains people who choose to live in Riverside County and commute on SR-91. This smaller group is likely to be more homogeneous than the larger group of those who live in a large MSA.

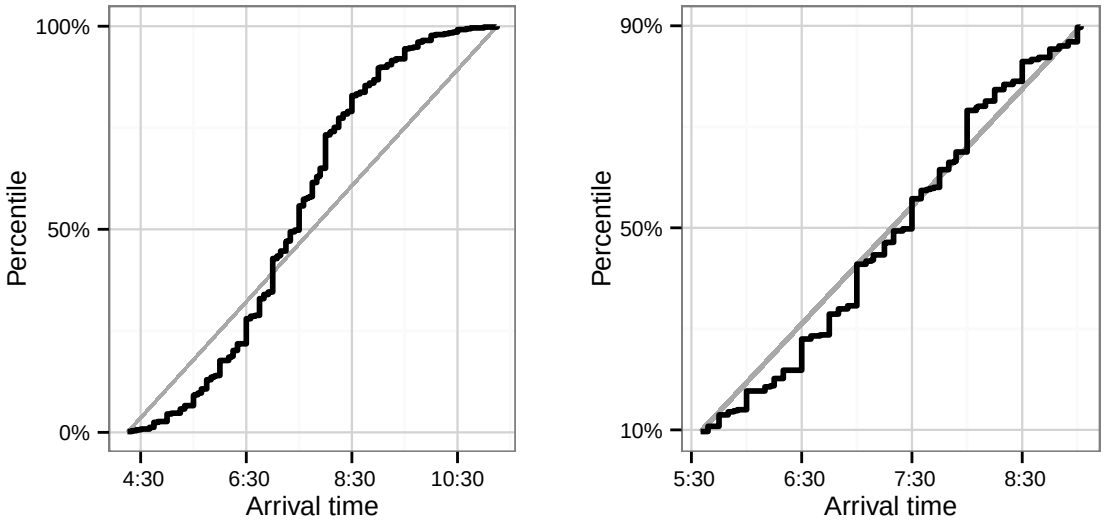
As a benchmark, we can compare my estimates of the distribution of the value of time to those of Small, Winston, and Yan (2005), who use more detailed data and more sophisticated methods to measure the distribution of value of time for road users on SR-91. While they do not estimate the distribution separately for flexible and inflexible agents, I can compare their results to those from my relatively simple method when pooling the flexible and inflexible agents. Adjusting for inflation, they find that the median value of time is \$29.54 and the interquartile range is \$10.47, while I find a median of \$23.58 and an interquartile range of \$17.06. As expected, I underestimate the median and overestimate the interquartile range.

8.4. **Distribution of desired arrival time.** In this subsection, I provide evidence that my assumption that desired arrival times are uniformly distributed is a reasonable approximation to the truth, as well as estimate the length of time over which agents desire to arrive.

What I care about is the time when agents *desire* to arrive at the *highway exit*, but what I observe in the data is when they *actually* arrive at their *destination*; this means the data

[50]To be more precise, multiplying everyone's income by the same scaling factor would not change our ability to obtain a Pareto improvement.

[51]This interpretation would be exact if there were no ties in the data.



(A) All data

(B) 10th–90th percentiles

FIGURE 2. Empirical cumulative distribution function of arrival time for agents who cannot choose when they arrive at their destination and arrive before noon. Data from SR-91 Impact Study.

is informative about the shape and spread in the distribution of desired arrival times for inflexible agents, but not for the position of the distribution and not for flexible agents.

Because I observe *actual* arrival times instead of *desired* arrival times, I must focus on those whose actual arrival times are their desired arrival times—that is, those in the inflexible category, who are unable to choose their arrival time.

Because I observe arrival times at agents' *destinations* rather than at the *highway exit*, the underlying distribution I want to recover (the distribution of inflexible agents' desired arrival times at the highway exit) is a smoothed and shifted version of the distribution I observe (the distribution of inflexible agents' actual arrival times at their destination). The distribution is smoothed because the distance agents must travel from the end of the highway to their destinations varies, and so among those who want to reach their destination at 7:00 a.m., there are some who want to reach the end of the highway at 6:40 a.m. and others who want to reach it at 6:55 a.m. The distribution is shifted because agents want to exit the highway earlier than they want to arrive at work. To correct for this shift, I estimate the first desired arrival time, t_s, as part of the structural estimation of the distribution of flexibility.

To test whether the distribution of desired arrival times is uniformly distributed, I compare the cumulative distribution function (CDF) of a uniform distribution to the empirical CDF of desired arrival times for the inflexible agents using SR-91. I do so in

Figure 2a. If the distribution were uniform then the empirical CDF would lie along the 45 degree line; it is clear that the distribution of desired arrival times is not uniform. However, when we remove the first and last 10 percent of road users to arrive, as in Figure 2b, then the distribution is close to being uniform. A remaining difference is that the empirical CDF is not as smooth as that of a uniform distribution, but, as discussed in the last paragraph, this is exactly what we expected to find. These patterns are robust, and hold within the NHTS data as well.[52]

Truncating the extreme deciles is relatively innocuous. By doing so I am ignoring agents who want to arrive extremely early or late. Some of these agents are arriving outside of rush hour, and so they are not relevant for my analysis. The rest are among those who are least harmed by congestion pricing; they are already traveling at undesirable times, and so will not be displaced. Furthermore, because congestion pricing can reduce the length of rush hour, they may find that after pricing they are traveling outside of rush hour and so face no congestion or toll. Should congestion pricing help those who want to arrive at the peak of rush hour, it almost certainly helps those who want to arrive at the tails.

I estimate the spread in the distribution of desired arrivals times by matching the largest and smallest remaining observation to the expected value of their order statistics. I show in Appendix H.5 that this procedure gives me an unbiased estimate of the length of time over which inflexible agents wish to arrive, $t_e - t_s$. I estimate that $t_e - t_s$ is 4.40 hours, as is reported in the first row of the first column of Table 4.

As I do not observe the distribution of desired arrival times for those who are flexible, I assume it is the same as the distribution for those who are inflexible. This assumption is relatively harmless. When an agent is marginal, his desired arrival time determines whether he is early or late, but not his actual arrival time. Ascribing the wrong desired arrival time to an agent who is always marginal will not affect the equilibrium or the change in the agents' trip costs due to congestion pricing.[53] Fortunately, most agents in the flexible category are always marginal.

8.5. **Distribution of inflexibility.** The bottleneck model provides a mapping between model parameters and the travel time profile (TTP). By inverting this mapping, I estimate the remaining parameters: the distribution of inflexibility, $N_\delta(\delta)$; the ratio of the cost of being early to late, ξ; the length of rush hour on a free route, $1/s$; the first desired arrival time at the highway exit, t_s; and free flow travel times, T_f.

I am only able to estimate the distribution of inflexibility for those road users who do not arrive on-time. For all other road users I only obtain a lower bound. This follows from Lemma 2 and is due to the fact that the TTP does not reflect the preferences of the inframarginal road users.

[52]In Appendix H.4 I recreate Figure 2 for all of the United States, Los Angeles, and New York City, and find the same results.

[53]It will affect the level of their trip costs, but in a consistent way so that it differences out.

Because I am unable to observe a portion of the distribution of inflexibility, I make assumptions about its shape, and then test the sensitivity of my results to these assumptions. I assume that the distribution of inflexibility for those agents who are in the flexible category, N_δ^f, is uniform on $\left[0, \bar{\delta}\right]$; and the inflexibility of those in the inflexible category, N_δ^i, has a beta$(5, 0.5)$ distribution transformed to have support $\left[\tilde{\delta}, 1\right]$. This means I am assuming most of the inflexible agents are very inflexible, as most of the weight of N_δ^i is near one and its mode is one. The assumed form of N_δ^i does not affect the estimation of the other parameters; however, it does affect the counterfactual results, and so Appendix I reports the counterfactual results for a wide variety of assumptions about N_δ^i.

I estimate the remaining parameters $\vec{\theta} = \left\{\bar{\delta}, \xi, s, t_s, T_f\right\}$ by using the Generalized Method of Moments (GMM) to choose $\vec{\theta}$ to best match the model-predicted TTP to the empirical TTP calculated from the PeMS data set. The estimation uses my estimates of the fraction of agents who are flexible, ϕ, and the length of desired arrivals, $t_e - t_s$, from Sections 8.2 and 8.4. While I could estimate ϕ and $t_e - t_s$ as part of the GMM routine, and doing so would allow me to match the TTP better, I estimate them separately because I have natural measures of them and want to match those particular "moments" exactly.

Letting x_{ij} be the observed travel time on SR-91W from the center of Corona to the junction of SR-91 and I-605 for arrival time i on day j, I can then write the moment conditions as

$$T\left(t_i; \bar{\delta}, \xi, s, t_s, \phi, t_e - t_s\right) + T_f = \sum_j x_{ij}/250 \quad \text{for } t_i \in \{\text{4:00, 4:05}, \ldots, \text{ 10:00}\},$$

where $T\left(t_i; \bar{\delta}, \xi, s, t_s, \phi, t_e - t_s\right)$ is defined by (G.19) in Appendix G.

While each parameter in $\vec{\theta}$ is chosen to best match the model's predicted TTP to the empirical TTP, each parameter also directly maps into a particular feature of the predicted TTP, and thus the estimate for each parameter is strongly affected by a particular feature of the empirical TTP. These relationships are reported in Table 3.

To test the restrictiveness of my functional form assumptions, I fit a relaxed version of the model to the data non-parametrically. After relaxing the functional form assumptions for the distributions of inflexibility and desired arrival times, as well as the assumption that the ratio of the cost of being late to early is the same for all agents, the theory still imposes three sets of constraints on travel times. Letting $i^{\max}$ index the start of the five-minute period in which the peak of rush hour occurs, the constraints are as follows:

(1) Travel times are positive:

$$T_i > 0 \ \forall i.$$

(2) Travel times are increasing before the peak and decreasing after:

$$T_i \geq T_{i-1} \quad \forall i \leq i^{\max} \quad \text{and}$$

TABLE 3. Which features of the data identify which parameters

Parameter	Notation	Feature of data which identifies parameter
Distribution of inflexibility	$n_\delta(\delta)$	Distribution of the slope of the travel time profile
Ratio of schedule delay costs late-to-early	ξ	Ratio of the average slope after the peak to the average slope before the peak
Free flow travel time	T_f	Average travel time between 4 a.m. and when travel times start climbing
Length of rush hour	$1/s$	Length of time when travel times are above T_f
First desired arrival time	t_s	Time when the slope of the travel time profile stops changing because the marginal type becomes constant at t_s when desired arrival times are uniformly distributed

$$T_i \leq T_{i-1} \quad \forall i > i^{\max}.$$

(3) Travel times are convex before the peak and convex after the peak:

$$\frac{T_i - T_{i-1}}{t_i - t_{i-1}} \geq \frac{T_{i-1} - T_{i-2}}{t_{i-1} - t_{i-2}} \qquad \forall i \notin \{i^{\max}+1, i^{\max}+2, i^{\max}+3\}.$$

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

To fit the relaxed model to the data non-parametrically, I find the travel times, T_i, as as well as the index of the five-minute window in which the peak of rush hour occurs, $i^{\max}$, which minimize the GMM criterion subject to the three sets of constraints above. I can then use the predicted TTP from the non-parametric estimation and the relationships from Table 3 to non-parametrically estimate $t_s, t_e - t_s, 1/s, \xi$, and T_f.[54]

Table 4 reports the GMM and non-parametric estimates, which are very similar. In particular, the non-parametric estimate of the length of desired arrivals is almost the same the estimate from Section 8.4, despite the fact that it is estimated from the TTP rather than survey data.

The *length of rush hour* is the period of time when travel times are higher than they would be in free flow conditions, not just when they are exceptionally long. Using this definition, I estimate that rush hour is more than seven and a half hours long, starting before five in the morning and not ending until a little after noon.

I estimate that the inflexibility of those in the flexible category is uniformly distributed on $[0, 0.228]$ and that the ratio of the cost of being late to the cost of being early is 0.4. This last result means the cost of being late is *less* than the cost of being early; while this appears unreasonable, it is largely a result of how I estimate this ratio, and is best interpreted as

[54]See Appendix H.6 for details on the non-parametric estimation.

TABLE 4. Remaining parameter estimates

	GMM	Non-parametric
Length of desired arrivals (hours) $(t_e - t_s)$	4.40	4.33
	(0.22)	(0.22)
First desired arrival time (hours) (t_s)	5.556	5.333
	(0.063)	(0.068)
Length of rush hour on free route (hours) $(1/s)$	7.74	8.00
	(0.40)	(0.33)
Maximum inflexibility of flexible agents $(\bar{\delta})$	0.228	—
	(0.042)	
Ratio of schedule delay costs late to early (ξ)	0.411	0.403
	(0.033)	(0.039)
Free flow travel time (minutes) (T_f)	36.71	36.69
	(0.90)	(0.77)

Note: Bootstrapped standard errors in parentheses. The estimate in the first element of the first row comes from fitting the largest and smallest observations of the trimmed sample of the inflexible agents' desired arrival times to the expected value of their order statistics (N = 488). The last five rows of the first column report the GMM estimates (N = 250). The second column reports non-parametric estimates, which come from finding the predicted travel times that best meet a mimimum set of restrictions implied by the model, and then estimating parameters from these predicted travel times (N = 250). The non-parametric estimates of t_s and $t_e - t_s$ require assuming desired arrival times are uniformly distributed.

saying that the marginal driver who is late pays lower schedule delay costs than the marginal driver who is early; there is nothing unreasonable about this. Furthermore, being late does not necessarily mean literally arriving late to an appointment, but can mean you would prefer to go to the doctor at 9 a.m. but instead schedule the appointment for 11 a.m. to avoid traffic. You arrive exactly on-time to your 11 a.m. appointment, but still have schedule delay costs.

The empirical TTP along with the predicted TTPs from both methods are shown in Figure 3. The two predicted TTPs both match the data well and are difficult to tell apart.[55] Making the additional assumptions about functional forms only increases the root GMM criterion by 7.7 percent. The small difference in the root GMM criterion of the non-parametric and GMM estimates, as well as the similarity in their parameter estimates, suggests that it is innocuous to make these additional assumptions.

9. COUNTERFACTUALS

Given the estimated distribution of driver preferences from Section 8, we can use the results of Section 7 to solve for the equilibrium under counterfactual congestion

[55]The two predicted TTPs differ the most at 5:00 and 10:00 a.m; the difference at 5:00 is due to the assumption that the marginal distribution of inflexibility of those in the flexible category is uniformly distributed, and the difference at 10:00 largely results from not imposing the assumption that $\gamma_i = \xi\beta_i$ for all groups i.

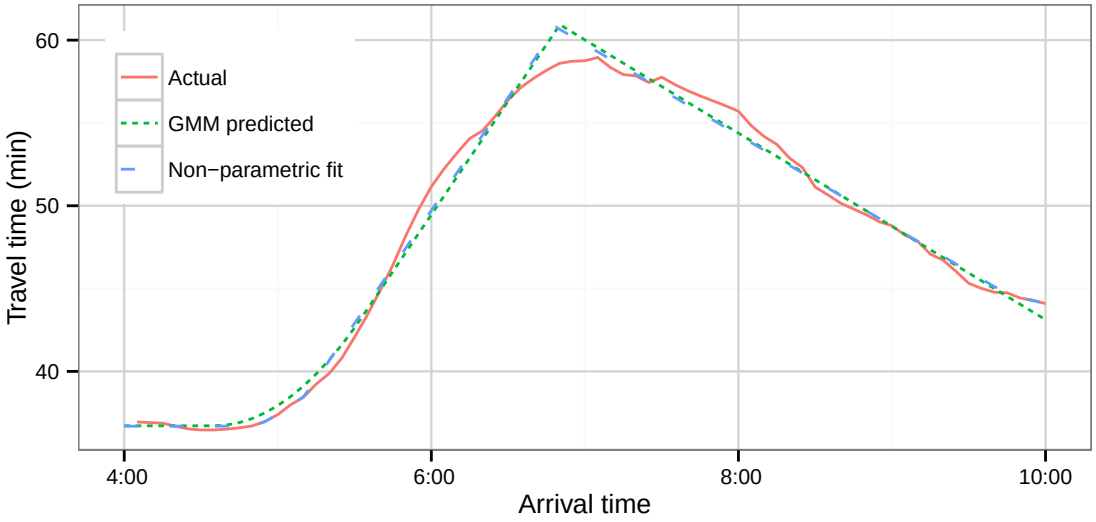


FIGURE 3. Actual versus predicted travel times. Data from PeMS.

pricing regimes. This allows me to estimate the aggregate welfare effects as well as the distributional effects of pricing a portion or all of the lanes. I conclude this section with a variety of sensitivity checks on these results.

The final parameter needed to evaluate counterfactuals is the amount throughput falls once a queue forms. The traffic engineering literature summarized in Section 2 estimates that queues at bottlenecks reduce throughput by roughly 10 percent, while queue spillovers reduce throughput by 25 percent. I report results for both of these levels of throughput drop, as well as the midpoint between them. For the sake of the discussion, I focus on the middle case, though the same patterns in the results hold for all three levels of the throughput drop.

9.1. **Aggregate welfare effects.** Table 5 reports the largest annual welfare loss, the average annual welfare effects, and the decomposition of the welfare effect of pricing all or part of the lanes of the highway. The headline result is that pricing a portion of the lanes helps all road users, while pricing all of them significantly hurts some road users. Pricing a portion of the lanes generates a Pareto improvement, while pricing all of the lanes does not.

While pricing all of the lanes raises significant revenue, it is not enough for a uniform rebate to make pricing the entire road Pareto improving. The worst-off agent is hurt by $2,390 per year, which is 70 percent larger than annual toll revenue per capita. This means using the revenue to make pricing the entire road generate a Pareto improvement requires spending it in a way that targets those who are harmed.

Value pricing generates a Pareto improvement even before using the revenue; however, doing so requires giving up some of the potential social welfare gains. By not pricing

TABLE 5. Average annual welfare effects of congestion pricing

Size of throughput drop (%)	10		17.5		25	
Fraction of lanes priced	1	0.25	1	0.5	1	0.5
Largest welfare loss ($)†	3420	0	2390	0	1590	0.0
	(420)	(120)	(320)	(55)	(290)	(0.0021)
	[0.0]	[0.74]	[0.0]	[0.94]	[0.01]	[1.0]
Welfare gains ($)						
Social	2270	1010	2400	1740	2510	1910
	(280)	(140)	(290)	(230)	(290)	(250)
Private	490	310	1080	760	1580	1020
	(240)	(85)	(240)	(150)	(290)	(200)
Reduction in travel time (hours)	76.5	18.1	76.5	40.5	76.5	45.9
	(9.1)	(3.5)	(9.1)	(6.0)	(9.1)	(7.0)
Reduction in travel time costs ($)	1960	820	1960	1390	1960	1490
	(250)	(130)	(250)	(200)	(250)	(210)
Reduction in schedule delay (hours)	113.6	30.7	199	108.9	284	162
	(7.3)	(2.0)	(13)	(7.0)	(18)	(10)
Reduction in schedule delay costs ($)	305	195	438	342	543	417
	(41)	(28)	(52)	(42)	(63)	(50)
Tolls Paid ($)	1780	700	1320	970	930	888
	(200)	(82)	(160)	(110)	(150)	(98)

Notes: Bootstrapped standard errors in parentheses. The fraction of bootstrapping iterations for which pricing a given fraction of the road yields a Pareto improvement is in brackets. I assume two trips per working day and 250 working days per year. Social welfare gains are the sum of the reduction in travel time costs and the reduction in schedule delay costs; and they do not include the value of saving gasoline or reducing pollution. Private welfare gains are social welfare gains minus the private cost of the tolls paid. Numbers in the table do not add up exactly due to rounding.

† The largest welfare loss is not an average, but the maximum annual welfare loss.

all of the lanes we leave some lanes congested and with lower throughput; this costs 30 percent of the social welfare gains available from congestion pricing. That said, if by making congestion pricing yield a Pareto improvement we are able to actually implement congestion pricing then we are trading $660 per person per year of potential, unrealized, welfare gains for $1,740 per person per year of actual welfare gains.

The welfare gains available from congestion pricing are large; even in the conservative case they are over $1,000 per agent per year. In the middle case pricing half of the lanes would be equivalent to increasing the median income of these agents by over 3.5 percent, and pricing all of the lanes would increase median income by over 5 percent. Most of the welfare gain comes from changing the currency used to pay for desired arrival times from time to money. The time spent in traffic is a social loss while the money spent on tolls is just a transfer. Most of this portion of the welfare gains accrues to whomever gets to keep

TABLE 6. Travel times and tolls

Size of throughput drop (%)		10		17.5		25	
Fraction of lanes priced	0	1	0.25	1	0.5	1	0.5
Excess travel times (min)							
Average	9.2	0	9.6	0	9.6	0	8.58
	(1.1)	(0)	(1.1)	(0)	(1.0)	(0)	(0.93)
Peak	23.3	0	23.1	0	22.2	0	19.7
	(2.6)	(0)	(2.5)	(0)	(2.4)	(0)	(2.2)
Toll ($)							
Average	0	3.56	5.18	2.64	3.55	1.86	3.11
	(0)	(0.41)	(0.60)	(0.32)	(0.39)	(0.31)	(0.34)
Peak	0	9.2	12.6	6.48	8.55	4.36	7.32
	(0)	(1.2)	(1.4)	(0.87)	(0.93)	(0.77)	(0.85)

Notes: Bootstrapped standard errors in parentheses. Averages are calculated over agents, not over time.

the toll revenue. However, a significant amount of the welfare gains goes to the road users themselves. Even if the toll revenue is wasted the average road user will be $760 better off each year due to value pricing.

Value pricing captures a large portion of the welfare gains available, even though we are pricing only half of the lanes. This contrasts with Verhoef, Nijkamp, and Rietveld (1996), Liu and McDonald (1998, 1999), and Verhoef and Small (2004) who find that pricing part of the highway yields a less-than-proportional share of the welfare gains available from congestion pricing. The logic behind their result is that if congestion pricing reduces throughput, then by pricing agents out of the priced lanes we make traffic worse in the free lanes. This leads to lower tolls and more congestion on the priced route, and so to a smaller share of the welfare gains. In contrast, when pricing increases throughput we do not have this additional concern.

It is perhaps surprising that pricing half of the lanes captures more than half of the available social welfare gains. As we would expect, pricing half of the lanes reduces travel times by roughly half the amount as pricing all of the lanes. However, because those with a high value of time are choosing to travel on the priced lanes we save the most valuable half of the travel time, and so capture over 70 percent of the value of the reduction in travel time. The same pattern repeats itself with the reduction in schedule delay. This allows us to capture a more-than-proportional share of the available social welfare gains when pricing a portion of the lanes.

9.2. **Effect on travel times and tolls.** Table 6 reports peak and average excess travel times (i.e., the additional travel time due to congestion) and tolls in a variety of counterfactuals. These are averaged over agents rather than time, which is why the table shows that average

travel times are higher when pricing a portion of the lanes.[56] Value pricing increases the typical agent's travels times, but also reduces their schedule delay, and so they are better off. If we instead compare travel times for any given arrival time we find that value pricing reduces travel times at every point in time, on average by between 1.3 and 22 percent, depending on what size of throughput drop we use. This means we pass the simple test for generating a Pareto improvement we constructed at the start of Section 6.

Tolls are higher when we price only some of the lanes. This occurs because when there are fewer agents in the priced lanes, the marginal agent has a higher value of time. The tolls reflect the marginal agents' preferences and so are higher. This is also why tolls paid (i.e., annual per capita toll revenue) in Table 5 are more than proportional to the fraction of lanes priced.

9.3. **Distributional effects.** Figure 4 shows the annual welfare changes due to pricing all (Panel A) or half (Panel B) of the lanes, averaged by group.[57] The agents harmed the most by pricing all of the lanes are the inflexible poor (in the bottom right of Panel A)—those who need to arrive to work exactly on-time and who would strongly prefer to pay with their time to do so instead of their money. The curve of darkest red in the lower right of Panel A lies along the curve $\alpha = \hat{\beta} \cdot \delta$; these are the agents who were able to arrive exactly on time when the road was free, but when the road is priced they are displaced by flexible rich agents who start arriving during the peak. The inflexible rich (in the upper right) are the best off; when the road is free they arrive on-time but bear large travel time costs, and they are delighted to pay with their money instead of their time. The flexible (on the left) are not very affected by adding tolls; they avoided paying with travel time by arriving off-peak and they will avoid paying with money by continuing to arrive off-peak. They are better off since they have a little less schedule delay, but as they are flexible they do not value the reduction highly.

By pricing just half of the lanes we preserve the ability of the poor to pay with their time, and so, as Panel B shows, avoid hurting the inflexible poor. We reduce the benefits to the inflexible rich, but we have generated a Pareto improvement.

Panel B also shows which agents are on which route. The black lines are the maximum and minimum values of $\hat{\alpha}(\delta, t^*)$ for each δ, and so separate the space of groups into those on the priced route and those on the free route. Those above both lines are on the priced route, those below both are on the free route, and those groups between the two lines have members on both routes.

In both panels of Figure 4 the change in trip cost is constant for a given value of time across a range of high levels of inflexibility. This occurs for the same reason $\hat{\alpha}(\delta, t^*)$ is

[56]The difference in average excess travel time when we weight by arrival time or agent occurs because there are now times on the free route when travel times are zero, but as no one travels at these times they are not included in the average travel time experienced by agents.

[57]Figure 4 does not show the 8 percent of agents with a value of time above fifty dollars an hour.

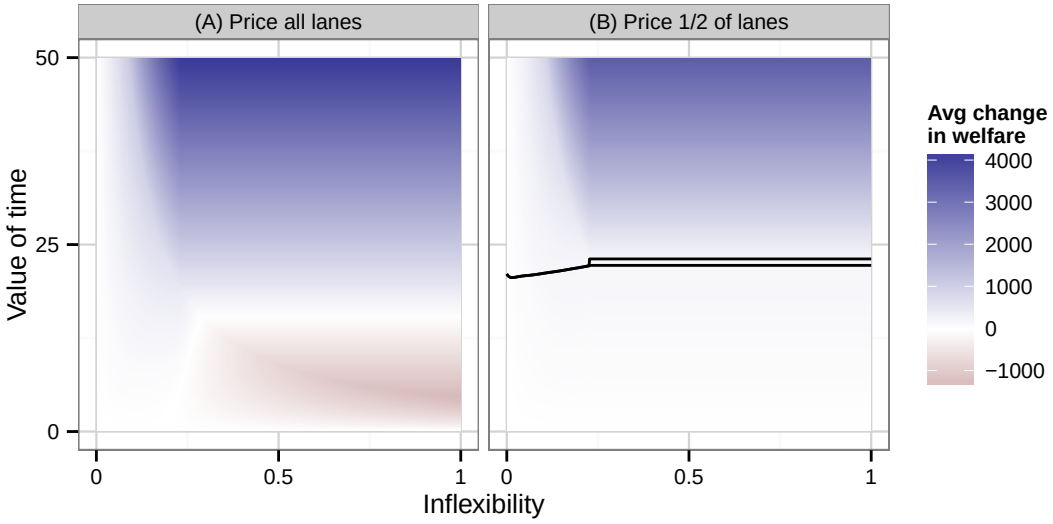


FIGURE 4. Annual change in welfare, averaged by group, when the throughput drop is 17.5 percent. The black lines in Panel B are the maximum and minimum values of $\hat{\alpha}(\delta, t^*)$ for each δ.

flat when $\delta > \hat{\delta}$ and $\hat{\alpha}(\delta, t^*) > \hat{\beta}/\hat{\delta}$: if an agent is inframarginal regardless of whether the road is free or priced, then he arrives exactly on time and so his actual inflexibility does not affect his trip cost or the change in his trip cost.

If we are willing to relax the requirement that pricing must generate a Pareto improvement and instead put some bound on the maximum harm done, then we can reap a greater portion of the potential welfare gains. Figure 5 shows this trade off when the throughput drop is only 10 percent. The largest drop in the maximum harm comes from leaving at least some of the lanes unpriced, because the inflexible poor would prefer to have a more congested free option where they can pay with their time instead of needing to pay with their money. By pricing 75 percent of the lanes we enjoy 80 percent of the social welfare gains while inflicting only 50 percent of the maximum harm.

9.4. **Extrapolating to the rest of the United States.** I use results from Margiotta, Cohen, Morris, Trombly, and Dixson (1994) and Schrank, Eisele, and Lomax (2012) and data from U.S. Department of Transportation (2009) to extrapolate my estimates of the private and social welfare gains, both per road user and in total, from value pricing to other cities in the United States. These results are reported in Appendix Tables I.1 and I.2. I find that pricing half the lanes on all urban highways would increase social welfare by over $30 billion per year. By way of comparison, recent estimates of the cost of congestion have included $45 billion (Winston, 2013), $82 billion (Couture, Duranton, and Turner, 2014), and $160 billion (Schrank, Eisele, Lomax, and Bak, 2015); while these estimates are larger

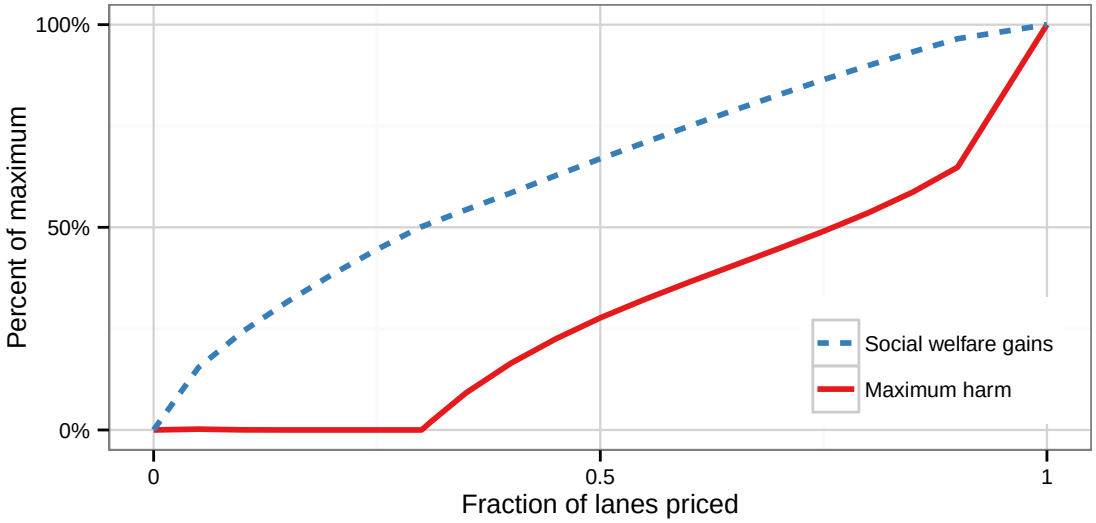


FIGURE 5. Trade-off between maximum harm and social welfare gains when throughput drop is 10 percent.

than mine, they are measuring the cost of congestion on all the lanes of all roads for all hours.

The social welfare gains for the typical commuter average $850 per year, with slightly more than half the welfare gains accruing directly to the commuters. The social welfare gains are smaller for the typical urban road user than for those on SR-91 because SR-91 is among the most congested highways in America and those who use it have longer-than-average commutes.

9.5. **Sensitivity checks.** I conduct a variety of sensitivity checks in Appendix I. Table I.3 recreates Table 5 using the typical-trip measure of flexibility. The results are largely unchanged: value pricing generates a Pareto improvement while pricing all of the lanes does not, the largest welfare loss from pricing all of the lanes is 30 percent lower, and social welfare gains are similar. The composition of the social welfare gains does differ: tolls revenue are lower and private welfare gains are larger.

Table I.5 and Figure I.1 compare results when using different assumptions on the distribution of the inflexibility of those agents in the inflexible category. The estimates of the size of the social and private welfare gains are unchanged, but in the most conservative case (specific-trip measure of flexibility with a small throughput drop) whether value pricing generates a Pareto improvement depends on the distribution assumed.

10. Conclusion

This paper has shown that a carefully designed toll applied to a portion of the lanes of a highway can generate a Pareto improvement, even before the toll revenue is spent. Specifically, I first show that a time-varying toll that smooths the rate at which drivers enter the highway can increase both speeds and throughput, generating a Pareto improvement when agents are homogeneous. I then show that when agents are heterogeneous a Pareto improvement can still be generated, but we will typically be limited to pricing a portion of the lanes. By pricing a portion of the lanes, we increase total highway throughput while preserving the ability of the poor to pay with time instead of money. I derive an intuitive sufficient condition for value pricing to yield a Pareto improvement: we simply need some rich drivers to be using the highway at the peak of rush hour.

To confirm the practical relevance of this theoretical possibility result, and to measure the size of the social welfare gains, I estimate the joint distribution of agent preferences and use these estimates to evaluate the effects of congestion pricing. I find that pricing half the lanes would generate a Pareto improvement, and increase social welfare by over $1,700 per road user per year.

There are at least four ways to make it even more likely value pricing generates a Pareto improvement. First, we can use the revenue to help those whom congestion pricing harms. Second, we can include in our analysis other ways for the poor to pay with time instead of money to use the priced lanes. Riding a bus that uses the priced lanes and car pooling both take extra time, but provide access to the priced lanes at reduced financial cost. Third, we can recognize that everyone is in a hurry sometimes (i.e., agents face shocks to their preferences), and so even if some drivers are worse off on some days, they main gain enough value from taking the faster priced lanes on days they are in a hurry such that value pricing yields a Pareto improvement. Fourth, we can include in our analysis the benefits from reducing fuel usage as decreasing pollution helps everyone.

The potential welfare gains from value pricing are large and obtainable. Extrapolating my results to the rest of the United States suggests that pricing half the lanes on urban highways would increase social welfare by over $30 billion per year, without hurting any road users. Furthermore, implementing the tolls which generate a Pareto improvement is straightforward. The technology Vickrey (1963) envisioned for the electronic collection of tolls is in use today and since highway throughput is observable, writing pricing algorithms to maximize throughput is relatively straightforward.

References

AASHTO (2005): *A Policy on Design Standards-Interstate System*. American Association of State Highway and Transportation Officials, Washington D.C.

Abrantes, P. A., and M. R. Wardman (2011): "Meta-Analysis of UK Values of Travel Time: An Update," *Transportation Research Part A: Policy and Practice*, 45(1), 1–17.

Arnott, R. J. (2013): "A Bathtub Model of Downtown Traffic Congestion," *Journal of Urban Economics*, 76, 110–121.

Arnott, R. J., A. J.-L. de Palma, and C. R. Lindsey (1990): "Economics of a Bottleneck," *Journal of Urban Economics*, 27(1), 111–130.

——— (1993): "A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand," *American Economic Review*, 83(1), 161–179.

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

Belenky, P. (2011): "Revised Departmental Guidance on Valuation of Travel Time in Economic Analysis," Discussion paper, U.S. Department of Transportation, Washington, D.C.

California Department of Transportation (2014): *Performance Measurement System*. Sacramento, California.

Chu, X. (1995): "Endogenous Trip Scheduling: The Henderson Approach Reformulated and Compared with the Vickrey Approach," *Journal of Urban Economics*, 37(3), 324–343.

Couture, V., G. Duranton, and M. Turner (2014): "Speed," *Working Paper*.

Currie, J., and R. Walker (2011): "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*, pp. 629–646, Lyon, France. Pergamon.

de Meza, D., and J. R. Gould (1987): "Free Access versus Private Property in a Resource: Income Distributions Compared," *Journal of Political Economy*, 95(6), 1317–1325.

de Palma, A., and R. Lindsey (2002): "Comparison of Morning and Evening Commutes in the Vickrey Bottleneck Model," *Transportation Research Record: Journal of the Transportation Research Board*, 1807(1), 26–33.

Duranton, G., and M. A. Turner (2011): "The Fundamental Law of Road Congestion: Evidence from US Cities," *American Economic Review*, 101(6), 2616–2652.

Fosgerau, M., and K. A. Small (2013): "Hypercongestion in Downtown Metropolis," *Journal of Urban Economics*, 76, 122–134.

Halvorson, R., and K. R. Buckeye (2006): "High-Occupancy Toll Lane Innovations: I-394 MnPASS," *Public Works Management & Policy*, 10(3), 242–255.

Hendrickson, C., and G. Kocur (1981): "Schedule Delay and Departure Time Decisions in a Deterministic Model," *Transportation Science*, 15(1), 62–77.

Hotelling, H. (1929): "Stability in Competition," *Economic Journal*, 39(153), 41–57.

Hymel, K. (2009): "Does Traffic Congestion Reduce Employment Growth?," *Journal of Urban Economics*, 65(2), 127–135.

Johnson, M. B. (1964): "On the Economics of Road Congestion," *Econometrica*, 32(1/2), 137–150.

Knight, F. H. (1924): "Some Fallacies in the Interpretation of Social Cost," *Quarterly Journal of Economics*, 38(4), 582–606.

Light, T. (2009): "Optimal highway design and user welfare under value pricing," *Journal of Urban Economics*, 66(2), 116–124.

Liu, L. N., and J. F. McDonald (1998): "Efficient Congestion Tolls in the Presence of Unpriced Congestion: A Peak and Off-Peak Simulation Model," *Journal of Urban Economics*, 44(3), 352–366.

——— (1999): "Economic Efficiency of Second-Best Congestion Pricing Schemes in Urban Highway Systems," *Transportation Research Part B: Methodological*, 33(3), 157–188.

Margiotta, R., H. Cohen, R. Morris, J. Trombly, and A. Dixson (1994): "Roadway Usage Patterns: Urban Case Studies," Discussion paper, Volpe National Transportation Systems Center and Federal Highway Administration.

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

Newell, G. F. (1987): "The Morning Commute for Nonidentical Travelers," *Transportation Science*, 21(2), 74–88.

Patterson, T. M., and D. M. Levinson (2008): "Lexus Lanes or Corolla Lanes? Spatial Use and Equity Patterns on the I-394 MnPASS Lanes," *Working Paper*.

Perez, B. G., and G.-C. Sciara (2003): "A Guide for HOT Lane Development," Discussion Paper FHWA-OP-03-009, Federal Highway Administration, Washington, D.C.

Pigou, A. C. (1920): *The economics of welfare*. Macmillan and co., ltd., London, 1 edn.

Rotemberg, J. J. (1985): "The efficiency of equilibrium traffic flows," *Journal of Public Economics*, 26(2), 191–205.

Schrank, D., B. Eisele, and T. Lomax (2012): "2012 Urban Mobility Report," Discussion paper, Texas A&M Transportation Institute, College Station, Texas.

Schrank, D., B. Eisele, T. Lomax, and J. Bak (2015): "2015 Urban Mobility Scorecard," Discussion paper, Texas A&M Transportation Institute, College Station, Texas.

Small, K., and E. Verhoef (2007): *The Economics of Urban Transportation*. Routledge, New York.

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*, pp. 53–96.

Small, K. A. (1983): "The Incidence of Congestion Tolls on Urban Highways," *Journal of Urban Economics*, 13(1), 90–111.

——— (1992): "Using the Revenues From Congestion Pricing," *Transportation*, 19(4), 359–381.

——— (2015): "The bottleneck model: An assessment and interpretation," *Economics of Transportation*, 4(1–2), 110–117.

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

Small, K. A., C. Winston, and J. Yan (2005): "Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability," *Econometrica*, 73(4), 1367–1382.

Stiglitz, J. (1998): "Distinguished Lecture on Economics in Government: The Private Uses of Public Interests: Incentives and Institutions," *Journal of Economic Perspectives*, 12(2), 3–22.

Sullivan, E. (1999): *State Route 91 Impact Study Datasets*. California Polytechnic State University, San Luis Obispo, California.

——— (2002): "State Route 91 Value-Priced Express Lanes: Updated Observations," *Transportation Research Record: Journal of the Transportation Research Board*, 1812(-1), 37–42.

Sullivan, E., and M. Burris (2006): "Benefit-Cost Analysis of Variable Pricing Projects: SR-91 Express Lanes," *Journal of Transportation Engineering*, 132(3), 191–198.

Sullivan, E., and J. Harake (1998): "California Route 91 Toll Lanes Impacts and Other Observations," *Transportation Research Record: Journal of the Transportation Research Board*, 1649(-1), 55–62.

U.S. Department of Transportation (2006): "Congestion Pricing: A Primer," Discussion Paper FHWA-HOP-07-074, Washington, D.C.

——— (2009): *2009 National Household Travel Survey*. Washington, D.C.

van den Berg, V., and E. T. Verhoef (2011): "Winning or Losing from Dynamic Bottleneck Congestion Pricing?: The Distributional Effects of Road Pricing with Heterogeneity in Values of Time and Schedule Delay," *Journal of Public Economics*, 95(7–8), 983–992.

Verhoef, E., P. Nijkamp, and P. Rietveld (1996): "Second-Best Congestion Pricing: The Case of an Untolled Alternative," *Journal of Urban Economics*, 40(3), 279–302.

Verhoef, E. T., and K. A. Small (2004): "Product Differentiation on Roads: Constrained Congestion Pricing with Heterogeneous Users," *Journal of Transport Economics and Policy*, 38(1), 127–156.

Vickrey, W. S. (1963): "Pricing in Urban and Suburban Transport," *American Economic Review*, 53(2), 452–465.

——— (1969): "Congestion Theory and Transport Investment," *American Economic Review*, 59(2), 251–260.

——— (1987): "Marginal and Average Cost Pricing," in *The New Palgrave Dictionary of Economics*, ed. by S. N. Durlauf, and L. E. Blume. Palgrave Macmillan, Basingstoke, 1 edn.

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

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

WINSTON, C. (2013): "On the Performance of the U.S. Transportation System: Caution Ahead," *Journal of Economic Literature*, 51(3), 773–824.