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

JONATHAN D. HALL
UNIVERSITY OF TORONTO

ABSTRACT. Though economists have long advocated road pricing as an efficiency-enhancing solution to traffic congestion, it has rarely been implemented, primarily because it is thought to create losers as well as winners. This paper shows that a judiciously designed toll applied to a portion of the lanes of a highway can generate a Pareto improvement before using the revenue, a sufficient condition being that drivers with a high value of time travel at the peak of rush hour. I obtain these new theoretical results by extending a standard dynamic congestion model to reflect an important additional traffic externality: extra traffic does not simply increase travel times, but also introduces frictions that reduce throughput. The analysis draws attention to a practical policy that may help overcome the widespread opposition to road pricing.

JEL Classification: D62, H41, R41, R48.

*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, Jan Brueckner, Ian Fillmore, Mogens Fosgerau, Edward Glaeser, Brent Hickman, William Hubbard, Kory Kroft, Ethan Lieber, Robin Lindsey, Robert McMillan, Peter Morrow, John Panzar, Devin Pope, Mark Phillips, Allen Sanderson, Ken Small, Chad Syverson, George Tolley, Matt Turner, 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

Traffic congestion is a major problem facing large cities worldwide. In the United States, for example, congestion consumes 42 hours per urban commuter annually—nearly an entire work week (Schrank et al. 2015), as well as imposing a host of other social costs.[1] At least since Pigou (1920), economists have advocated solving traffic congestion using tolls. Adding tolls would help drivers internalize the externalities they impose on others and would greatly increase social welfare. Yet tolling is rarely used in practice, in large part because of the received wisdom that such tolls impose losses on many, if not most, road users.[2] As Lindsey and Verhoef (2008) state, "most likely, these losses are the root of the longstanding opposition to congestion tolling in road transport," a view echoed widely.[3]

A long literature has been concerned with the distributional consequences from tolling. Prior research has identified specific situations where it is possible for tolling to generate a Pareto improvement with homogeneous agents,[4] by using the toll revenue,[5] or charging negative tolls off-peak.[6] In practice, however, road users are not homogeneous, it is difficult to target the spending precisely enough to actually generate a Pareto improvement,[7] and charging negative tolls is impractical. The literature has also shown that pricing a portion of the lanes (a practice generally called "value pricing") reduces, but does not eliminate, the harm done

[1] Congestion also wasted 3.1 billion gallons of fuel in 2014 (Schrank et al. 2015), releasing an additional 28 million metric tons of carbon dioxide into the atmosphere. This additional pollution amounts to more than six times the annual emissions saved by the current fleet of hybrid and electric vehicles, 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).

[2] See Small and Verhoef's (2007) classic textbook for an explanation of this standard result (pp. 120–127) and see Appendix A for a brief discussion of other barriers to congestion pricing.

[3] For further examples, see Beesley (1973), Starkie (1986), Cohen (1987), Giuliano (1992), Arnott et al. (1994), Lave (1994), Small et al. (2005), and Small and Verhoef (2007).

[4] See Johnson (1964), De Meza and Gould (1987), Arnott and Inci (2010), Arnott (2013), and Fosgerau and Small (2013).

[5] See Foster (1975) and Arnott et al. (1994).

[6] See Braid (1996) and van den Berg and Verhoef (2011).

[7] Foster (1975) first noted the difficulty in targeting the spending. 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. In addition, even if we can design transfers that make a policy Pareto-improving, they can still be difficult to implement. As Stiglitz (1998) points out, 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.

by congestion pricing.[8] Additionally, Arnott, de Palma, and Lindsey (1994) find that if only agents with the highest value of time (i.e., the rich) travel at the peak of rush hour, then it is possible to generate a Pareto improvement with heterogeneous agents. Nevertheless, as one of these authors argues, "economists came to appreciate . . . the practical impossibility of designing a tolling scheme that leaves everyone better off" (Lindsey 2006, p. 332).

In this paper, I show it is possible to design such a tolling scheme. The main result of this paper is that a carefully designed, time-varying toll on a portion of the lanes of a highway can generate a Pareto improvement, even before the toll revenue is spent and even with realistic heterogeneity. I obtain this new result by extending the bottleneck congestion model of Vickrey (1969) and Arnott, de Palma, and Lindsey (1993) to reflect an important additional traffic externality that has been identified by the transportation engineering literature but that has largely been ignored in the economics literature: not only does each additional vehicle slow others down, but in heavy enough traffic, additional vehicles can create frictions that reduce throughput (the number of trips per unit time).

By combining this empirically relevant additional externality with pricing a portion of the lanes we can obtain a Pareto improvement. The second externality pushes traffic off the production possibilities frontier (PPF). By using a time-varying toll to prevent the second externality from occurring, we can move back to the PPF, increasing both speeds and throughput (this is different than is typically assumed, and I explain how it is possible in Section 5). Were agents homogeneous, this would be enough to conclude everyone is better off.[9] However, road users are heterogeneous, and so pricing all the lanes likely hurts some of them. Adding tolls increases the financial costs of traveling while reducing the time costs. As not all road users value their time equally, this probably leaves some worse off.

Pricing a portion of the lanes can overcome these negative distributional effects. Adding tolls increases speeds and throughput in the priced lanes, allowing them to carry more traffic than they did before. This means the free lanes carry less traffic than before, improving travel times in the free lanes. Since travel times

[8]For example, see Braid (1996), Small et al. (2006), Light (2009), and van den Berg and Verhoef (2011). Light (2009) finds using the revenue to pay for new capacity generates a Pareto improvement while Braid (1996) and van den Berg and Verhoef (2011) find that charging a negative toll off-peak generates a Pareto improvement.

[9]Johnson (1964) and De Meza and Gould (1987) show this result in a static model, and Arnott and Inci (2010), Arnott (2013), and Fosgerau and Small (2013) derive it in dynamic models of downtown, rather than highway, congestion.

in the free lanes are lower, those who continue to use these 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, even without using the revenue.

Obtaining a Pareto improvement, however, often comes at a cost. By only pricing a portion of the lanes, we leave the other lanes congested, with all the resulting social costs. That said, inasmuch as generating a Pareto improvement makes it politically feasible to implement tolling, then we are trading potential, but unrealized, welfare gains for actual welfare gains.

My model focuses on two groups of agents: one rich, with high value of time, and the other poor, with low value of time. The groups also differ in the flexibility of their schedules and, within each group, desired arrival times are continuously distributed. Agents choose when to arrive at work and which route to take to minimize their total cost of traveling.

Using my model, I characterize the set of parameter values for which it is possible to generate a Pareto improvement without using the revenue. I find that pricing is more likely to generate a Pareto improvement the larger the reduction in throughput due to queuing, the greater the correlation between income and schedule inflexibility, and as income inequality decreases. Furthermore, I provide an intuitive sufficient condition for pricing a portion of the lanes to yield a Pareto improvement that holds even when there is an arbitrary number of groups: we simply need some rich drivers to be using the highway at the peak of rush hour. I then use the 2009 National Household Travel Survey to show that this sufficient condition is satisfied empirically—drivers with household incomes above $100,000 make up over 25 percent of those traveling at the peak. Finally, I characterize the trade-off between efficiency and equity by deriving the differences in social welfare gains and maximum harm done when pricing all the lanes or generating a Pareto improvement. I find that it is often the case that for parameter values such that pricing all of the lanes does the most harm, generating a Pareto improvement requires the least sacrifice of potential social welfare gains.

Central to the results of this paper is the notion that too many vehicles on the road reduces throughput. Over fifty years ago, Walters (1961) conjectured this additional externality existed, which Vickrey (1987) named "hypercongestion." As the theoretical arguments for hypercongestion did not carry over to dynamic

models of highway congestion, the economics literature cast doubt on hypercongestion's existence.[10,11] Since then, the transportation engineering literature has identified the causal mechanisms and provided extensive empirical evidence that hypercongestion is an important real-world phenomenon—one that should be incorporated into models of congestion. I review this evidence in Section 2.

I build most closely on two important papers. First, in a valuable contribution, Arnott et al. (1994) extend the bottleneck model to allow agents to be heterogeneous. They work through several possible cases of heterogeneity, showing the distributional effects both before and after the revenue is either rebated lump sum or used to build new capacity. Most relevant to this paper, they show that when there are two groups with different values of time and schedule flexibility but homogeneous desired arrival time, pricing all of the lanes can help all agents before using the revenue if only those with a high value of time (i.e., the rich) travel at the peak of rush hour.[12]

I augment Arnott et al. (1994) by adding hypercongestion, value pricing, and heterogeneity in desired arrival times. I build on their results by showing how adding tolls can generate a Pareto improvement even when the poor travel at the peak of rush hour.

Second, an innovative paper by van den Berg and Verhoef (2011) extends Arnott et al. (1994) to allow agent preferences to vary continuously on two dimensions: value of time and schedule inflexibility. They show numerically that for intuitively reasonable parameter values, pricing all the lanes does not always hurt the majority of agents and that it is possible to generate a Pareto improvement by pricing a third of the lanes and forgoing revenue by charging a negative toll off-peak.

[10]The argument in favor of hypercongestion was that throughput (vehicles/hour) is the product of speed (miles/hour) times density (vehicles/mile), $T = S \times D$, and since speed is decreasing in density, $dS/dD < 0$, then it is possible for throughput to be decreasing in density, $dT/dD = D \cdot dS/dD + S \gtrless 0$ (e.g., Walters 1961, Johnson 1964, De Meza and Gould 1987). The literature correctly responded that hypercongestion is a dynamic phenomenon, and showed that in dynamic models, the mathematical relationships above were not enough to generate hypercongestion (e.g., Newell 1988, Evans 1992, Verhoef 1999, 2001, 2005, May et al. 2000, Small and Chu 2003).

[11]There is some research arguing hypercongestion is possible for urban centers. See Small and Chu (2003), Arnott and Inci (2010) and Fosgerau and Small (2013).

[12]Vickrey (1973) derives a similar result also using the bottleneck model. He shows that if everyone is indifferent between all chosen arrival times (so we could assign arrival times so that only the rich are traveling at the peak), then adding tolls generates a Pareto improvement before using the revenue. The intuition for these results is that if the poor are willing to arrive at the worst equilibrium arrival time before a toll is added, then they can avoid any harm from tolling by choosing to do so once a toll is added.

My model differs from van den Berg and Verhoef (2011) by adding hypercongestion and by allowing agent preferences to vary discretely along two dimensions (value of time and schedule inflexibility) and continuously along a third (desired arrival time). This allows me to build on their results by showing how it is possible to generate a Pareto improvement before using the revenue. I further build on their work by deriving analytical solutions, the first for a dynamic model of value pricing with heterogeneous agents. Deriving analytical solutions allows me to (1) fully characterize the parameter values for which adding tolls generates a Pareto improvement before using the revenue, (2) show which factors make it easier to do so, and (3) let the social planner choose the fraction of lanes priced, rather than considering pricing a specific fixed fraction.

2. Evidence for hypercongestion

In this section, I draw on the transportation engineering literature to explain the two causal mechanisms behind the additional externality, hypercongestion, and the empirical evidence for these mechanisms. Both of these causal mechanisms occur at bottlenecks.[13]

Before going into the evidence for hyerpcongestion, it is important to be clear about the distinction between the two externalities. Consider a two-lane highway that merges into one lane, 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. This lengthening of the queue is the standard externality. However, this simple externality fails to capture the fact that a queue creates additional frictions that *reduce throughput* (the rate at which vehicles can pass through the bottleneck), 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.

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

[13] A bottleneck can occur at any place where the capacity of a highway decreases, generally because of a reduction in lanes. While the most noticeable bottlenecks are 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.

freeway and blocking through traffic; similarly, a queue on the highway can block upstream exits. Vickrey (1969) labeled the second situation a triggerneck 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; there will typically 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 reduces throughput because the vehicle that forced its way over will be moving very slowly and so goes through the bottleneck at a slow speed. Equivalently, it opens up a gap in front of itself; this will be a period of time during which 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 Table 1.

3. Model

I use the bottleneck model of Vickrey (1969), which was formalized by Arnott et al. (1990, 1993). I make three 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]

[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.

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

Paper	Throughput drop (%)	Location
Hurdle and Datta (1983)†	0	Queen Elizabeth Way, Toronto
Banks (1990)	2.8	I-8, San Diego
Hall and Agyemang-Duah (1991)	5.8	Queen Elizabeth Way, Toronto
Banks (1991)	−1.2–3.2	4 sites in San Diego
Elefteriadou et al. (1995)†	10	I-290, Chicago
Persaud et al. (1998)	10.6–15.3	3 site/time pairs in Toronto
Cassidy and Bertini (1999)	7.4–8.7	2 sites in Toronto
Bertini and Malik (2004)	4	US-169, Minneapolis
Zhang and Levinson (2004)	2–11	27 sites in Minneapolis–St. Paul
Bertini and Leal (2005)	9.7	M4, London
	12	I-494, Minneapolis
Cassidy and Rudjanakanoknad (2005)	11.7	I-805, San Diego
Rudjanakanoknad (2005)	13.2	SR-22, Orange County, California
Chung et al. (2007)	12.3	I-805, San Diego
	6.2	SR-24, San Francisco
	5.8	Gardiner Expressway, Toronto
Guan et al. (2009)	15	Fourth Ring Road, Beijing
Leclercq et al. (2011)†	25	M6, Manchester, UK
Oh and Yeo (2012)	8.9–16.3	16 sites in California
Srivastava and Geroliminis (2013)	15	US-169, Minneapolis

† This paper does not expressly test the two-capacity hypothesis, but does contain figures reporting throughput at an isolated bottleneck present both before and after a queue forms. The throughput drop reported in this table is based on these figures.

The second modification is to allow the social planner to choose the fraction of the lanes that are priced, as in Braid (1996). This modification was dropped in following papers, such as Liu and McDonald (1998), Small et al. (2006), and van den Berg and Verhoef (2011), who consider 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. 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.[15] However, once agents are heterogeneous along other dimensions then allowing for agents' desired arrival time to be continuously distributed has significant effects on equilibrium outcomes.

Allowing for a continuum of desired arrival times affects equilibrium outcomes because it allows a positive measure of agents to arrive exactly on time. When agents are heterogeneous those who arrive on time will often be inframarginal with regard to the choice of when to arrive, meaning that if the cost of their chosen arrival time increases by a small amount, holding all else constant, they do not change when they arrive. I call these agents *inframarginal*, suppressing the specification of which dimension of choice they are inframarginal on.

Allowing for inframarginal agents changes the results, in particular, making it harder to generate a Pareto improvement when pricing all the lanes. Inframarginal agents strictly prefer their current arrival time, but adding tolls often changes the allocation of agents to arrival times. While this reallocation is efficiency enhancing, previously inframarginal agents who end up with a different arrival time are generally worse off, and often significantly so.

Allowing for inframarginal agents is also necessary for the model to make realistic predictions about travel times. As Hall (2017) shows, 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 someone with a fixed work schedule. 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

[15]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 both when the road is completely free or completely priced, and when only a portion of the lanes are priced.

free. Let λ_{toll} and λ_{free} denote the fraction of capacity devoted to each route, where $\lambda_{\text{toll}} + \lambda_{\text{free}} = 1$.[16] 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

(3.1)

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

I then simplify by taking the limit as $\epsilon \to 0$, so throughput on a congested route is constant.[17]

Travel time on route r for an agent arriving at t is

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

where T^f is fixed travel time—the amount of time it takes to travel the road absent any congestion—and $T_r^v(t)$ is variable travel time for route r. For simplicity, and without loss of generality, let $T^f = 0$. Throughout the rest of this paper, when we

[16]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, 39-40) and on I-394 in Minnesota there is a two foot buffer without any barrier (Halvorson and Buckeye 2006, 246). As federal standards call for twelve foot lanes on interstates (AASHTO 2005, 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.

[17]This allows me to keep the model simple, by not needing to model the short period where there is minor congestion but throughput has yet to fall, while avoiding existence of equilibrium problems which can occur when evaluated at $\epsilon = 0$. If $\epsilon = 0$ then it is possible for, 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.

discuss travel time we are only referring to the variable congestion-related travel time.[18]

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

$$p\left(t, r; i, t^*\right) = \alpha_i T_r\left(t\right) + \tau_r\left(t\right) + D_i\left(t^* - t\right) \tag{3.2}$$

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,

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

where β is the cost per unit time early to work, and γ is the cost per unit time late to work. Each of these parameters represents how much an agent is willing to pay in money to reduce travel time or schedule delay by one unit of time.

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. It is needed to prevent the departure rate from being infinite.

To simplify, 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.[19]

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. We will primarily be concerned with the case where there are two groups.

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

[18]I define travel times as a function of when an agent *arrives* rather than *departs* for simplicity. Because this model is deterministic, there is a one-to-one mapping between departure times and arrival times, and thus doing so is innocuous. See Appendix B.1 for more details.

[19]Relaxing this assumption would only affect my results if there are agents who switch from arriving early to arriving late, or vice-versa, when tolls are added to the road.

other sources of heterogeneity in agents' value of time,[20] 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 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, and provide a measure of how inflexible his schedule is, so let $\delta_i = \beta_i/\alpha_i$ be group i's *inflexibility*. 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 factory worker is late he generally faces penalties and when he is early he passes the time talking with co-workers. Since there is not much difference for the factory worker between spending time traveling or being at work early, his δ is close to one (the largest possible δ). Similarly, due to the penalty when late, $\xi\delta = \gamma/\alpha$ 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 δ is closer to zero. Thus variation in δ is driven by variation in schedule flexibility, where jobs that are more flexible lead to a lower δ.

Within each group, agents' desired arrival times are uniformly distributed over $[t_s, t_e]$. 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 $N_i = (t_e - t_s)\, n_i$ denotes 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.

[20] 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).

The density 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. Without this assumption the distribution of desired arrival times would no longer be uniform once tolls were added to the highway because tolling changes different types' trip costs by different amounts, and the model would lose its analytical tractability.[21]

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.

I show that an equilibrium exists by construction, and show that almost everywhere in the parameter space equilibrium trip prices, travel times, and tolls are unique, and the travel time profile and toll schedule have a single local maximum in Appendix E.[22]

4. Finding the equilibrium

In this section, I solve for equilibrium trip cost. I start by making some observations that allow me to simplify the notation and analysis. Then I show how to find the equilibrium allocation of arrival times to agents on a free or tolled route, given the set of agents on that route, and how to back out equilibrium travel times and tolls from this allocation. Following this, I solve for the equilibrium when the road is free, when all the lanes are priced, and when a portion of the lanes are priced.

[21]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 the lanes, regardless of the effect it has on throughput, never hurts any agents even before the revenue is used.

[22]The exception is when $n_i = (1 - \lambda_{\text{toll}})\, s$ for $\delta_i > \delta_j$ on a free route and $n_i = \lambda_{\text{toll}} s^*$ for $\beta_i > \beta_j$ on a priced route. When uniqueness fails, there exists an equilibrium with trip prices, travel times, and tolls that are the limit of those same objects as $n_i \to (1 - \lambda_{\text{toll}})\, s$ for $\delta_i > \delta_j$ on a completely free route and $n_i \to \lambda_{\text{toll}} s^*$ for $\beta_i > \beta_j$, and so I use those values.

4.1. **Arrival rates.** 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.

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 a 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.

4.2. **Assigning agents to arrival times.** 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 E.

Lemma 1. *If group i is more inflexible than group j (i.e., $\delta_i > \delta_j$) 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.*

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 2. *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.*

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. 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' $t_i^{\max}$.

Given these definitions and Lemma 1, we can assign agents to arrival times on a free route as follows. First, assume we know $t^{\max}$ and $t_i^{\max}$ for all $i \in \mathcal{G}$. Then starting at $t^{\max}$ and working our way backward, assign to each arrival time t the most inflexible agents of those who want to arrive early or on-time at t and are not yet assigned an arrival time until we have filled the available capacity. Likewise start at $t^{\max}$ and work forward, assigning the most inflexible agents who want to arrive late or on-time at t. Break ties by allowing those with an earlier desired arrival time to arrive earlier.

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

We can assign agents to arrival times on a priced route similarly, but using agents' β to order them rather than their inflexibility.

4.3. **Travel times and tolls.** Once we have assigned agents to arrival times on a free route, 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 3.

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

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

Likewise, on a priced route, we can use agents' preferences to back out the toll schedule (i.e., the function τ). Using similar logic as above, we obtain the following lemma:

Lemma 4.

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

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.

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

4.4. **Equilibrium trip prices.** Now that we know when agents arrive, and thus their schedule delay, and their travel time or toll, we can derive agents' trip costs. I do so for one of two cases when the road is completely free or priced, and one of eight cases when a portion of the lanes are priced, leaving the remaining cases for Appendix C.[25]

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 road $\beta_A/\alpha_A < \beta_B/\alpha_B$ and Lemma 2 implies that on a toll road $\beta_A < \beta_B$. When the entire road is either free or priced, one of two cases apply: either $n_B \leq s$ or $n_B > s$. I solve the first case below, and save the second for Appendix C.

4.4.1. *Equilibrium when road completely free.* 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.

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 3 to define the equilibrium travel time profile as the solution to

$$\frac{dT_I}{dt}(t) = \begin{cases} \beta_A/\alpha_A & t_{0A} \leq t < t_A^{\max} \\ -\gamma_A/\alpha_A & t_A^{\max} \leq t < t_{A0} \\ 0 & \text{otherwise} \end{cases}, \tag{4.1}$$

$$T_I(t_{0A}) = 0. \tag{4.2}$$

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

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

[25]When the road is completely free or priced the two cases are whether or not any agents are inframarginal. When pricing a portion of the lanes there are eight possible cases. The three dimensions in which the cases differ are (1) which group is not arriving off-peak, (2) whether some agents from this group are inframarginal or not, and (3) whether they are on one or two routes.

the start of rush hour and the peak. In this period of time $\left(t_A^{\max} - t_{0A}\right) s$ agents can arrive. However, we need to account for the capacity used by agents in group B. Since they arrive on-time, $\left(t_A^{\max} - t_s\right) n_B$ 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 $\left(t_A^{\max} - t_s\right) n_A$. Thus in equilibrium

$$(t_A^{\max} - t_{0A})\, s - (t_A^{\max} - t_s)\, n_B = (t_A^{\max} - t_s)\, n_A. \tag{4.3}$$

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

$$(t_{A0} - t_A^{\max})\, s - (t_e - t_A^{\max})\, n_B = (t_e - t_A^{\max})\, n_A. \tag{4.4}$$

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

$$T_I\left(t_{A0}\right) = 0. \tag{4.5}$$

We now know enough to find equilibrium trip costs. Solving (4.3), (4.4), and (4.5) yields the equilibrium travel time profile. We then find each type's equilibrium trip cost $\bar{p}\left(i, t^*\right) = \min_{t,r} p\left(t, r; i, t^*\right)$, and do so in Appendix C. The equilibrium trip costs for agents in group A are

$$\bar{p}_{I,\text{free}}\left(A, t_A^{\max}\right) = \beta_A \left(N_A + N_B\right) \frac{1}{s} \frac{\xi}{1+\xi}, \tag{4.6}$$

$$\bar{p}_{I,\text{free}}\left(A, t^*\right) = \bar{p}_{I,r}\left(A, t_A^{\max}\right) - \left(t_A^{\max} - t^*\right) \begin{cases} \beta_A & t^* \leq t_A^{\max} \\ -\xi\beta_A & t^* > t_A^{\max} \end{cases}. \tag{4.7}$$

For group B agents equilibrium trip prices are

$$\bar{p}_{I,\text{free}}\left(B, t^*\right) = \frac{\alpha_B}{\alpha_A} \bar{p}_{I,\text{free}}\left(A, t^*\right). \tag{4.8}$$

While (4.7) and (4.8) can be calculated directly, they are also fairly intuitive. First, note that due to the slope of the travel time profile 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 (4.7) 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 (4.8) 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 road 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.

4.4.2. *Equilibrium when road completely priced.* We find the equilibrium when the road is priced similarly. 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

$$\frac{d\tau_I}{dt}(t) = \begin{cases} \beta_A & t_{0A} \leq t < t_A^{\max} \\ -\gamma_A & t_A^{\max} \leq t < t_{A0} \\ 0 & \text{otherwise} \end{cases}, \tag{4.9}$$

$$\tau_I(t_{0A}) = 0. \tag{4.10}$$

Again we have three variables still to determine. Because capacity on toll road increases to s^*, we replace s with s^* in (4.3) and (4.4), as well as changing subscripts to denote that we are considering a toll road. Finally, we replace (4.5) with

$$\tau_I(t_{A0}) = 0. \tag{4.11}$$

We now know enough to find the equilibrium toll schedule and trip costs. As before, the details are in Appendix C. The equilibrium trip costs for agents in group A are

$$\bar{p}_{I,\text{toll}}(A, t_A^{\max}) = \beta_A (N_A + N_B) \frac{1}{s^*} \frac{\xi}{1+\xi}, \tag{4.12}$$

$$\bar{p}_{I,\text{toll}}(A, t^*) = \bar{p}_{I,r}(A, t_A^{\max}) - (t_A^{\max} - t^*) \begin{cases} \beta_A & t^* \leq t_A^{\max} \\ -\xi\beta_A & t^* > t_A^{\max} \end{cases}. \tag{4.13}$$

For group B agents equilibrium trip prices are

$$\bar{p}_{I,\text{toll}}(B, t^*) = \bar{p}_I(A, t^*). \tag{4.14}$$

The only difference between (4.12) and (4.6) is that highway capacity is higher, and (4.13) and (4.7) are identical. This is because the equilibrium trip price for agents in group A when group B is is inframarginal is pinned down by the cost of arriving at the start or end of rush hour. Adding tolls only changes this by changing highway capacity, and thus changing the length of rush hour.

The intuition behind (4.13) and (4.14) is similar to that for (4.7) and (4.8). The only slight difference is as follows. For (4.14), consider two agents with desired arrival time t^*, one from each group. Both are willing to arrive at t^*. When arriving at t^* they face the same toll and have no schedule delay or travel time, and so their trip costs are identical.

4.4.3. *Equilibrium when value pricing.* We now solve for the equilibrium when pricing a portion of the lanes. Solving for the equilibrium is now more complicated because agents choose which route they take as well as their arrival time. To solve for the equilibrium when there are two groups, I first derive two results that make assigning agents to routes simpler, and then setup and solve a system of linear equations for each of the eight value pricing cases.[26] I solve one case below, and solve the remaining cases in Appendix C.2.

The first result which simplifies assigning agents to routes is as follows:

Lemma 5. *The same group arrives off-peak on both routes, or at least is indifferent about doing so.*

This first result follows from the fact that for all agents the cost of arriving at the very start of rush hour is the same on both routes because at those times there is no toll or travel time, just schedule delay. Likewise, the cost of arriving at the very end of rush hour is the same on both routes.

The second result formalizes the intuition that the rich prefer to be on the priced route and the poor prefer the free route:

[26] In two of the cases the toll schedule or travel time profile is not completely defined by Lemmas 3 and 4 and so I use another indifference relation to characterize part of the toll schedule or travel time profile.

Lemma 6. *If there are two families and two routes, one priced and one free, then the rich will never be on the free route unless the poor are too, and the poor will never be on the priced route unless the rich are too.*

Now I solve the case when the poor group is not arriving off-peak, they are inframarginal, and are traveling on one route. The subscript $1R, I,$ poor denotes objects which below to this case. To keep subscripts from being unwieldy, define group 1 as rich and group 2 as poor (i.e., $\alpha_1 \geq \alpha_2$).

In this case the rich agents travel on both routes and are always marginal while the poor agents travel only on the free route and are inframarginal. For this to be possible there must be enough capacity on the free route for all poor agents to arrive on-time, i.e., $n_2 \leq (1 - \lambda_{\text{toll}})\, s$.

Tolls and travel times are the same as when the entire road is free or priced and poor agents are inframarginal. The travel time profile is defined by (4.1), (4.2), and (4.5), and the toll schedule is defined by (4.9)–(4.11).

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

$$(t_1^{\max} - t_{01})\,(\lambda_{\text{toll}} s^* + (1 - \lambda_{\text{toll}})\, s) - (t_1^{\max} - t_s)\, n_2 = (t_1^{\max} - t_s)\, n_1, \text{ and}$$
$$(t_{10} - t_1^{\max})\,(\lambda_{\text{toll}} s^* + (1 - \lambda_{\text{toll}})\, s) - (t_e - t_1^{\max})\, n_2 = (t_e - t_1^{\max})\, n_1.$$

Solving this system of equations gives us the equilibrium travel time profile and toll schedule, from which we find the trip costs, which are

$$\bar{p}_{1R,I,\text{poor}}\,(1, t_1^{\max}) = \beta_1 \frac{N_1 + N_2}{\lambda_{\text{toll}} s^* + (1 - \lambda_{\text{toll}})\, s} \frac{\xi}{1 + \xi}, \tag{4.15}$$

$$\bar{p}_{1R,I,\text{poor}}\,(1, t) = \bar{p}_{1R,I}\,(1, t_1^{\max}) - (t_1^{\max} - t) \begin{cases} \beta_1 & t \leq t_1^{\max} \\ -\xi\beta_1 & t > t_1^{\max} \end{cases}, \text{ and} \tag{4.16}$$

$$\bar{p}_{1R,I,\text{poor}}\,(2, t) = \frac{\alpha_2}{\alpha_1} \bar{p}_{1R,I}\,(1, t)\,. \tag{4.17}$$

5. Results

We now use the model to understand how 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. We start with the case where every agent is identical to highlight

the importance of hypercongestion to my results.[27] Then I show how heterogeneity in agent preferences makes it difficult to generate a Pareto improvement, even in the presence of hypercongestion, and how value pricing can overcome these difficulties. Next I derive a simple, and general, sufficient condition for when congestion pricing leaves all agents 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 holds even when there are an arbitrary number of groups. I end the section by exploring the trade-offs between equity and efficiency inherent in pricing a portion of the lanes to generate a Pareto improvement.

If congestion pricing can undo the effects of hypercongestion and increase both speeds and throughput, then, as Johnson (1964) first showed using a static model, congestion pricing generates a Pareto improvement when agents are identical. A similar result has also been derived in dynamic models of downtown congestion in Arnott and Inci (2010), Arnott (2013), and Fosgerau and Small (2013).

Proposition 1. *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 then a carefully designed, time-varying toll can smooth the rate at which vehicles get on the highway, eliminating the queue with its attendant frictions, and thereby increase both speeds and throughput. Because rush hour is shorter, all agents are better off.

Figure 5.1 gives a numerical example of how this occurs.[28] 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

[27] These agents will be identical in every dimension, so $t_s = t_e$ and n_1 is a point mass.
[28] This example sets $\delta = 1/3, \xi = 9, s = 32, s^* = 40, t^* = 9{:}00$, and $N = 4,480$.

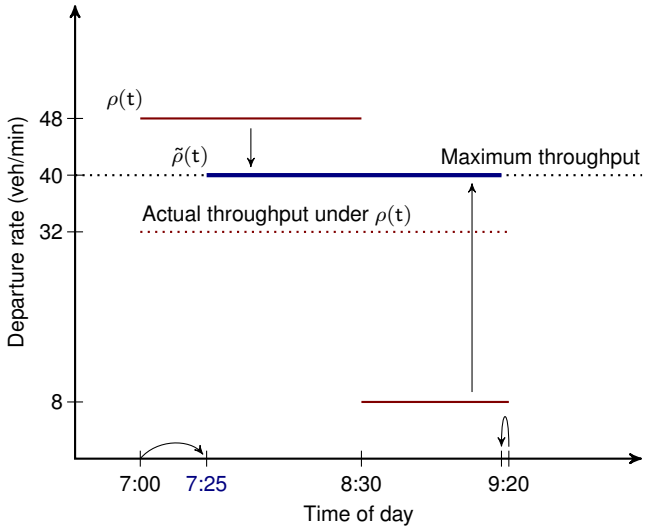


FIGURE 5.1. Tolls can smooth the departure rate, preventing queuing and increasing throughput.

early (or late) to avoid 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 numerical example, rush hour can start 25 minutes later and end 3 minutes earlier.

We can use the effect of pricing on the first driver to depart in the morning as a sufficient statistic for 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, and therefore is better off. As 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.

Allowing for heterogeneous agents makes it difficult for congestion pricing to generate a Pareto improvement. While charging time-varying tolls can increase speeds and 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 increase social welfare, changing the currency used typically hurts poor agents, and in particular it hurts poor inflexible agents.

Changing the currency hurts agents who are both inflexible and poor because the direct effect of changing the currency is to make 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, even a tiny amount of heterogeneity can make it essentially impossible to generate a Pareto improvement when pricing all of the lanes, as Proposition 2 shows.

Proposition 2. *If all agents are homogeneous except for a single agent who is poorer and must arrive on-time at the peak of rush hour, then in the bottleneck model with a throughput drop (i.e., $s < s^*$),* pricing all the lanes *helps all agents before the toll revenue is spent if, and only if,*

$$\frac{s}{s^*} \leq \frac{\alpha_{poor}}{\alpha_{rich}}.$$

Thus if there exists just a single inflexible agent whose value of time is half that of the rest of the agents, then adding tolls must double throughput in order for adding tolls to all of the lanes to generate a Pareto improvement. Inflexible poor agents do exist, and as we saw in Section 2, the largest estimates for s/s^* are 25 percent.

However, by pricing just a portion of the lanes, we can preserve the ability of the poor to pay with their time instead of their money to travel at the peak, while still increasing highway throughput in some of the lanes. While the increase in total throughput is smaller than when pricing all the lanes, and so the social welfare gains are smaller too, doing so makes it easier to obtain a Pareto improvement.

We can formalize this intuition most cleanly by considering 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 the lanes there is no guarantee they are not worse off; however, when we price just a portion of the lanes we can know that they are better off.

Proposition 3. *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 *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 1 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 agent.

While there is no guarantee that value pricing will generate a Pareto improvement when agents are heterogeneous, even when pricing increases throughput, we will shortly derive an intuitive sufficient condition which suggests it typically will. First, consider an example of when it does not generate a Pareto improvement. In this example there are two groups: 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, this is not much different from getting to work early and wasting time waiting for their shift to 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 the order of arrival reverses on any lanes we toll. 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.[29] The cashiers are worse off.

That said, even in this example there are two ways value pricing can still generate a Pareto improvement. First, the increase in capacity due to pricing could be large enough to off-set the harm to the cashiers from being displaced from the peak. Second, we may be able to avoid the cashiers being displaced at all. If the highway capacity was large enough when the road was free for all 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), and if we leave enough of the lanes unpriced so that all the cashiers can continue to travel on an unpriced route and arrive on-time, then the finance professors who already had been traveling at the peak will travel on the priced lanes, and 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.

Furthermore, by considering the dynamics of rush hour it is even possible for pricing all the lanes to help all road users. Consider an example where the rich are more inflexible than the poor, so that instead we have relatively poor yet flexible humanities professors, and rich yet inflexible lawyers. 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.

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 4 (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 traveling at the peak of rush hour when the road is free, then there exists*

[29] 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).

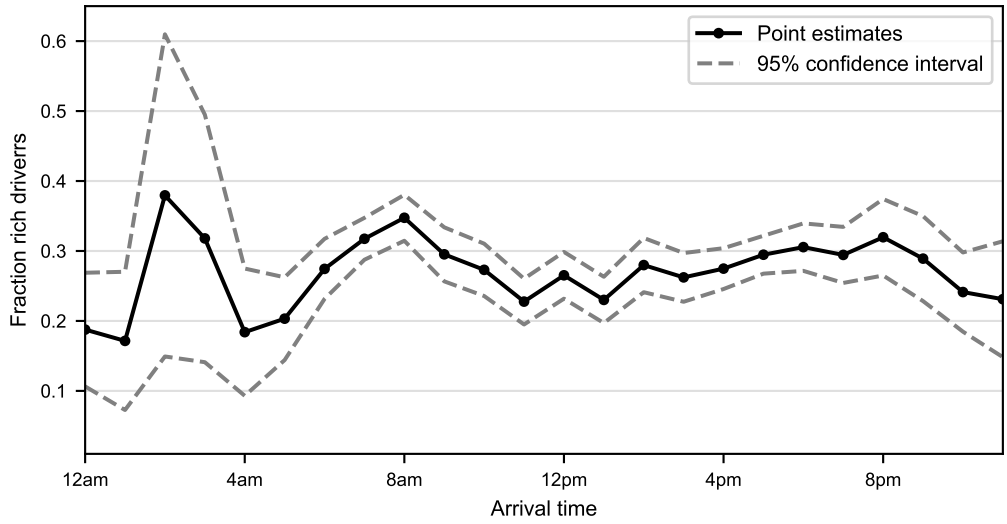


FIGURE 5.2. Fraction of drivers on highway with household income $\geq \$100,000$

Source: 2009 National Household Travel Survey
Notes: Figure plots the fraction of drivers using the interstate, on a weekday, who live in a metropolitan statistical area (MSA), who have household income over $100,000 for each hour of the day. Sample weighted to reflect the population of MSAs. Confidence intervals calculated using Jackknife-2 replicate weights.

a $\lambda_{toll} \in (0, 1]$ such that pricing λ_{toll} of the lanes generates a Pareto improvement even before the revenue is spent. Furthermore, if x percent of those using the road at the peak are rich, then pricing x percent of the lanes generates a Pareto improvement.

Using data from the 2009 National Household Travel Survey I confirm the requirements of this proposition hold. Figure 5.2 plots the fraction of drivers using the highway who are have household incomes greater than $100,000, the highest incomes reported on the survey, and shows that the rich make up more than 25 percent of the vehicles on the road during both the morning and afternoon rush hours. This implies pricing a fourth of the lanes will generate a Pareto improvement even before the revenue is spent.[30]

Proposition 4 can be generalized beyond two groups. To generalize, we require that for all times t the richest agents arriving are at least as rich as those who have

[30]This empirical test is imperfect as I must lump all drivers with household incomes above $100,000 into one group. Ideally I would have finer gradations of value of time. That said, the fact that drivers in this income category are using the road at all times suggests that even with finer gradations I would find a similar pattern.

arrived between t and $t^{\max}$. The logic continues to hold that when we price a small fraction of the lanes the rich who had been arriving at t (or closer to the peak) will use the priced capacity on the priced lane and so no one will be displaced, allowing us to generate a Pareto improvement. Figure 5.2 shows that, if we are comfortable with grouping all those earning more than \$100,000 into the group of "richest agents", this condition is met.

The examples above also highlighted the importance of the correlation between value of time and inflexibility in whether we are able to generate a Pareto improvement. In the first example, the rich were also more flexible than the poor, and this made it more difficult to generate a Pareto improvement, while when the poor were more flexible than the rich pricing all of the lanes generated a Pareto improvement.[31]

I fully characterize the set of parameters for which pricing part or all of the lanes generates a Pareto improvement when there are two groups in Appendix D. Both sets of results are shown visually in Figure 5.3. Because equilibrium trip costs take a different form depending on whether the rich or poor are more inflexible, whether any agents are inframarginal, and whether both groups are on one or two routes, doing so requires solving 19 different cases. Proposition 5 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.*

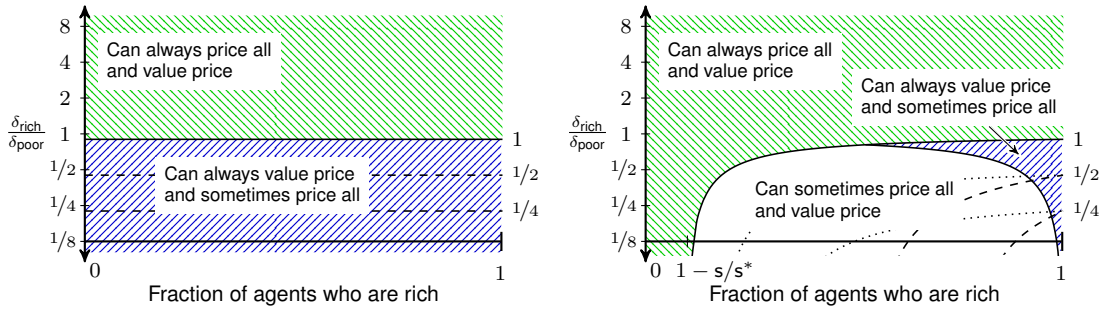
Proposition 5. *If there are two groups then pricing all or part of the lanes is more likely to generate a Pareto improvement prior to spending the toll revenue as*

- *the throughput drop* $(1 - s/s^*)$ *increases,*
- *the ratio of inflexibility of rich to poor* $\left(\delta_{rich}/\delta_{poor}\right)$ *increases, and*
- *income inequality* $\left(\alpha_{rich}/\alpha_{poor}\right)$ *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*

[31]This result also generalizes to an arbitrary number of groups. If the rank correlation between value of time and inflexibility is -1, and $s^* \geq s$, then pricing all of the lanes generates a Pareto improvement before the revenue is used.



(A) $n_{\text{poor}} < s$ (B) $n_{\text{poor}} \geq s$

FIGURE 5.3. Parameter values where pricing leads to a Pareto improvement. In the areas we can only sometimes price our ability to achieve a Pareto improvement depends on whether $\beta_{\text{rich}}/\beta_{\text{poor}}$ is small enough. Several threshold levels of $\beta_{\text{rich}}/\beta_{\text{poor}}$ are drawn with dashed lines for when $n_{\text{poor}} \leq s^*$ or dotted lines for when $n_{\text{poor}} > s^*$. Figures drawn with $s/s^* = 0.9$.

- *the set of parameter values such that pricing all 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.*

The intuition for these results is as follows. First, as the throughput drop increases, it becomes more likely that the efficiency gains from adding tolls offset any harm done by pricing to the poor. Furthermore, if the throughput drop is ridiculously large, then adding tolls makes it possible for everyone to arrive at their desired arrival time while paying a minimal toll, and so regardless of the parameters we can generate a Pareto improvement given a large enough throughput drop.

Second, as the ratio of inflexibility of rich to poor increases or income inequality decreases, the harm done to the poor falls. The harm done to the poor falls when the ratio of inflexibility increases because the poor are less likely to be displaced from their previous arrival time. The harm done to the poor falls when income inequality decreases because changing the currency used to allocate arrival times from time to money has less of an effect.

Finally, except for a zero measure set of knife-edge cases, if pricing all of the lanes generates a Pareto improvement, then so does pricing some of them. This follows from the trip cost functions being continuous almost everywhere, and

implies that by considering pricing just a portion of the lanes we can expand the set of situations where we are able to generate a Pareto improvement.

Figure 5.3 also shows how these results build on prior results. Arnott et al. (1994) show adding tolls to all the lanes generates a Pareto improvement when parameters put us in the top half of Figure 5.3b. Furthermore, this paper's Proposition 4 simplifies Figure 5.3 by saying we can price part or all of the lanes if anywhere in Figure 5.3a or in the top half of Figure 5.3b.

In many cases, obtaining a Pareto improvement requires leaving some of the lanes unpriced, and thus requires leaving some of the potential welfare gains from congestion pricing unrealized. In these cases, policy makers face a trade-off between equity and efficiency: by pricing all of the lanes they hurt some road users but maximize efficiency, while pricing some of the lanes is more equitable but less efficient. I explore this trade-off in two ways. First, in Appendix Proposition D.3, I characterize the ratio of maximum harm to toll revenue per capita when pricing all the lanes. Normalizing by toll revenue per capita helps in two ways; first, it removes the dependence on the levels of the parameters (rather than ratios); second, it quantifies how much of the revenue would need to be rebated on a lump sum basis to generate a Pareto improvement. Second, I calculate the share of total welfare gains that must be sacrificed to obtain a Pareto improvement. Figure 5.4 plots both these objects, and Appendix Figures D.1 and D.2 show the same plots for five sets of alternate parameter values. To a large degree the figures in the appendix are simply shifted versions of Figure 5.4. None of the plots show results for $\delta_{\text{rich}}/\delta_{\text{poor}} > 1$ since, as Figure 5.3 shows, in these cases pricing all of the lanes generates a Pareto improvement.

The left side of Figure 5.4 shows the ratio of the maximum harm done to toll revenue per capita. The harm done often exceeds toll revenue per capita, suggesting that using the revenue to turn congestion pricing into a Pareto improvement may require targeting the use of the revenue very precisely. For the parameters plotted, the harm done increases sharply as the fraction of agents who are rich exceeds a third, because at that point the poor agents become inframarginal. When the poor are inframarginal, they strictly prefer their ex-ante arrival times, and so are severely hurt when adding tolls either causes them to be displaced from the peak or requires them to outbid rich agents to maintain their current arrival time. The

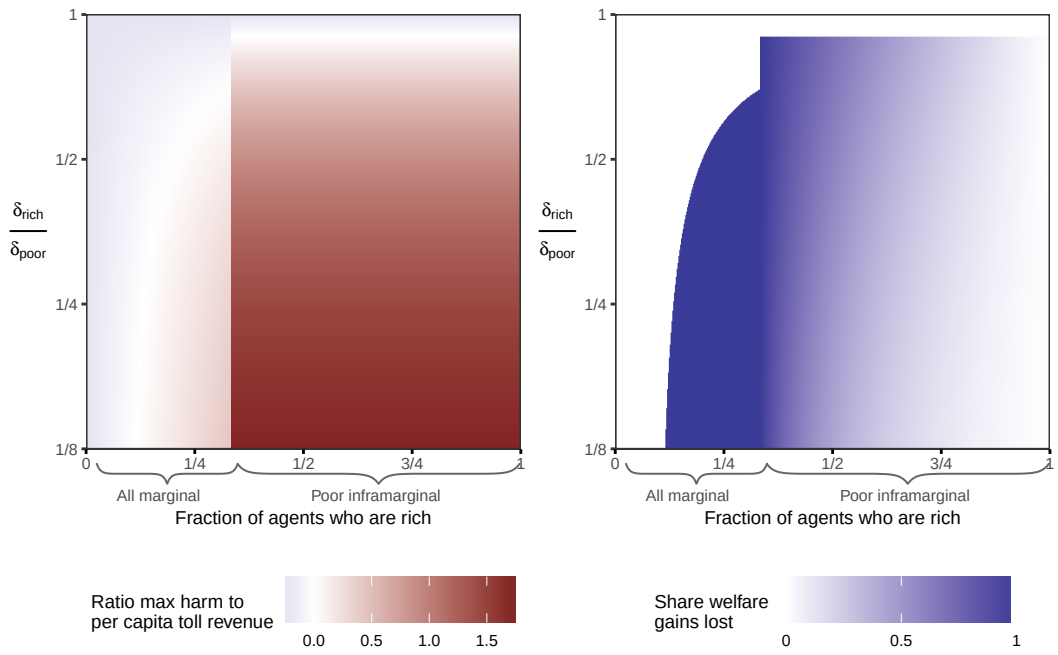


FIGURE 5.4. The ratio of the maximum harm done to toll revenue per capita, *left*, and the share of social welfare gains sacrificed to generate a Pareto improvement, *right*. Figures drawn with $s/s^* = 0.9$, $\left(n_{\text{poor}} + n_{\text{rich}}\right)/s = 1.5$, and $\beta_{\text{rich}}/\beta_{\text{poor}} = 1$.

harm done to the poor is increasing as the ratio of inflexibilities decreases because this increases the harm from being displaced.[32]

The right side of Figure 5.4 shows the share of social welfare gains sacrificed to generate a Pareto improvement. Consistent with Figure 5.3b, pricing all of the lanes generates a Pareto improvement when very few agents are rich or the ratio of inflexibility is high. Once the fraction of agents who are rich exceeds a third, so that the poor are inframarginal when the road is free, then, consistent with Figure 5.3a, pricing a portion of the lanes generates a Pareto improvement, and typically captures a large share of the welfare gains.[33] Once the poor are inframarginal when the road is free, the share of welfare gains lost is decreasing as the fraction of agents who are rich is increasing. This occurs because the larger the fraction of

[32]In general, a lower ratio of inflexibility also increases the probability of displacement, since if the ratio is greater than one then the poor are traveling off-peak when the road is free, and so cannot be displaced from the peak, while if the ratio is less than one the poor are traveling at the peak and so may be displaced (depending $\beta_{\text{rich}}/\beta_{\text{poor}}$).

[33]In Figure 5.4, the median share of the welfare gains captured when pricing a portion of the lanes is 71 percent.

agents who are rich, the smaller the fraction of agents who are poor, and thus the smaller the fraction of lanes that must remain free to preserve the ability of the poor to travel at the peak on unpriced lanes. The smaller the share of lanes that remain free, the smaller the share of welfare gains lost.

Two observations worth highlighting from Figure 5.4 are that, first, it is precisely when pricing all of the lanes does the most harm to the poor that value pricing captures a large share of the possible welfare gains. This occurs when the poor are inframarginal when the road is free. Second, there exist parameter values, such as when 10–20 percent of agents are rich, for which it is impossible to generate a Pareto improvement before using the revenue, but at the same time the harm done from pricing all of the lanes is very small (averaging 6 percent of the toll revenue). For such parameter values, it seems reasonable to expect that the use of the revenue could generate a Pareto improvement, or even that it would be politically feasible to implement congestion pricing without compensating those who are hurt by it.

6. 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. Further, I show that this sufficient condition is satisfied. Finally, while obtaining a Pareto improve often requires sacrificing some of the potential social welfare gains from congestion pricing, it is when the harm done to the poor is at its largest that value pricing captures a large share of the potential welfare gains from congestion pricing.

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 carpooling both take extra time, but provide access to the priced lanes at reduced financial cost.[34] 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 may 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.

References

AASHTO. 2005. *A Policy on Design Standards-Interstate System*. Washington D.C.: American Association of State Highway and Transportation Officials. http://dx.doi.org/10.4135/9781483346526.n55.

Abrantes, Pedro A.L. and Mark R. Wardman. 2011. "Meta-Analysis of UK Values of Travel Time: An Update." *Transportation Research Part A: Policy and Practice* 45: 1–17. doi:10.1016/j.tra.2010.08.003.

Arnott, Richard, André de Palma, and Robin Lindsey. 1990. "Economics of a Bottleneck." *Journal of Urban Economics* 27:111–130. doi:10.1016/0094-1190(90)90028-L.

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

———1994. "The Welfare Effects of Congestion Tolls with Heterogeneous Commuters." *Journal of Transport Economics and Policy* 28:139–161. http://www.jstor.org/stable/20053032.

Arnott, Richard J. 2013. "A Bathtub Model of Downtown Traffic Congestion." *Journal of Urban Economics* 76:110–121. doi:10.1016/j.jue.2013.01.001.

Arnott, Richard J. and Eren Inci. 2010. "The Stability of Downtown Parking and Traffic Congestion." *Journal of Urban Economics* 68:260–276. doi:10.1016/j.jue.2010.05.001.

Banks, James H. 1990. "Flow Processes at a Freeway Bottleneck." *Transportation Research Record*:20–28. http://onlinepubs.trb.org/Onlinepubs/trr/1990/1287/1287-003.pdf.

[34]We could further reduce the cost of taking the bus or carpooling by using the toll revenue to subsidize them. Note that carpooling reduces the financial cost of using the priced lanes because the driver and passenger can split the cost of the toll.

———. 1991. "Two-Capacity Phenomenon at Freeway Bottlenecks: A Basis for Ramp Metering?" *Transportation Research Record* 1320:83–90. http://onlinepubs.trb.org/Onlinepubs/trr/1991/1320/1320-011.pdf.

Beesley, Michael. 1973. "Road Pricing: Economics, Techniques and Policy." In *Urban Transport: Studies in Economic Policy*. London: Butterworths:223–286.

Bertini, Robert L. and Monica T. Leal. 2005. "Empirical Study of Traffic Features at a Freeway Lane Drop." *ASCE Journal of Transportation Engineering* 131:397–407. doi:10.1061/(asce)0733-947x(2005)131:6(397).

Bertini, Robert and Shazia Malik. 2004. "Observed Dynamic Traffic Features on Freeway Section with Merges and Diverges." *Transportation Research Record: Journal of the Transportation Research Board* 1867:25–35. doi:10.3141/1867-04.

Braid, Ralph M. 1996. "Peak-Load Pricing of a Transportation Route with an Unpriced Substitute." *Journal of Urban Economics* 40:179–197. doi:10.1006/juec.1996.0028.

Cassidy, Michael J. and Robert L. Bertini. 1999. "Some Traffic Features at Freeway Bottlenecks." *Transportation Research Part B: Methodological* 33:25–42. doi:10.1016/S0191-2615(98)00023-X.

Cassidy, Michael J. and Jittichai Rudjanakanoknad. 2005. "Increasing the Capacity of an Isolated Merge by Metering Its On-Ramp." *Transportation Research Part B: Methodological* 39:896–913. doi:10.1016/j.trb.2004.12.001.

Chung, Koohong, Jittichai Rudjanakanoknad, and Michael J. Cassidy. 2007. "Relation Between Traffic Density and Capacity Drop at Three Freeway Bottlenecks." *Transportation Research Part B: Methodological* 41:82–95. doi:10.1016/j.trb.2006.02.011.

Cohen, Yuval. 1987. "Commuter Welfare under Peak-Period Congestion Tolls: Who Gains and Who Loses." *International Journal of Transport Economics* 14:239–266. http://www.jstor.org/stable/42748188.

Currie, Janet and Reed Walker. 2011. "Traffic Congestion and Infant Health: Evidence from E-ZPass." *American Economic Journal: Applied Economics* 3:65–90. doi:10.1257/app.3.1.65.

Daganzo, Carlos. 1996. "The Nature of Freeway Gridlock and How to Prevent It." In *Traffic and Transportation Theory*:629–646. Lyon, France: Pergamon.

De Meza, David and J. R. Gould. 1987. "Free Access versus Private Property in a Resource: Income Distributions Compared." *Journal of Political Economy* 95:1317–1325. doi:10.1086/261518.

de Palma, André and Robin Lindsey. 2002. "Comparison of Morning and Evening Commutes in the Vickrey Bottleneck Model." *Transportation Research Record: Journal of the Transportation Research Board* 1807:26–33. doi:10.3141/1807-04.

Duranton, Gilles and Matthew A. Turner. 2011. "The Fundamental Law of Road Congestion: Evidence from US Cities." *American Economic Review* 101:2616–2652. doi:10.1257/aer.101.6.2616.

Elefteriadou, Lily, Roger P. Roess, and William R. McShane. 1995. "Probabilistic Nature of Breakdown at Freeway Merge Junctions." *Transportation Research Record* 1484:80–89. https://trid.trb.org/view.aspx?id=451862.

Evans, Alan W. 1992. "Road Congestion: The Diagrammatic Analysis." *Journal of Political Economy* 100:211–217. doi:10.1086/261814.

Fosgerau, Mogens and Kenneth A. Small. 2013. "Hypercongestion in Downtown Metropolis." *Journal of Urban Economics* 76:122–134. doi:10.1016/j.jue.2012.12.004.

Foster, C. D. 1975. "A Note on the Distributional Effects of Road Pricing." *Journal of Transport Economics and Policy* 9:186–187. http://www.jstor.org/stable/20052404.

Giuliano, Genevieve. 1992. "An Assessment of the Political Acceptability of Congestion Pricing." *Transportation* 19:335–358. doi:10.1007/BF01098638.

Guan, Yu, Jishuang Zhu, Ning Zhang, and Xiaobao Yang. 2009. "Traffic Flow Characteristics of Bottleneck Segment with Ramps on Urban Expressway." In *International Conference on Transportation Engineering 2009*:1679–1684: American Society of Civil Engineers. doi:10.1061/41039(345)278.

Hall, Fred L. and Kwaku Agyemang-Duah. 1991. "Freeway Capacity Drop and the Definition of Capacity." *Transportation Research Record* 1320:1–98.

Hall, Jonathan D. 2017. "Improving the Fit of Structural Congestion Models." *Working Paper*.

Halvorson, Randy and Kenneth R. Buckeye. 2006. "High-Occupancy Toll Lane Innovations: I-394 MnPASS." *Public Works Management & Policy* 10:242–255. doi:10.1177/1087724X06288331.

Hendrickson, Chris and George Kocur. 1981. "Schedule Delay and Departure Time Decisions in a Deterministic Model." *Transportation Science* 15:62–77. doi:10.1287/trsc.15.1.62.

Hotelling, Harold. 1929. "Stability in Competition." *Economic Journal* 39:41–57. doi:10.2307/2224214.

Hurdle, Van F. and Pradip K. Datta. 1983. "Speeds and Flows on an Urban Freeway: Some Measurements and A Hypothesis." *Transportation Research Record* 905:127–137.

Hymel, Kent. 2009. "Does Traffic Congestion Reduce Employment Growth?" *Journal of Urban Economics* 65:127–135. doi:10.1016/j.jue.2008.11.002.

Johnson, M. Bruce. 1964. "On the Economics of Road Congestion." *Econometrica* 32:137–150. doi:10.2307/1913739.

Lave, Charles. 1994. "The Demand Curve Under Road Pricing and the Problem of Political Feasibility." *Transportation Research Part A: Policy and Practice* 28:83–91. doi:10.1016/0965-8564(94)90030-2.

Leclercq, Ludovic, Jorge A. Laval, and Nicolas Chiabaut. 2011. "Capacity Drops at Merges: An Endogenous Model." *Procedia - Social and Behavioral Sciences* 17: 12–26. doi:16/j.sbspro.2011.04.505.

Light, Thomas. 2009. "Optimal Highway Design and User Welfare under Value Pricing." *Journal of Urban Economics* 66:116–124. doi:10.1016/j.jue.2009.05.003.

Lindsey, Robin. 2006. "Do Economists Reach A Conclusion on Road Pricing? The Intellectual History of an Idea." *Econ Journal Watch* 3:292–379.

Lindsey, Robin and Erik Verhoef. 2008. "Congestion Modeling." In David Hensher and Kennth Button eds. *Handbook of Transportation Modelling*. New York: Elsevier. 2nd edition:417–441. `https://doi.org/10.1108/9780857245670-021`.

Liu, Louie Nan and John F. McDonald. 1998. "Efficient Congestion Tolls in the Presence of Unpriced Congestion: A Peak and Off-Peak Simulation Model." *Journal of Urban Economics* 44:352–366. doi:10.1006/juec.1997.2073.

May, Anthony D, Simon P. Shepherd, and John J. Bates. 2000. "Supply Curves for Urban Road Networks." *Journal of Transport Economics and Policy* 34:261–290. `http://www.jstor.org/stable/20053846`.

Muñoz, Juan Carlos and Carlos F. Daganzo. 2002. "The Bottleneck Mechanism of a Freeway Diverge." *Transportation Research Part A: Policy and Practice* 36:483–505. doi:10.1016/S0965-8564(01)00017-9.

Newell, Gordon F. 1987. "The Morning Commute for Nonidentical Travelers." *Transportation Science* 21:74–88. doi:10.1287/trsc.21.2.74.

_____1988. "Traffic Flow for the Morning Commute.." *Transportation Science* 22: 47. doi:10.1287/trsc.22.1.47.

Oh, Simon and Hwasoo Yeo. 2012. "Estimation of Capacity Drop in Highway Merging Sections." *Transportation Research Record: Journal of the Transportation*

Research Board 2286:111–121. doi:10.3141/2286-13.

Perez, Benjamin G. and Gian-Claudia Sciara. 2003. "A Guide for HOT Lane Development." Federal Highway Administration. Washington, D.C.

Persaud, Bhagwant, Sam Yagar, and Russel Brownlee. 1998. "Exploration of the Breakdown Phenomenon in Freeway Traffic." *Transportation Research Record: Journal of the Transportation Research Board* 1634:64–69. doi:10.3141/1634-08.

Pigou, Arthur Cecil. 1920. *The Economics of Welfare*. London: Macmillan and co., ltd.. 1st edition.

Rudjanakanoknad, Jittichai. 2005. "Increasing Freeway Merge Capacity Through On-Ramp Metering." Dissertation. University of California at Berkeley. Berkeley, CA. `https://doi.org/10.1016/j.trb.2004.12.001`.

Schrank, David, Bill Eisele, Tim Lomax, and Jim Bak. 2015. "2015 Urban Mobility Scorecard." Texas A&M Transportation Institute. College Station, Texas.

Small, Kenneth A. 1983. "The Incidence of Congestion Tolls on Urban Highways." *Journal of Urban Economics* 13:90–111. doi:10.1016/0094-1190(83)90047-5.

_____1992. "Using the Revenues From Congestion Pricing." *Transportation* 19: 359–381. doi:10.1007/BF01098639.

Small, Kenneth A. and Xuehao Chu. 2003. "Hypercongestion." *Journal of Transport Economics and Policy* 37:319–352. `http://www.jstor.org/stable/20053940`.

Small, Kenneth A. and Erik T. Verhoef. 2007. *The Economics of Urban Transportation*. New York: Routledge.

Small, Kenneth A., Clifford Winston, and Jia Yan. 2005. "Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability." *Econometrica* 73: 1367–1382. doi:10.1111/j.1468-0262.2005.00619.x.

_____2006. "Differentiated Road Pricing, Express Lanes, and Carpools: Exploiting Heterogeneous Preferences in Policy Design." *Brookings-Wharton Papers on Urban Affairs*:53–96. doi:10.1353/urb.2006.0027.

Srivastava, Anupam and Nikolas Geroliminis. 2013. "Empirical Observations of Capacity Drop in Freeway Merges with Ramp Control and Integration in a First-Order Model." *Transportation Research Part C: Emerging Technologies* 30:161–177. doi:10.1016/j.trc.2013.02.006.

Starkie, David. 1986. "Efficient and Politic Congestion Tolls." *Transportation Research Part A: General* 20:169–173. doi:10.1016/0191-2607(86)90044-0.

Stiglitz, Joseph. 1998. "Distinguished Lecture on Economics in Government: The Private Uses of Public Interests: Incentives and Institutions." *Journal of Economic*

Perspectives 12:3–22. doi:10.1257/jep.12.2.3.

van den Berg, Vincent and Erik 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* 95:983–992. doi:10.1016/j.jpubeco.2010.12.003.

Verhoef, Erik T. 1999. "Time, Speeds, Flows and Densities in Static Models of Road Traffic Congestion and Congestion Pricing." *Regional Science and Urban Economics* 29:341–369. doi:10.1016/S0166-0462(98)00032-5.

———2001. "An Integrated Dynamic Model of Road Traffic Congestion Based on Simple Car-Following Theory: Exploring Hypercongestion." *Journal of Urban Economics* 49:505–542. doi:10.1006/juec.2000.2203.

———2005. "Speed-Flow Relations and Cost Functions for Congested Traffic: Theory and Empirical Analysis." *Transportation Research Part A: Policy and Practice* 39:792–812. doi:10.1016/j.tra.2005.02.023.

Vickrey, William S. 1969. "Congestion Theory and Transport Investment." *American Economic Review* 59:251–260. http://www.jstor.org/stable/1823678.

———1973. "Pricing, Metering, and Efficiently Using Urban Transportation Facilities." *Highway Research Record* 476:36–48. http://pubsindex.trb.org/view.aspx?id=92531.

———1987. "Marginal and Average Cost Pricing." In Steven N. Durlauf and Lawrence E. Blume eds. *The New Palgrave Dictionary of Economics*. Basingstoke: Palgrave Macmillan. 1st edition. http://www.dictionaryofeconomics.com/article?id=pde1987_X001391.

von Thünen, Johann Heinrich. 1930. *Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*: Jena: Gustav Fischer.

Walters, Alan A. 1961. "The Theory and Measurement of Private and Social Cost of Highway Congestion." *Econometrica* 29:676–699. doi:10.2307/1911814.

Zhang, Lei and David Levinson. 2004. "Some Properties of Flows at Freeway Bottlenecks." *Transportation Research Record: Journal of the Transportation Research Board* 1883:122–131. doi:10.3141/1883-14.