Tolling Roads to Improve Reliability*

Jonathan D. Hall
University of Toronto

Ian Savage
Northwestern University

July 23, 2019

Abstract

A significant cost of traffic congestion is unreliable travel times. A major source of this unreliability is that when roads are congested, interactions between drivers can lead to capacity unexpectedly falling. For example, collisions can close lanes and aggressive lane changers can slow traffic. This paper analyzes how tolls should be set when accounting for such endogenous reliability. We find tolls should be higher and maximum flow lower than we might naïvely expect; and that such tolls make homogeneous drivers better off, even before the toll revenue is used. Simulations suggest the socially optimal maximum departure rate is fifteen percent below that which maximizes expected throughput, and that tolling reduces private costs by almost ten percent.

*We are grateful for helpful comments from John Cairncross, Mogens Fosgerau, Robin Lindsey, Kenneth Small, a Co-Editor (Gilles Duranton), two anonymous referees, and participants at the annual meetings of the International Transportation Economics Association. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

‘

1 Introduction

Traffic congestion is a serious problem for cities worldwide. In the United States, 41 percent of mayors report that traffic congestion is one of their top three problems (Bloomberg Philanthropies, 2018). More than half of the time lost to traffic congestion is due to non-recurring congestion; that is, congestion caused by crashes, bad weather, and other shocks (Dowling et al., 2004; Kwon et al., 2006). Non-recurring congestion leads to unpredictable travel times. Due to this unpredictability, drivers have to depart earlier than they would prefer. On good days they arrive with time to spare and on bad days they suffer the consequences of arriving late for work or an important appointment. This lack of reliability accounts for between 30 and 70 percent of the total cost of congestion (Small et al., 2005; Bento et al., 2017).

While some of the shocks which cause non-recurring congestion are exogenous, such as bad weather, others shocks are endogenous. For endogenous shocks, their probability of occurring increases with traffic flow (the number of vehicles passing a given point per lane per hour). For example, the rate at which crashes occur per vehicle-mile traveled more than doubles as the flow increases from 1,500 to 2,000 vehicles per hour per lane (vphpl) (Kononov et al., 2012).[1] Furthermore, as traffic flow increases the probability that a small shock leads to a collapse in outflow also increases. Transportation engineers call this collapse “flow breakdown” or a “capacity drop”. Numerous types of small shocks can cause flow breakdown, including vehicles weaving between lanes, excessively slow vehicles, aggressive driving, tailgating, and sharp braking. Lorenz and Elefteriadou (2001) estimate that as flow increases from 1,900 to 2,200 vphpl the probability of breakdown increases by 50 percentage points. The magnitude of the breakdown appears to

[1]Much of the literature on the relationship between traffic flow and crash risk finds a U-shaped relationship, with the minimum risk when flow is around half of capacity (around 1,000 vphpl). Zhou and Sisiopiku (1997) and Kononov et al. (2012) find evidence that the high risk of a crash at low levels of flow is driven by observations from late at night when the majority of crashes involve alcohol, drugs, or falling asleep.

be quite large, with documented declines in capacity of 25 percent (Persaud et al., 1998).[2]

This paper's contribution is to analyze how to implement tolls in the face of this endogenous non-recurring congestion. We do so by extending the bottleneck model of dynamic congestion (Vickrey, 1969; Arnott et al., 1993) to allow for an endogenous probability of breakdown, where the probability is increasing in the flow.

To keep the analysis tractable, we limit the space of possible toll schedules to a commonly observed class of tolls: preset toll schedules that vary by time of day to keep traffic flow from exceeding a target maximum. For example, California's SR-91 Express Lanes have a stated maximum average flow of 1,600 vphpl.[3] Other facilities implicitly have a maximum flow target because they aim to achieve a set minimum average vehicle speed. The majority of facilities in the United States with time-varying tolls have a preset schedule.[4] These tolls are anticipatory, and differ from dynamic tolls which respond in real time to realized traffic levels. Dynamic tolls, with one exception, have only been used when pricing a portion of

[2]For further discussion and evidence of flow breakdown, see (among others) Dong and Mahmassani (2009), Kim et al. (2010), Chen et al. (2014), Chen and Ahn (2015), Luo et al. (2015), Qian et al. (2017), Kontorinaki et al. (2017), and Geistefeldt and Shojaat (2019). There is a parallel literature on "phantom traffic jams" that explores how variations in speeds of cars following each other can lead to breakdown (Sugiyama et al., 2008; Wilson and Ward, 2011). While there is broad acceptance that breakdown occurs and is probabilistic, there is debate over the magnitude of the breakdown and the probability of it occurring. For example, Doig et al. (2013) cautions that sometimes what appears to be flow breakdown is actually the result of unobserved congestion on a downstream link.

[3]Tolls vary by day and by hour. Tolls are reviewed and adjusted every six months. Tolls are increased or decreased depending on whether observed volumes are greater than or less than the target flow. See the detailed policy at `https://www.91expresslanes.com/wp-content/uploads/2014/04/TollPolicy.pdf`.

[4]Data is from a database maintained by the U.S. Transportation Research Board's Standing Committee on Managed Lanes, as of July 2018. An earlier report by the Federal Highway Administration (2016) found that 47 (23%) of the 210 toll highway facilities in the United States as of January 1, 2015, excluding bridges and tunnels, had toll schedules that varied by time of day or traffic conditions. Of the 47 with tolls that varied by time of day, two-thirds had prices that were on a preset schedule.

the lanes, because for dynamic tolls to be effective drivers need to have the ability to immediately choose an alternate route, or cancel or delay their trip.[5,6]

We find two important results. First, tolls should be higher, and maximum flow lower, than we might naïvely expect. We use as the naïve benchmark the toll schedule which maximizes expected throughput.[7] Increasing the toll to reduce the maximum flow rate below that which maximizes expected throughput has a cost: it increases the length of the peak period. However, it has two benefits: it reduces the probability of breakdown, and it reduces the severity of congestion on days breakdown occurs by spreading out when drivers depart.

Second, tolling leaves homogeneous drivers better off, even before the resulting revenue is used. Furthermore, providing that the probability of breakdown when the road is untolled is not too high, the improvement in drivers' private costs occurs despite a reduction in the average flow rate relative to when no tolls were charged. This is due to drivers valuing the reduction in uncertainty and travel time more than they dislike paying the toll. Leaving drivers better off is important as a major barrier to implementing tolling is the concern that it hurts drivers.[8] While drivers are, in reality, heterogeneous, this result highlights that accounting for the value drivers place on reliability improves the welfare consequences of tolling.

[5]The exception is I-66 in Virginia. This facility was originally only for high-occupancy vehicles during peak periods. Solo drivers can now pay dynamic tolls to use the facility.

[6] The purpose of our tolls is to smooth traffic flow and reduce the chance that interactions between drivers leads to breakdown. Another tool for doing so is ramp metering. Ramp metering and tolling are complements rather than substitutes. Tolling smooths traffic flow at a macro level by incentivizing drivers to modify the times they leave home, while ramp metering smooths traffic flow at the micro level by physically preventing drivers from entering the road. Tolling solves two problems with ramp metering. The first is that ramp metering can lead to large queues off the highway, which while likely better than large queues on the highway, still has a large social cost. The second is that ramp metering inefficiently penalizes drivers entering the highway in the urban core relative to the suburbs.

[7]We use this as our benchmark for two reasons. First, it is regularly discussed when setting policy, and second, in the standard bottleneck model, the toll schedule which maximizes social welfare also maximizes throughput.

[8]For example, Lindsey and Verhoef (2008) argue "most likely, these losses are the root of the longstanding opposition to congestion tolling". See Hall (2018) for a longer discussion of this.

We simulate our model using parameter values from Lorenz and Elefteriadou (2001), U.S. Department of Transportation (2016), and Hall (2019b). Our simulations suggest that the socially optimal maximum departure rate is fourteen percent below that which maximizes expected throughput, with the average toll more than three times those which maximize expected throughput. These tolls reduce private costs by almost ten percent.

While our model is of a single link, the problem of endogenous non-recurrent congestion applies equally to networks. Charging time-varying tolls smooths entry to a network, reducing the probability of flow breakdown, as well as its consequences, throughout the network. This is especially applicable to downtown traffic congestion.[9]

2 Literature Review

The paper contributes to two literatures. First, it builds on a literature analyzing congestion mitigation policies, such as tolling or information, in the face of uncertainty. Within this literature, we relate most closely to an influential paper by Arnott et al. (1999), and an innovative paper by Fosgerau and Lindsey (2013).[10] The former uses the bottleneck model with exogenous supply and demand shocks to show that providing imperfect information can reduce social welfare. The latter analyzes the effect of traffic crashes that are modeled as exogenous supply shocks that can happen at any point during the day. Tolling can improve social welfare by reducing the cost of a crash, since there are fewer drivers on the road at a given time. We build on this work by allowing for endogenous supply shocks, which allows us to show tolling helps by re-arranging traffic flow to both reduce the probability and consequences of the supply shock.

[9]For examples of papers directly modeling downtown traffic congestion, see Arnott (2013) and Fosgerau (2015).

[10]Other papers in this literature include Noland and Small (1995), Noland (1997), and Lindsey (1999). Empirical papers estimating the value of reliability include Small et al. (2005) and Bento et al. (2017).

The paper is also closely related to Zhu et al. (2017). Their model has a random element to the queuing time at a bottleneck that increases with the total number of drivers. They analyze implementation of a uniform toll using numerical examples. A higher uniform toll reduces total overall demand and the variability in travel time. The current paper differs by solving for equilibrium analytically, demonstrating that flow should be reduced below that which maximizes expected throughput, and concluding that it is possible to make all drivers better off.

The paper also contributes to a second literature concerning the welfare consequences of tolling. This literature finds that changes in the departure rate and changes in private cost are negatively correlated. For example, in the traditional static model adding tolls increases private costs while reducing departures (e.g. Walters, 1961), and in the standard bottleneck model adding tolls leaves the average departure rate and private costs unchanged (Arnott et al., 1993). In models with hypercongestion, adding tolls increases the average departure rate and reduces private costs (Fosgerau and Small, 2013; Hall, 2018). Our contribution is to highlight the value of reliability, and show that this makes it possible to reduce private costs while reducing the average departure rate.

3 Model

Our model introduces probabilistic flow breakdown into the standard bottleneck congestion model of Vickrey (1969) and Arnott et al. (1990, 1993). This model is dynamic, and since drivers have preferences over arrival times, disliking arriving early or late, it allows drivers to be risk averse. Having risk averse drivers matters because a major cost of probabilistic flow breakdown is that it increases uncertainty.

In common with the standard model, a single link connects where people live to where they work. There are no alternative routes or modes. The only source of congestion is a bottleneck. For simplicity, travel time before and after the bottleneck and vehicle operating costs are normalized to zero. Consequently, the departure rate of drivers from home is identical to the inflow rate into the bottleneck. The terms "departure rate" and "inflow" are used synonymously, using the former

term in the modeling, and the latter when describing policy options for toll road operators.

Each morning the maximum bottleneck capacity value, denoted by s, is drawn from a distribution with a continuous cumulative distribution function of $P(s)$ and a probability density function of $p(s)$. This distribution is non-degenerate and has a lower bound of s_B. The subscript B indicates a "bad" or post-breakdown state. Let $r(t)$ denote the departure rate of drivers from home at time t. As soon as $r(t)$ surpasses s, interaction between drivers causes the flow to break down and the capacity of the bottleneck falls to s_B. It remains at this level until the resulting queue dissipates, at which point the highway capacity reverts to s.

The paper is concerned with equilibria where breakdown is not an everyday occurrence. If breakdown were a daily occurrence, the model would be equivalent to the standard bottleneck model but with decreased bottleneck capacity when a queue forms (as in Hall (2018)). To make the model interesting $P(s)$ should only equal 1 for values of s greater than the largest departure rate. This restriction is not imposed on the derivations, although discussion of the results often takes it as given.

There is a mass N of homogeneous drivers in single-occupant vehicles. The number of drivers is perfectly inelastic. Drivers have a common desired arrival time at work, denoted as t^*. Drivers choose when to depart from home to minimize their expected trip costs.

Let t_S and t_E denote the start and end of the period of departures. Further define $\bar{r}(t) = \max_{x \in [t_S, t]} r(x)$ as the largest departure rate that has already occurred. If $\bar{r}(t) \leq s$, breakdown has not occurred, there has been no queuing, and travel time, denoted as T, is zero. However, if $\bar{r}(t) > s$, breakdown has occurred, and the queue evolves according to

$$\frac{dQ(t,s)}{dt} = r\left(t\right) - s_B,$$

where $Q(t,s)$ represents the number of vehicles in the queue, and travel time is given by

$$T(t,s) = \frac{Q(t,s)}{s_B}. \tag{1}$$

Drivers' expected trip costs are the sum of the cost of travel time, the cost of an early or late arrival at work, and any toll payments. The expected cost is described by the following function

$$c(t) = \int_{s_B}^{\infty} \left[\alpha T(t,s) + D(t + T(t,s))\right] dp(s) + \tau(t), \tag{2}$$

where α is the hourly cost of travel time, D represents the cost of arriving early or late, and $\tau(t)$ is the toll. We follow the standard bottleneck model in assuming schedule delay costs are piecewise-linear, so

$$D(x) = \begin{cases} -\beta(x - t^*) & x \leq t^*, \\ \gamma(x - t^*) & x > t^*; \end{cases}$$

with β and γ being the hourly cost of being early or late. Consistent with the literature, we assume $\beta < \alpha$ to avoid a mass of drivers departing at the same time.

Drivers have rational expectations, and in deciding when to depart are aware of the probability of breakdown and its effects.[11] However, on any given day, drivers do not observe the maximum bottleneck capacity that is drawn. They cannot decide to stay home or deviate from their chosen departure time if breakdown either has occurred or is about to occur.

4 Equilibrium without tolls

This section describes the stochastic user equilibrium in the absence of tolls. The superscript U (for untolled) indicates equilibrium values. As is usual in these models, there are two equilibrium conditions. The first is that supply equals demand. The number of people who want to travel equals the number who actually travel. The second is that no driver can find a profitable time deviation. This implies that the trip cost is the same for all departure times that people choose, and is not any lower at any times that people do not choose.

[11] It may be that the distribution of capacity varies with some observable signal, such as the weather or the season. In this case, $P(s)$ is the cumulative distribution function conditional on that signal, and all of our results carry through within each information state.

To preview the results, the departure rate from home is non-increasing across the peak period, the period of time when drivers are on the road. Consequently, when the value of s is drawn each morning, either breakdown happens immediately or it does not happen at all. A non-increasing departure rate further implies it is not possible for the highway to recover and then break down again. As a result, drivers have either a good day when breakdown does not occur at all, or a bad day when the bottleneck is at capacity s_B for the entire peak period.

In the following subsections, we characterize the departure rate, the interval over which drivers depart, and equilibrium trip costs.

4.1 Equilibrium departure rates

In this subsection we prove that departure rates are non-increasing over the period when drivers depart and characterize the departure rate. The intuition for departure rates being non-increasing starts with the equilibrium requirement that all drivers must be indifferent between all departure times that are actually chosen. Because different departure times lead to different expected schedule delay costs, expected travel times must vary to keep drivers indifferent. Expected schedule delay costs are U-shaped, being high for very early departures, low for departures that are early sometimes and late sometimes, and high for very late departures. To keep drivers indifferent between departure times, expected travel time costs must have an inverse U-shape, reaching their peak when expected schedule delay costs are the lowest. Thus, the first derivative of travel times is non-increasing, and since the departure rate is proportional to the first derivative of travel times, this implies the departure rate is non-increasing. This is formalized in the following lemma.

Lemma 1. *The departure rate is non-increasing after the first departure:*

$$r'(t) \leq 0 \quad \forall \quad t \geq t_S.$$

Proof. The proof is in three steps. The first step is deriving an expression for the second difference of trip costs during the time that agents are departing. Consider

four times at which drivers depart, $t_a < t_b < t_c < t_d$. To reduce notational clutter, assume $t_b - t_a = t_d - t_c$. To simplify notation, define

$$\begin{aligned}
\Delta D_b(s) &= D(t_b + T(t_b, s)) - D(t_a + T(t_a, s)), \\
\Delta D_d(s) &= D(t_d + T(t_d, s)) - D(t_c + T(t_c, s)), \\
\Delta T_b(s) &= T(t_b, s) - T(t_a, s), \text{ and} \\
\Delta T_d(s) &= T(t_d, s) - T(t_c, s).
\end{aligned}$$

With this notation, the change in the trip cost between t_a and t_b can be written as

$$c(t_b) - c(t_a) = \int_{s_B}^{\infty} \left[\alpha \Delta T_b(s) + \Delta D_b(s)\right] dp(s).$$

Comparing the change in trip costs between t_c and t_d to that between t_b and t_a produces

$$\begin{aligned}
\left[c(t_d) - c(t_c)\right] - \left[c(t_b) - c(t_a)\right] = \\
\int_{s_B}^{\infty} \left\{\alpha \left[\Delta T_d(s) - \Delta T_b(s)\right] + \Delta D_d(s) - \Delta D_b(s)\right\} dp(s). \quad (3)
\end{aligned}$$

The second step is to impose the equilibrium constraint that the expected cost at each departure time actually chosen is the same. Hence (3) equals zero.

The third step is to show that the convexity of D means that for (3) to equal zero, the departure rate must be non-increasing after the first departure. Because D is weakly convex, its average slope between t_c and t_d is weakly greater than its average slope between t_a and t_b. Letting $\omega(s)$ be the average slope between t_a and t_b,

$$\omega(s) = \frac{\Delta D_b(s)}{t_b - t_a + \Delta T_b(s)},$$

and

$$\Delta D_d(s) \geq \omega(s) \left(t_d - t_c + \Delta T_d(s)\right).$$

These two equations imply that

$$\Delta D_d(s) - \Delta D_b(s) \geq \omega(s) \left[\Delta T_d(s) - \Delta T_b(s)\right]. \quad (4)$$

Imposing that (3) equals zero and substituting in (4) yields

$$0 \geq \int_{s_B}^{\infty} \{\alpha [\Delta T_d(s) - \Delta T_b(s)] + \omega(s) [\Delta T_d(s) - \Delta T_b(s)]\} dp(s) = \int_{s_B}^{\infty} (\alpha + \omega(s)) [\Delta T_d(s) - \Delta T_b(s)] dp(s) \quad (5)$$

By assumption, $\alpha + D'(x) > 0$ for all x, so $\alpha + \omega(s) > 0$. Thus, for (5) to be non-positive, we need $\Delta T_d(s) - \Delta T_b(s) \leq 0$. However, if $r(t)$ is increasing over (t_a, t_d) then $\Delta T_d(s) - \Delta T_b(s) \geq 0$ for all s, with the inequality strict for $s \leq \bar{r}(t_d)$. Thus, $r(t)$ cannot be increasing. ■

Once we know the departure rate is non-increasing, then we know that breakdown either happens at the start of the peak period, or not at all (i.e., $\bar{r}(t)$ is constant for $t \geq t_S$). Given this, we simplify our notation for travel time by defining $T(t) = T(t, s) \forall s < r(t_S)$, the travel time that occurs on days when breakdown occurs. This allows us to write the expected trip cost as

$$c(t) = P(r(t_S)) [\alpha T(t) + D(t + T(t) - t^*)] + [1 - P(r(t_S))] D(t - t^*) \quad (6)$$

We can now derive conditions for the timing of the first (t_S) and last (t_E) departures. Since drivers prefer to arrive at t^*, in equilibrium $t_S \leq t^* \leq t_E$. Furthermore, the last departure occurs at the first time (weakly) later than t^* such that, even in the absence of any other departures, the trip cost is increasing in departure time. This can occur for two reasons. The first is that once travel times on bad days return to zero, there is no benefit to drivers from further delaying their departure. The second is that marginal expected schedule delay costs have grown large enough that they are greater than the expected time savings from leaving later. An unfortunate implication is that there are two equilibrium cases, depending on which condition for the last departure time applies. These conditions are formalized in the following lemma.

Lemma 2. *When there is no toll,*

$$t_S \leq t^*, \text{ and}$$
$$t_E = \min \left\{ t | t \geq t^* \text{ and } \left(T(t) = 0 \text{ or } P(r(t_S)) < \frac{D'(t)}{\alpha + D'(t)} \right) \right\}.$$

Proof. First, we prove $t_S \leq t^*$. Assume by way of contradiction that the first departure occurs after t^*, then a driver departing at t^* has no travel time on a bad day, and no schedule delay on either a good or bad day. Thus, $c(t^*) = 0$ and since $c(t_S) \geq 0$, departing at t^* would be a profitable deviation.

Next, we prove $t_E \geq t^*$. Assume by way of contradiction that $t_E < t^*$. This implies $T(t^*) < T(t_E)$, $D(t^*) < D(t_E)$, and $D(t^* + T(t^*)) < D(t_E + T(t^*))$, and so $c(t^*) < c(t_E)$, and thus departing at t^* is a profitable deviation.

Finally, we derive conditions on the last departure time. The last departure occurs at the first time where $c'(t) > 0$ when $r(t) = 0$ and $t \geq t^*$. To find when $c'(t) > 0$ when $r(t) = 0$ and $t \geq t^*$, differentiate (6) and note that

$$r(t) = 0 \Rightarrow T'(t) = \begin{cases} 0 & \text{if } T(t) = 0, \text{ and} \\ -1 & \text{if } T(t) > 0. \end{cases}$$

Doing so yields

$$\begin{aligned} c'(t) &= \begin{cases} P(r(t_S))\left[-\alpha\right] + \left[1 - P(r(t_S))\right] D'(t) & \text{if } T(t) > 0, \text{ and} \\ D'(t) & \text{if } T(t) = 0; \end{cases} \\ \Rightarrow c'(t) > 0 &\Leftrightarrow \left(T(t) = 0 \text{ or } P(r(t_S)) < \frac{D'(t)}{\alpha + D'(t)} \right). \end{aligned}$$

■

Given a non-increasing departure rate and the conditions for the timing of the first and last departures, we can derive equilibrium departure rates.

Lemma 3. *When there is no toll, the departure rate exists, is unique, and, for departure times actually chosen, is defined by*

$$r(t) = s_B \left(1 - \frac{P(r(t_S)D'(t + T(t)) + \left[1 - P(r(t_S))\right] D'(t - t^*)}{P(r(t_S)[\alpha + D'(t + T(t))]} \right). \tag{7}$$

Proof. Equilibrium requires that $c'(t) = 0$ for all departure times actually chosen. Differentiating (6) gives

$$c'(t) = P(r(t_S)) \left[\alpha T'(t) + D'(t + T(t))(1 + T'(t)) \right] + \left[1 - P(r(t_S))\right] D'(t)$$

$$\Rightarrow T'(t) = -\frac{P(r(t_S)D'(t+T(t)) + [1-P(r(t_S))]\,D'(t)}{P(r_{0E})[\alpha + D'(t+T(t))]}.$$

Using the technology of the bottleneck, described in equation (1), when $r'(t) > s_B$ or $T(t) > 0$,

$$T'(t) = \frac{r(t) - s_B}{s_B}$$

and so

$$r(t) = s_B\left(1 - \frac{P(r(t_S)D'(t+T(t)) + [1-P(r(t_S))]\,D'(t)}{P(r(t_S)[\alpha + D'(t+T(t))]}\right). \tag{8}$$

This equation implicitly defines $r(t_S)$ and explicitly defines $r(t)$ for $t > t_S$.

To show that a unique solution to (8) exists for t_S, note that the first driver to depart never faces any congestion, and so (8) simplifies to

$$r(t_S) = s_B\left(1 - \frac{D'(t)}{P(r(t_S))[\alpha + D'(t)]}\right).$$

Note that the left-hand side is a continuous, unbounded, and increasing function of $r(t_S)$, and the right-hand side is a continuous and decreasing function of $r(t_S)$. Thus, by the intermediate value theorem, a solution exists, and furthermore it is unique. ■

Note that (7) simplifies to the standard departure rates in the standard bottleneck model when the probability of breakdown is one (cf. Arnott et al., 1993).

Further note that by Lemma 2 and Lemma 3,

$$r(t_S) = s_B\left(1 + \frac{\beta}{P(r(t_S))[\alpha - \beta)]}\right) > s_B.$$

Consequently, the probability of breakdown is greater than zero in equilibrium.

4.2 Equilibrium trip costs and total social cost

We now solve for equilibrium trip costs and total social costs. Given the departure rate, determined using the equilibrium requirement that the cost is the same for all

departure times, we now use the requirement that supply equals demand on both good and bad days to find the first and last departure times. With piecewise-linear schedule delay costs, a straightforward linear system of equations determines these departure times. After solving this system of equations, we find the equilibrium trip cost by evaluating the trip cost at the time of the first departure.

Proposition 1. *When there is no toll, the first departure occurs at*

$$t_S^U = t^* - \frac{N}{s_B}\frac{\hat{\gamma}}{\beta+\hat{\gamma}}, \tag{9}$$

the equilibrium trip cost is

$$\bar{c}^U = \frac{N}{s_B}\frac{\beta\hat{\gamma}}{\beta+\hat{\gamma}}, \tag{10}$$

and the total social cost is

$$TSC^U = \frac{N^2}{s_B}\frac{\beta\hat{\gamma}}{\beta+\hat{\gamma}}; \tag{11}$$

where

$$\hat{\gamma} = \min\{P(r(t_S))(\alpha+\gamma), \gamma\}.$$

The proof for this proposition is in Appendix A.

These results have an important implication: the equilibrium travel cost, as well as the time of the first and last departure, do not depend on the probability of breakdown when we are in the case where the last departure occurs when travel times on bad days have returned to zero. This case occurs when

$$P(r(t_S)) \geq \frac{\gamma}{\alpha+\gamma} = \frac{\gamma/\alpha}{1+\gamma/\alpha},$$

so that the probability of breakdown is large relative to drivers' willingness to exchange arriving late for travel time. When this case applies, the only way the technology of the highway matters is through the capacity after breakdown occurs. Consequently, the duration of the peak period and the equilibrium trip cost are the same whether breakdown occurs every day (as would occur in the standard bottleneck model) or when the probability is large, but less than one. In this case, the existence of "good days" does not reduce trip costs or shrink the period over which drivers depart.

5 Equilibrium with tolls

Tolling is now introduced. Specifically, equilibrium is characterized for a toll road operator that chooses a maximum inflow rate to the facility, $\hat{r}$. Within this model, a maximum inflow rate is the same as a maximum departure rate from home. As discussed in the introduction, this second-best toll scheme is analyzed because it accords with actual practice. First-best tolling is less analytically tractable, and less amenable to practical application. That said, it is likely that the first-best departure rate would increase the social welfare gains by having the departure rate be increasing at the start of the peak period. This would mean that breakdown, should it occur, is likely to happen later in the peak period and thus affect fewer drivers.

The toll road operator chooses $\hat{r}$ to maximize social welfare. It achieves this by minimizing drivers' expected travel and schedule delay costs. The tolls paid by drivers are treated as transfers from drivers to the road operator.

As was the case in the no-toll equilibrium, breakdown either occurs immediately at the start of the peak period or not at all. Given this, the total social cost of travel (TSC) can be written as

$$\text{TSC} = P(\hat{r}) \int_{t_S}^{t_E} \left[\alpha T_B(t) + D(t + T_B(t))\right] r(t)\, dt + \left[1 - P(\hat{r})\right] \int_{t_S}^{t_E} D(t) r(t)\, dt \quad (12)$$

The toll road operator charges a toll if, and only if, it is needed to keep the departure rate from going above its target maximum. We assume the toll at the start of the peak period is zero, $\tau(t_S) = 0$. Once the toll returns to zero, it stays at zero. Let t_0 be the time when tolls return to zero.

The following lemma allows us to reduce the number of cases we need to consider from eight to two.[12]

Lemma 4. *If the toll is chosen to minimize total social cost then the maximum departure rate is binding for all $t \in [t_S, t^*]$. Furthermore $t_S < t^*$.*

[12]The eight possible cases differ along four dimensions: first, either $t_S < t^*$ or $t_S \geq t^*$, second, if $t_S < t^*$, either $t_E < t^*$ or $t_E \geq t^*$, third, if $t_E \geq t^*$ either $t_0 < t^*$ or $t_0 > t^*$, and fourth, either $T(t_E) = 0$ or $T(t_E) > 0$.

Proof. Assume by way of contradiction that $t_S > t^*$. Shifting all departure times earlier so $t_S = t^*$ reduces all drivers' schedule delay on both good and bad days, and thus reduces social costs.

Assume, by way of contradiction, that the maximum departure rate isn't binding for all $t \in [t_S, t^*]$. Consider the alternate maximum departure rate equal to the average departure rate during $[t_S, t^*]$. This alternate departure rate reduces travel time on bad days and schedule delay on good days for all drivers except the first and, perhaps, the last. Furthermore, it reduces the probability of breakdown. Therefore, the alternate maximum departure rate reduces total social cost. This is a contradiction. ■

The two remaining cases differ by whether or not there is a period of time when the departure rate is below $\hat{r}$. The logic is exactly the same as that used in Lemma 2. When the toll is zero, drivers stop departing either (1) when travel times on bad days return to zero or (2) when the marginal expected schedule delay costs outweigh the expected travel time savings from leaving later. Given piecewise-linear schedule delay costs and knowing that the maximum departure rate is binding at least until t^* (Lemma 4), the second reason for drivers to stop departing binds either immediately once tolls return to zero, or does not bind at all. Thus, if $P(\hat{r}) < \gamma/(\alpha + \gamma)$, the departure rate is never below $\hat{r}$. Otherwise, there will be a period of time when it is.

5.1 Toll schedule

For a given maximum departure rate, the toll schedule is determined by the equilibrium requirement that all drivers are indifferent between departure times that are actually chosen. Drivers' trip costs are

$$c(t) = P(\hat{r})\left[\alpha T(t) + D(t + T(t))\right] + \left[1 - P(\hat{r})\right] D(t) + \tau(t).$$

Solving $c'(t) = 0$ for $\tau'(t)$ yields

$$\tau'(t) = -\Bigg[\left[1 - P(\hat{r})\right] D'(t) + P(\hat{r}) \left(D'(t + T(t)) + \left[\alpha + D'(t + T(t))\right] \frac{\hat{r} - s_B}{s_B} \right) \Bigg]. \quad (13)$$

This yields a concave toll schedule, with tolls climbing at the start of the peak period (as long as $\hat{r} < r^U(t_S)$) and falling at the end of the peak period.

We can compare this toll to that in the standard bottleneck model. In the standard bottleneck model the toll is set to eliminate congestion, and so $\tau'(t) = -D'(t)$, which is the same as (13) when the probability of breakdown is zero or one.[13]

When the probability of breakdown is strictly between zero and one, the toll climbs slower and falls faster than in the standard bottleneck model. The reason it does so is that the toll varies in order to keep drivers indifferent between departure times that are actually chosen. In the standard bottleneck model, the toll is lower away from t^* to compensate drivers for their schedule delay costs. When the probability of breakdown is strictly between zero and one, there is congestion on days breakdown occurs, and the amount of congestion is higher for later departure times. As a result, the toll climbs slower and falls faster in order to compensate drivers for their expected travel time costs.

Further note that as the maximum departure rate is reduced, tolls climb at a faster rate, and fall at a slower rate.

Lemma 5. *When the toll is non-zero, the slope of the toll schedule is decreasing in the maximum departure rate.*

[13] When the probability of breakdown is one the optimal $\hat{r} = s_B$, and $T(t) = 0$.

Proof. Differentiating (13) with respect to $\hat{r}$ yields

$$\frac{d\tau'(t)}{d\hat{r}} = -\Bigg[p(\hat{r})\left(\left[D'(t+T(t)) - D'(t)\right] + \left[\alpha + D'(t+T(t))\right]\frac{\hat{r}-s_B}{s_B}\right) + P(\hat{r})\left(\left[\alpha + D'(t+T(t))\right]\frac{1}{s_B} + D''(t+T(t))\frac{dT(t)}{d\hat{r}}\frac{\hat{r}}{s_B}\right)\Bigg]. \quad (14)$$

Because D is weakly convex, $D'(t+T(t)) - D'(t) > 0$ and $D''(t+T(t)) > 0$ and because $\beta < \alpha$, $\alpha + D'(t+T(t)) > 0$. Therefore (14) is less than zero. ■

This implies that when holding the start of the peak period fixed, a reduction in the maximum departure rate increases the average and maximum toll, and the length of time a toll is charged.

5.2 Trip costs and total social cost

We now solve for private trip costs and total social costs. We do so by integrating the expression for total social costs, while imposing the supply equals demand constraint. This gives us the time period when the maximum departure rate is binding. We then solve for t_S and $\hat{r}$ that maximize social welfare. We denote by superscript W objects associated with the welfare-maximizing toll.

Proposition 2. *When a toll is charged to impose a maximum departure rate, the first departure occurs at*

$$t_S^W = t^* - \frac{N}{\hat{r}}\frac{\gamma}{\beta+\gamma}\omega_1(\hat{r}),$$

the equilibrium trip cost is

$$\bar{c}^W = \frac{N}{\hat{r}}\frac{\beta\gamma}{\beta+\gamma}\omega_1(\hat{r}),$$

the equilibrium total social cost is

$$TSC^W = \frac{N^2}{2\hat{r}}\frac{\beta\gamma}{\beta+\gamma}\omega_2(\hat{r}),$$

and the maximum departure rate, $\hat{r}$, is the solution to

$$\hat{r} = \omega_2(\hat{r})\left(\frac{d\omega_2(\hat{r})}{d\hat{r}}\right)^{-1};$$

where

$$\omega_1(\hat{r}) = \begin{cases} \xi_1 & \text{if } P(\hat{r}) < \frac{\gamma}{\alpha+\gamma} \\ \xi_1/(1-\xi_2) & \text{if } P(\hat{r}) \geq \frac{\gamma}{\alpha+\gamma} \end{cases},$$

$$\omega_2(\hat{r}) = \begin{cases} \xi_1(1+\xi_2) & \text{if } P(\hat{r}) < \frac{\gamma}{\alpha+\gamma} \\ \xi_1/(1-\xi_2) & \text{if } P(\hat{r}) \geq \frac{\gamma}{\alpha+\gamma} \end{cases},$$

$$\xi_1 = 1 + \left(\frac{P(\hat{r})(\alpha+\gamma)}{\gamma}\right)\frac{\hat{r}-s_B}{s_B}, \text{ and}$$

$$\xi_2 = P(\hat{r})\frac{\hat{r}-s_B}{s_B}\left(\frac{\hat{r}-s_B}{\hat{r}}\left[\frac{\gamma - P(\hat{r})[\alpha+\gamma]}{\beta}\right] + \frac{\alpha}{\beta}\right).$$

The proof for this proposition is in Appendix B.

Note that if either $P(\hat{r}) = 0$ or $P(\hat{r}) = 1$, and $\hat{r} = s_B$, the expressions for $\bar{c}^W$, t_S^W, and TSC^W all simplify to their values in the standard bottleneck model.

6 Optimal departure rate does not maximize expected throughput

The paper has two important theoretical results. In this section we show that the social welfare-maximizing maximum departure rate is lower than that which maximizes expected throughput. Then in the next section, we show that even with non-negative tolls, the welfare of all drivers is improved by the charging of tolls.

We characterize the socially optimal maximum departure rate by comparing it to a benchmark. This benchmark is the maximum departure rate that results in the highest expected throughput while the maximum is binding. Expected throughput is given by the weighted average of outflow from the bottleneck on good and bad days, and is given by

$$[1 - P(r)]\, r + P(r) s_B.$$

Using the superscript F to denote expected throughput (or flow) maximizing

equilibrium values, the maximum departure rate is implicitly defined by

$$r^F = s_B + \frac{1 - P(r^F)}{p(r^F)}.$$

For the purposes of this equilibrium, expected throughput is defined over the period when the maximum departure rate is binding. In the event that any drivers depart later when the maximum is not binding, they do not enter into the calculation of the expected throughput.[14]

The first key result of this paper is that the road operator should restrict the departure rate below that which would maximize expected throughput. While reducing the departure rate increases the length of the peak period, it has two benefits. First, it reduces the probability of breakdown. Second, it reduces the negative consequences of breakdown by spreading out when drivers depart. Congestion is less on days when breakdown occurs.

Proposition 3. *The maximum departure rate that maximizes social welfare is less than that which maximizes expected throughput, and is greater than s_B.*

Proof. We first show the maximum departure rate which maximizes social welfare is less than that which maximizes expected throughput.

First note that if the expected throughput maximizing departure rate is high enough that neither of the cases solved for in this paper applies, then by Lemma 4, we know the socially optimal maximum departure rate is lower than that which maximizes expected throughput.

Next, we evaluate the first-order condition for $r = r^F$. If it is positive, then reducing r below r^F reduces total social cost. Evaluating (34) when $P(r^F) <$

[14] When $P(\hat{r}) < \gamma/(\alpha + \gamma)$, the maximum departure rate is binding for all drivers, and so r^F maximizes throughput over the entire peak period. If there is a period when the toll is not binding, we presume that the tolling authority is interested in maximizing throughput during the period when tolls apply.

$\gamma/(\alpha+\gamma)$:

$$\frac{d\text{TSC}^W(\hat{r})}{d\hat{r}}\Big|_{r=r^F} = \frac{N^2}{2}\left\{\left([1-P(\hat{r}]^2+s_Bp(\hat{r})\right)\alpha+[1-P(\hat{r})]^2\gamma\right\}$$
$$\times\left(\alpha\cdot 2P(\hat{r})\left\{[1-P(\hat{r})]\left[1-2P(\hat{r})+[P(\hat{r})]^2+p(\hat{r})s_B\right]\right\}\right.$$
$$+\beta\cdot s_Bp(\hat{r})\left\{1-P(\hat{r})+p(\hat{r})s_B\right\}$$
$$+\gamma\left\{-2\left[P(\hat{r})\right]^4+6\left[P(\hat{r})\right]^3-6\left[P(\hat{r})\right]^2\right.$$
$$\left.\left.+P(\hat{r})[2-p(\hat{r})s_B]+p(\hat{r})s_B[1+p(\hat{r})s_B]\right\}\right)$$
$$\times\left\{s_B^2\left[1-P(\hat{r})+p(\hat{r})s_B\right]^3(\beta+\gamma)\right\}^{-1} \quad (15)$$

Each term in brackets is positive, so (15) is positive.

Next, evaluating (34) when $P(r^F)\geq\gamma/(\alpha+\gamma)$:

$$\frac{d\text{TSC}^W(\hat{r})}{d\hat{r}}\Big|_{r=r^F} = \frac{N^2s_B\beta^2\left[p(r^F)\right]^3}{2}$$
$$\times\left[\alpha\left(\left[1-P(r^F)\right]^2+p(r^F)s_B\right)+\gamma\left[1-P(r^F)\right]^2\right]$$
$$\times\left\{p(r^F)s_B(1+p(r^F)s_B)\beta+\left[P(r^F)\right]^3\left(P(r^F)-3\right)(\alpha+\gamma)\right.$$
$$-P(r^F)\left[\alpha+\gamma+p(r^F)s_B(\alpha+\beta)\right]$$
$$\left.+\left[P(r^F)\right]^2\left[\left(3+p(r^F)s_B\right)\alpha+3\gamma\right]\right\}^{-2}. \quad (16)$$

Each term is positive, so (16) is positive.

Next, for any $\hat{r}>r^F$, reducing $\hat{r}$ increases both expected throughput and reduces the probability of breakdown, both of which reduces total social cost. Thus, it is never optimal to have $r>r^F$.

Finally, we show the departure rate which maximizes social welfare is greater

than s_B. When $r = s_B$,

$$\frac{d\text{TSC}^W(r)}{dr}\Big|_{r=s_B} = -\frac{1}{2}\left(\frac{N}{s_B}\right)^2 \frac{\beta\gamma}{\beta+\gamma} < 0,$$

and so increasing r above s_B reduces social cost. Note that having r below s_B increases schedule delay while yielding no benefit in reduced travel times or reduced probability of departure, and so is never optimal. ■

Reducing the maximum departure rate increases the period of time the maximum departure rate is binding, and thus the period of time a non-zero toll is charged. Lemma 5 tells us that reducing the maximum departure rate increases the slope of the toll schedule at any given point in time. As a result, as long as the start of the peak period does not change too much, the social welfare maximizing toll schedule has a larger average and maximum toll than the expected throughput maximizing toll schedule.[15]

7 Adding tolls reduces private costs

We now evaluate the welfare consequences of tolling. We find that by accounting for the value of reliability in a structural model of congestion, there is a non-negative toll schedule that reduces drivers' costs, even if the toll revenue is not used productively.[16] Moreover, the welfare benefits will be even larger if the toll revenues are used to pay for highway maintenance, perhaps offsetting gas taxes or license fees, or are transferred and used productively outside of the road operator. This is even possible when tolling reduces the average departure rate.

[15]If the period of time when the toll is charged differs greatly, then the result that the slope of the toll schedule is greater when maximizing social welfare at any given point in time does not provide a meaningful bound on the toll schedules. Given that drivers wish to arrive close to t^*, such a difference in when tolls are charged between welfare-maximizing and expected-throughput-maximizing is unlikely. In our numerical simulations in Section 8, we find that the social welfare maximizing toll is on average more than triple the expected throughput maximizing toll.

[16]While in principle toll road operators could set negative toll rates at certain times, the only example of negative tolls we are aware of is in the Netherlands, where they have experimented with paying drivers to avoid the peak when there is road construction (Knockaert et al., 2012).

Proposition 4. *There exist non-negative tolls that make all drivers better off before using the toll revenue. Furthermore, this is possible while* decreasing *the average departure rate if, when the road was untolled, the probability of breakdown is low enough:* $P(r^U(t_S^U)) < \gamma/(\beta+\gamma)$.

Proof. The proof proceeds by considering both cases for equilibrium when the road is untolled.

Case 1. First, consider the case where $T\left(t_E^U\right) = 0$. Let $t_S = t_S^U + \epsilon$ for an arbitrarily small $\epsilon > 0$. The first driver thus has less schedule delay, and since the first driver never faces any congestion, and since his toll is, by assumption, zero, he is better off.

We next show the tolls are non-negative. Set a constant departure rate, so that $r = N/(t_E - t_S)$, by (13),

$$\tau(t_E) = \int_{t_S}^{t_E} \tau(t)dt = \left(t_E - t_E^U\right)\left[(\alpha+\gamma)\,P(r) - \gamma\right] - \left[(\alpha+\gamma)\,P(r) + \beta\right]\epsilon.$$

Choosing t_E such that $\tau(t_E) = 0$ yields

$$t_E = t_E^U - \frac{\beta + (\alpha+\gamma)P(r)}{\gamma - (\alpha+\gamma)\,P(r)} \times \epsilon.$$

Because (13) is concave, this implies the toll schedule is non-negative. Thus, there exists a non-negative toll that makes all drivers better off when $T\left(t_E^U\right) = 0$.

Case 2. Next, consider the case where $T\left(t_E^U\right) > 0$, which occurs when $P(r^U(t_S^U)) < \gamma/(\beta+\gamma)$. Let $t_S = t_S^U + \epsilon$ for an arbitrarily small $\epsilon > 0$, and let $t_E = t_E^U + 2\epsilon$. The first driver thus has less schedule delay, and since the first driver never faces any congestion, and since her toll is, by assumption, zero, she is better off. Furthermore, we have increased the length of the period of departures, and so thus have decreased the average departure rate.

Next, we show tolls are non-negative. Set a constant departure rate, so that $r = N/(t_E - t_S)$, by (13),

$$\tau(t_E) = \int_{t_S}^{t_E} \tau(t)dt = \left(P(r^U(t_S^U)) - P(r)\right)\frac{N}{s_B}\frac{\beta(\alpha+\gamma)}{\beta + P(r^U(t_S^U))(\alpha+\gamma)} - \left[\beta + 2\gamma - P(r)(\alpha+\gamma)\right]\epsilon.$$

This is greater than zero for an arbitrarily small $\epsilon > 0$. Because (13) is concave, this implies the toll schedule is non-negative. Thus, there exists a non-negative toll that makes all drivers better off when $T\left(t_E^U\right) > 0$. ■

This result stands in contrast to the existing literature, where the socially optimal departure rate and private costs are negatively correlated. In the traditional static model adding tolls increases private costs while reducing departures (e.g. Walters, 1961). In the standard bottleneck model adding tolls leaves the average departure rate and private costs unchanged (Arnott et al., 1993). In models with hypercongestion, adding tolls increases the average departure rate and reduces private costs (Fosgerau and Small, 2013; Hall, 2018). Our result shows it is possible to decrease both the average departure rate and private costs, because of the value that drivers derive from a reduction in uncertainty.

An important caveat on this second result is that it is derived assuming drivers are homogeneous, and Hall (2018) shows that allowing for heterogeneous preferences makes it difficult for pricing all of the lanes of the highway to help all drivers prior to using the toll revenue. This result does, however, point to the value of accounting for reliability in assessing the distributional consequences of tolling. Furthermore, in Hall (2018) the set of drivers most hurt by tolling are the inflexible poor, drivers who are willing to tolerate a large amount of congestion to arrive on-time, but have low values of time. There are two reasons to believe that accounting for reliability especially helps these drivers, and so makes it easier to generate a Pareto improvement. First, it is the inflexible who suffer the most from the lack of reliability, since arriving early or late is especially costly for them. This means they benefit the most from the improvement in reliability caused by tolling. Second, the main reason these drivers are hurt by tolling is that they are displaced from the peak by flexible drivers with higher values of time. However, once we account for uncertainty in travel times, the inflexible poor are likely to be departing early to avoid the risk of arriving late on days when breakdown occurs. This means there is less scope for displacement, and that any displacement will be less damaging.

8 Simulations

To illustrate our results, we take parameter values from the literature and simulate equilibrium with and without tolls.

8.1 Setup

We make the following assumptions for driver preferences. For the cost of travel time we use the U.S. Department of Transportation (2016) recommended value of $14.10, adjusted using the consumer price index to be in 2019 dollars. This gives us $\alpha = \$15.19$ per hour.

For the cost of time early, β, we follow the recommendation of Hall (2019b) and assume $\beta = 0.1 \times \alpha$.

For the cost of time late, γ, we consider two possible values so we can illustrate equilibrium outcomes for both cases when the road is untolled. First, we follow the recommendation of Hall (2019b) and assume $\gamma = \beta$. As a result, when the road is untolled, $T(t_E) = 0$. Second, we choose $\gamma = 6 \times \beta$ so that in the untolled equilibrium $T(t_E) > 0$.

Without loss of generality, we let the desired arrival time, t^*, be zero. As a result, we can interpret time as $t - t^*$.

We assume the mass of drivers is such that on a bad day it takes three hours for all drivers to use the highway: $N = 3 \times s_B$. In making this assumption, we are focusing our attention on less congested road segments. Three hours may sound long, however, in many congested cities the length of the peak period is closer to eight hours (cf. Hall, 2019a).

We base the probability of breakdown on the estimates of Lorenz and Elefteriadou (2001). They use data from two isolated bottlenecks on Highway 401 in Toronto, Canada, to estimate the probability of breakdown as a function of flows in the prior fifteen minutes non-parametrically, as well as throughput after

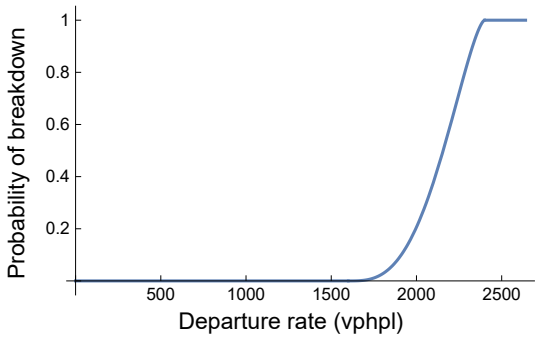


Figure 1: Probability of breakdown

Notes: Based on estimates from Lorenz and Elefteriadou (2001). Departure rate measured in vehicles per hour per lane.

breakdown occurs.[17] We approximate Lorenz and Elefteriadou's (2001) estimates of the probability of breakdown reported in their Figure 7 using a beta distribution and choose s_B to match their estimate of throughput after breakdown reported in their Figure 8. Specifically, we fit a beta distribution over the range from 1,600 to 2,400 vehicles per hour per lane (vphpl). The lower end of this range marks the traffic volume below which breakdown cannot occur, and is the same as s_B. The upper end marks the volume at which breakdown is certain to occur. In addition, we choose the parameters of this distribution so that $P(1,900) = 0.09$, and $P(2,200) = 0.60$. The resulting distribution is similar to what we would get if we based the probability of breakdown on estimates from Geistefeldt and Shojaat (2019). Our probability of breakdown function is plotted in Figure 1.

8.2 Results

Given these parameter values, we solve for equilibrium when the road is free and priced. For both sets of parameter values, the tolled equilibrium is in the case where $T(t_E) > 0$, and so the maximum departure rate is always binding. This is because adding the toll reduces the probability of breakdown so much that the

[17] An isolated bottleneck is one where traffic at the bottleneck is not affected by a downstream bottleneck. They also estimate probabilities based on 1 and 5 minute flows.

marginal expected schedule delay costs from leaving a little bit later outweigh the expected travel time savings from leaving later. As mentioned earlier, when $\gamma = \beta$ the untolled equilibrium is in the $T(t_E) = 0$ case, and when $\gamma = 6 \times \beta$ the untolled equilibrium is in the $T(t_E) > 0$ case.

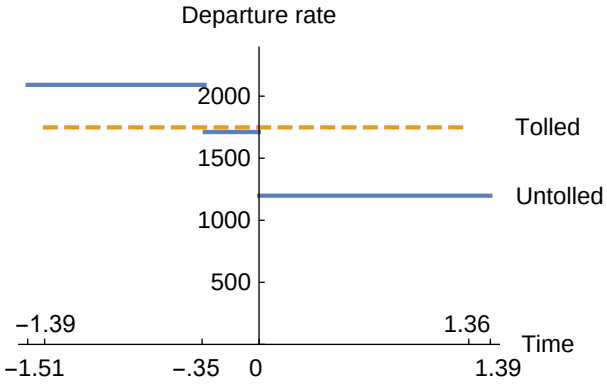
Figure 2 plots the equilibrium departure rates both when the road is untolled and when it is tolled. The upper part of the figure is for when $\gamma = \beta$, and the lower part for when $\gamma = 6 \times \beta$. In both cases, adding tolls smooths the rate at which drivers depart from home, decreasing it at the start and increasing it at the end.

More details on the simulation results are presented in Table 1. The first row in the table reports that the reduction in the departure rate at the start of the peak period reduces the probability of breakdown by 35 percentage points. As a result, breakdown goes from occurring once or twice a week to occurring quarterly.[18]

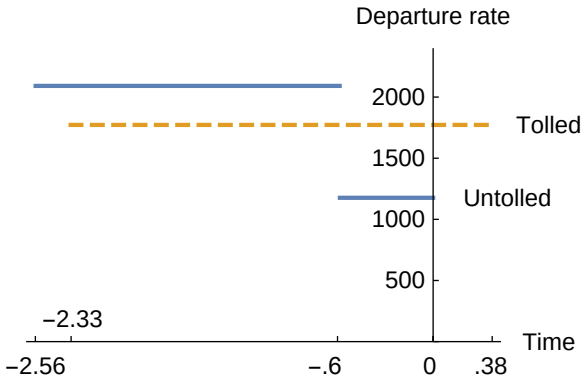
As Proposition 4 proved, this reduction in the probability of breakdown, as well as the reduction in the consequences of breakdown, helps drivers. We can evaluate the welfare effects of tolling graphically using Figure 2 and focusing on the first driver to depart. When the road is free, the first driver to depart faces no congestion (by virtue of being the first to depart) and pays no toll. Her only cost is due to arriving earlier than desired. When the road is tolled, she continues to face no congestion and pays no toll (by virtue of being the first to depart), but now departs later and so has less schedule delay than before. She is better off. Since all drivers are identical, if the first driver is better off, all drivers are better off, and so adding the toll helps all drivers, even before using the toll revenue.

The combined effects of the reduction in the probability of breakdown and the reduction in the consequences of breakdown can be seen by comparing the travel times in Table 1. Tolling reduces the consequences of a bad day by reducing the average and maximum travel time on a bad day by 32–56 percent. Since it also reduces the probability of breakdown, average travel time is reduced by 98

[18]Recall that we have chosen our parameter values to focus attention on less congested road segments. On the most congested roads, breakdown occurs almost daily, while on less congested roads it is less common.



(a) Case 1: $\gamma = \beta$



(b) Case 2: $\gamma = 6 \times \beta$

Figure 2: Departure rates

Notes: Time measured in hours from desired arrival time. Departure rate measured in vehicles per hour per lane.

Table 1: Simulation results

	Case 1: $\gamma = \beta$			Case 2: $\gamma = 6 \times \beta$		
	Free	Tolled	Change	Free	Tolled	Change
Probability of breakdown	0.362	0.011	-97%	0.362	0.017	-95%
Average departure rate (vphpl)	1,600	1,748	+9.3%	1,876	1,772	-5.6%
Average throughput (vphpl)	1,600	1,747	+9.2%	1,776	1,769	-0.40%
Per trip costs						
Social	\$2.78	\$1.06	-53%	\$3.89	\$1.81	-54%
Private	\$2.78	\$2.11	-7.5%	\$3.89	\$3.54	-8.8%
Travel time (minutes)						
Average	4.44	0.08	-98%	7.23	0.15	-98%
Average on a bad day	12.27	7.64	-38%	19.98	8.74	-56%
Maximum on a bad day	22.61	15.27	-32%	26.51	17.49	-34%
Toll						
Average	—	\$1.04	—	—	\$1.74	—
Maximum	—	\$2.08	—	—	\$3.44	—

Notes: Average departure rate and throughput are the averaged across time and breakdown state. Average travel time and average toll are the averages measured across drivers. The average travel time and toll across time and across drivers differ by less than five percent.

percent.[19]

Further consistent with Proposition 4, Figure 2b shows visually that when the untolled equilibrium is such that $T(t_E) > 0$, adding these tolls *reduces* the average departure rate and lengthens the period over which drivers depart. In this case, tolling reduces the average departure rate by 104 vphpl, and even reduces average throughput by 7 vphpl. Despite this, and in contrast to existing results in the literature, tolling still reduces private costs.

Consistent with Proposition 3, the welfare-maximizing maximum departure rate is lower than the departure rate which maximizes expected throughput. The departure rate which maximizes expected throughput is 2,039 vphpl, and achieves an expected throughput (across good and bad days) of 1,921 vphpl. In contrast, when the road is free the maximum departure rate is 2,091 vphpl, and when the road is tolled the maximum (and constant) departure rate is 1,748 or 1,772 vphpl (depending on the case). While increasing the maximum departure rate above the social optimum would increase expected throughput, doing so reduces social welfare since drivers are risk averse over arrival times.

Figure 3 shows this visually. It plots the total social cost per trip as a function of the maximum departure rate.[20] The figure shows that while setting the toll to maximize expected throughput reduces total social cost relative to the road being untolled, we more than triple the social welfare gains by charging the toll that maximizes total social welfare.

Figure 4 compares the toll schedule which maximizes social welfare to that which maximizes expected throughput. Consistent with our discussion in Section 6, the tolls that maximize social welfare are higher. The average and maximum tolls are both more than three times higher when maximizing social welfare rather than expected throughput. For these parameter values, the expected throughput maximizing departure rate is not binding for all departures, and the toll returns to

[19] In interpreting this number it is useful to remember that this is the excess travel time due to congestion.

[20] The figure is for the case where $\gamma = 6 \times \beta$, and is fundamentally the same for the case where $\gamma = \beta$.

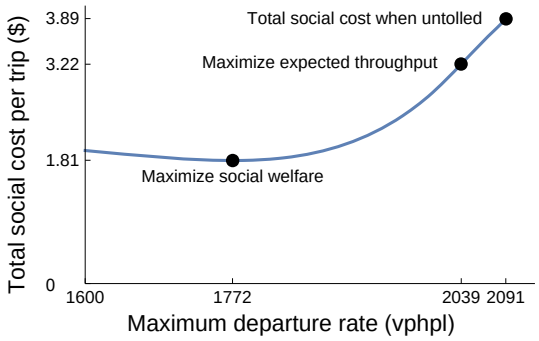


Figure 3: Total social cost per trip vs. maximum departure rate

Notes: This figure is plotted for Case 2: $\gamma = 6 \times \beta$. Maximum departure rate measured in vehicles per hour per lane. Creating this figure requires solving for equilibrium in cases further from the optimum which were not reported in the paper but are available in the replication files.

zero prior to the last departure, which in this case occurs at t^* (i.e., 0).

Figure 4 also shows that when expected throughput maximizing tolls are charged, the earliest departure is 6 minutes earlier than when welfare maximizing tolls are charged. In addition, the last departure is 23 minutes earlier than when welfare maximizing tolls are charged. The result of maximizing expected throughput is that everyone departs over a shorter interval, however, to help mitigate the uncertainty, they also leave earlier.

Table 1 also reports the magnitudes of the social and private welfare gains from tolling. Given our parameter values, private costs are reduced by 7.5-8.8 percent. If the typical commuter makes two trips on each of the 250 working days in a year, this amounts to 175–335 dollars per commuter per year, depending on the case. The social welfare gains are much larger. Total social costs of congestion fall by 53–54 percent, and are equivalent to 860–1,040 dollars per commuter per year.

9 Conclusion

Unpredictable travel times impose significant costs on drivers, and are estimated to account for 30–70 percent of the total cost of congestion (Small et al., 2005; Bento

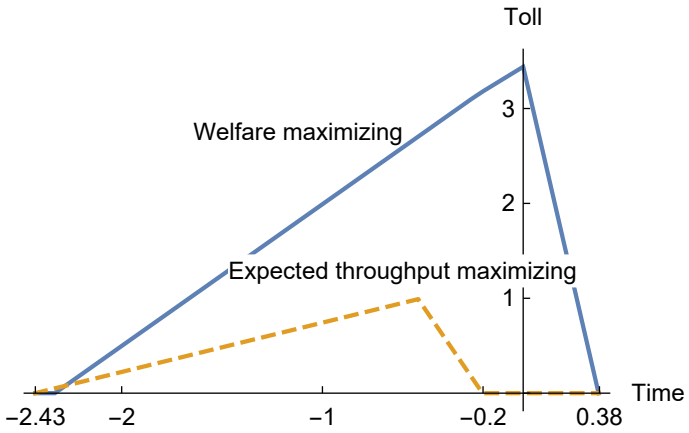


Figure 4: Toll schedule when maximizing social welfare vs. maximizing expected throughput

Notes: This figure is plotted for Case 2: $\gamma = 6 \times \beta$.

et al., 2017). Due to this unpredictability, drivers choose to depart earlier to reduce the risk of being late. On good days they arrive unnecessarily early and on bad days they suffer the consequences of arriving late. While some sources of the unpredictability are exogenous, such as weather, others are endogenous, such as crashes and flow breakdown, and the probability that they occur is increasing in the traffic flow.

This paper analyzes how to implement tolls in the presence of endogenous non-recurring congestion. We find two important results. First, tolls should be implemented to restrict inflow to the facility below the rate that maximizes expected throughput. Our simulations suggest that expected throughput should be a little more than ten percent lower than the maximum possible, which requires charging tolls that are, on average, more than three times as large. This is worthwhile because by smoothing the rate at which drivers enter a congestible link, decreasing it at the start of the peak period and increasing it at the end, these tolls reduce both the probability of breakdown and the severity of congestion when breakdown occurs. Smoothing entry in this way increases traffic reliability. Because drivers value reliability, it is worthwhile to sacrifice some expected capacity by reducing inflow to the facility.

Second, weakly positive tolls exist that make drivers better off. In our simulations, private costs decrease by almost ten percent. Drivers are better off since the increase in reliability means they no longer need to leave as early. Social and private welfare would be further enhanced if the toll revenues are used productively or offset other funding instruments such as the tax on gasoline. Making drivers better off matters because it can help overcome the political barriers to congestion pricing.

References

Arnott, Richard, André de Palma, and Robin Lindsey (1990) "Economics of a Bottleneck," *Journal of Urban Economics*, Vol. 27, No. 1, pp. 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*, Vol. 83, No. 1, pp. 161–179, URL: `https://www.jstor.org/stable/2117502`.

——— (1999) "Information and Time-of-Usage Decisions in the Bottleneck Model with Stochastic Capacity and Demand," *European Economic Review*, Vol. 43, No. 3, pp. 525–548, DOI: 10.1016/S0014-2921(98)00013-0.

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

Bento, Antonio M., Kevin Roth, and Andrew Waxman (2017) "Avoiding Traffic Congestion Externalities? The Value of Urgency," *Working Paper*.

Bloomberg Philanthropies (2018) *2018 American Mayors Survey*, New York, NY, URL: `https://www.bbhub.io/dotorg/sites/2/2018/04/American-Mayors-Survey.pdf`.

Chen, Danjue and Soyoung Ahn (2015) "Variable Speed Limit Control for Severe Non-Recurrent Freeway Bottlenecks," *Transportation Research Part C: Emerging Technologies*, Vol. 51, pp. 210–230, DOI: 10.1016/j.trc.2014.10.015.

Chen, Danjue, Soyoung Ahn, Jorge Laval, and Zuduo Zheng (2014) "On the Periodicity of Traffic Oscillations and Capacity Drop: The Role of Driver Characteristics," *Transportation Research Part B: Methodological*, Vol. 59, pp. 117–136, DOI: 10.1016/j.trb.2013.11.005.

Doig, Jean C., Vikash V. Gayah, and Michael J. Cassidy (2013) "Inhomogeneous Flow Patterns in Undersaturated Road Networks: Implications for Macroscopic Fundamental Diagram," *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2390, No. 1, pp. 68–75, DOI: 10.3141/2390-08.

Dong, Jing and Hani S. Mahmassani (2009) "Flow Breakdown and Travel Time Reliability," *Transportation Research Record*, Vol. 2124, No. 1, pp. 203–212, DOI: 10.3141/2124-20.

Dowling, Richard, Alexander Skabardonis, Michael Carroll, and Zhongren Wang (2004) "Methodology for Measuring Recurrent and Nonrecurrent Traffic Congestion," *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1867, pp. 60–68, DOI: 10.3141/1867-08.

Federal Highway Administration (2016) "Toll Facilities in the United States: Bridges, Roads, Tunnels, Ferries," FHWA Report FHWA-PL-16-011, Government Printing Office, Washington, D.C., URL: `http://www.fhwa.dot.gov/policyinformation/tollpage/`.

Fosgerau, Mogens (2015) "Congestion in the Bathtub," *Economics of Transportation*, Vol. 4, No. 4, pp. 241–255, DOI: 10.1016/j.ecotra.2015.08.001.

Fosgerau, Mogens and Robin Lindsey (2013) "Trip-Timing Decisions with Traffic Incidents," *Regional Science and Urban Economics*, Vol. 43, No. 5, pp. 764–782, DOI: 10.1016/j.regsciurbeco.2013.07.002.

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

Geistefeldt, Justin and Siavash Shojaat (2019) "Comparison of Stochastic Estimates of Capacity and Critical Density for U.S. and German Freeways," *Transporta-*

tion Research Record: Journal of the Transportation Research Board, pp. 1–9, DOI: 10.1177/0361198119843471.

Hall, Jonathan D. (2018) "Pareto Improvements from Lexus Lanes: The Effects of Pricing a Portion of the Lanes on Congested Highways," *Journal of Public Economics*, Vol. 158, pp. 113–125, DOI: 10.1016/j.jpubeco.2018.01.003.

——— (2019a) "Can Tolling Help Everyone? Estimating the Size and Distribution of Welfare Gains from Value Pricing," *Working Paper*.

——— (2019b) "Improving Structural Models of Congestion," *Working Paper*.

Kim, Jiwon, Hani S. Mahmassani, and Jing Dong (2010) "Likelihood and Duration of Flow Breakdown: Modeling the Effect of Weather," *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2188, No. 1, pp. 19–28, DOI: 10.3141/2188-03.

Knockaert, Jasper, Yin-Yen Tseng, Erik T. Verhoef, and Jan Rouwendal (2012) "The Spitsmijden Experiment: A Reward to Battle Congestion," *Transport Policy*, Vol. 24, pp. 260–272, DOI: 10.1016/j.tranpol.2012.07.007.

Kononov, Jake, David Reeves, Catherine Durso, and Bryan K. Allery (2012) "Relationship between Freeway Flow Parameters and Safety and Its Implication for Adding Lanes," *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2279, No. 1, pp. 118–123, DOI: 10.3141/2279-14.

Kontorinaki, Maria, Anastasia Spiliopoulou, Claudio Roncoli, and Markos Papageorgiou (2017) "First-Order Traffic Flow Models Incorporating Capacity Drop: Overview and Real-Data Validation," *Transportation Research Part B: Methodological*, Vol. 106, pp. 52–75, DOI: 10.1016/j.trb.2017.10.014.

Kwon, Jaimyoung, Michael Mauch, and Pravin Varaiya (2006) "Components of Congestion: Delay from Incidents, Special Events, Lane Closures, Weather, Potential Ramp Metering Gain, and Excess Demand," *Transportation Research Record*, Vol. 1959, No. 1, pp. 84–91, DOI: 10.3141/1959-10.

Lindsey, Robin (1999) "Effects of Driver Information in the Bottleneck Model," in Emmerink, Richard and Peter Nijkamp eds. *Behavioural and Network Impacts of Driver Information Systems*: Routledge, 1st edition, pp. 15–51, DOI: 10.4324/9781351119740-2.

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

Lorenz, Matt R. and Lily Elefteriadou (2001) "Defining Freeway Capacity as Function of Breakdown Probability," *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1776, No. 1, pp. 43–51, DOI: 10.3141/1776-06.

Luo, Ying, M. Hadiuzzaman, Jie Fang, and Tony Z. Qiu (2015) "Assessing the Mobility Benefits of Proactive Optimal Variable Speed Limit Control During Recurrent and Non-Recurrent Congestion," *Canadian Journal of Civil Engineering*, Vol. 42, No. 7, pp. 477–489, DOI: 10.1139/cjce-2013-0427.

Noland, Robert B. (1997) "Commuter Responses to Travel Time Uncertainty Under Congested Conditions: Expected Costs and the Provision of Information," *Journal of Urban Economics*, Vol. 41, No. 3, pp. 377–406, DOI: 10.1006/juec.1996.2006.

Noland, Robert B. and Kenneth A. Small (1995) "Travel-Time Uncertainty, Departure Time Choice, and the Cost of Morning Commutes," *Transportation Research Record*, Vol. 1493, pp. 150–158.

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*, Vol. 1634, No. 1, pp. 64–69, DOI: 10.3141/1634-08.

Qian, Wei-Liang, Adriano F. Siqueira, Romuel F. Machado, Kai Lin, and Ted W. Grant (2017) "Dynamical Capacity Drop in a Nonlinear Stochastic Traffic Model," *Transportation Research Part B: Methodological*, Vol. 105, pp. 328–339, DOI: 10.1016/j.trb.2017.09.017.

Small, Kenneth A., Clifford Winston, and Jia Yan (2005) "Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability," *Econometrica*, Vol. 73, No. 4, pp. 1367–1382, DOI: 10.1111/j.1468-0262.2005.00619.x.

Sugiyama, Yuki, Minoru Fukui, Macoto Kikuchi, Katsuya Hasebe, Akihiro Nakayama, Katsuhiro Nishinari, Shin-ichi Tadaki, and Satoshi Yukawa (2008) "Traffic Jams Without Bottlenecks—Experimental Evidence for the Physical Mechanism of the Formation of a Jam," *New Journal of Physics*, Vol. 10, No. 3, DOI: 10.1088/1367-2630/10/3/033001.

U.S. Department of Transportation (2016) "The Value of Travel Time Savings: Departmental Guidance for Conducting Economic Evaluations Revision 2,"Technical report, U.S. Department of Transportation, Washington, D.C., URL: `https://www.transportation.gov/sites/dot.gov/files/docs/2016%20Revised%20Value%20of%20Travel%20Time%20Guidance.pdf`.

Vickrey, William S. (1969) "Congestion Theory and Transport Investment," *American Economic Review*, Vol. 59, No. 2, pp. 251–260, URL: `http://www.jstor.org/stable/1823678`.

Walters, Alan A. (1961) "The Theory and Measurement of Private and Social Cost of Highway Congestion," *Econometrica*, Vol. 29, No. 4, pp. 676–699, DOI: 10.2307/1911814.

Wilson, R. E. and J. A. Ward (2011) "Car-Following Models: Fifty Years of Linear Stability Analysis—A Mathematical Perspective," *Transportation Planning and Technology*, Vol. 34, No. 1, pp. 3–18, DOI: 10.1080/03081060.2011.530826.

Zhou, Min and Virginia P. Sisiopiku (1997) "Relationship Between Volume-to-Capacity Ratios and Accident Rates," *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1581, No. 1, pp. 47–52, DOI: 10.3141/1581-06.

Zhu, Shanjiang, Gege Jiang, and Hong K. Lo (2017) "Capturing Value of Reliability through Road Pricing in Congested Traffic under Uncertainty," *Transportation Research Procedia*, Vol. 23, pp. 664–678, DOI: 10.1016/j.trpro.2017.05.037.

A Proof of Proposition 1

Proof. The requirement that supply equals demand on good days and bad gives us the following two equilibrium requirements:

$$\int_{t_S}^{t_E} r\left(t\right) dt = N, \text{ and} \tag{17}$$

$$s_B\left(t_E + T(t_E) - t_S\right) = N. \tag{18}$$

Once we have solved for t_S we find equilibrium trip costs by evaluating the trip cost at the start of the peak period:

$$\bar{c}^U = c(t_S) = D(t_S). \tag{19}$$

With the assumption of piecewise-linear schedule delay costs, by Lemma 3:

$$r(t) = s_B \begin{cases} 1 + \frac{\beta}{P(r(t_S))(\alpha-\beta)} & \text{if } t_S \leq t < t_M, \\ 1 + \frac{(1-P(r(t_S)))\beta - P(r(t_S))\gamma}{P(r(t_S))(\alpha+\gamma)} & \text{if } t_M \leq t < t^*, \text{ and} \\ 1 - \frac{\gamma}{P(r(t_S))(\alpha+\gamma)} & \text{if } t^* \leq t \leq t_E, \text{ and} \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

Equation (20) introduces a new variable, t_M, which is the time when drivers go from always arriving early to only arriving early on good days. As such, t_M is defined by

$$t_M + T(t_M) = t^*, \tag{21}$$

$$\Leftrightarrow t_M + (t_M - t_S)\,\frac{r(t_S) - s_B}{s_B} = t^*. \tag{22}$$

Starting with the case where $T(t_E) = 0$,

$$(17) \Leftrightarrow r(t_S)(t_M - t_S) + r(t_M)(t^* - t_M) + r(t_{t^*})(t_E - t^*) = N, \tag{23}$$

$$(18) \Leftrightarrow s_B\left(t_E - t_S\right) = N. \tag{24}$$

Equations (19), (20), (22), and (24) define a linear system of equations, which yields, in part, the following:

$$\bar{c} = \frac{N}{s_B}\frac{\beta\gamma}{\beta+\gamma}, \tag{25}$$

$$t_S = t^* - \frac{N}{s_B}\frac{\gamma}{\beta+\gamma}, \text{ and} \tag{26}$$

$$t_E = t^* + \frac{N}{s_B}\frac{\beta}{\beta+\gamma}.$$

A piecewise-linear schedule delay function implies that in the case where $T(t_E) > 0$, the last departure must occur at t^*, and that

$$(17) \Leftrightarrow r(t_S)(t_M - t_S) + r(t_M)(t^* - t_M) = N, \tag{27}$$

$$(18) \Leftrightarrow s_B\left(t^* + T(t^*) - t_S\right) = N. \tag{28}$$

Likewise, (19), (20), (22), (27), and (28) define a linear system of equations, which yields, in part, the following:

$$t_S = t^* - \frac{N}{s_B}\frac{P(r(t_S))\,(\alpha+\gamma)}{\beta + P(r(t_S))\,(\alpha+\gamma)} \tag{29}$$

$$\bar{c} = \frac{N}{s_B}\frac{P(r(t_S))(\alpha+\gamma)\beta}{\beta + P(r(t_S))(\alpha+\gamma)}. \tag{30}$$

Lemma 2 implies the last departure occurs when travel times, on bad days, are still positive when

$$P(r(t_S)) < \frac{\gamma}{\alpha+\gamma}$$

$$\Rightarrow P(r(t_S))\,(\alpha+\gamma) < \gamma.$$

As a result, (10) nests (25) and (30), and (9) nests (26) and (29).

Total social cost is the sum of private costs, and so $\text{TSC}^U = N \cdot \bar{c}^U$. ■

B Proof of Proposition 2

Proof. We have two cases to consider.

Case 1. First consider the case where $t_E = t_0$, and so there is not a period when the departure rate is below the maximum. This occurs when $P(\hat{r}) < \gamma/(\alpha + \gamma)$.

As when solving for equilibrium when the road is free, we use the requirement that supply equals demand on good days to solve for t_E. This requirement is given by (17).

Since the maximum departure rate is always binding:

$$r(t) = \begin{cases} \hat{r} & \text{if } t_S \leq t \leq t_0, \text{ and} \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

To use (31), we must solve for t_0, which is the time the toll returns to zero, and is the solution to

$$0 = \int_{t_S}^{t_0} \tau(t)\, dt. \tag{32}$$

Solving this equation requires knowing the time when drivers go from always arriving early to only arriving early on good days, t_M, which is defined by (21).

Equations (17), (21), (32), and $t_E = t_0$ are a linear system of equations which define t_S, t_M, t_0, and t_E. Solving this system of equations gives

$$\begin{aligned} t_S &= t^* - \frac{N}{\hat{r}} \frac{\gamma}{\beta + \gamma} \xi_1, \\ t_M &= t^* - \frac{N}{\hat{r}} \frac{\gamma}{\beta + \gamma} \xi_1 \left(1 - \frac{s_B}{\hat{r}}\right), \text{ and} \\ t_E &= t^* + \frac{N}{\hat{r}} \left(1 - \frac{\gamma}{\beta + \gamma} \xi_1\right). \end{aligned}$$

Substituting these into (12) and solving the integral yields

$$\text{TSC}(\hat{r}) = \frac{N^2}{2\hat{r}} \frac{\beta\gamma}{\beta + \gamma} \xi_1 (1 + \xi_2).$$

We find the equilibrium trip cost by calculating the trip cost of the first driver to depart. Since all drivers face the same trip cost in equilibrium, this is everyone's trip cost. The first driver to depart does not pay a toll, does not incur travel time, and always arrives at t_S, thus

$$\bar{c} = \frac{N}{\hat{r}} \frac{\beta\gamma}{\beta + \gamma} \xi_1.$$

Case 2. Next, consider the case where $t_E > t_0$, and so there is a period when the departure rate is below the maximum. This occurs when $P(\hat{r}) \geq \gamma/(\alpha+\gamma)$.

Solving for equilibrium in this case is largely the same. We continue to use the requirement that supply equals demand on good days to solve for t_E.

Given the assumption of piecewise-linear schedule delay costs, Lemma 4, which tells us the maximum departure rate is binding at least till t^*, and the departure rate when there is no toll from Lemma 3:

$$r(t) = \begin{cases} \hat{r} & \text{if } t_S \leq t \leq t_0, \\ s_B\left(1 - \frac{\gamma}{P(\hat{r})(\alpha+\gamma)}\right) & \text{if } t_0 < t \leq t_E \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

As before we must solve for t_0 and t_M using (21) and (32).

Equations (17), (21), (32), are a linear system of equations which define t_M, t_0, and t_E. Solving this system of equations gives

$$\begin{aligned} t_M &= t_S + \frac{s_B}{\hat{r}}(t^* - t_S), \\ t_0 &= t_S - (t^* - t_S)\frac{\beta+\gamma}{\gamma}\xi_1^{-1}, \text{ and} \\ t_E &= t_S + \frac{\frac{N}{s_B}P(\hat{r})(\alpha+\gamma) - (\beta+\gamma)(t^* - t_S)}{P(\hat{r})\alpha - [1 - P(\hat{r})]\,\gamma}. \end{aligned}$$

Substituting these into (12) and solving the integral yields

$$\begin{aligned} \text{TSC}(t_S, \hat{r}) = (t^* - t_S)\beta N + (t^* - t_S)^2\frac{\beta+\gamma}{2} \\ \times \frac{\hat{r}s_B\gamma + P(\hat{r})(\hat{r} - s_B)[P(\hat{r})[r - s_B](\alpha+\gamma) - [\hat{r} - s_B]\gamma - \hat{r}\alpha]}{s_B\gamma + P(\hat{r})[\hat{r} - s_B](\alpha+\gamma)}. \end{aligned} \quad (33)$$

Taking the first-order condition with respect to t_S and solving yields

$$t_S = t^* - \frac{N}{\hat{r}}\frac{\gamma}{\beta+\gamma}\left(\frac{\xi_1}{1-\xi_2}\right).$$

Once again we find the equilibrium trip cost by calculating the trip cost of the first driver to depart, which yields

$$\bar{c} = \frac{N}{\hat{r}}\frac{\beta\gamma}{\beta+\gamma}\left(\frac{\xi_1}{1-\xi_2}\right).$$

Substituting t_S into (33) and simplifying gives

$$\text{TSC} = \frac{N^2}{2\hat{r}} \frac{\beta\gamma}{\beta+\gamma} \left(\frac{\xi_1}{1-\xi_2} \right).$$

Finally, to find the optimal $\hat{r}$ we solve the first-order condition for total social costs,

$$\begin{aligned} \frac{d\text{TSC}}{d\hat{r}} &= 0 \\ \Leftrightarrow \frac{N^2}{2\hat{r}^2} \frac{\beta\gamma}{\beta+\gamma} \left(\omega_2(\hat{r}) - \hat{r} \frac{d\omega_2(\hat{r})}{d\hat{r}} \right) &= 0 \\ \Leftrightarrow \omega_2(\hat{r}) &= \hat{r} \frac{d\omega_2(\hat{r})}{d\hat{r}}. \end{aligned} \tag{34}$$

■