This paper investigates the conditions under which dominant airlines internalize self-imposed delays in a deterministic bottleneck model of airport pricing, complementing the similar analysis of Brueckner and Van Dender (BVD, 2008) for the standard congestion-pricing model. A unified model of bottleneck tolling includes untolled, uniform-, coarse-, multi-step-, and continuous-toll equilibria as special cases. Explicit closed-form solutions for aircraft scheduling and tolling provide a theoretical foundation for Daniel's (1995, 2008) empirical findings that dominant airlines often ignore self-imposed delays to preempt additional entry by fringe aircraft. Optimal tolls for dominant and fringe airlines have identical step increments but they may differ by a uniform amount, rather than varying in inverse proportion to market share as in BVD. The model is the price-setting dual of Daniel's (2014) quantity-setting slot constraint model, and achieves identical social costs. Optimal tolling pays for optimal airport capacity, under constant returns for building airport capacity.

Keywords: airport congestion, slot constraints, pricing, bottleneck, queuing (JEL R4, H2, L5, L9).

1. Introduction

Congestion pricing models were originally developed to address the problem of highway congestion. In that context, it is reasonable to treat traffic as composed of atomistic units that operate independently of one another. In adapting congestion pricing models to airports, however, a number of researchers note that the atomistic traffic assumption is not satisfied, see e.g., Daniel (1995), Brueckner (2002), Mayer and Sinai (2003), and Brueckner and Van Dender (2008). Nevertheless, Daniel argues on the basis of empirical evidence using a stochastic bottleneck model that dominant airlines often appear to behave atomistically. He hypothesizes that Stackelberg dominant airlines anticipate that atomistic behavior by fringe airlines will offset any benefit dominant airlines might obtain from internalization of self-imposed delays. His earlier empirical and computational models, however, do not provide closed-form analytical solutions demonstrating the optimality of internalizing or atomistic behavior.

Brueckner (2002) and Mayer and Sinai (2003) find that more concentrated airports have less delay than less concentrated airports, ceteris paribus. They interpret this as evidence of internalization by dominant firms. Brueckner and Van Dender (2008) develop a formal model of dominant and fringe aircraft behavior based on the standard (steady-state) congestion-pricing model. They show that the model can generate internalizing or atomistic behavior depending on the elasticity of fringe aircraft demand. When fringe demand is sufficiently inelastic, dominant airlines can increase their surplus by reducing their traffic and congestion below atomistic levels; but when fringe demand is sufficiently elastic, new fringe entry drives traffic and congestion levels back up, so dominant airlines behave atomistically. This paper develops similar results for the bottleneck model.
Congestion pricing can optimize traffic in two ways; by adjusting the total traffic volume so that the marginal social cost of trips equal their marginal social benefit, and by adjusting the scheduling of trips to reduce social cost. The standard congestion-pricing model includes the first but not the second objective, while the bottleneck model includes both. The standard model applies the theory of the firm to the production of highway trips by interpreting the individual travel time function as the average cost curve and the social travel time function as the marginal cost curve. The intersection of the individual travel time function and demand curves determines the untolled equilibrium, which involves too many trips. The vertical difference between individual and social travel times is the external time an individual trip imposes on other travelers. Imposing a toll equal to this difference implements an optimal decentralized equilibrium, with marginal social costs (supply) equal to marginal social benefits (demand), by “tolling off” travelers whose willingness to pay is below their social cost. The standard model is essentially steady state, with travel time being a function of instantaneous traffic volume alone. Delays do not carry over from one period to the next. Travelers do not optimize the choice of when to travel. There is no cost of schedule delay, i.e., deviating from a preferred operating time. Applying the standard model over time requires dividing time into multiple steady-state periods (possibly with cross elasticities of demand between periods). Dividing time into shorter periods captures more of the variation in demand, and allows for exogenous demand peaks, but limits the effect of previous traffic levels on current congestion.

The bottleneck model has several significant advantages over the standard model, particularly as applied to airports. First, it explicitly models the optimal choice of when to travel by trading off schedule delays against queuing delays and any congestion tolls. Second, it has a queuing system that depends on the current state (length) of the queue as well as the current traffic rate. The state-dependent queuing system accounts for the entire history of traffic rates and carries accumulated queues forward in time. Third, the scheduling of traffic and evolution of queues occur in continuous time. Fourth, the bottleneck model provides a richer environment for analyzing interactions between dominant and fringe airlines than the standard model because it models the dynamics of airport queuing delay and the optimal scheduling of aircraft arrivals and departures that adjust in response to congestion levels and tolls. This paper extends the deterministic bottleneck model to include a dominant airline that determines the optimal level of internalizing or atomistic behavior by choosing the number and timing of its aircraft operations. It develops a general step-tolling model for which the no-toll and continuously varying toll equilibria are limiting cases as the number of tolling steps goes from zero to infinity. This unified model of tolling demonstrates how dominant-fringe equilibria interact with the tolling structures.

Following the literature review, the theory section presents the unified model of bottleneck tolling equilibria. Next, the paper derives three sets of conclusions concerning airline behavior in untolled and tolled equilibria, under the assumption that dominant and fringe airlines have homogenous operating time preferences and valuations of time. The conclusions explain when the dominant airline internalizes or behaves atomistically and how the tolls interact with airline behavior in establishing an optimal equilibrium. A brief preview of these conclusions is as follows: 1) The dominant airline goes from fully internalizing to fully atomistic behavior as fringe demand goes from perfectly inelastic to perfectly elastic. 2) Dominant and fringe tolls vary by a uniform amount that approaches zero as dominant aircraft approach fully atomistic behavior, or in the limit as step tolls approach continuous variation. 3) An optimally tolled airport with dominant and fringe airlines is self-financing under reasonable assumptions. The paper concludes with a discussion of the model’s policy implications.
2. Review of the literature

William Vickrey (1969) originally developed the bottleneck model to provide a dynamic model of congestion with travelers that adjust their times of travel optimally to minimize the sum of their trip duration and schedule delay costs. Vickrey's congestion technology is a deterministic queue that develops at a highway bottleneck that prevents travelers from all arriving at their destinations at their most preferred times. The queue length depends on the entire traffic pattern starting from the most recent time it was empty and affects future travel delay until the queue is empty again. The no-toll equilibrium traffic and queuing patterns adjust endogenously over time so that identical travelers have the same total costs of queuing and early or late arrival times. The optimal tolls adjust continuously throughout the peak period to shift traffic and reduce queuing delay. With deterministic queuing, the toll completely replaces the queue and converts all queuing costs into revenues. The model’s improvements over the standard model include: dynamic treatment of congestion, explicit modeling of travelers' choices of travel times, endogenous peaking of traffic and delay, and inclusion of schedule delays associated with travel time decisions.

Unfortunately, the economics literature largely ignored the bottleneck model until the late 1980's and even now the standard steady-state model still appears to be the preferred framework for modeling congestion. Richard Arnott, André de Palma, and Robin Lindsey (1990) revived the bottleneck model in the economics literature by formalizing it, extending it to include a coarse (single-step) toll, and determining the optimal highway capacity. Arnott, et al., (1993) subsequently determined the optimal uniform, coarse, and continuous tolls with elastic demand. They demonstrate that applying the standard model to subintervals of the peak period is conceptually unsound, but that the standard model can represent a “semi-reduced form” of the entire peak period. They also show that efficiency gains are substantially greater when accounting for endogenously chosen travel times than those estimated with the standard model. They demonstrate that the self-financing properties of Herbert Mohring and Mitchell Harwitz (1962) and Robert Strotz (1965) apply to the bottleneck model whenever the pricing regime optimizes traffic levels under whatever constraints on the form of pricing that the airport faces. Ralph Braid independently extends the bottleneck model to cover elastic demand. Arnott, et al., (1989) and Yuval Cohen (1987) extend the bottleneck model to heterogeneous travelers.

Daniel (1991) develops a bottleneck model with deterministic congestion that is non-linear in traffic rates and applies the model to a stylized hub-and-spoke airline network. Daniel (1995) develops a bottleneck model with stochastic queuing and includes Nash and Stackelberg dominant airlines with atomistic or non-atomistic traffic. He implements the stochastic bottleneck model empirically using tower log data from Minneapolis-St. Paul airport and performs specification tests that suggest Northwest Airlines does not internalize its self-imposed delays. In particular, Northwest flights apparently do not adjust their operating times to account for delays they impose on other Northwest flights. Daniel (2001) extends the stochastic bottleneck model to include elastic demand, heterogeneous aircraft costs, and fringe aircraft with uniformly distributed preferred operating times.

Brueckner (2002) develops a congestion-pricing model with a dominant airline and a competitive fringe using the standard congestion technology in which the dominant airline internalizes its self-imposed delays so that its optimal congestion fee is inversely proportional to its market share. Brueckner also uses on-time performance data aggregated by airport per annum to show that more concentrated
airports have less delay. Mayer and Sinai (2003) perform an extensive empirical study of the concentration-delay relationship using data on excess flight times over minimum flight times by city-pair routes. They use dichotomous variables to control for level of hubbing activity by airport. They find a statistically significant, inverse relationship between airport concentration and excess flight time, which they interpret as evidence of internalization by the dominant airlines. They also find a much stronger direct relationship between hubbing activity and delays.

Daniel and Harback (2008) applies Daniel's (1995) specification test of internalization versus non-internalization to twenty-seven major hub airports, finding that dominant airline flight schedules are more consistent with minimizing individual aircraft costs rather than joint costs. They argue that in the stochastic bottleneck model, the atomistic fringe's adjustment of its flight times in response to peak spreading by the dominant airline (in an attempt to internalize self-imposed delays) offsets any reduction in peak traffic. The internalizing dominant airline realizes more schedule and queuing delay than it expects under the Nash assumption that fringe schedules do not change. Knowing this, a Stackelberg dominant airline behaves atomistically. Brueckner and van Dender (2008) seek to unify the internalization versus non-internalization debate using a simple transparent model with two periods and the standard steady state congestion technology. They criticize the stochastic bottleneck model of Daniel (1995) and Daniel and Harback (2008) as opaque because the stochastic queuing system prevents closed-form solution of the model. Brueckner and van Dender obtain either internalization or non-internalization as the dominant airline's optimal solution depending on the fringes’ demand elasticity for aircraft operations during the congested period. Daniel (2014) applies a deterministic bottleneck model with dominant and fringe traffic to airport slot constraints. That paper develops similar equilibria as developed here, but addresses quantity restrictions rather than congestion pricing. The models are dualistic and produce identical traffic patterns and social costs when the number of slot windows and toll periods are identical.

This paper parallels that of Brueckner and van Dender (2008) by using a deterministic bottleneck model with explicit closed form solutions to show the conditions under which either internalization or non-internalization is an optimal strategy for the dominant airline. Although the deterministic bottleneck model is more complicated than a two-period model with the standard congestion technology, it provides a richer environment for modeling the effects of congestion and pricing on aircraft scheduling. The state-dependent queuing system captures the dynamic nature of congestion. The model also includes schedule delay among the private and social costs. The dominant airline may adopt atomistic, internalizing, or mixed behavior in each specification depending on the fringe demand elasticities. In all cases of fully atomistic behavior and for all continuous tolls, the step tolling increments for the dominant and fringe aircraft are the same. The internalizing and mixed cases generally require an additional uniform toll that differentiates between dominant and fringe aircraft, but not simply on the basis of traffic shares. The paper also determines the optimal airport capacities and demonstrates that tolling revenues equal optimal capacity costs under constant returns to construction of airport capacity when there are dominant and fringe operations. Unlike Daniel’s (1995) stochastic bottleneck model, the deterministic version produced here has closed form solutions for all specifications.
3. The model

We now develop a general multi-step tolling model based on the framework of Vickrey (1969) and Arnott, et al. (1990). The model includes the uniform and coarse tolls of Arnott, et al. as special cases. As the number of steps in the toll structure increases, the model approaches Vickrey’s continuously varying toll in the limit. This unified treatment of the tolling structures covers all the relevant cases of time-dependent tolling in a deterministic bottleneck model.

3.1 Atomistic aircraft scheduling decisions

In this section, we reinterpret the bottleneck framework of Vickrey and Arnott, et al., in the context of airport traffic arriving and departing from an airport with limited capacity. Let $m$ be the total number of atomistic aircraft wanting to operate during an atomistic peak. Let $t_L^*$ and $t_T^*$ be the most preferred landing and takeoff times. Runway capacity limits the landing and takeoff rates to $s$ operations of each type per minute. Let $r_L[t]$ and $r_T[t]$ be the rate at which aircraft join the landing and takeoff queues at time $t$. Air traffic control alternates landing and takeoffs in such a way that aircraft do not impose delays on each other. This justifies using two identical models with separate queues and different time-cost parameters to treat the arrival peak and departure peaks separately. Throughout the paper, we consider a generic peak period that can apply to either landings or takeoffs.

Deterministic queues develop at the runway bottlenecks from time $t_{ab}$, the beginning of an atomistic peak, according to the equation:

$$q(t) = \int_{t_{ab}}^{t} (r[u] - s)du.$$  

Limited capacity prevents all aircraft from exiting the queue at precisely time $t^*$, so the operations must spread out before or after the most-preferred operating time. An aircraft that joins the landing or takeoff queue at time $t$ will spend $q[t]/s$ minutes in the queue at a cost of $\alpha$ per minute, and it will complete service at time $t + q[t]/s$. An early aircraft (completing service before $t^*$) will have $t^* (t + q[t]/s)$ minutes of early time before $t^*$. Early time has a cost of $\beta$ per minute. A late aircraft (completing service after $t^*$) will have $(t + q[t]/s) - t^*$ minutes of late time. The cost of late time is $\gamma$ per minute. While the values of $\alpha$, $\beta$, and $\gamma$ are different on landing and takeoff, the subscripts $L$ or $T$ that differentiate landing or takeoffs are suppressed because the models are otherwise identical. Assume that all aircraft have identical time costs. The sum of an aircraft’s operating costs during a peak period is:

$$C(t) = \alpha \frac{q[t]}{s} + \beta \max[0, t^* - t - \frac{q[t]}{s}] + \gamma \max[0, t + \frac{q[t]}{s} - t^*].$$

In a no-fee atomistic bottleneck equilibrium, traffic adjusts to maintain constant costs $C(t) = C^*$ across all times in which aircraft operate. Solve for $r[t]$ separately when $t$ is early, $t \leq t^* - q[t]/s$, or late, $t > t^* - q[t]/s$, by substituting (1) into (2) and differentiating with respect to $t$ while imposing the constant cost condition. This gives the aggregate traffic rates during fringe operating times:

---

1. actual airports operating under balanced traffic (landing and takeoff) conditions. The author's observations of traffic counts indicate that somewhat higher rates of takeoff are possible when there are no landings, but that no additional landings are possible when there are no takeoffs.
Equations (1)-(3) represent the essential structure of the deterministic bottleneck model that Vickrey (1969) and Arnott, et al. (1990) developed. We now extend these models to develop a unified bottleneck model with no tolls, uniform tolls, coarse tolls, multiple-step tolls, and continuous tolls.

3.2 Airport tolling decisions: multiple toll periods with atomistic traffic

To solve the multiple-step tolling problem, the airport authority takes the number of early and late toll periods parametrically and chooses a toll (and implicitly a duration) for each period. Tolls are assessed at the time aircraft join the queue, not when they complete service. For notational purposes, it is convenient to include a dummy toll \( \tau_0 \) and instantaneous period 0 between periods -1 and 1, so that these periods may be treated similarly to the others. The dummy toll \( \tau_0 \) represents the cost of an aircraft that arrives precisely on time (which is also the equilibrium operating cost \( C^* \) for all aircraft in the peak). The duration of the toll period is determined by the difference between tolls in adjacent periods stepping away from time \( 0 \). Figure 1 illustrates how to set up the airport authority’s optimization problem for choosing the tolls (and their durations) subject to the atomistic traffic rates as determined in Equation (3).

Figure 1 depicts two early and two late toll periods. The horizontal axis represents the time of the operations relative to \( t^* = 0 \). The vertical axis indexes the aircraft that participate in the peak period. Aircraft 0 is the one that completes service precisely on time. Aircraft with negative indices arrive early, and those with positive indices arrive late. The cumulative service function is the straight diagonal line with slope \( s \) equal to the service rate. The remaining line segments above and to the left of the cumulative-service function represent the cumulative-arrival function for the toll-equilibrium. Starting at times \( t \) on the far left of the time axis, early time cost \( \beta(t^*-t) \) is such that the cost of operating exceeds the equilibrium cost \( C^* \), therefore the traffic rate given by Equation (3) is 0. At time \( t \), such that early time cost \( \beta(t^*-t) \) equals the equilibrium level \( C^* \), the traffic rate jumps to \( \alpha s / (\alpha - \beta) \). The cumulative-arrival function begins increasing with slope equal to the traffic rate, which exceeds the service rate \( s \). The vertical distance between the cumulative-service function and the cumulative-arrival function is the number of aircraft that have arrived for service but have not yet been served, i.e., the number of aircraft in the queue. For any point on the cumulative-arrival function, the horizontal distance to the cumulative-service function is the time required to get through the queue. The queue grows over time until the queuing time cost \( \alpha q_t \) equals the toll increment \( (\tau_{i-1} - \tau_i) \), i.e., the cost of waiting until the next period to operate. So the maximal queue in toll period \( i \) will be \( (\tau_{i-1} - \tau_i) / \alpha \), and will occur exactly as the next toll period begins. The new toll (in addition to the long queue) raises the operating cost above \( C^* \). The atomistic traffic rate is 0 until the queue empties. As soon as the queue empties, the process repeats for the next early toll period. The first aircraft to operate in early toll periods \(-i \) and \(-i+1 \) experience no queuing cost, so the difference in early time costs must just offset the difference in tolls. The total time required to service the aircraft in early toll period \(-i \) is \( (\tau_{i-1} - \tau_i) / \beta \) and the number of aircraft served is \( (\tau_{i-1} - \tau_i)s / \beta \).
The first late toll period in Figure 1 is the same as a late arrival period in an untolled or uniform tolled bottleneck equilibrium of Vickrey (1969) or Arnott, et al. (2003). The queue decreases gradually from the previous period’s peak because the traffic rate for late arrivals given in Equation (3) is below the service rate. The second late toll period appears different because the queue empties just before the toll period begins. The lower toll and empty queue causes a sudden decrease in operating costs below $C^*$. Equation (3) requires an instantaneous group of arrivals sufficient to raise the expected queuing cost to the equilibrium level, so the actual queuing cost must be twice that which is necessary to raise operating costs to $C^*$. Immediately after these simultaneous arrivals, Equation (3) requires the traffic rate remain at zero until queuing costs reduce sufficiently to bring operating costs down to $C^*$. At this point, traffic rates increase to $\alpha s/\left(\alpha + \gamma\right)$ until the queue empties. As soon as the queue empties, the process repeats for the next late toll period. It turns out that for any given toll increment, the first and subsequent late toll periods have the same costs, as shown below.\(^3\)

To derive the total social cost of all the operations, it is necessary to determine the areas of the triangles and rectangles labeled in Figure 1. The area between the cumulative-arrival and cumulative-service functions is equal to the total queuing time. For early toll periods, we saw above that the maximum queuing time is $\left(\tau_{-(i-1)} - \tau_{-i}\right)/\alpha$ and the number of aircraft operating in the period is $\left(\tau_{-(i-1)} - \tau_{-i}\right)s/\beta$. The traffic and service rates are constant, so queuing time varies uniformly from zero to its maximum. The total queuing time is

$$
\left(\frac{\tau_{-(i-1)} - \tau_{-i}}{\alpha}\right) \left(\frac{(\tau_{-(i-1)} - \tau_{-i})}{\beta}\right) s/2.
$$

The area between the cumulative-service function and the vertical axis (to the left of $t^*$) is equal to the total early time. As shown above, the change in early time during an early period is $\left(\tau_{-(i-1)} - \tau_{-i}\right)/\beta$ and the number of aircraft operating in the period is $\left(\tau_{-(i-1)} - \tau_{-i}\right)s/\beta$, so the early time is

$$
\left(\frac{\tau_{-(i-1)} - \tau_{-i}}{\beta}\right) \left(\frac{(\tau_{-(i-1)} - \tau_{-i})}{\beta}\right) s/2\text{ plus the total time these aircraft have to wait on the ground while aircraft in subsequent toll periods operate, i.e., }\left(\tau_{-(i-1)} - \tau_{-i}\right)/\beta \sum_{j=0}^{i-1} (\tau_{-(j-1)} - \tau_{-j})/\beta.
$$

For late toll periods, the mass of arrivals at the beginning of the period must increase expected queuing costs sufficiently to offset the toll reduction. The number of aircraft required to do so is $2\left(\tau_{i} - \tau_{(i-1)}\right)s/\left(\alpha + \gamma\right)$. The total number of aircraft operating in the toll period, as shown above is $\left(\tau_{i} - \tau_{(i-1)}\right)s/\gamma$, so the number of aircraft that operate after the initial mass is $\left(\alpha - \gamma\right)(\tau_{i} - \tau_{(i-1)})s/\gamma\left(\alpha + \gamma\right)$. The length of the queue when these operations resume is $\left(\alpha - \gamma\right)(\tau_{i} - \tau_{(i-1)})s/\gamma\left(\alpha + \gamma\right)$, so the total queuing time during a late toll period is

$$
2\left(\tau_{i} - \tau_{(i-1)}\right)^2 s/\left(\alpha + \gamma\right)^2 + \left(\left(\alpha - \gamma\right)(\tau_{i} - \tau_{(i-1)})\right)^2 s/\left(2\alpha \gamma\right)\left(\alpha + \gamma\right)^2.
$$

This expression reduces to $\left(\tau_{i} - \tau_{(i-1)}\right)^2 s/\left(2\alpha \gamma\right)$, which is the same total queuing cost as for toll period 1.

The area between the vertical axis and the cumulative-service function (to the right of $t^*$) is equal to total late time. As shown above, the change in late time during a late period is $\left(\tau_{i} - \tau_{(i-1)}\right)/\gamma$ and the number of aircraft operating in the period is $\left(\tau_{i} - \tau_{(i-1)}\right)s/\gamma$, so the total time is $\left(\tau_{i} - \tau_{(i-1)}\right)/\gamma$.

---

2 I assume here, that $\gamma<\alpha$ that involves shifting the entire peak relative to $t^*$. Daniel and Harback’s (2008) extensive empirical estimates uniformly show that queuing time is more expensive than late time, so this issue is unlikely to arise in the airport context.

3 If we extend the dummy period 0 to cover the entire time aircraft 0 spends in the queue, then period 1 would have the same traffic pattern as the subsequent periods.
\( \tau_{(i-1)}/y \) s/2 plus the late time these aircraft experience while aircraft in previous late toll periods operate, i.e., ((\( \tau_i - \tau_{(i-1)} \))s/y) \( \sum^{i-1}_{j=1} (\tau_j - \tau_{(j-1)})/y \).

The airport authority’s optimization problem also involves three constraints. The first requires that the airport serve a total of \( m \) aircraft. In subsequent sections, \( m \) is determined endogenously using a demand function and treating the solution of this cost-minimization problem as the marginal social cost or the airport’s supply of aircraft operations. The second constraint requires that the first and last aircraft in the busy period have the same cost. Aircraft adjust the timing of their operations to satisfy this condition, so it acts as a constraint on airport’s ability to set tolls. Similar constraints apply to all the toll periods, but they are redundant because the aircraft response functions guarantee they are satisfied so long as the first and last aircraft satisfy the second constraint. The third constraint requires operating costs to be the same in toll periods \(-1\) and \(1\). Unlike the other toll periods, periods \(-1\) and \(1\) are linked by the length of the queue when the toll period changes and by sharing the instantaneous toll, \( \tau_0 \), which is actually the equilibrium operating cost \( C^* \). The airport cannot enforce different tolls in these periods without causing aircraft to shift between the periods to reestablish the condition, so it acts as a constraint on the airport. The last two constraints always hold, but they only bind strictly when the numbers of early and late periods are different. Their Lagrange multipliers shift aircraft between periods \(-1\) and \(1\) and make offsetting shifts between aircraft in the remaining late and early toll periods.

The airport authority’s problem is to choose \( \tau_{-y}, \ldots, \tau_z \) to minimize the total cost of aircraft in a step toll regime with a given number of early and late tolling periods, \( y \) and \( z \), subject to the constraints that all aircraft are served, and that all aircraft have the same total operating costs. Its objective function is:

\[
\text{Minimize} \quad \tau_{y}, \ldots, \tau_z, \lambda, \mu, \eta \\
\alpha \left( \sum^{y}_{i=1} \frac{(\tau_{i-1})^{2}}{2 \beta} s + \sum^{z}_{j=1} \left( \frac{(\tau_{j})^{2}}{2 \alpha} + \frac{(\beta - \gamma)^{2}}{2 \alpha \gamma} \right) \right) s + \\
\beta \left( \sum^{y}_{i=1} \frac{(\tau_{i,0})^{2}}{\beta} s + \sum^{z}_{j=1} \left( \frac{(\tau_{j,0})^{2}}{\beta} s \right) \right) - \\
\gamma \left( \sum^{y-1}_{i=1} \frac{(\tau_{i,0})^{2}}{\beta} s + \sum^{z-1}_{j=1} \frac{(\tau_{j,0})^{2}}{\beta} s \right) - \\
\lambda \left( \sum^{y}_{i=1} \frac{(\tau_{i,0})^{2}}{\beta} s + \sum^{z}_{j=1} \frac{(\tau_{j,0})^{2}}{\beta} s - m \right) - \\
\mu \left( \sum^{y}_{i=1} \frac{(\tau_{i,0})^{2}}{\beta} s - y \frac{(\tau_{i,0})^{2}}{\beta} \right) - \\
\eta \left( \sum^{z}_{j=1} \frac{(\tau_{j,0})^{2}}{\beta} s - y \frac{(\tau_{j,0})^{2}}{\beta} \right).
\]

The first order necessary conditions for (4) are as follows:

\[
\frac{\alpha}{\beta} s + \frac{(\tau_{i,y})^{2}}{\beta} s - \sum^{y}_{i=1} \frac{(\tau_{i,0})^{2}}{\beta} s = 0, \\
(\tau_{i,i+1})^{2} \beta s - \frac{(\tau_{i,0})^{2}}{\beta} s = 0, \text{ for } i = 2, \ldots, y-1, \\
\eta s + \frac{(\tau_{i,0})^{2}}{\beta} s - \frac{(\tau_{i,0})^{2}}{\beta} s = 0, \\
- \frac{(\beta + \gamma)}{\beta} y s + \frac{(\tau_{i,0})^{2}}{\beta} s + \frac{(\tau_{i,0})^{2}}{\beta} s + \sum^{y}_{i=1} \frac{(\tau_{i,0})^{2}}{\beta} s + \sum^{z}_{j=1} \frac{(\tau_{j,0})^{2}}{\beta} s = 0, \\
- \eta s + \frac{(\tau_{i,0})^{2}}{\beta} s - \frac{(\tau_{i,0})^{2}}{\beta} s = 0, \\
- \frac{(\tau_{i,0})^{2}}{\beta} s - \frac{(\tau_{i,0})^{2}}{\beta} s = 0, \text{ for } i = 2, \ldots, z-1,
\]
\[
\frac{\lambda}{\gamma} x - \mu x - \frac{(\tau_{x} - \tau_{x-1})}{\gamma} x - \sum_{i=1}^{\gamma} \frac{(\tau_{i} - \tau_{i+1})}{\gamma} x = 0,
\]
\[
\left( \sum_{i=1}^{\gamma} \frac{(\tau_{i-1} - \tau_{i})}{\beta} y \right) + \left( \sum_{j=1}^{\gamma} \frac{(\tau_{j} - \tau_{j+1})}{\gamma} y \right) - m = 0,
\]
\[
\beta \left( \sum_{i=1}^{\gamma} \frac{(\tau_{i-1} - \tau_{i})}{\beta} y \right) - y \left( \sum_{j=1}^{\gamma} \frac{(\tau_{j} - \tau_{j+1})}{\gamma} y \right) = 0,
\]
\[
\beta \left( \frac{(\tau_{z} - \tau_{z-1})}{\beta} y \right) - y \left( \frac{(\tau_{z} - \tau_{z+1})}{\gamma} y \right) = 0.
\]

Equation (6) implies that the early tolls (with negative indices) have common increments, and Equation (10) implies that late tolls (with positive indices) also do. Let \(\Delta_{y}\) and \(\Delta_{z}\) denote these toll increments:

\[
\Delta_{y} = (\tau_{y-1} - \tau_{y-2}) = (\tau_{y-2} - \tau_{y-3}) = \cdots = (\tau_{y-y} - \tau_{y-1}),
\]
\[
\Delta_{z} = (\tau_{z} - \tau_{z-1}) = (\tau_{z+1} - \tau_{z+2}) = \cdots = (\tau_{z} - \tau_{z-1}).
\]

Substituting \(\Delta_{y}\) and \(\Delta_{z}\) into Equations (7) and (9) implies that initial toll increments, \(\Delta_{1}\) and \(\Delta_{1}\), which involve \(\tau_{0}\) may be expressed as:

\[
\Delta_{1} = (\tau_{1} - \tau_{0}) = \Delta_{y} + \eta \beta, \quad \text{and} \quad \Delta_{1} = (\tau_{1} - \tau_{0}) = \Delta_{z} - \eta \gamma.
\]

Now we can evaluate all the summation terms in the first order conditions by substituting the expressions in Equations (15), (16), and (17) to obtain

\[
\sum_{i=1}^{\gamma} \frac{(\tau_{i-1} - \tau_{i})}{\beta} s = \frac{y \Delta_{y} + \eta \beta}{\beta} s \quad \text{and} \quad \sum_{j=1}^{\gamma} \frac{(\tau_{j} - \tau_{j+1})}{\gamma} s = \frac{z \Delta_{z} - \eta \gamma}{\gamma} s.
\]

Substituting these expressions for the summations in (12) and then cancelling the terms in \(\eta\) provides:

\[
\left( \frac{y \Delta_{y} + \eta \beta}{\beta} \right) + \left( \frac{z \Delta_{z} - \eta \gamma}{\gamma} \right) = \frac{m}{s} = \left( \frac{y \Delta_{y}}{\beta} \right) + \left( \frac{z \Delta_{z}}{\gamma} \right).
\]

Equations (13) and (18) imply:

\[
\Delta_{z} = y \Delta_{y}
\]

Substituting \(y \Delta_{y}\) for \(z \Delta_{z}\) in Equation (19), solving for \(\Delta_{y}\), and then substituting \(\Delta_{y}\) back into Equation (20) to solve for \(\Delta_{z}\) yields:

\[
\Delta_{y} = \frac{\beta y m}{(\beta + \gamma) s y} \quad \text{and} \quad \Delta_{z} = \frac{\beta y m}{(\beta + \gamma) s z}.
\]

Subtracting Equation (5) from (11) to eliminate \(\lambda\) using (18) to substitute for the summations, using (15) and (16) to substitute for the extra toll increment terms, solving for \(\mu\) and \(\Delta_{y}\) and \(\Delta_{z}\), yields:

\[
\mu = \frac{(z+1) \Delta_{z} - (y+1) \Delta_{y}}{\beta + \gamma} = - \frac{\beta y (y+1) m}{(\beta + \gamma) s y}. \quad \text{where} \quad 
\]

Substituting \(\mu\) and then \(\Delta_{y}\) and \(\Delta_{z}\) back into Equation (11) to solve for \(\lambda\) gives:

\[
\lambda = \frac{y (y+1) \Delta_{y} + \beta (z+1) \Delta_{z}}{\beta + \gamma} = \frac{\beta y (y+1) m}{(\beta + \gamma) s y z}.
\]

Subtracting Equation (9) from (7) and using 14 gives:

\[
\lambda = \frac{y (y+1) \Delta_{y} + \beta (z+1) \Delta_{z}}{\beta + \gamma} = \frac{\beta y (y+1) m}{(\beta + \gamma) s y z}.
\]
\[
\eta = \frac{\Delta z - \Delta y}{\beta + \gamma} = \frac{\beta y (y - z) m}{(\beta + \gamma)^2}.
\]

Finally, substituting \( \eta \) into Equation (17), yields the central toll increments:

\[
\Delta_{-1} = (\tau_0 - \tau_{-1}) = \Delta y + \eta \beta = \frac{\beta y (\beta + \gamma) m}{(\beta + \gamma)^2} \Delta y - \eta \gamma = (\tau_1 - \tau_0) = \Delta_1. \tag{25}
\]

Note that Equation (12) requires that the total number of aircraft equal \( m \), and that its Lagrange multiplier represents the marginal social cost of adding the \( m \)th aircraft to the peak. The value of \( \tau_0 \) represents the equilibrium operating cost (including the toll) of aircraft during the peak period. It is not determined by the optimization problem (4) and may take any value so long as the toll increments satisfy Equations (21) and (25). To obtain the optimal number of aircraft, however, requires that the marginal social benefit (i.e., travel demand or willingness to pay) of the \( m \)th aircraft equal the marginal social cost of its operation. Therefore, we set the equilibrium cost of operating in the peak (\( \tau_0 \)) equal to the marginal social cost (\( \lambda \)).

\[
\tau_0 = \frac{\beta y (\beta + \gamma) m}{(\beta + \gamma)^2} \Delta y - \eta \gamma = (\tau_1 - \tau_0) = \Delta_1. \tag{26}
\]

Equations (26) reveal the general structure of optimal step tolls, as illustrated in Figure 2 for the case of four early periods, three late periods, plus the central peak. The tolls during the central period (steps \(-1\) and \(1\)) equal the marginal social cost minus the equilibrium (average) cost of operating during the those periods. All other tolls and intervals follow from the peak toll and the times that the peak toll periods begin and end. The tolls step up and down in equal increments on either side until they reach zero at the beginning and end of the peak period. It follows that the early intervals are of equal length \((\tau_{i+1} - \tau_{i-1})/\beta\). The late intervals are of equal length \((\tau_{n+1} - \tau_{n-1})/\gamma\). The central periods \(-1\) and \(1\), work like an untolled bottleneck equilibrium. The point of this optimal step pattern is to reduce the accumulation of traffic in the queue during the early periods. The tolls provide incentives for aircraft to shift operating times to periods in which the traffic rates would otherwise be too low. The effectiveness of peak spreading improves as the number of toll steps increases.

In Figure 2, the saw-tooth line along the top of the diagram shows how operating costs \( C(t) \) vary over time. Traffic only operates during the minimized flat regions between the teeth when \( C(t) \) equals \( C^* \). Each early tolling period begins with a sharp spike in operating costs \( C(t) \) as toll schedule steps up. This increases costs above the equilibrium level \( C^* \) and stops traffic from flowing into the queueing system. The saw-toothed function at the bottom of the diagram represents the queuing costs. The queue empties during the cost spike and then builds again once traffic resumes, until there is another toll increment. In the late periods there is a rush to join the queue each time the toll schedule steps down. Queuing costs

---

4 Note that this solution is equivalent to the slot-constrained solution in Daniel (2014) where the optimal number of slot permits issued by the airport authority were:

\[
q_{-1} = \frac{y m}{(\beta + \gamma) y} = \frac{\Delta y}{\beta} s, \quad q_{-1} = \frac{\beta (y + y z) m}{(\beta + \gamma)^2 y z} = \frac{\Delta_1}{\beta} s, \quad q_1 = \frac{\beta m}{(\beta + \gamma) y} = \frac{\Delta_1}{\gamma} s.
\]

Consequently, the average aircraft operating cost, marginal social cost of operations, and total social costs of the aircraft operations are identical in the two specifications.
must jump by twice the toll increment so that the average increase just offsets the reduction in the tolls. This causes another cost spike that stops the traffic flow until queue diminishes sufficiently to reestablish the equilibrium costs, i.e., $C[t]$ equals $C^*$. If there were no tolls, the queuing costs would continue to build throughout the early period, eventually peaking at the equilibrium cost level for the aircraft completing service precisely at the most preferred time.

Integrating the marginal cost ($\lambda$) of the $m^{th}$ aircraft in the peak period from Equation (26) with respect to the number of aircraft ($m$) gives the total social cost. Dividing that result by $m$ by gives the average operating-time cost experienced by an aircraft, not including the toll. Separating out the multiple $\Gamma$, by which these costs differ from the unconstrained equilibrium costs in Vickrey and Arnott, et al.,

\begin{equation}
\begin{aligned}
    MC &= 2\Gamma \delta \frac{m}{z}, \\
    TC &= \Gamma \delta \frac{m^2}{z}, \\
    AC &= \Gamma \delta \frac{m}{z}.
\end{aligned}
\end{equation}

Step tolling recovers some of the deadweight loss from queuing in the form of airport revenues. The amount of the efficiency gain depends on the value of $\Gamma$ in Equation (27), which depends on the number of tolling periods. Table 1 shows how the efficiency of the step toll system varies with the number of early and late steps. The table uses cost parameters, $\alpha$, $\beta$, and $\gamma$, that are typical of those Daniel and Harback (2014) estimate for major hub airports in the US, but the overall efficiency results in the table are not particularly sensitive to variations in the cost parameter. The parameters $\beta$ and $\gamma$ affect the relative advantage of early versus late tolling periods. Efficiency improves rapidly as the number of steps increases; one step on each side of central period recovers half the efficiency loss from congestion, while five early and three late steps plus the central period recovers eighty percent of the loss.

When the numbers of early and late steps, $y$ and $z$, equal one, there is only one toll level and the model produces the no toll or uniform toll equilibria. These specifications have the same operating costs because no toll or a uniform toll has no effect on traffic schedules. The value of $\Gamma$ is one. As the values of $y$ and $z$ go to infinity, $\Gamma$ approaches one-half. The general solution for any atomistic step-toll equilibrium has the total (social) congestion given in Equation (27) as $\Gamma \delta m^2/s$. The equilibrium operating cost (not including tolls) is the average cost $\Gamma \delta (m)/s$. The marginal social cost of a landing or takeoff is $2 \Gamma \delta m/s$ or twice the average cost of the operation. The external cost of an operation is $MC-AC = 2 \Gamma \delta m/s$. Arnott, et al. (1993) notes that even though there is a dynamic structural model underlying Equations (27), all travelers face the same untolled operating cost (AC) or optimally tolled social cost (MC) regardless of when they operate in the peak. The entire peak period is represented by airport supply functions, $p = \delta m/s$ (untolled) or $p = 2 \Gamma \delta m/s$ (optimally tolled). In this sense, the bottleneck model and the standard steady state model produce similar results.

### 3.3 Airline choices: scheduling dominant and fringe aircraft, the no toll case

Now consider a dominant airline that provides hub-and-spoke service during a particularly busy period in which there is also atomistic traffic operated by many non-dominant airlines. The dominant airline schedules large groups (banks) of aircraft jointly, while the atomistic airlines schedule each aircraft independently of the others. The dominant airline wants to schedule a bank of arriving flights to land close to the optimal time $t^*_L$ so that passengers can connect with a bank of departing flights that takeoff close to the optimal time $t^*_T$. The optimal operating times, $t^*_L$ and $t^*_T$ are also desirable to the atomistic
aircraft because passengers prefer flights scheduled at these times. The dominant airline seeks to minimize the joint cost of its aircraft, recognizing that the other traffic will adjust atomistically. As we saw above, the bottleneck model captures the effects of both the timing of operations and the number of operations on congestion delays. A dominant airline internalizes its self-imposed delays. It does so primarily by spreading its traffic to even out the arrival rates, rather than by reducing the number of flights.\(^5\)

Let \(d=d[p_d]\) denote the dominant airline’s demand for operations during a single bank of arriving or departing aircraft, where \(p_d\) is the full price (operating cost plus toll, if any) of a dominant aircraft operation, as determined below. Suppose \(x\) of these dominant aircraft operate during the atomistic peak and \(d-x\) operate during the service intervals before and after the atomistic peak. There are non-dominant aircraft that also operate during the atomistic peak\(^6\). The demand for operations by non-dominant airlines is \(f=f[p_f]\) where \(p_f\) is the full price of a fringe operation. Assume that all aircraft share the same time values and the most-preferred operating time \(t^*\), which is a popular travel time for passengers.

Equation (3) now gives the aggregate traffic rates that are necessary to satisfy the equilibrium condition in Equation (2), which still determines the atomistic traffic rates. The dominant airline has no incentive to exceed these rates during the atomistic peak because doing so would increase its queuing delay without reducing its schedule delay. The best (scheduling) responses of the atomistic traffic as a function of the dominant airline’s arrival rates are:

\[
(28) \quad r_f(t; r_d(t)) = \begin{cases} 0 & \text{for } C[t] > C^*, \\ \text{Max} \left[0, \frac{s\alpha}{\alpha - \beta} - r_d[t] \right] & \text{for } C[t] = C^* \text{ and } t \leq t^* - \frac{q[t]}{s}, \\ \text{Max} \left[0, \frac{s\alpha}{\alpha + \gamma} - r_d[t] \right] & \text{for } C[t] = C^* \text{ and } t^* - \frac{q[t]}{s} \leq t, \text{ and} \\ \infty^+ & \text{for } C[t] < C^*. \end{cases}
\]

\(\dagger\) i.e., simultaneous arrival of \(2(C^*-\beta \text{ Max}[s\alpha - t - t^*\gamma]/\alpha + \text{ Max}[s\alpha + s\gamma - t^*\gamma]/\alpha)\) aircraft.

In the no-toll case where \(\Gamma=1\),\(^7\) the full price for a fringe aircraft operation as given by Equation (27) is \(p_f = \delta (f+x)/s\). To obtain explicit solutions for the fringe demand, it is useful to assume linear demand. Let the supply be given by Equation (27), and demand fringe demand be \(f = \eta - \pi p_f\) substitute the supply price into the demand function to solve for the optimal number of the fringe aircraft, \(f\), as a function of the number of dominant aircraft scheduled during the atomistic peak, \(x[d]\).

\[
(29) \quad p_f[f, x[d]] = \frac{\delta (f+x)}{s}, \quad f^e[p_f] = \eta - \pi p_f[f, x[d]], \quad \text{and} \quad \hat{f}^e = \frac{\eta s - \pi x[d]}{s + \pi}.
\]

Now consider the dominant airline’s problem of scheduling aircraft before or after the atomistic peak. These aircraft cannot obtain service more rapidly than the service rate \(s\), and will have no queue if they

---

\(^5\) Peak spreading can reduce the full price of operating so much that the dominant airline actually increases its number of aircraft relative to the atomistic equilibrium.

\(^6\) The “atomistic peak” refers to the period of time during which the airport is congested, i.e., there is a queue. The “busy” period refers to the period of time during which the arrival rate is positive. The busy period includes the atomistic peak and the internalizing dominant operations.

\(^7\) The no tolling case has two “tolling periods” (\(y=1, z=1\)) in which the toll is zero (\(\tau_1 = 0, \tau_2 = 0\)).
operate at or below rate $s$. It follows that $s$ is the least costly operating rate. Let $t_{db}$ and $t_{de}$ be the beginning and ending times of the dominant aircraft operations, and $t_{ab}$ and $t_{ae}$ be the beginning and ending times of the atomistic peak. Applying the second constraint of (4) to determine, $t_{db}$ and $t_{de}$, gives the dominant airline’s traffic pattern for aircraft not scheduled during the atomistic peak:

\[
\begin{align*}
    r_d(t) &= s, \quad \text{for } t \in [t_{db}, t_{ab}] \text{ or } t \in [t_{ae}, t_{de}], \\
    t_{db} &= t^* - \frac{d + f}{s} \frac{1}{\beta + \gamma} \beta + \gamma, \quad \text{and} \\
    t_{de} &= t^* + \frac{d + f}{s} \frac{1}{\beta + \gamma} \beta + \gamma.
\end{align*}
\]

By Equation (13), the fractions of aircraft operating early and late are $\gamma / (\beta + \gamma)$ and $\beta / (\beta + \gamma)$ respectively. The $\gamma / (\beta + \gamma) (d - x)$ dominant aircraft operating before the atomistic peak will experience early time on average of $\gamma / (\beta + \gamma) ((d - x)/(2s) + (f + x)/s)$. The $\beta / (\beta + \gamma) (d - x)$ dominant aircraft operating after the atomistic peak will experience late time on average of $\beta / (\beta + \gamma) ((d - x)/(2s) + (f + x)/s)$. Multiplying the numbers of aircraft by their time values and average delay times gives the total cost of the internalizing dominant aircraft. The terms in $\beta$ and $\gamma$ factor to equal $\delta = \beta \gamma / (\beta + \gamma)$. Adding the total cost of the dominant aircraft in the atomistic peak, $\delta x (f + x)/s$, and substituting the fringe demand gives the dominant airline’s objective function. The dominant airline acts as a Stackelberg leader and chooses the number of aircraft to schedule during the atomistic peak subject to fringe’s optimal number of aircraft as determined in Equation (29):

\[
\begin{align*}
    \text{Minimize}_{\delta x} \left\{ \delta (d - x) \left( \frac{(d - x)}{s} + \frac{(f + x)}{s} \right) + \delta x f \right\}, \quad \text{s.t.} \quad \frac{\delta x}{s + \delta x} = \frac{s - \delta x}{s} \Rightarrow x[d] = \frac{s \delta x}{s + \delta x}. \quad \text{d}.
\end{align*}
\]

The solution for $x[d]$ is the essential result in this section. It shows that a Stackelberg dominant airline schedules a fraction $\delta \pi / (s + \delta \pi)$ of its aircraft during the atomistic peak. These aircraft behave atomistically because the bottleneck equilibrium conditions (Equations (2) and (3)) determine the aggregate traffic pattern, and dominant aircraft cannot change the traffic pattern by adjusting their operating time within the atomistic peak. The dominant airline reduces its self-imposed delays in the atomistic peak by shifting aircraft to operating times before and after the atomistic peak. The fraction of dominant aircraft that behave atomistically depends on $\pi$, the slope of the fringe inverse demand function. When $\pi$ is zero, fringe demand is perfectly inelastic and the dominant airline internalizes all self-imposed delays by scheduling them before and after the atomistic peak. When $\pi$ is infinite, fringe demand is perfectly elastic and the dominant airline schedules all of its aircraft atomistically.

This leads to the following conclusion:

Conclusion 1: When dominant and fringe airlines have identical time values and operating-time preferences, the untolled equilibrium has an atomistic bottleneck equilibrium surrounding $t^*$ that includes all of the fringe aircraft and a fraction of the dominant aircraft that varies from zero (the perfectly inelastic case) to one (if fringe demand is sufficiently elastic). The dominant airline internalizes the self-imposed delays of its remaining aircraft by scheduling them to operate before and after the atomistic peak at exactly the rate of service. These internalizing dominant aircraft do not create or experience any queuing delay.

To understand the intuition behind this result, suppose that fringe demand were inelastic but there were some dominant aircraft in the atomistic peak. All of the periods during the atomistic peak have the
same equilibrium cost, so the dominant airline could always reschedule any of its aircraft from the atomistic peak to the edges of the peak without increasing their cost. Atomistic aircraft would shift to fill the gaps in traffic left by the dominant aircraft. The dominant airline would set the traffic rates of the rescheduled aircraft equal to the service rate so that they would not impose delays on one another. The length of the atomistic peak would decrease by one service period for each rescheduled dominant aircraft. The equilibrium cost in the atomistic peak would decrease to equal that of the dominant aircraft at the edge of the peak. This process would continue until no dominant aircraft remained in the atomistic peak.

With elastic demand, moving dominant aircraft out of the atomistic peak reduces the average cost (full price) of atomistic operations at the rate of $\delta/s$ per aircraft, which induces additional fringe aircraft to enter the peak, driving the cost up. As new fringe aircraft enter, the peak period re-expands, pushing the internalizing dominant aircraft away from their preferred operating time. If new fringe entry only partially offsets the cost reduction, then dominant aircraft remaining in the atomistic peak benefit from internalizing. The dominant airline balances the reduction in cost for its atomistic aircraft against the increase in cost of its internalizing aircraft. If fringe demand is sufficiently elastic, then $x$ approaches $d$; i.e., the dominant airline leaves all its aircraft in the atomistic peak to preempt entry by the fringe.

### 3.4) Airport decisions with dominant and fringe airlines, the tolling equilibria

The basic principle of optimal tolling requires that every aircraft face its full social cost to optimize both the number and scheduling of its operation. The multiple-step toll structure achieves the optimal scheduling of aircraft operating in the atomistic peak. It does not, however, optimize the number of atomistic aircraft, because they impose additional external costs by pushing dominant aircraft operating before or after the atomistic peak further from their most preferred time. An additional uniform toll is necessary to internalize this cost and optimize the number of atomistic aircraft.

The airport authority’s problem is to set the full price for atomistic aircraft equal to their marginal social costs. Recall from Equation (31) that the total cost of the dominant aircraft that operate before and after the atomistic peak is $\delta(d-x)((d-x)/(2s) + (f+x)/s)$. From Equation (27), the social cost of a step-tolled atomistic peak is $\Gamma \delta (f+x)^2/s$. Differentiating the total social cost and subtracting the cost that a fringe aircraft experiences, $\Gamma \delta (f+x)/s$, determines the optimal fee for fringe aircraft.

\[
p_f^* = \partial_f \left\{ \delta (d-x) \left( \frac{(d-x)}{2s} + \frac{(f+x)}{s} \right) + \Gamma \delta \left( \frac{(f+x)}{s} \right)^2 \right\} = \delta (d-x) + \Gamma \delta \left( \frac{(f+x)}{s} \right).
\]

It follows that the optimal fringe toll structure consists of the step tolls in Equation (26) plus an additional uniform toll $T_f^U = \delta(d-x)/s$ to internalize the additional delay fringe aircraft impose on the internalizing dominant aircraft. Substituting the full price in the fringe demand function gives the fringe’s optimal demand under the step toll as a function of the dominant airline’s demand and number of aircraft scheduled during the atomistic peak: $f_\ast[x,d] = \frac{\eta + \delta \pi (d+(f-1)x)}{s+\delta \pi}$. The social cost minimizing number of dominant aircraft operating in the atomistic peak is the value of $x$ that minimizes the bracketed term in Equation (32). The Stackelberg dominant firm, however, treats fringe demand as a function of $x$ and $d$, and minimizes this expression subject to fringe demand, which distorts its solution relative to the social optimum. An additional toll on $x$ is necessary to prevent the dominant airline from preemptively scheduling aircraft during the atomistic peak to reduce the number of
fringe aircraft. The amount of this distortion is difference between the first order conditions for the socially optimal and Stackelberg solutions of x. Introducing non-negativity constraints on x, we obtain:

$$
\begin{align}
\frac{\partial}{\partial x} \left\{ \delta (d - x) \left( \frac{(d-x)}{2x} + \frac{(f(x+d)+x)^2}{s} \right) + \Gamma \frac{\delta (f(x+d)+x)^2}{s} - \lambda x \right\} - \frac{\partial}{\partial x} \left\{ \delta (d - x) \left( \frac{(d-x)}{2x} + \frac{(f+x)^2}{s} \right) + \Gamma \frac{\delta (f+x)^2}{s} - \psi x \right\} = \\
\frac{\delta (d-x) + 2 \delta \Gamma (f[x+d]+x) f_d[x,d]}{s} + \frac{\delta + \lambda x}{s} - \psi = 0.
\end{align}
$$

The dominant firm experiences $\delta (d-x)/s + \delta \Gamma (f[x,d] + x) f_d[x,d]/s$ of this difference in social cost, so the external part is $\delta \Gamma (f[x,d] + x) f_d[x,d]/s$. This is the average cost of an aircraft operating in the atomistic peak multiplied by the rate of change in fringe demand with respect to additional dominant aircraft in the atomistic peak. Imposing an additional uniform toll of this amount on dominant aircraft operating in the atomistic peak results in the full price for additional dominant aircraft before or after the atomistic peak. Imposing an additional uniform toll of this amount on dominant aircraft operating in the atomistic peak multiplied by the rate of change in fringe demand with respect to additional dominant aircraft in the atomistic peak being the same as the social optimum:

$$
\frac{\delta (2 \Gamma - 1) (f[x+d]+x) + \psi - \lambda = 0 \text{ and } x (\psi - \lambda) = 0 \Rightarrow \lambda = 0.}
$$

The solution is to set x equal to zero because it is socially cheaper to schedule dominant aircraft before and after the atomist peak where they perfectly internalize delays rather than in the atomistic peak where they imperfectly internalize delays.

Similarly, an additional toll on the d=x dominant aircraft scheduled before and after the atomistic peak is necessary to prevent the dominant airline from preemptively scheduling aircraft before or after the atomistic peak to reduce the number of fringe aircraft. The amount of this distortion is difference between the first order conditions for the socially optimal and Stackelberg solutions for d.

$$
\begin{align}
\frac{\partial}{\partial d} \left\{ \delta (d - x) \left( \frac{(d-x)}{2x} + \frac{(f[x,d]+x)^2}{s} \right) + \Gamma \frac{\delta (f[x,d]+x)^2}{s} \right\} - \frac{\partial}{\partial d} \left\{ \delta (d - x) \left( \frac{(d-x)}{2x} + \frac{(f+x)^2}{s} \right) + \Gamma \frac{\delta (f+x)^2}{s} \right\} = \\
\frac{\delta (d+x(2 \Gamma - 1) + 2 \Gamma (f[x,d]+x)) f_d[x,d]}{s}.
\end{align}
$$

Substituting x=0 and noting that the dominant firm experiences $\delta (d + \Gamma (f[x,d])) f_d[x,d]/s$ of the difference in social cost, the external effect is $\delta \Gamma (f[x,d]) f_d[x,d]/s$. This is the average cost of an aircraft operating in the atomistic peak multiplied by the rate of change in fringe demand with respect to additional dominant aircraft before or after the atomistic peak. Imposing an additional uniform toll of this amount on dominant aircraft operating before or after the atomistic peak results in the full price for dominant aircraft operations equaling their marginal social cost in Equation (35).

The dominant firm schedules all its flights before and after the atomistic peak and pays a uniform toll of $T_d^U = \delta \Gamma (f[x,d]) f_d[x,d]/s$ per aircraft. It does not schedule any aircraft in the atomistic peak because the uniform toll of $\delta \Gamma (f[x,d] + x) f_d[x,d]/s$ on such aircraft in addition to the step toll would assure that scheduling them before or after the atomistic peak would be cheaper. The dominant airline faces the full social cost of all its operations and fully internalizes their self-imposed delay. The fringe aircraft pay step tolls depending on their operating times plus a uniform toll of $T_f^U = \delta (d-x)/s$.

The top panels of Figure 3 illustrate the untolled and step tolled equilibria for homogenous fringe and dominant and aircraft respectively. As Arnott, et al. (1993) observed, the reduced form of the bottleneck equilibrium appears similar to the standard model applied to the entire peak period, but it provides a structural model of airport supply based on the underlying congestion technology and optimal aircraft
scheduling. Figure 3 illustrates that determination of equilibrium full prices and quantities reduces to standard supply and demand diagrams, with the bottleneck model determining the shape of airport supply functions. A critical difference between previous atomistic bottleneck models and the dominant-fringe specification developed here is that the supply functions must account for the interaction between dominant and fringe operations. Each type of operation imposes a negative externality in production on the other. The supply curves depicted in the graph are actually projections of supply surfaces in \( f \times d \) space that account for the number of both types of operations. The tolled and untolled supply curves are projections of these surfaces holding the other output constant at the corresponding tolled or untolled equilibrium level. The position of the untolled supply surface at the tolled equilibrium output levels is different from its position at the untolled equilibrium level, so the toll is not the vertical difference between the depicted supply projections—as it would be in both the standard model and the atomistic bottleneck model. The actual supply surfaces do not cross (except at the origin) contrary to the appearance of their projections in the graph of dominant supply curves.

Conclusion 2: When dominant and fringe airlines have the same time values and preferences, imposing the same atomistic step-toll schedule achieves the constrained-optimal scheduling of aircraft. Different uniform tolls are necessary to account for the effects of internalizing dominant aircraft that operate outside the peak. These uniform tolls optimize the number of dominant and fringe operations.

4. Capacity Implications

Arnott, et al., (1993) compares the full prices, efficiency losses, traffic volumes, and efficient capacity levels for atomistic bottleneck equilibria with homogeneous traffic facing uniform, coarse, continuous, and no tolls. It also shows that optimal capacity is self financing, extending the result of Mohring and Harwitz (1962) to the atomistic bottleneck model with homogeneous traffic. The extensions to the bottleneck model developed here have several sources of heterogeneity. These include, dominant and fringe airlines, different time values, and different scheduling time preferences. This section examines whether the properties hold for these extensions.

First consider the multiple-step tolling equilibrium derived above. Equations (7) give the total social cost function for the step tolls in the form \( TC^*[m,s] = \Gamma^* \delta m^*/s \), where \( \Gamma^* \) depends only on the cost parameters and the number of steps in the toll structure. Summarizing the derivation of Arnott, et al., (1993), we have \( ATC^*[m,s] = \Gamma^* \delta m/s \), \( MSC^*[m,s] = 2\Gamma^* \delta m/s \), and \( p^*[m,s] = ATC^*[m,s] + \tau^*[m,s] = MSC^*[m,s] \), which imply that \( \tau^*[m,s] = ATC^*[m,s] \). It follows that the supply function for the step tolled equilibrium is \( p^*[m,s] = MSC^*[m,s] = 2\Gamma^* \delta m/s \). The optimal prices and quantities \( p^* \) and \( m^* \) satisfy the supply \( p^*[m,s] \) and demand functions \( m[p^*,s] \) simultaneously. Since the corresponding values for \( \Gamma^* \) and \( \Gamma^* \) in the untolled and uniform equilibria are equal one, and \( \Gamma^* \) for the continuous toll is one half, it follows that \( p^* > p^0 > p^* = p^* \) and \( m^* < m^* < m^* = m^* \). Defining the social surplus function, \( SS^*[s] = \int_{p^*}^{\infty} m^*[p]dp + m^*[s]\tau^*[s] \), and the efficiency loss, \( EL^*[s] = SS^*[s] - SS^*[s] \), then because the uniform toll is a special case of the step toll, the efficiency losses are ranked, \( EL^*[s] > EL^*[s] > EL^*[s] > 0 \). Now define the marginal social benefit of capacity as the derivative of social surplus with respect to capacity:
\( MB^s[s] = -m^s[s] \frac{dp^s[s]}{ds} + \tau^s[s] \frac{dm^s[s]}{ds} + m^s[s] \frac{d\tau^s[s]}{ds}, \)

\[ = \tau^s[s] \frac{dm^s[s]}{ds} - m^s[s] \frac{d\tau^s[s]}{ds} \]

\[ = m^s[s] \tau^s[s] \left( \frac{1}{m^s[s]} \frac{dm^s[s]}{ds} + \frac{1}{\tau^s[s]} \frac{d\tau^s[s]}{ds} \right) \]

\[ = \frac{m^s[s] \tau^s[s]}{s} = \frac{m^s[s] ATC^s[s]}{s} = \frac{TC^s[s]}{s} \]

The bracketed term equals \( 1/s \) whenever, as here, the total cost for fixed traffic volume is proportional to \( 1/s \). The envelope theorem applies so that total cost is also proportional to \( 1/s \) with traffic volume varying optimally. Now with inelastic demand (which is the likely case for derived demand for landings and takeoffs) and the ordering of full prices as above, the full cost of operations and the total tolls also have the same ranking. Applying this to the last line of (36), ranks the marginal benefits as \( MB^u[s] > MB^s[s] > MB^o[s] \). This in turn implies that the optimal capacity has the order, \( su^* > ss^* > so^* \).

For similar reasons, using the assumption that capacity construction costs are inelastic with respect to capacity, the optimal price and quantities at the optimal capacities are also ordered \( pu^* > ps^* > po^* \) and \( mu^* < ms^* < mo^* \).

The problem of finding the optimal long run price and capacity for the multiple step toll case is to choose price and capacity to maximize the sum of consumer surplus and toll revenue less capacity construction costs, \( K[s] \):

\[ \max_{p^s, s^s} \left\{ \int_{p}^{\infty} m^s[p] dp + (p^s - ATC^s[m[p^s], s])m[p^s] - K[s] \right\}, \]

which has first order conditions:

\[ -m^s[p^s] + (p^s - ATC^s[m[p^s], s]) \frac{dm[p^s]}{ds} + m[p^s] \left( 1 - \frac{d ATC^s}{dm} \frac{dm}{dp^s} \right) = 0 \]

\[ \frac{d K[s]}{ds} - m^s[p] \frac{d ATC^s}{ds} = 0 \]

These simplify to the following rules: set the average toll equal to the marginal capacity cost, \( \tau[m, s] = p^s - ATC^s[m[p^s], s^s] = m[p^s] dATC/dm, \) and set capacity so that marginal construction costs equal the marginal benefit of additional capacity, \( dK[s]/ds = m[p^s] dATC/ds. \) Finally, assuming for convenience that \( ATC \) and \( K \) are homogeneous of degree \( h^C \) and \( h^K \), combining the first order conditions, and applying Euler’s Theorem produces:

\[ m^s[p^s] \left( m^s[p] \frac{d ATC^s}{dm} \right) = h^K K + m h^C ATC. \]

With the toll equal to the marginal congestion externality, the LHS of (39) is revenue, and the second term on the RHS is zero because \( ATC \) is homogeneous of degree zero. It follows that multiple step tolls cover the construction cost of the optimal capacity if \( K \) exhibits constant returns to scale, i.e., \( h^K = 1 \).

There are two key properties on which the above comparisons of the tolling structures and the self-financing result depend: that average total costs are inversely proportional to the service rate, \( s \), and that
costs are homogeneous of degree zero in traffic levels and service capacities, \( m \) and \( s \). We now consider whether these properties apply in the cases of dominant and fringe traffic. Note that all the full prices for the uniform and continuous toll cases where the dominant firm behaves fully atomistically or fully internalizes are inversely proportional to \( s \) and homogeneous of degree zero in quantities \( f \) and \( d \) and service rate \( s \). Further note that Equations (5) and (6) imply that the full prices for the step toll cases are \( p^f[m, s] = \Gamma_{y,TC} m/s \), where \( m = f + x[d] \) is the sum of fringe demand and the atomistically scheduled dominant aircraft and \( \Gamma_{y,TC} \) is constant with respect to \( m \) and \( s \). The full prices for the step toll cases are also inversely proportional to \( s \) and homogeneous of degree zero in quantities \( f \) and \( d \) and service rate \( s \). Thus, the dominant-fringe equilibria satisfy Equation (39), and the relationships apply as before. To see that it is sufficient for these properties to hold jointly that they hold individually for the dominant and fringe aircraft, define the social surplus and marginal social benefit as: \( d^s^* \)

\[
SS^s[s] = \int_p^\infty f^s^*[p]dp + f^s^*[s]\tau_f^s^*[s] + \int_p^\infty d^s^*[p]dp + d^s^*[s]\tau_d^s^*[s],
\]

\[
MB = f^s^*[s]\tau_f^s^*[s]\left(\frac{\partial f^s^*[s]}{\partial s} + \frac{\partial \tau_f^s^*[s]}{\partial s}\right) + d^s^*[s]\tau_d^s^*[s]\left(\frac{\partial d^s^*[s]}{\partial s} + \frac{\partial \tau_d^s^*[s]}{\partial s}\right)
\]

\[
= \frac{f^s^*[s]\tau_f^s^*[s]}{s} + \frac{d^s^*[s]\tau_d^s^*[s]}{s} = \frac{\tau C}{s}
\]

For the self-financing result, consider the program:

\[
\max_{p_f^s, s_f^s, p_d^s, s_d^s} \int_p^\infty f^s^*[p]dp + \int_p^\infty d^s^*[p]dp + (p_f^s - ATC_f^s[f^s^*[p_f^s], d^s^*[p_d^s]s])f^s^*[p_f^s] +
\]

\[
(p_d^s - ATC_d^s[f^s^*[p_f^s], d^s^*[p_d^s]s])d^s^*[p_d^s] - K[s]
\]

which has first order conditions that simplify to:

\[
\tau_f^s^*[f^s^*, s] = p_f^s - ATC_f^s[f^s^*[p_f^s], d^s^*[p_d^s]s] = f^s^*[p_f^s]\frac{\partial ATC_f^s}{\partial f^s^*} + d^s^*[p_d^s]\frac{\partial ATC_d^s}{\partial f^s^*},
\]

\[
\tau_d^s^*[m_f^s, s] = p_d^s - ATC_d^s[f^s^*[p_f^s], d^s^*[p_d^s]s] = f^s^*[p_f^s]\frac{\partial ATC_f^s}{\partial d^s^*} + d^s^*[p_d^s]\frac{\partial ATC_d^s}{\partial d^s^*}, \text{ and}
\]

\[
\frac{\partial K}{\partial s} = f^s^*[p_f^s]\frac{\partial ATC_f^s}{\partial s} + d^s^*[p_d^s]\frac{\partial ATC_d^s}{\partial s}.
\]

As before, the rules for optimal tolls and capacity are to set the average toll equal to the marginal capacity cost and choose capacity to equate marginal construction costs and the marginal benefit of additional capacity. Since the full price of operations is homogeneous of degree zero in \( m_f, m_d, \) and \( s \), in the tolling cases with dominant and fringe airlines, it follows that the optimal tolls will pay for the optimal capacity if construction costs are homogeneous of degree one.

Conclusion 3: Step-tolled airports subject to constant returns to scale with dominant and fringe traffic are self-financing.
5. Policy discussion and conclusions

The preceding sections extend the deterministic bottleneck model by developing the multiple-step pricing rules for an atomistic fringe and a dominant airline that controls a significant share of the traffic. The bottleneck model is more appropriate than the standard model for application to airport traffic because it has a dynamic model of congestion and it models the aircraft choice of when to operate. Unlike the standard model, delays depend on the entire prior pattern of traffic during the busy period and current traffic affects delays in subsequent periods. The bottleneck model captures the effects of aircraft schedules on delays and the effects of delays and tolling on scheduling. Unlike the standard model in which reducing traffic volume is the only means of lowering congestion, the bottleneck model includes internalization of delay by rescheduling aircraft.

The model demonstrates that dominant airlines may or may not internalize their self-imposed delays depending on fringe airline demand elasticity and shows the traffic patterns associated with either behavior. Dominant firms are less likely to internalize when they face a fringe with more elastic demand. Generally speaking, there is no queue during periods when a dominant firm internalizes, and there is a queue in periods when the dominant firm is behaving atomistically. Periods with long queues and mixed dominant and fringe traffic are indicative of atomistic behavior by dominant aircraft. In the standard model, the dominant airline should pay a fraction of the atomistic tolls equal to the fringe’s share of traffic. In the bottleneck model with dominant and fringe airlines, fully internalizing dominant aircraft should optimally pay a fee equal to the marginal cost they impose on each other. This toll does not cause double internalization and has no effect on the dominant airlines’ behavior because it only brings the full price for infra marginal aircraft up to their willingness to pay, which then equals their marginal social cost. Tolling these internalizing dominant aircraft transfers producer surplus to the airport that would otherwise go to the dominant airline, making the surplus available to fund the costs of capacity. Dominant aircraft that behave atomistically should face the same step tolls as the fringe, but pay a different uniform toll that accounts for differences in marginal social cost (if any).

The results of the model have a number of important policy implications for real-world implementation of airport congestion pricing. The problems of calculating efficient tolls are overstated. Optimal continuous (fine) tolling involves valuing the queuing delay experienced in the current untolled equilibrium and imposing a toll structure with the same pattern over time as the untolled queuing costs. While such tolls would be exactly correct only in a deterministic world, they would be a reasonable approximation of the correct continuous tolling structure in a stochastic world. Daniel’s stochastic bottleneck model calculates equilibrium tolls that are similar to those in a deterministic model. Another rule-of-thumb is to calculate the delay cost of the most delayed aircraft in each peak. This is also equal to the marginal social cost of aircraft operations in a deterministic model. The schedule delay costs of the first and last aircraft are also equal to the marginal social cost. The rule-of-thumb for the optimal multiple-step tolling structure has constant toll increments as the tolls increase or decrease over the peak. All of the increasing steps have the same duration and all of the decreasing steps have the same duration, while the central peak is somewhat longer. The step durations are determined by the toll increment, so the entire toll structure can be determined from an estimate of the cost of the most delayed aircraft and the given number of steps.

The problem of differential tolls between the dominant and fringe aircraft has also been overstated. Dominant and fringe aircraft should face the same incremental tolls to optimize their scheduling during
the atomistic peak. Different flat tolls may be necessary to obtain the appropriate numbers of aircraft of each type. Other policy instruments are available to optimize the number of aircraft. The internalizing dominant aircraft may be exempted from tolling with no effect on the optimality of aircraft scheduling. The marginal internalizing aircraft faces the correct full price, because it is the furthest from its preferred operating time and it imposes no delay other than what it experiences. Other internalizing aircraft (infra-marginal) have full prices below their willingness to pay, so they retain consumer surplus in the no-toll equilibrium. The airport authority may obtain efficient scheduling of operations by tolling only the atomistic peak with an undifferentiated toll structure. The traffic levels, however, would not necessarily be optimized.

Mohring and Harwitz’s (1962) and Strotz’s (1965) self-financing of optimal capacity results largely survive the model’s extension to dominant and fringe aircraft, provided that all aircraft pay the marginal social cost of their operations. These results support imposition of common multi-step or continuous tolling structures on dominant and fringe airlines, possibly with differentiated flat tolls per operation. Differentiated tolls are more important when dominant airlines internalize, when toll structures have fewer steps, and when demand for operations is more elastic. Moderately graduated toll schedules could recapture most of the efficiency loss from congestion and pay for optimally sized airports.
REFERENCES


Table 1--Percent of Minimum Cost Achieved by Number of Toll Periods

<table>
<thead>
<tr>
<th>Early\Late</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>10</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>200</td>
<td>184</td>
<td>178</td>
<td>175</td>
<td>174</td>
<td>172</td>
<td>170</td>
<td>167</td>
<td>167</td>
</tr>
<tr>
<td>1</td>
<td>167</td>
<td>150</td>
<td>145</td>
<td>142</td>
<td>140</td>
<td>139</td>
<td>137</td>
<td>134</td>
<td>134</td>
</tr>
<tr>
<td>2</td>
<td>156</td>
<td>139</td>
<td>134</td>
<td>131</td>
<td>129</td>
<td>128</td>
<td>125</td>
<td>123</td>
<td>123</td>
</tr>
<tr>
<td>3</td>
<td>150</td>
<td>134</td>
<td>128</td>
<td>125</td>
<td>124</td>
<td>122</td>
<td>120</td>
<td>117</td>
<td>117</td>
</tr>
<tr>
<td>4</td>
<td>147</td>
<td>130</td>
<td>125</td>
<td>122</td>
<td>120</td>
<td>119</td>
<td>117</td>
<td>114</td>
<td>114</td>
</tr>
<tr>
<td>5</td>
<td>144</td>
<td>128</td>
<td>123</td>
<td>120</td>
<td>118</td>
<td>117</td>
<td>114</td>
<td>112</td>
<td>111</td>
</tr>
<tr>
<td>10</td>
<td>139</td>
<td>123</td>
<td>118</td>
<td>115</td>
<td>113</td>
<td>112</td>
<td>109</td>
<td>107</td>
<td>106</td>
</tr>
<tr>
<td>50</td>
<td>135</td>
<td>118</td>
<td>113</td>
<td>110</td>
<td>108</td>
<td>107</td>
<td>104</td>
<td>102</td>
<td>102</td>
</tr>
<tr>
<td>100</td>
<td>134</td>
<td>118</td>
<td>112</td>
<td>109</td>
<td>108</td>
<td>106</td>
<td>104</td>
<td>101</td>
<td>101</td>
</tr>
</tbody>
</table>

*Note: Assumes 100 aircraft with $\tau = $*
Figure 1—Geometric derivation of step toll equilibrium
Figure 3—Simultaneous Determination of Supply and Demand for Dominant and Fringe Operations in Tolled and Untolled Equilibria