Airlines' strategic interactions and airport pricing in a dynamic bottleneck model of congestion

Hugo E. Silva*, Erik T. Verhoef, Vincent A.C. van den Berg

Department of Spatial Economics, VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands.
Tinbergen Institute, Gustav Mahlerplein 117, 1082 MS, Amsterdam, The Netherlands.

This version: 08/08/2013. This is a post-print of the paper to be published in the Journal of Urban Economics (http://www.journals.elsevier.com/journal-of-urban-economics/)

Abstract

This paper analyzes efficient pricing at a congested airport dominated by a single firm. Unlike much of the previous literature, we combine a dynamic bottleneck model of congestion and a vertical structure model that explicitly considers the role of airlines and passengers. We show that a Stackelberg leader interacting with a competitive fringe partially internalizes congestion, and that there are various toll regimes that induce the welfare maximizing outcome, widening the set of choices for regulators. In particular, charging the congestion toll that would apply for fully competitive carriers and that ignores any internalization, to both the leader and the fringe, yields the first-best outcome.

Keywords: Airport pricing, Congestion, Bottleneck model

1. Introduction

As congestion at major airports worldwide continues to increase and traffic approaches existing capacities, implementing policies aimed at reducing delays effectively is becoming essential. For example, in the first half of 2007, 30 percent of commercial flights in U.S. arrived more than 15 minutes late, and similar figures hold for European airports (Rupp, 2009; Santos and Robin, 2010). Policies to solve the congestion problem have been extensively discussed during the last decades. One alternative is capacity enlargements, but these have the drawback of bringing benefits only after a long period of time, and at a relatively high cost (see Jorge and de Rus (2004) for a cost-benefit analysis). Another option is congestion pricing, perhaps the most discussed policy in the academic economics literature, often heavily inspired by the road pricing literature.[1] However, governments, regulators and airports have not followed this path. The current practice at many airports is to levy weight-based landing fees, a rule that has been criticized since early contributions by Levine (1969) and Carlin and Park (1970), who were the first to argue that these charges provide wrong incentives and lead to inefficiencies. Despite of four decades of theoretical and empirical contributions calling for implementation

*Corresponding author
Email addresses: h.silvamontalva@vu.nl (Hugo E. Silva), e.t.verhoef@vu.nl (Erik T. Verhoef), vberg@feweb.vu.nl (Vincent A.C. van den Berg)

[1] Quantity-based approaches to congestion management are also being discussed as an alternative. See Brueckner (2009), Basso and Zhang (2010) and Verhoef (2010) for analyses on slot sales and slot trading.

of efficient landing and takeoff charges based on economic principles, airport pricing schemes have been kept remarkably unchanged. But, as delays are reaching critical levels and other negative externalities, such as pollution and noise, are becoming more important, congestion pricing is likely to turn into a serious option for governments and regulators.[2] This policy may be specially appealing because landing fees are already in place, and only changes are needed in the way that they are charged. Moreover, in some countries, such as the U.S., landing fees are allowed to vary by time of the day, a fundamental feature of an efficient congestion pricing scheme.

It is now widely agreed that the vast literature on road congestion pricing may not be directly applicable to airports, because airlines are non-atomistic players, in contrast to road drivers. Carriers have market power and have non-negligible shares of the overall traffic and, as a consequence, they can be expected to internalize the congestion imposed on themselves. Daniel (1995) was the first to recognize this, and Brueckner (2002) and Pels and Verhoef (2004) analyzed the problem assessing the internalization of congestion with theoretical models. Subsequent works by Brueckner (2005), Zhang and Zhang (2006), and Basso and Zhang (2007) extend the analysis. The main conclusion regarding congestion pricing, based on static models of congestion, is that carriers competing in a Cournot-Nash fashion internalize self-imposed congestion and, therefore, should be charged for the fraction of congestion that they impose on others. This leads to a congestion charge that depends on the rivals' market share at the congested airport, and, therefore, may be perceived as inequitable, as dominant airlines should face lower charges than small carriers.

The contribution of this paper is to provide clear-cut insights into and understanding of airlines' strategic interactions and airport congestion pricing in a model of dynamic congestion. We recognize the vertical nature of aviation markets, thus explicitly including the role of airport's tolls on airlines' behavior, and incorporating that airlines compete taking these into account, while facing the passengers' demand for trips. We use the deterministic bottleneck model of congestion developed by Vickrey (1969) and Arnott et al. (1990, 1993). This allows for an analysis that balances analytical tractability and the inclusion of behavioral decisions that we believe are essential: airlines endogenously adjust departure or arrival rates, trading off queuing delays and schedule delays, and passengers dislike queuing and schedule delays in a different manner than airlines (i.e. at different shadow prices). By combining these two modeling features, we have a structural model of dynamic congestion that allows for an analysis of the firms' inefficiency in terms of the number of flights as well as the scheduling, and, as a consequence, allows for a derivation of the optimal policy that deals with both. We focus on sequential competition between a Stackelberg leader and a competitive fringe. The model set-up is consistent with the empirical findings of Daniel and Harback (2008), who show that observed traffic patterns at most of the major U.S. airports are consistent with the dynamic bottleneck model of congestion, and that most of the U.S. hub airports seem best described by competition between a Stackelberg leader and a competitive fringe.

Our main result is that, while the (untolled) equilibrium is fully consistent with what previous literature

[2] Congestion pricing can be a second-best solution for environmental externalities. See, for example, Carlsson (2003) for an analysis of airport pricing with congestion and emissions, and Brueckner and Girvin (2008) for an investigation of airport noise regulation.

with static congestion suggests, first-best congestion pricing is not. In particular, when a Stackelberg leader faces a competitive fringe, the equilibrium is fully consistent with static models in that the fringe does not internalize any congestion, and in that the leader's ability to exert market power and to internalize self-imposed congestion depends critically on the assumed substitution pattern (just as in Brueckner and Van Dender (2008)). On the other hand, we find that the first-best optimum can be decentralized with a pricing policy that consists of a market power subsidy for the leader, that is indeed a function of the assumed substitution pattern, and a congestion toll for both agents that is independent of whether internalization occurs in the untolled setting. We show that charging the congestion toll that is derived for the fully atomistic carriers to both leader and fringe always yields the first-best outcome. This is because the subsidy deals with the leader's overpricing due to market power, and the time-varying congestion toll eliminates queuing and provides the right incentives to take into account the delays imposed on the rival airlines. We further show that there are various alternative toll regimes that also attain the first-best, dealing with the congestion inefficiency in yet different ways, while still correcting for the market power exertion. Again, the congestion component of all toll regimes is independent of the degree of internalization by the leader in the unregulated equilibrium.

The results of this paper suggest that optimal congestion pricing may have a more significant role on airports than what has been suggested in the literature before. The congestion pricing scheme that is obtained for fully atomistic carriers induces the first-best outcome, and results in a revenue for the airport that restores the well known self-financing result for congested facilities: the ratio between first-best capacity investment costs and total revenue from congestion pricing equals the degree of economies of scale in capacity provision (Mohring and Harwitz, 1962).[3] In addition, our results suggest that the political feasibility of optimal congestion pricing would be enhanced, as the (first-best) atomistic congestion charges do not vary across airlines and therefore are less likely to be perceived as inequitable. Finally, the fact that there are several tolling regimes that yield the social welfare maximizing outcome widens the set of choices for regulators.

Our analysis contributes to the policy analysis on congested airports and extends previous literature that considers dynamic congestion at airports. Works such as Daniel (1995, 2001) and Daniel and Harback (2008, 2009) focus on cost minimization of scheduling flights, hence ignoring the passengers' role in the problem, or at least treating that role only implicitly. Moreover, most of these papers aim at testing whether the observed patterns of arrivals and departures of flights support the internalization hypothesis. Daniel (2009) analytically studies the conditions under which dominant airlines internalize self-imposed congestion with a deterministic bottleneck model, focusing on Stackelberg-fringe competition, but omits the passengers in the model, hence ignoring the fact that airlines use the airport as an input to sell an output in a downstream market. By combining the bottleneck congestion model with the explicit consideration of two groups of agents (airlines and passengers) in a theoretical model, we are able to study key elements

[3] We also show how the market-power exertion has to be corrected, finding insights that are consistent with those in the previous literature, and that this overturns the self-financing result if market-specific subsidies are drawn from the airport budget.

that were not present in previous exercises with dynamic congestion. These include an analysis on how airlines set the ticket price according to the time of departure, a derivation of an explicit relation between the internalization of congestion and the assumed passengers' demand substitution pattern between airlines, and a clear comparison between the results derived in models of static congestion and the results obtained with dynamic congestion. We are also able to study the implications, for the optimal pricing policy, of the strategic interaction between the leader and the fringe, finding that there is a set of various pricing schemes that maximize social welfare, as opposed to a single optimal congestion toll.[4] Finally, our analysis complements the findings of Brueckner and Van Dender (2008) and Silva and Verhoef (2013) who show that congestion charges can be optimally close to the atomistic charges depending on the assumptions on the prevailing market structure.

Our results have to be qualified according to our assumptions. Naturally, the dynamic bottleneck model is not directly applicable when queuing is not necessary or helpful for airlines in order to obtain a certain arrival time, as in fully slot-constrained airports. This is because the airport's regulator directly controls the timing through slot allocations. For this case, more common in European airports, an analysis of slot sales and slot trading is more pertinent (see Brueckner 2009). We also assume that airlines and passengers share a most desired time of arrival or departure, and that airlines are homogeneous in values of time. The model can be straightforwardly extended in these directions following the road pricing literature.[5] Lastly, we use the deterministic version of the bottleneck model for analytical simplicity. A stochastic version that does not require attempted inflows at or above capacity to yield queues would be more realistic. However, as the trade off between expected queuing and expected schedule delays will be driving airlines' interactions, general results may not change significantly, while detailed results such as equilibrium delays, traffic rates and queue lengths will change.

The paper is organized as follows. Section 2 introduces the model and the assumptions that are necessary for the analysis. We illustrate the main features of the model by characterizing the untolled equilibrium and deriving first-best and time-invariant second-best tolls for perfectly competitive airlines. We then study a monopoly carrier in the market. Section 3 extends the analysis to competition where a Stackelberg leader faces a group of competitive carriers, focusing on the untolled equilibrium and on first-best tolling. We study the case of imperfectly elastic demand and imperfectly substitutable airlines, and also look at the special cases of perfect substitution, independent markets and perfectly elastic demand. Finally, Section 4 concludes.

[4] Daniel (2009) recognizes that the dynamic atomistic toll charged to all airlines induces the welfare maximizing output in his scheduling model, but he does not analyze the leader's response to the fringe behavior when facing the toll, and therefore does not find alternative schemes. He also omits the passengers' role in the analysis, and our behavioral model seems to match his set of assumptions only when leader and fringe serve independent markets whose demands are related only through congestion.

[5] The original model by Vickrey (1969) analyses heterogeneity in desired arrival time. For heterogeneity in values of time see e.g. Vickrey (1973), Arnott et al. (1994) and Van den Berg and Verhoef (2011).

2. The model

2.1. The basics and perfect competition

This section describes our model of dynamic airport congestion and shows how it incorporates the main features of airlines and passengers behavior, by looking at the case of perfect competition. Subsequent sections look at the extension to different market structures. We base our analysis in the work of Vickrey (1969) and Arnott et al. (1990, 1993), extending their dynamic congestion modeling for atomistic users (road users), to the case where congestion occurs in a facility used by carriers with market power (airlines), who sell their output (trips) in a downstream market of passengers. We focus our analysis on arrivals, thus the bottleneck is the airport's runway and queuing takes place in the air, before landing. The analytical results would apply for departures as well, and can in principle be extended to a network setting with multiple airports and delays in both arrivals and departures.[6]

This model considers "pure" bottleneck congestion behind a bottleneck of finite capacity, implying that in absence of a queue and, as long as the arrival rate of flights at the bottleneck is below its capacity, there are no travel delays. Under other conditions, the queuing delay experienced by a flight and its passengers depends on the length of the queue at the moment of joining it. We assume that free-flow travel time is zero,[7] so that a flight that departs from the origin at t_d, arrives at the bottleneck at the same time. As a consequence, in absence of queuing, the time of arrival at the destination (landing), t, also matches the time of departure from the origin. When there is queuing, the arrival time, t, is the time of departure (t_d) plus the queuing delay $T(t)$, and the length of the queue, $Q(t)$, grows or shrinks at a rate $\dot{Q} = r_d - K$, where r_d is the aggregate airlines departure rate and K the bottleneck capacity. The travel delay of a flight arriving at destination at t, is the length of the queue at the moment of joining it, divided by the bottleneck's capacity:

$$T(t) = \frac{Q(t_d)}{K} \quad with \quad t_d = t - T(t) \tag{1}$$

Note that this definition, due to the perfect information assumption, allows us to write time costs as a function of the arrival time at the destination, instead of the departure time from the origin.

We follow Small's (1982) model of scheduling behavior for both passengers and airlines, so that their time costs are the sum of travel delay cost and schedule delay cost. Passengers, in a nutshell, face travel delays in the form of queuing delays to land, and have a preferred arrival time t^* from which any deviation (early or late) induces a schedule delay cost. The passengers' schedule delay cost, that arises from the difference between desired and actual arrival (or departure) time, was introduced in the context of aviation by Douglas and Miller (1974) and estimated by Morrison and Winston (1989) as part of a passengers' discrete choice model of airline.[8] A natural interpretation for this cost is that people, everything else constant, want to

[6] Note that our bottleneck model is relevant when the airport's operational conditions for arrivals (or departures) follow the first-in first-out (FIFO) discipline, and is not directly applicable when the airport is managed with slots, because the airport's regulator directly controls the timing through slot allocations.

[7] In a single origin-destination pair, we can assume zero free-flow travel time without loss of generality, but this is generally different with multiple origin-destination pairs.

[8] Using a reduced-form for the schedule delay cost in models of static congestion is common in the aviation literature (e.g. Oum et al., 1995; Brueckner, 2004).

arrive at their destination at a certain moment, that can be, for instance, the start of the working day in order to make the most out of it.[9] The schedule delay costs for airlines is a less studied matter. However, the scheduling of crew and coordination of arrivals and departures (specially in hub-and-spoke networks), are possible interpretations for including early and late schedule delays costs for airlines. In addition, as we show in Appendix A, our analysis and results hold in absence of airlines' schedule delay costs. Phrased differently, an airline's own schedule delay costs enters its maximization problem in the same way as its passengers' schedule delay costs do, as the latter imply a decreased willingness to pay a ticket fare.

Let g be a sub-index that denotes agent-type (p for passengers and a for airlines), α_g the value of travel time for agent type g, β_g the value of early schedule delay, and γ_g the value of late schedule delay. Then, the time cost of arriving at t, for an agent type g, $C_g(t)$, can be written as:

$$C_g(t) = \alpha_g \cdot T(t) + \begin{cases} \beta_g \cdot (t^* - t) & \text{if } t \leq t^* \\ \gamma_g \cdot (t - t^*) & \text{if } t \geq t^* \end{cases} \tag{2}$$

The airline's delay cost differs from user's time cost only in the values of time, which reflects our assumption that airlines share the desired arrival time t^* with the passengers.[10]

Having described the congestion modeling, we can turn to the passengers' demand specification, and the airlines' costs and profit. We assume, for the perfectly competitive case, that passengers perceive airlines as perfect substitutes, and that the demand for an airline follows a linear inverse demand function:

$$D\left(\sum_i q_i\right) = A - B \cdot \sum_i q_i \tag{3}$$

which gives the marginal willingness to pay for traveling; q_i is the number of passengers traveling with airline i; A represents the maximum reservation price, and B is the demand sensitivity parameter. We use the linear specification for analytical simplicity, but our results do not depend crucially on this.

The full price p_i for a passenger traveling with airline i is the sum of the fare (ρ_i) and the generalized cost experienced by the passenger. As we consider dynamic congestion, the various components of the generalized cost are generally not constant over time (see Eq. (2)). The condition for an equilibrium, where all flights are used by passengers and where passengers are indifferent between all the flights, is given by:

$$\rho_i(t) + C_p(t) = A - B \cdot \sum_i q_i \tag{4}$$

[9] For example, this directly applies to business travel. It can be argued that for leisure passengers this also hold as well, as, everything else constant, they prefer to arrive at a certain time during the day. Another possible interpretation is related to transfers at hub airports. Although our model does not consider network effects, one can think of passengers using the flight for a transfer and, in that case, t^* would represent their most preferred moment to arrive at the hub airport (the time that makes the transfer possible without experiencing undesired waiting). Clearly, in this case, a late arrival would be significantly more costly than an early arrival because it may imply loosing the connection.

[10] Although the preferred arrival time for airlines may be endogenous, following from desired arrival times for passengers, the analysis of this issue is beyond the scope of this paper. With endogenous t^*, it can be expected that the airlines' preference is significantly affected by the passengers' preferred arrival time and will be close in practice. For example, in hub-and-spoke networks, airlines coordinate arrivals and departures to facilitate passenger connections. Cost advantages because of high passenger density may also drive airlines to adopt the passengers' preferred arrival time.

which is simply the full price of taking any airline i's flight, that arrives at destination at time t, equals marginal willingness to pay. Recall that the generalized cost experienced by the passenger does not depend on the identity of the airline, but only on the time of arrival. The equilibrium condition in Eq. (4) implies that airlines charge different fares for flights scheduled at different times, except for flights whose users experience the same generalized cost. Forbes (2008) provides empirical evidence that airlines indeed charge lower fares when they face higher delays.

As usual in the airport pricing literature, we assume that the product of the load factor and the seat capacity is constant, so that the number of passengers per flight is given. Airlines' cost consists of a time-invariant operating cost per flight c_1, a time-invariant operating cost per passenger c_2, and the time-variant cost $C_a(t)$ described in Eq. (2). Denoting the constant product between seat capacity and load factor as s, time-invariant costs can be expressed as a constant cost per flight $c = c_1 + s \cdot c_2$.[11] With the cost structure defined, we can analyze the equilibrium in the airline market and then study the regulator's problem. This section looks at the perfectly competitive case, to illustrate the main features of the model.

In the case of imperfect competition, airlines would have as decision variables the number of flights (or prices),[12] and the departure time of each flight. In order to analyze the perfectly competitive case, we assume that there is a continuum of small competitive airlines that can enter the market by scheduling a single flight at any time. Therefore, each competitive airline's decision variable is the time of arrival t, and the aggregate number of flights will be given by the zero-profit condition.[13] The profit of an airline, that schedules its only flight to arrive at t, is revenues minus costs:

$$\pi(t) = s \cdot \rho_i(t) - C_a(t) - c - \tau(t) \tag{5}$$

where $\tau(t)$ is the time-variant per-flight toll (in this case, landing fee) that the regulator might charge to airlines. Denoting f as the aggregate number of flights, the total number of passengers is $s \cdot f$, and using the interior equilibrium condition in Eq. (4), airline's profit is:

$$\pi(t) = s \cdot [A - B \cdot sf - C_p(t)] - C_a(t) - c - \tau(t) \tag{6}$$

where the term between square brackets is the fare. Using Eq. (2) and defining $\overline{\alpha} = s \cdot \alpha_p + \alpha_a$, $\overline{\beta} = s \cdot \beta_p + \beta_a$ and $\overline{\gamma} = s \cdot \gamma_p + \gamma_a$, the profit of an airline whose flight arrives at time t can be simplified as:

$$\pi(t) = s\,[A - B \cdot sf] - c - \tau(t) - \overline{\alpha} \cdot T(t) - \begin{cases} \overline{\beta} \cdot (t^* - t) & \text{if } t \leq t^* \\ \overline{\gamma} \cdot (t - t^*) & \text{if } t \geq t^* \end{cases} \tag{7}$$

[11] Because of the fixed-proportions assumption, constant costs per passenger and per flight have the same effect, and can be aggregated. The same occurs when airlines are charged a landing fee; it does not matter if it is a per-passenger fee or a per-flight fee.

[12] Choosing the number of flights is equivalent to setting the number of passengers (quantity) because the fixed-proportion assumption implies $q_i = f_i \cdot s$.

[13] An alternative interpretation of the perfectly competitive case is that airlines are not necessarily small, but they view price and congestion level as parametric. This would imply that their fare and time of departure is given by the zero-profit condition and they choose volumes.

This reduced form shows that airlines take into account the generalized cost of its own passengers, because the lower the passengers' generalized cost is, the higher the fare can be (see Eq. (4)), on a dollar-by-dollar basis. Therefore, we can interpret the airline's problem as if they face a *generalized cost per flight*, that is the sum of its own delay costs, $C_a(t)$, and the generalized cost of all the passengers on its flight, $s \cdot C_p(t)$.

The dynamic equilibrium is such that an airline cannot improve its profit by changing the schedule of its single flight, for a given scheduling behavior of the other airlines. By looking at Eq. (7), this can only be achieved when every airline, i.e. every flight, faces the same sum of toll and generalized cost per flight (the travel and schedule delay cost terms on the right-hand side of Eq. (7)), because all other terms are time-invariant. This generalized cost per flight from the airlines' perspective, is similar to the generalized costs typically found in the bottleneck road pricing literature for individual drivers (e.g., Arnott et al., 1990, 1993). A difference is that the values of time considered by the airline, for a single flight, are its own values of time plus the summed passengers' values of time in that flight. But, through the use of the composite shadow prices $\overline{\alpha}$, $\overline{\beta}$, and $\overline{\gamma}$, this difference disappears from the formal model. This enables us to describe the equilibrium in schedules following the road pricing literature, and keep the discussion concise.

We first characterize the untolled equilibrium. The equilibrium condition then is that the generalized cost per flight $(C_a(t) + s \cdot C_p(t))$ must be constant over time during the period of operation; otherwise an airline would have an incentive to reschedule its flight and increase its profit. As shown in the road pricing literature, there is a unique aggregate queuing pattern that satisfies this equilibrium property, and this pattern defines the (equilibrium) scheduling behavior of the competitive airlines (see Appendix A for the calculations and derivation of this result). Denote t_s as the (endogenous) first moment of operation, i.e. the time where the first flight arrives at destination, and t_e as the (endogenous) end of the operation period. The first airline's flight departs at t_s and arrives at the same time, as there is no queue, incurring only early schedule delay cost. The same holds for the last flight, at t_e, incurring only late schedule delay costs (if the last flight incurred queuing, its costs could be reduced by departing later and still arriving at the same moment). Arrivals are continuous in this model, and as a consequence, the duration of the peak period has to be f/K, the total number of flights divided by the capacity of the bottleneck. From t_s onward, the queue evolves, growing up to a maximum level (just when a flight arrives at t^*) and then decreasing until it dissipates completely at t_e, in the unique way that makes the generalized cost per flight constant over time. The resulting constant generalized cost per flight can be found by determining the equilibrium timing of the peak of duration f/K, such that the schedule delay costs are the same for the first and last flight. This gives two conditions $(\overline{\beta} \cdot (t^* - t_s) = \overline{\gamma} \cdot (t_e - t^*) \wedge te - ts = f/k)$ that are sufficient to determine the equilibrium generalized cost:

$$C_a(t) + s \cdot C_p(t) = \frac{\overline{\delta} \cdot f}{K} \quad \forall \quad t \in [t_s, t_e] \tag{8}$$

where $\overline{\delta} = (\overline{\beta} \cdot \overline{\gamma})/(\overline{\beta} + \overline{\gamma})$.[14] The equilibrium departure rates can be derived from equating the time derivative of Eq. (7) to zero. Note that the aggregate scheduling pattern is unique, but an individual airline's scheduling is undefined due to the perfect competitive assumption. This yields an equilibrium profit (superscript e) for

[14] See Appendix A for a derivation of this result, and Arnott et al. (1990, 1993) for a detailed discussion.

any airline of:

$$\pi^e = s\left[A - B \cdot sf\right] - c - \frac{\overline{\delta} \cdot f}{K} \tag{9}$$

Recall that airlines are indifferent between any arrival time t between t_s and t_e, and passengers are indifferent between any flight, because the full price of all flights is constant, equal to $A - B \cdot sf$, and given by:

$$p_i = \rho_i(t) + C_p(t) = A - B \cdot sf = \frac{1}{s} \cdot \left(c + \frac{\overline{\delta} \cdot f}{K}\right) \tag{10}$$

where the last equality comes from the zero-profit condition of the perfectly competitive case ($\pi^e = 0$). The passengers' full price in the no-toll equilibrium equals the airlines' constant operating cost per passenger (c/s) plus the generalized cost per flight divided by the number of passengers. The total generalized costs (or travel delay plus schedule delay costs) are the generalized costs per flight times the number of flights, $\overline{\delta} \cdot f^2/K$, as in the road case. In Appendix A we extend the analysis by looking how the equilibrium fare varies over time.

Figure 1 illustrates the no-toll equilibrium for the competitive case. The equilibrium is represented by the constant generalized costs per flight (from Eq. (8)), and the depiction of $s[A - B \cdot sf]$ satisfying Eq. (10). The only conditions on the values of time that are needed for this equilibrium to exist are that $\overline{\alpha} > \overline{\beta} > 0$ and $\overline{\gamma} > 0$. As these values of time are made up of a combination of the passengers' and the airline's values of time, the interpretation for the condition is not immediately straightforward. In the case of passengers' values of time, empirical evidence indicates that the conditions are satisfied, i.e. that the value of travel time is higher than the value of schedule delay early ($\alpha_p > \beta_p$), and that the value of schedule delay late is above zero ($\gamma_p > 0$) (see Morrison and Winston, 1989; Lijesen, 2006).[15] In the case of airlines, to the best of our knowledge, there is no empirical evidence for the values of schedule delay. However, given that the passengers relation is intuitive and has empirical support, the only additional assumptions that we need on the airlines' values of time are that $\alpha_a \geq \beta_a$ and that $\beta_a \geq 0 \wedge \gamma_a \geq 0$. The requirement on the relation between value of travel time (α_a) and early schedule delay (β_a) is consistent with the plausible assumption that, when a flight is set to arrive early, the airline prefers landing over extending the trip by making a detour; the other requirement only states that values of schedule delay are not negative.

With the untolled equilibrium characterized, we analyze the regulator's problem of maximizing social welfare through a per-flight toll. First, consider the case of a time-invariant toll. As that toll does not vary over time, the airlines treat it as a constant operating cost and, for a given number of flights, it does not alter the scheduling decisions: the toll can only affect the number of flights. The regulator's optimization problem follows:

$$\max SW = \int_0^{sf} (A - Bx)dx - \int_{t_s}^{t_e} (K \cdot s \cdot C_p(t))dt - \int_{t_s}^{t_e} (K \cdot c + K \cdot C_a(t))dt \tag{11}$$

where the first term is gross benefits for sf travelers, the second is total passengers' generalized costs (at t, a flow of K flights will serve s passengers each), and the third term is total airlines' costs that includes

[15] In fact, Lijesen (2006) finds evidence that $\gamma_p > \beta_p$, something that is usually found for road users.

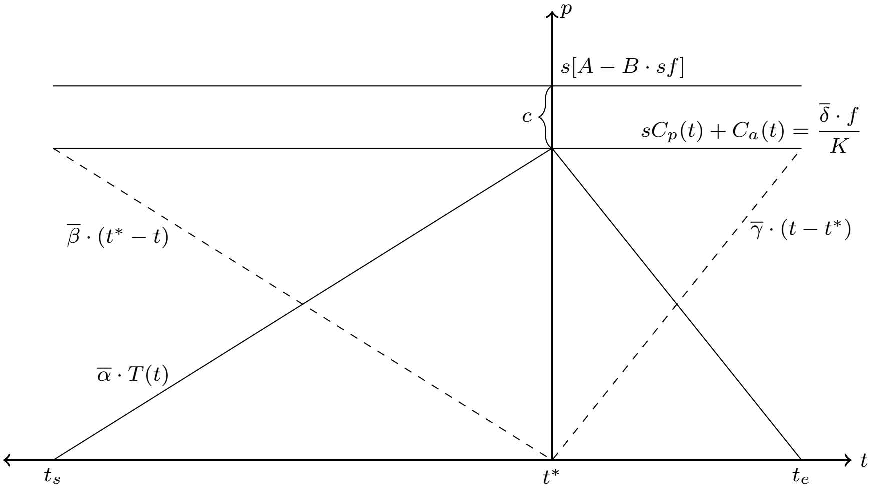
p
$s[A - B \cdot sf]$
c
$sC_p(t) + C_a(t) = \frac{\overline{\delta} \cdot f}{K}$
$\overline{\beta} \cdot (t^* - t)$
$\overline{\gamma} \cdot (t - t^*)$
$\overline{\alpha} \cdot T(t)$
t_s
t^*
t_e
t

Figure 1: Competitive no-toll equilibrium.

constant and generalized costs (fares and tolls cancel out). Rewriting,

$$\begin{aligned} SW &= \int_0^{sf} (A - Bx)dx - K \cdot \int_{t_s}^{t_e} (s \cdot C_p(t) + C_a(t))dt - K \cdot c \int_{t_s}^{t_e} dt \\ &= \int_0^{sf} (A - Bx)dx - \frac{\overline{\delta} \cdot f^2}{K} - f \cdot c \end{aligned} \tag{12}$$

where the second equality uses that the duration of the peak is f/K, and that, in equilibrium, $s \cdot C_p(t) + C_a(t)$ is constant (condition in Eq. (8)).

Let $\widehat{\tau}$ be the time-invariant toll. Comparing the first-order conditions for welfare maximization and the airline zero-profit condition, we then obtain:

$$\frac{\partial SW}{\partial f} - \pi^e = s(A - B \cdot sf) - 2\frac{\overline{\delta} \cdot f}{K} - c - \left[s(A - B \cdot sf) - c - \frac{\overline{\delta} \cdot f}{K} - \widehat{\tau} \right] \tag{13}$$

As a consequence, the welfare maximizing time-invariant toll per flight is:

$$\widehat{\tau} = \frac{\overline{\delta} \cdot f}{K} \tag{14}$$

This toll matches the flat toll for the road bottleneck (Arnott et al., 1993), because without altering the flights' schedule, average (per flight) generalized costs are $\overline{\delta} \cdot f/K$, and marginal social generalized costs are therefore $2 \cdot \overline{\delta} \cdot f/K$, which is fully consistent with the road case. As a consequence, it is straightforward that the second-best flat toll is the difference between the two. The flat-toll in Eq. (14) is equal to the marginal delay cost that a flight imposes on all airlines' flights (including their passengers). This time-invariant toll induces an aggregate number of flights $f^{'}$, which is second-best optimal, given that queuing is not eliminated. The fares will keep the dynamic structure that they have in the no-toll equilibrium (see Appendix A for details), but the beginning and the end of the peak (t_s and t_e) will be different, as the total number of flights is lower and the peak period shorter.

As queuing delay is a pure loss in this model, welfare can be improved further. The reason is that, any number of flights in an equilibrium with queues can be served in the same time interval, without queuing while incurring the same schedule delay costs. This requires an arrival rate equal to capacity throughout the peak, which cannot be achieved spontaneously in equilibrium as the flights closer to t^* would face a lower generalized cost. The first-best charge is the time-variant toll $\tau(t)$ that decentralizes this queue-free configuration. The toll is equal to the value of queuing delay per flight in the no-toll equilibrium, $\overline{\alpha} \cdot T(t)$ in Figure 1, as it makes the sum of generalized cost and toll constant over time (the dynamic equilibrium condition) only when there is no queue and solely schedule delay costs are experienced. As travel delays are eliminated, the toll reduces aggregate generalized cost by 50%, so that marginal cost becomes equal to the generalized price. That is, total cost becomes $\overline{\delta} \cdot f^2/2K$, so marginal cost is $\overline{\delta} \cdot f/K$, which is equal to the price faced by the first and last flight, and therewith by all flights. Denoting f^* as the optimal aggregate number of flights, the optimal toll is,

$$\tau(t) = \frac{\overline{\delta} f^*}{K} - \begin{cases} \overline{\beta} \cdot (t^* - t) & \text{if } t \leq t^* \\ \overline{\gamma} \cdot (t - t^*) & \text{if } t \geq t^* \end{cases} \tag{15}$$

We call this toll structure the *dynamic atomistic toll*, in contrast to the second-best flat toll of this problem, in Eq. (14). With $\tau(t)$, the flight that arrives at t^* faces no generalized costs and pays a toll equal to the marginal social cost. The first and last flight face a schedule delay equal to the marginal social generalized cost and therefore do not pay any toll. In addition, the part of the toll that reflects passenger valuation of delays is transferred to them through the fare to maintain passenger equilibrium. The fare will thus show a stronger time variation than in the no-toll equilibrium.

This section extended the arguably most used model of dynamic congestion in the transport pricing literature to a case with a vertical structure, where passengers have a demand for trips, offered by atomistic, perfectly competitive carriers that make the scheduling decisions. In the rest of the paper, we relax this assumption and allow for different degrees of market power through the analysis of various market structures.

2.2. The monopoly case

Here, we consider a market with a single airline facing a linear inverse demand as in Eq. (3). The monopoly carrier has as decision variables the number of flights F and how to schedule them, i.e. the time of departure of each flight. Let t_s be the time when the carrier schedules its first flight and t_e the time when the last flight is scheduled, then the airline's profit is:

$$\begin{aligned} \pi &= \int_{t_s}^{t_e} K \cdot s \cdot \rho(t) - K \cdot c - K \cdot C_a(t) dt = K \int_{t_s}^{t_e} s[A - B \cdot sF] - sC_p(t) - c - C_a(t) dt \\ &= s \cdot F \cdot [A - B \cdot sF] - F \cdot c - K \int_{t_s}^{t_e} s \cdot C_p(t) + C_a(t) dt \end{aligned} \tag{16}$$

The second equality uses Eq. (4), and the third equality, that the peak lasts F/K. We have shown that the last term on the right hand side of Eq. (16) reflects the road case with composite values of time $\overline{\alpha}$, $\overline{\beta}$, and $\overline{\gamma}$. Since the airline faces no competition,[16] the flights will be scheduled to minimize delays. The

[16] We are abstracting from potential entry in this setting, but we address this question in Section 3.

airline realizes that by choosing a departure rate equal to the runway capacity, it will achieve the minimum possible generalized cost, only facing schedule delay costs and no travel delay cost through queuing.

This allows us to write the generalized costs per flight as schedule delay costs that diminish linearly from $\overline{\delta}F/K$ at t_s to zero at t^* and then grow to $\overline{\delta}F/K$ at t_e. Taking this into account, and considering a per-flight time-invariant toll $\widehat{\tau}$ (that is seen as parametric by the airline), the profit in Eq. (16) can be expressed as:

$$\pi = s \cdot F \cdot [A - B \cdot sF] - F \cdot c - \frac{\overline{\delta} \cdot F^2}{2K} - F \cdot \widehat{\tau} \tag{17}$$

The airline first-order condition for profit maximization is,

$$\frac{\partial \pi}{\partial F} = s[A - B \cdot sF] - B \cdot s^2 F - c - \frac{\overline{\delta} F}{K} - \widehat{\tau} = 0 \tag{18}$$

which means that the (constant) full price paid by passengers is:

$$p = \rho(t) + C_p(t) = A - B \cdot sF = \frac{1}{s} \cdot \left(c + \frac{\overline{\delta} \cdot F}{K}\right) + B \cdot sF + \frac{\widehat{\tau}}{s} \tag{19}$$

implying that the fare, that maintains equilibrium, is:

$$\rho(t) = \frac{1}{s} \cdot \left(c + \frac{\overline{\delta} \cdot F}{K}\right) + B \cdot sF + \frac{\widehat{\tau}}{s} - C_p(t) \tag{20}$$

In contrast to the competitive case, this condition shows that the monopoly carrier charges to the passengers a markup of $B \cdot sf$. This is simply the number of passengers times the own-demand price sensitivity, the traditional market power effect first described, in the aviation context, by Pels and Verhoef (2004). Note that the fare is time-dependent, as the passengers' generalized cost ($C_p(t)$) is not constant in this setting. Figure 2 depicts the time-invariant-toll equilibrium for a monopoly. There is no queue, and the first and last flight (at t_s and t_e, respectively) experience a generalized cost of $\overline{\delta}F/K$. The fulfillment of the first-order condition for profit maximization is represented in the vertical axis, where $s[A - B \cdot sF] = c + \overline{\delta}F/K + B \cdot sF + \widehat{\tau}$. The time-variant per-flight fare, $s \cdot \rho(t)$ in Eq. (20), is also depicted in Figure 2. The slopes of passengers' generalized cost ($C_p(t)$) and airline's delay costs ($C_a(t)$) are the same as in the optimum of the competitive case, and therefore the slope of the per-flight fare is also the same.

Now, the regulator's maximization problem is:

$$SW = \int_0^{sF} (A - Bx)dx - K \cdot \int_{t_s}^{t_e} (s \cdot C_p(t) + C_a(t))dt - F \cdot c \tag{21}$$

but, in contrast with the competitive case, the airline is scheduling the flights in such a way that there is no queue. Hence, the second term on the right-hand side of Eq. (21) is the same as derived in Eq. (17), shaping social welfare in the following way:

$$SW = \int_0^{sF} (A - Bx)dx - \frac{\overline{\delta} \cdot F^2}{2K} - F \cdot c \tag{22}$$

This is gross benefits of the sF passengers minus total social costs; when there are no queuing delays, total generalized costs is $\overline{\delta} \cdot F^2/2K$. Taking the derivative with respect to F, we get the first-best condition:

$$s[A - B \cdot sF] = c + \frac{\overline{\delta} \cdot F}{K} \tag{23}$$

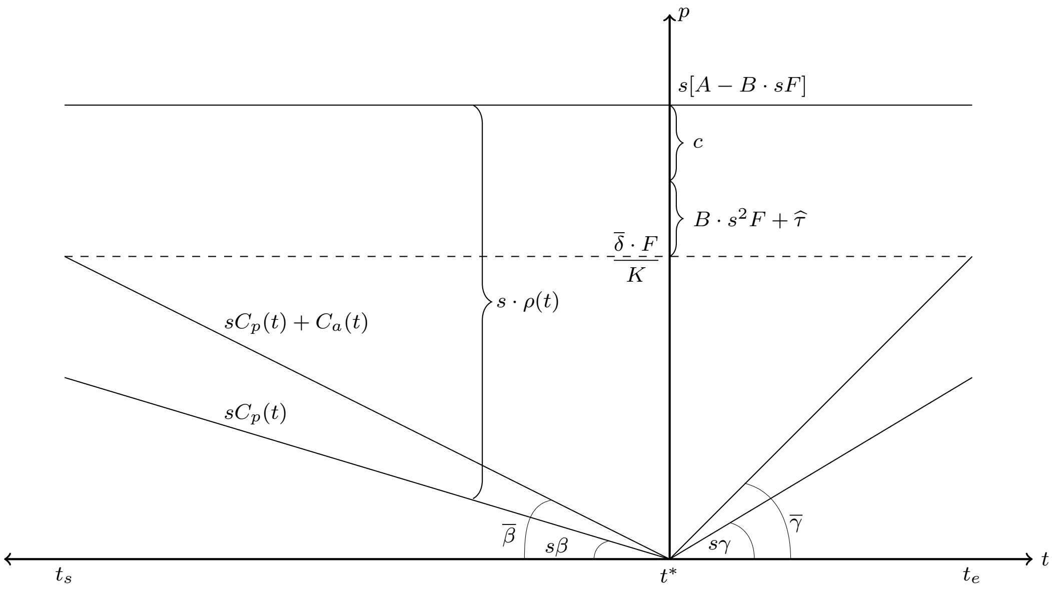

Figure 2: Monopoly time-invariant-toll equilibrium.

At the optimum, full price equals marginal social cost, which is the sum of the marginal operating cost plus marginal total generalized costs (including airlines and passengers through $\overline{\delta}$).[17] Comparing the monopolist's first-order condition in Eq. (18) and the first-order condition for welfare maximization in Eq. (23), it is straightforward that the first-best toll is:

$$\widehat{\tau} = -B \cdot s^2F \tag{24}$$

The regulator corrects the market power exertion by subsidizing the airline, and does not have to give an incentive to the monopolist to internalize congestion. This subsidy ($-B \cdot sF$ per passenger) induces the optimal number of passengers, and is analogous to the one obtained in the static model (Pels and Verhoef, 2004). A monopoly airline internalizes all the congestion costs by scheduling the flights efficiently: there is no queuing and therefore there is no need for congestion pricing. In Figure 2, when the optimal subsidy is applied, the term $B \cdot s^2F + \widehat{\tau}$ disappears and the first-best condition in Eq. (23) is satisfied. Moreover, the per-flight fare (s multiplied by the per-passenger fare) at the first and last flight is simply the airline's costs per flight, as Figure 2 shows.

Despite the fact that in the monopoly case only a subsidy that decreases price is needed, it is important to emphasize that the dynamic atomistic toll in Eq. (15) *could* also be charged to the monopoly airline without altering social welfare, but transferring part of the monopoly carrier profits to the regulator. This is the result of the congestion technology: the monopoly airline cannot do better than setting the arrival rate equal to capacity in time windows where it has arrivals, so as to face only schedule delay costs, regardless of the dynamic toll schedule it faces. To see this, it is enough to add the time-variant toll to the monopoly

[17] We look at the full price of a flight (the full price of a trip $A - B \cdot sF$ times the number of passengers in a flight s) and the marginal social cost of a flight, but there is no loss of generality. The condition also implies that the full price of a trip equals the marginal social cost of a seat.

profit in Eq. (16):

$$\pi = s \cdot F[A - B \cdot sF] - F \cdot c - K \int_{t_s}^{t_e} s \cdot C_p(t) + C_a(t) + \tau(t)\, dt \tag{25}$$

By charging the dynamic atomistic toll in Eq. (14) to the monopoly, it is straightforward that $s \cdot C_p(t) + C_a(t)$ in Eq. (25) cancels out with the time-variant part of the toll, and the airline will set full price of a flight equal to c plus the constant part of the toll per flight ($\overline{\delta} F^*/K$) and the market power mark-up. By including the subsidy, the outcome will be the first-best.

After having discussed monopoly and before moving to the leader-fringe setting, a comment on the simultaneous competition between a small number of airlines in a Cournot fashion is needed. Space is lacking to provide a detailed discussion, but we can assure that, within the framework of the deterministic dynamic bottleneck model described above and with symmetric airlines, there is no equilibrium in pure strategies for a general set of values of time.[18] It is not unlikely that, and currently under study whether, under different conditions and with additional modeling assumptions, it may be possible to find an equilibrium. However, the required changes make the model to depart far from the current one and, therefore, would place the analysis outside the scope of this paper.

3. A Stackelberg leader with a competitive fringe

We now turn to the case of competition between a Stackelberg leader and a follower that behaves competitively. This market structure was shown to be empirically relevant by Daniel (1995), and was studied further by Daniel and Harback (2008). They show that most of the U.S. airports have queuing patterns that are consistent with a stochastic bottleneck model, and exhibit evidence that the Stackelberg-fringe market structure is the one that fits best the observed queuing patterns. The purpose of this section is to assess the degree of internalization of congestion by the leader, and to derive the first-best tolls. To the best of our knowledge, there are two papers that study this type of set-up from a theoretical point of view. Brueckner and Van Dender (2008)—with a static congestion model—show that the internalization of self-imposed congestion by a Stackelberg leader facing a competitive follower can approach the atomistic levels, depending on the assumed substitution pattern, and that the first-best congestion toll can also approach the atomistic toll. On the other hand, Daniel (2009), with a dynamic bottleneck model of congestion, argues the need for atomistic tolls for both the leader and the competitive fringe with an analytical model that includes only the airlines, and therefore omits the vertical structure and passengers' role in the analysis.

The competitive follower can be interpreted as a group of competitive airlines, as in Section 2.1 with a free-entry condition. These airlines do not need to be small in general, but only to have a small share of flights at the airport under consideration, where a single airline acts as a leader. Following the aviation literature, we use the term "fringe" for this group of airlines that behaves competitively, regardless of the

[18] For the conventional case ($\overline{\gamma} > \overline{\alpha} > \overline{\beta} > 0$), some intuition can be obtained by realizing that the only candidate symmetric equilibrium that has firm-internal marginal cost equalized over time, would provide an incentive to an individual firm to deviate by moving all queuing flights to the end of the peak and schedule them at a rate equal to the spare capacity.

temporal location of its flights. We assume that both the leader and the fringe treat the tolls that the regulator sets as parametric, and that when the Stackelberg leader makes its decisions, it is aware of the toll that the regulator applies to the fringe.[19]

3.1. Untolled equilibrium

To study the airlines' interactions and assess the internalization of congestion, we first look at the no-toll equilibrium, following the framework proposed in Section 2. We extend the demand model to account for various substitution patterns between the leader and the fringe, by using the representative consumer model proposed by Dixit (1979). Demands are assumed to arise from the following strictly concave quadratic utility function: $U(q_l, q_f) = A \cdot (q_l + q_f) - (B \cdot q_l^2 + 2 \cdot E \cdot q_l \cdot q_f + B \cdot q_f^2)/2$, where A, B, and E are positive, and q_l and q_f are the number of passengers of the leader and the fringe respectively. This implicitly assumes that fringe carriers are perceived as perfect substitutes, and gives rise to the following inverse demand structure:

$$D_i(q_i, q_j) = A - B \cdot q_i - E \cdot q_j \quad i \in \{l, f\} \wedge j \neq i \tag{26}$$

where A represents the maximum reservation price, B is the own-demand sensitivity parameter, and E is the cross-demand sensitivity parameter. We assume $B \geq E \geq 0$ in general, and usually $B > E > 0$ so that outputs are imperfect substitutes. Perfect substitutability is a special case of our specification ($E = B$), while $E = 0$ has airlines serving independent markets. This specification allows us to account for horizontal product differentiation that may come from particular aspects that may differ across carriers and make passengers perceive airlines as imperfect substitutes (e.g. food and language).[20] The passengers' equilibrium condition, that stipulates that marginal willingness to pay has to be equal to the per-trip generalized cost, implies the following fare:

$$\rho_i(t) = A - B \cdot q_i - E \cdot q_j - C_p(t) \quad i \in \{l, f\} \wedge j \neq i \tag{27}$$

This again implies that all carriers, in general, charge a fare that depends on the time of departure, as $C_p(t)$ does.

In this game, each fringe carrier has the departure time of its flight as a decision variable, and the aggregate volume of the fringe is determined by the zero-profit condition. As in Section 2.1, in equilibrium, the generalized cost per flight $(s \cdot C_p(t) + C_a(t))$ must be constant in a period where the fringe operates (see Eq. (8)), otherwise a carrier will have an incentive to reschedule its flight. Moreover, in absence of the leader, this can only be possible by queuing in the center of the peak, i.e. around the desired time of arrival t^*, because that is where schedule delays are lower. In order to balance schedule delay costs and queuing delay costs, the queue must build up until t^* and, only then, start to dissipate until it disappears completely.

[19] Brueckner and Verhoef (2010) point out that assuming that agents are large enough to exert market power and to recognize the impact of their decisions on overall congestion, but that they do not take into account the impact of their actions on the tolls, is a strong assumption. We maintain this assumption to focus on the first-order effects and comparison with earlier literature, but discuss how the solution proposed by Brueckner and Verhoef (2010) applies to our case in Section 3.3.

[20] A model of (vertical) product differentiation where firms also choose quality would be more general, but it would divert attention from the implications of dynamic congestion on internalization and pricing.

As we describe in Section 2.1, and derive in Appendix A, there is a unique aggregate pattern of departures that makes the generalized cost per flight constant over time, that will be the equilibrium pattern during the time-window where the fringe operates.

On the other hand, the Stackelberg leader has as decision variables the number of flights and the departure time of each of its flights. Because it anticipates the behavior of the fringe, the leader's timing best response can be reduced to scheduling flights joining the queue of the fringe operators, and/or to schedule flights outside this congested period—in the peak shoulders—with a departure rate equal to capacity and bearing only schedule delay costs.[21] This is because the fringe carriers have the same (composite) values of time as the leader, and therefore the leader cannot benefit from taking over the center by causing higher queuing delays in this period. As a consequence, its best scheduling strategy for its flights in the center is to join the fringe's queue without exceeding the aggregate departure rate that makes generalized costs per flight constant over time. The leader also correctly perceives that scheduling flights in the shoulders of the peak, without queuing, may be attractive if the fringe's number of flights (that are queuing in the center) is sufficiently inelastic to its own decisions.

Let f be the number of flights that the fringe operates, l_c the number of flights that the leader schedules in the peak center along with the fringe, and l_s the leader's number of flights in the peak shoulders. The fringe equilibrium condition is the same as in the competitive case (see Eq. (9)), but includes the fact that the leader has $s(l_c + l_s)$ passengers, affecting its inverse demand. Furthermore, the generalized cost per flight in the peak center is constant, and will equal to $\overline{\delta} \cdot (f + l_c)/K$, as a result of the center's duration being $(f + l_c)/K$, and the first and last flight having the same generalized cost.[22] The fringe zero-profit equilibrium condition is then given by:

$$s\left[A - B \cdot sf - E \cdot s(l_c + l_s)\right] - c - \frac{\overline{\delta} \cdot (f + l_c)}{K} = 0 \tag{28}$$

This condition defines f as a function of l_c and l_s, and, therefore, it defines the fringe's response to a change in the number of flights set by the leader in both the center and the peak. The fringe's number of flights depends not only on the number of flights set by the leader in the peak center (l_c), but also on those in the shoulder (l_s), unless $E = 0$. This is an important point to stress, because it allows us to identify the condition that makes the fringe care only about what happens in the center. The latter is the assumption made by Daniel (2009). The full independence case of our model ($E = 0$) is thus the case where results may be comparable with Daniel's (2009) findings.

Straightforward calculations (see Appendix B for all derivations) yield the following conditions:

$$-1 \leq \frac{\partial f}{\partial l_c} < \frac{\partial f}{\partial l_s} \leq 0 \tag{29}$$

[21] The leader can set the departure rate equal to the capacity of the bottleneck in the peak shoulders and achieve the minimum time costs, because it does not face competition or potential entry in the peak shoulders.

[22] Denote t_{c1} the beginning of the center and t_{c2} the end. The conditions that determine the generalized cost per flight are: $\overline{\beta} \cdot (t^* - t_{c1}) = \overline{\gamma} \cdot (t_{c2} - t^*) \;\wedge\; t_{c2} - t_{c1} = (f + l_c)/k$. Solving for t_{c1} and t_{c2}, the costs at the borders will be $((\overline{\beta} \cdot \overline{\gamma})/(\overline{\beta} + \overline{\gamma})) \cdot (f + l_c)/K$, and denoting $\overline{\delta} = (\overline{\beta} \cdot \overline{\gamma})/(\overline{\beta} + \overline{\gamma})$ we get the result above. This also imply that a fraction $\overline{\gamma}/(\overline{\beta} + \overline{\gamma})$ of the flights will arrive early (between t_{c1} and t^*) and a fraction $\overline{\beta}/(\overline{\beta} + \overline{\gamma})$ of the flights will arrive late (between t^* and t_{c2}). The aggregate departure rate is obtained by equalizing the time-derivative of $s \cdot C_p(t) + C_a(t)$ to zero.

which imply that the leader anticipates that any reduction in quantities (through a reduction in frequency either in the center or in the shoulders) will be met by an increase in the fringe's number of passengers, or, equivalently, new entry, until the fringe profit is again zero.

There are two effects driving the fringe's response. First, as airlines are perceived as (imperfect) substitutes, any reduction in output by the leader will induce a shift in the inverse demand of the fringe, that will induce an increase in the fringe's output (this can be seen in the first term of Eq. (28)). Second, the airlines are imposing congestion on each other, and the leader predicts that any frequency reduction is partially offset by an increase of the number of flights set by the follower in response to reduced queuing (the third term in Eq. (28)). The substitutability effect is the same for changes in the number of flights in the center and in the shoulders, but the congestion effect happens only in the center, where there is congestion interaction. This is the reason why the fringe's response is stronger for changes in the center than in the shoulders ($\partial f/\partial l_c < \partial f/\partial l_s$).

When products are perfect substitutes ($E = B$), $\partial f/\partial l_c = -1$, which means that any change in the leader's number of flights in the center is fully offset by an opposite change of equal magnitude by the fringe. This is because the zero-profit condition in Eq. (28), determines a unique value for the aggregate number of flights when airlines are perfect substitutes. In the other extreme, when airlines serve independent markets ($E = 0$), the substitution effect disappears and only the congestion effect survives. This implies that $\partial f/\partial l_s = 0 \ \wedge \ -1 < \partial f/\partial l_c < 0$, because in the shoulders there is no congestion. The general case of imperfect substitutability ($B > E > 0$) is, naturally, in between the two cases above, satisfying Eq. (29) with strict inequalities.

With the response of the fringe defined, we can look at the first-order conditions for the Stackelberg leader and derive the equilibrium. In this untolled equilibrium, the leader's profit can be separated into two terms, the profit from the operations in the peak center and the profit from the peak shoulders. In the center, because of the fringe's presence, the generalized cost per flight must be constant and equal to $\overline{\delta} \cdot (f + l_c)/K$. In the shoulders, the leader's timing best response is to set the arrival rate equal to the bottleneck's capacity and experience only schedule delay costs. Since the duration of the entire peak has to be total number of flights over capacity, $(f + l_c + l_s)/K$, the schedule delay cost of the first and last flight is $\overline{\delta} \cdot (f + l_c + l_s)/K$. The reason is the same as for the peak center; equalizing schedule delay costs at the borders and knowing the duration, provide the necessary conditions to determine costs at the border (see footnote 22). Finally, with linear schedule delay costs, the average generalized cost per flight in the peak shoulders will be the average between the schedule delay cost at the exterior border of the shoulder (of the first and last flight) and the schedule delay cost at the interior border of the shoulder (at the beginning and end of the center): $[\overline{\delta} \cdot (f + l_c + l_s)/K + \overline{\delta} \cdot (f + l_c)/K]/2 = \overline{\delta} \cdot (f + l_c)/K + \overline{\delta} \cdot l_s/(2K)$. This shapes the profit in the following way:

$$\begin{aligned} \Pi = \quad & l_c \cdot \left(s[A - B \cdot s(l_c + l_s) - E \cdot sf] - c - \tfrac{\overline{\delta}(f+l_c)}{K} \right) + \\ & l_s \cdot \left(s[A - B \cdot s(l_c + l_s) - E \cdot sf] - c - \tfrac{\overline{\delta}(f+l_c)}{K} - \tfrac{\overline{\delta} l_s}{2K} \right) \end{aligned} \tag{30}$$

The leader's profit is a function only of l_s and l_c, because we are already taking into account the equilibrium strategy in departure times. Any amount of flights the leader sets in the shoulders, l_s, will be scheduled at a departure rate equal to capacity, and any number of flights in the center l_c, will be scheduled such that generalized costs are constant (taking into account that the fringe also schedules flights in the center). The first-order conditions, in Appendix B, show that the leader exerts market power through a markup that is less than the traditional monopoly markup, because of the fringe's offsetting behavior. We also find that the fraction of flights that the leader sets in the shoulders is:

$$\frac{l_s}{l_s + l_c} = 1 - \frac{Es^2 + \overline{\delta}/K}{Bs^2 + \overline{\delta}/K} < 1 \tag{31}$$

This shows that in the untolled equilibrium the leader always schedules flights in the peak center, a key result to understand the scheduling behavior of the leader. Recall that in our model there are two sources of inefficiency: the number of flights and the timing of flights. As explained above, the flights in the shoulders are scheduled without queuing and, in this sense, efficiently; on the other hand, the flights in the center are scheduled inefficiently as they share the queuing pattern with the competitive fringe. Therefore, the fraction $l_s/(l_s + l_c)$ is a direct measure of the leader's degree of efficiency in timing. As this ratio is always below 1, the leader never behaves fully efficiently in terms of timing.

When demand is imperfectly elastic and airlines are imperfect substitutes (i.e. $0 < E < B$), the leader schedules flights in both the peak center as well as in the peak shoulders ($l_s/(l_s+l_c) > 0$), behaving partially inefficient in terms of scheduling. This is also the case when the outputs are independent ($E = 0$). In the case of perfect substitution ($E = B$) and when demand is perfectly elastic ($B = E = 0$), the leader sets all of its flights in the peak center (so the peak center occupies the full peak), queuing along with the fringe, and being fully inefficient. The reason is that the leader knows that the fringe reacts to increases in l_c by offsetting them, so that the fringe will make room for the leader's flights; and conversely, if the leader decreases the number of flights in the peak center by shifting to the shoulders, the fringe will increase the number of flights raising the generalized costs. When the fringe fully offsets the changes in the leader's number of flights, the leader is better off setting all the flights in the peak center along with the fringe. When this effect is partial, the leader is better off setting part of the flights in the center.

The implications for congestion internalization are now straightforward to identify. The leader fails to fully internalize self-imposed congestion because the offsetting behavior of the fringe (shown in Eq. (29)) reduces its incentives to decrease output. When demands are perfectly elastic or when demand is imperfectly elastic and products are perfect substitutes, the leader does not internalize any congestion and behaves atomistically (consistent with its own demand becoming, in practice, perfectly elastic). In the case of full independence and imperfect substitutability, the leader internalizes only a fraction of the self-imposed congestion, because the offsetting behavior of the fringe is partial. These results reproduce previous findings, regarding internalization of self-imposed congestion in a Stackelberg-fringe competition, by Brueckner and Van Dender (2008), but now in a dynamic congestion model. Our result for full independence is also similar to the result by Daniel (2009), who finds that the leader sets a fraction of the flights in the peak center that ranges from 0 to 1 in the untolled equilibrium (Daniel's proposition 1). Daniel argues that the leader sets all of the flights in the peak shoulders when the number of flights by the fringe is fixed. This is also true in our model, and is obtained when market are independents and only the fringe faces a perfectly inelastic demand.

3.2. First-best tolls

Tolls are required to correct the inefficiency present in the two margins of choice of the airlines: number of flights and trip timing. The number of flights is not optimal as a result of market power exertion by the leader and the lack of full internalization of congestion by all airlines. The inefficiency in trip timing is due to the fact that queuing is a pure loss in this model: any amount of flights queuing in a certain period of time can be rescheduled to arrive during the same interval in such a way that there is no queuing while schedule delays do not increase, therefore reducing social costs.

As a result of this, the optimal timing decision by carriers must satisfy an aggregate departure rate equal to capacity, so that there are no queuing delays and no spare capacity. This allows us to drop the differentiation between the leader's flights in the center and shoulders, because there is no center with queuing. Let l be the number of flights of the leader and f the fringe's number of flights. Because, in the first-best optimum, the first and last flight must experience the same cost (only schedule delay cost) and the duration of the peak is $(l+f)/K$, delay costs will equal $\overline{\delta} \cdot (l+f)/K$ in the borders, and they will decrease linearly to zero at t^*. This yields a total social delay cost of $(l+f) \cdot \overline{\delta} \cdot (l+f)/2K$, and the first-best conditions, equating marginal social cost to full price for both the leader and the fringe, are given by:

$$s[A - B \cdot sl - E \cdot sf] = s[A - B \cdot sf - E \cdot sl] = c + \frac{\overline{\delta} \cdot (l+f)}{K} \tag{32}$$

Denote f^* and l^* the first-best number of flights that solve Eq. (32). As the fringe does not exert market power, it is inefficient only in the timing decisions (excessive queuing). This implies that the congestion toll that has to be charged to the fringe is the dynamic atomistic toll described in Section 2.1:

$$\tau(t) = \frac{\overline{\delta}(f^* + l^*)}{K} - \begin{cases} \overline{\beta} \cdot (t^* - t) & \text{if } t \leq t^* \\ \overline{\gamma} \cdot (t - t^*) & \text{if } t \geq t^* \end{cases} \tag{33}$$

This toll is the marginal social cost of the first-best equilibrium (first term on the RHS of Eq. (33)) minus the schedule delay cost at time t. It gives the incentive to each fringe carrier to schedule its single flight such that the aggregate departure rate equals capacity, because it is the only timing equilibrium that yields a constant generalized price over time (the experienced schedule delay cancels out with the time-variant part of the toll). A higher departure rate would generate queuing delays and, therefore, a higher and unbalanced generalized price over time. A lower aggregate rate will generate room for new entry of fringe carriers, that would occur until there is no spare capacity.

To derive the optimal toll that the leader has to pay, we need to derive its best response in timing and number of flights, when the fringe faces the dynamic atomistic toll in Eq. (33). With static model of congestion, this is straightforward, as the only decision variable is the number of flights. In the present setting, the leader also chooses the departure time of each of its flights. The main result of our analysis is that the first-best congestion toll for the leader is not unique. There is a time-invariant toll that can be charged to the leader in order to achieve the first-best outcome, a time-variant toll that also yields the efficient outcome, but there are also various other pricing schemes.

- Time-invariant toll

The Stackelberg leader has the potential to schedule its flights without incurring queuing delays, as we discussed in Section 2.2 for a monopoly, but it has reduced incentives to do so in the no-toll equilibrium of this game because of the fringe's presence. However, when the regulator imposes the dynamic atomistic toll in Eq. (33) to the fringe, the leader realizes that it can schedule flights efficiently (without queuing and operating at capacity), and knows that this keeps the fringe completely out of its own period of operation. This is because the fringe, when facing the atomistic toll in Eq. (33), experience a constant generalized price per flight equal to the marginal social cost ($\overline{\delta}(f^* + l^*)/K$), regardless of the time of operation, as long as the aggregate departure rate is, at most, equal to capacity. In any other case, the generalized price will be higher. Thus, the leader, by setting its departure rate equal to capacity, effectively prevents entry from the fringe to its period, because the fringe is always better off operating at times when departures do not exceed the capacity.

The leader realizes that it is better off operating in the peak center, around t^*, where the schedule delays are lower. For any amount of flights l, the profit maximizing timing strategy is to set the departure rate equal to capacity (to have only schedule delay costs and prevent the fringe from entering), from t_1 to t_2 such that the cost of the first flight and the last flight is the same ($\overline{\beta} \cdot (t^* - t_1) = \overline{\gamma} \cdot (t_2 - t^*)$). As $t_2 - t_1 = l/K$, the delay cost at the borders equals $\overline{\delta} \cdot l/K$, and as schedule delay costs are linear and zero at t^*, the average delay cost per flight will be $\overline{\delta} \cdot l/2K$. Then, the leader's profit and first-order condition, when facing a time invariant toll $\widehat{\tau}$, is:[23]

$$\Pi = l \cdot \left(s[A - B \cdot sl - E \cdot sf] - c - \frac{\overline{\delta} l}{2K} - \widehat{\tau} \right)$$

$$\frac{\partial \Pi}{\partial l} = 0 \Rightarrow \; s[A - B \cdot sl - E \cdot sf] = c + \frac{\overline{\delta} \cdot l}{K} + (B + E \cdot \frac{\partial f}{\partial l}) \cdot s^2 l + \widehat{\tau} \tag{34}$$

The leader fails to take into account the delays imposed on the fringe ($\overline{\delta} f/K$ is not in the full price), and exerts market power (third term on the RHS of Eq. (34)), which in this case is reduced compared to the monopolistic case, because of the (partial) offsetting behavior of the fringe. The fringe's full price depends on l only through the demand side (the substitutability effect) because there is no queuing interaction. Hence, any reduction of frequency by the leader will result in a lower full price for the fringe, that—because of the free-entry (zero-profit condition)—translates into an output expansion by the fringe. The first-best flat-toll for the leader is simply the toll that corrects market power and congestion effects, and makes the full price set by the leader (RHS of Eq. (34)) equal to the marginal social cost (RHS of Eq. (32)):

$$\widehat{\tau} = \frac{\overline{\delta} f^*}{K} - (B + E \cdot \frac{\partial f}{\partial l}) \cdot s^2 l^* \tag{35}$$

This first-best toll consists of a market power subsidy (second term on the RHS), that depends on the substitution pattern, and a congestion charge that is independent of the amount of internalization of the

[23] The derivation of the profit function is analogous to the monopoly case (see Eqs. (16) and (17)).

untolled equilibrium. Our result shows that the congestion side of the toll appears to be different from what has been found in the literature before. Brueckner and Van Dender (2008) find that the leader should pay a congestion toll that lies in between the congestion imposed on the fringe and total marginal congestion costs (depending on the substitution pattern). We find that, when the regulator charges the dynamic atomistic toll to the fringe, because of the sequential nature of the game and the congestion technology, the leader does not fail to internalize self-imposed congestion. As a consequence, the regulator can induce the first-best outcome by charging the delays imposed by the leader on the fringe (analogous to the so-called "Cournot" toll).

When airlines are perfect substitutes ($E = B$), there is no need for market power subsidy as the fringe fully offset any reduction of flights ($\partial f/\partial l = -1$). As both types of agent behave atomistically, only the aggregate number of flights is defined and, therefore, the toll $\overline{\delta} f^*/K$ will define the proportion of flights set by the leader and the fringe. In particular, the regulator can set the leader's congestion toll to zero, meaning that the leader will supply the optimal output making the optimal timing decisions. In the case of full independence, the fringe does not exhibit the offsetting behavior as there is no substitutability effect nor congestion effect ($\partial f/\partial l = 0$), and the congestion part of the toll is uniquely determined because f^* is unique (see Eq. (32)). In the general case of imperfect substitution ($0 < E < B$), the market power subsidy is lower because of the partial offsetting behavior of the fringe, and approaches zero as airlines become closer substitutes, while the congestion part of the toll remains uniquely defined by Eq. (32).

Figure 3 shows the (first-best) equilibrium that results from charging $\widehat{\tau}$ in Eq. (35) to the leader, and the dynamic atomistic toll in Eq. (33) to the fringe. The leader schedules its flights to arrive in the center, between $[t_1, t_2]$, the fringe operates outside, between $[t_s, t_1]$ and $[t_1, t_e]$, the first-best conditions in Eq. (32) are satisfied, and there are no queuing delays. The leader charges a fare that depends on the time of departure ($\rho(t)$), its profit is equal to the saved queuing costs $\overline{\delta} {l^*}^2/2K$ and the revenues from the market power effect ($l^* \cdot (B + E \cdot \partial f/\partial l) \cdot s^2 l^*$, not shown graphically). The congestion toll revenues (before subtracting the subsidy) per unit of capacity are equal to the shaded area in Figure 3: the sum of the revenues from the leader (the rectangle in the center) and from the fringe (the two triangles at the shoulders).

- Time-variant toll

In this section we show that the first-best can also be attained by charging the dynamic atomistic toll in Eq. (33) to both the leader and the fringe, if, in addition, the market power subsidy is given to the leader. To see this, consider a leader's flight that is scheduled to arrive at t to the destination. The profit that the leader gets from that flight is given by:

$$\pi(t) = s \cdot [A - B \cdot sl - E \cdot sf - C_p(t)] - [C_a(t) + c] - [\tau(t)] \tag{36}$$

where the first term in brackets is the fare that the leader charges for that flight (marginal willingness to pay minus passengers' generalized cost), the second bracketed term is the airline's cost from operating that flight, and the last term is the dynamic atomistic toll in Eq. (33). The negative component of the profit that depends on the time of arrival, that will determine the best response in timing, is the generalized cost

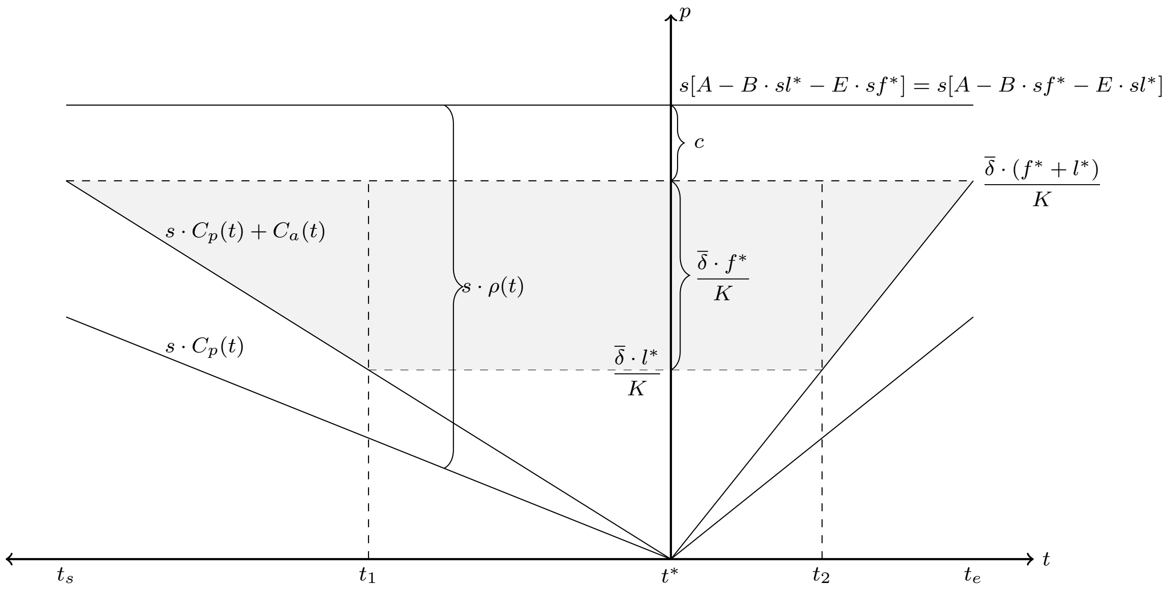
p
$s[A - B \cdot sl^* - E \cdot sf^*] = s[A - B \cdot sf^* - E \cdot sl^*]$
c
$\frac{\overline{\delta} \cdot (f^* + l^*)}{K}$
$s \cdot C_p(t) + C_a(t)$
$s \cdot \rho(t)$
$\frac{\overline{\delta} \cdot f^*}{K}$
$s \cdot C_p(t)$
$\frac{\overline{\delta} \cdot l^*}{K}$
t
t_s
t_1
t^*
t_2
t_e

Figure 3: First-best equilibrium with time invariant toll to the leader.

per flight minus the time dependent part of the toll:

$$[s \cdot C_p(t) + C_a(t)] - \begin{cases} \overline{\beta} \cdot (t^* - t) & \text{if } t \leq t^* \\ \overline{\gamma} \cdot (t - t^*) & \text{if } t \geq t^* \end{cases} \tag{37}$$

For any number of flights by the leader, the timing decisions must minimize the sum of the costs in Eq. (37) for all flights, and the unique way to do so is to schedule them such that the departure rate does not exceed the capacity of the bottleneck. This is because it minimizes the generalized cost per flight (first term in Eq. (37)), that will consist of only schedule delay costs and, as a consequence, the time varying component of the profit will be zero (the schedule delay costs cancel out with the time-varying part of the toll). This is also compatible with the competitive fringe's reaction, as fringe carriers facing the dynamic atomistic toll never schedule flights to exceed capacity. As the best response in timing makes the profit per flight constant ($s \cdot C_p(t) + C_a(t) + \tau(t) = \overline{\delta}(f^* + l^*)/K$), the leader's total profit can be written as the number of flights l times the profit per flight:

$$\Pi = l \cdot \left(s[A - B \cdot sl - E \cdot sf] - c - \frac{\overline{\delta}(f^* + l^*)}{K} \right) \tag{38}$$

The first-order condition gives the full price set by the leader:

$$\frac{\partial \Pi}{\partial l} = 0 \Rightarrow \; s[A - B \cdot sl - E \cdot sf] = c + \frac{\overline{\delta}(f^* + l^*)}{K} + (B + E \cdot \frac{\partial f}{\partial l}) \cdot s^2 l \tag{39}$$

This shows that the dynamic atomistic toll charged to the leader solves the externality inefficiency and only the market power exertion needs to be corrected with the same subsidy as in Eq. (35): $(B + E \cdot \partial f / \partial l) \cdot s^2 l^*$.

This result shows again that the congestion part of the first-best toll is not related with the degree of internalization in the untolled equilibrium. In this case, the dynamic atomistic toll for all carriers (leader and fringe) solves the externality inefficiency. This result also has an important implication for the financial situation of the airport. As Arnott et al. (1993) demonstrate, the self-financing results of Mohring and Harwitz (1962) for capacity investments hold for the bottleneck model with elastic demand. As we have shown, the results of our analysis parallel results for the road case regarding the toll; therefore, the self-financing result also holds when the toll in Eq. (33) is charged to both groups of airlines. If there are constant returns to scale in capacity provision, the revenues from the first-best toll then exactly cover the cost of providing the optimal capacity.[24] This differs from earlier results because first-best tolls are now not discounted by the fraction of congestion that is internalized by carriers; therefore, the self-financing result is not overturned by the internalization of congestion. However, the market-power subsidy does upset exact self-financing under neutral scale economies. This is because under marginal cost pricing and constant returns to scale, the surplus will be zero. When part of the revenues is used to subsidize the firm with market power, there will be insufficient revenue to cover capacity costs. The shortfall in self-financing equals the aggregate airlines' profit under constant returns to scale. With this time-varying pricing scheme, the leader makes a lower profit compared to the time-invariant case (this time only from the ability to exert market power), and revenues are equally higher. In fact, this is the pricing scheme that yields the highest revenue for the regulator.

- Alternative schemes

In addition to the two different ways to deal with the congestion inefficiency of the leader described above, namely a flat toll equal to the delays imposed on all the fringe's flights or the dynamic atomistic toll, there are other pricing schemes that can induce the first-best outcome. Although the market power distortion has to be corrected in any case with the subsidy in Eq. (35), $(B + E \cdot \partial f/\partial l) \cdot s^2 l^*$, the regulator can induce the leader to set the full price of its flights equal to the marginal social cost with regimes that correct the congestion effect differently.

First, note that if the leader is forced to give up a time-window $[t_1, t_2]$ around t^*, of duration f^*/K, and with t_1 and t_2 satisfying the condition of equal schedule delay costs ($\overline{\beta} \cdot (t^* - t_1) = \overline{\gamma} \cdot (t_2 - t^*)$), the first-best can again be attained (given that the market power distortion is being corrected). In this case the fringe will operate around t^*, between $[t_1, t_2]$, paying the dynamic atomistic toll, ensuring that full price equals marginal social cost. The generalized cost per flight at t_1 and t_2 equals $\overline{\delta} f^*/K$. The leader's best response is to schedule its flights with a departure rate equal to capacity outside the "forbidden period" $[t_1, t_2]$, and such that the first and last flight experience the same generalized cost. As a result, it schedules l^* flights from t_s to t_1 and from t_2 to t_e, without queuing, earning the saved queuing costs as profit. The leader's first and last flight face a generalized cost (per flight) of $\overline{\delta}(f^* + l^*)/K$, hence satisfying the first-best condition in Eq. (32).[25] The leader has no incentives to schedule more (nor less) flights, because the marginal revenue of the first and last flight is exactly equal to the marginal cost ($s \cdot \rho(t_s) - C_a(t_s) - c = s \cdot \rho(t_e) - C_a(t_e) - c = 0$ in Figure 3).

[24] In general, "the ratio of the revenue collected from the optimal toll to the costs of constructing optimal capacity equals the elasticity of construction cost with respect to capacity" (Arnott et al., 1993).

[25] This equilibrium is not shown graphically, but it is enough to see Figure 3 and change l^* for f^* (and vice versa). The duration of the center is $t_2 - t_1 = f^*/K$.

The toll regime that induces this outcome is the dynamic atomistic toll in Eq. (33) for the fringe, and for the leader the per-flight market power subsidy to correct dead-weight losses, and a toll arbitrarily higher than $\overline{\delta}(f^* + l^*)/K$ only during the period $[t_1, t_2]$. The latter works as a barrier for the leader to operate in the peak center, as it makes him better off operating outside it, not paying the toll. In fact, this is equivalent to restrict the interval of time where the leader can operate.

This configuration is similar to the previous time-invariant toll setting in the sense that the full price does not change, because the gain in costs by the fringe (resulting from operating closer to t^*) is offset by higher tolls, and the cost increase of the leader is offset by the absence of congestion tolls. This makes this setting identical to the time-invariant case analyzed above in social welfare, consumer surplus, profit per firm (hence total profit) and total revenue. The difference, besides the times of operation for each firm, is that the tax revenues are not the same for each individual firm, but total tax revenues remain unchanged. In fact, there is a continuum of configurations, where the leader faces a time restriction (or barrier-toll) and a flat toll, that follow these properties. These configurations are defined by more elaborate patterns of temporal separation of leader and fringe operations, and the congestion tolls become a more complicated matter.

Furthermore, there is also a continuum of alternative schemes that deal with the leader's congestion inefficiency with different time-varying tolls. This is due to the fact that the leader, knowing that the fringe faces the dynamic atomistic toll, will never schedule its flights in a way that causes the aggregate departure rate to exceed capacity. When the atomistic toll is charged to the leader, the schedule delay cost cancels out with the time-varying part of the toll at any time (see Eq. (37)). Thus, the marginal cost (to the firm) of a flight is only the fixed part of the toll, that is set equal to the marginal social cost by the regulator. If the regulator charges to the leader a toll with a time-varying component that is lower than the experienced schedule delay cost (i.e. less steep than the atomistic toll), they will no longer cancel out; as a result, the marginal cost will be the fixed part of the toll plus the time-varying cost. As the latter is now above zero, the former has to be reduced (with respect to the atomistic toll) in order to make the sum equal to marginal social cost. These toll schedules would produce toll revenues that are in between those from the flat toll, and those from the time varying atomistic toll.

3.3. Manipulable tolls

We assume in the analysis above that an airline, that is large enough to exert market power and to recognize the impact of its decisions on overall congestion (and followers), does not take into account the impact of its actions on the tolls. We are aware that this is a strong assumption, but it is common to most previous works. Brueckner and Verhoef (2010) propose a manipulable toll rule, designed to induce the social optimum when carriers predict the impact of their decisions on tolls, that can also be applied to our problem. They propose an adjustment such that the carriers' profit plus the (manipulable) toll liability varies perfectly in parallel with social surplus. In our problem, the welfare maximizing tolls can be straightforwardly adjusted with their methodology. For example, consider the toll regime where the leader pays the time-invariant toll in Eq. (35); the adjusted first-best time-invariant toll rule, designed to be "manipulated" by the leader, is given by:

$$T(l) = \frac{\bar{\delta} f^2}{2 \cdot K} \cdot \frac{1}{\partial f / \partial l} - (B + E \cdot \frac{\partial f}{\partial l}) \cdot \frac{s^2 \cdot l^2}{2} + T_c \tag{40}$$

where T_c is a constant and $\partial f / \partial l$ is independent of l, and it is obtained in a way analogous to the one in Appendix B for the untolled setting. By charging this toll rule, a leader that anticipates the effect of his decisions on a parametric toll (as in Eq. (35)) will have a pricing strategy that will lead to a social welfare maximizing number of flights and fares. This is because the marginal change in profit, from an increase in the number of flights, that is due to the toll rule ($\partial T(l)/\partial l$) is exactly the parametric toll in Eq. (35) when evaluated at the social optimum.

The adjustment for the time-variant tolling regime follows the same logic: the market power subsidy has to be the same as above (second term on the right-hand side of Eq. (40)); the time-invariant component of the congestion toll rule has to be adjusted in a similar way as above (in this case including a term involving the own number of flights); and the time-variant component of the congestion toll rule needs to have the same slope as before, i.e. to vary over time perfectly in line with schedule delay costs.

4. Conclusion

This paper studies airlines' interactions and scheduling behavior, together with airport efficient pricing, using a deterministic bottleneck model of congestion. We confirm that an airline acting as a Stackelberg leader, facing a competitive group of fringe carriers, partially internalizes self-imposed congestion in the sense that, without facing tolls, it schedules fewer flights than perfectly competitive carriers would, achieving lower social congestion costs. Consistent with findings in the earlier literature using static models of congestion (e.g., Brueckner and Van Dender, 2008), the degree of internalization of self-imposed congestion depends critically on the assumed demand substitution pattern. Nevertheless, and what is new, our results suggests that social welfare maximizing congestion tolls do not depend crucially on the degree of internalization, and that the time-variant tolls derived for perfectly competitive carriers apply also to a monopoly airline and to a setting where a Stackelberg leader interacts with a group of competitive carriers as followers.

Our analysis suggests that optimal congestion pricing may have a more significant role than what has been suggested in the earlier literature based on static models. Moreover, the efficient fully time-variant congestion toll regime results in a revenue for the airport that restores the well known self-financing result for congested facilities Mohring and Harwitz (1962). Still, if the market power distortion is corrected with a subsidy drawn from the airport's budget, the self-financing result is upset. Our results also suggest that the political feasibility of congestion pricing would be enhanced compared to earlier studies, as efficient congestion charges do not vary with market shares, and therefore are less likely to be perceived as inequitable.

We also find agreement with Daniel (1995, 2009) and Daniel and Harback (2008) in that dynamic atomistic tolls are efficient in markets well represented by an interaction between a leader and a competitive fringe as the follower, but we show that this is not the only efficient solution. The non-uniqueness of social welfare maximizing congestion tolls in this setting allows for other pricing schemes that also achieve the social optimum.

Incorporating heterogeneity and studying step-tolling are natural extensions of the present analysis, to complement Daniel's (2009) work. Our model allows for the inclusion of heterogeneity in values of time and preferences for both airlines and passengers. Certainly, the equilibrium and optimal toll will depend on the type of heterogeneity considered. Step-tolling, a relevant alternative in practice, may bring important benefits compared with the social optimum; as the number of steps is increased, it approaches the dynamic atomistic congestion toll, and, consequently, also its efficiency and consumer surplus increases, approaching the optimal values (see van den Berg (2012)). Finally, the analysis of simultaneous competition between airlines with market power is also a natural extension of this analysis.

Acknowledgments

Financial support from ERC Advanced Grant OPTION (#246969) is gratefully acknowledged. We are indebted to Achim Czerny, Sergejs Gubins, Eric Pels, the editor and two anonymous reviewers for helpful comments and suggestions.

Appendix A. Derivation of the equilibrium in the perfect competitive case

Appendix A.1. Equilibrium in scheduling

This Section determines the unique equilibrium values for the beginning (t_s) and the end (t_e) of the peak period, as well as the departure rate function $r(t)$ that defines the equilibrium (aggregate) queuing pattern. The travel delay, $T(t)$, and queue length, $Q(t)$, of a flight that departs at t are given by:

$$T(t) = \frac{Q(t)}{K} \ \wedge \ Q(t) = \int_{\widehat{t}}^{t} (r(u) - K)du \tag{A.1}$$

where $\widehat{t}$ is the most recent time at which there was no queue. Let $\tilde{t}$ be the departure time for an on-time arrival ($\tilde{t} + T(\tilde{t}) = t^*$), and consider a flight that departs at t and arrives early ($t < \tilde{t}$). The generalized cost of that flight is:

$$s \cdot C_p(t) + C_a(t) = \overline{\alpha} \cdot T(t) + \overline{\beta} \cdot (t^* - t - T(t)) \tag{A.2}$$

The equilibrium condition states that the generalized cost per flight has to be constant over time. By equating the time-derivative of Eq. (A.2) to zero, we obtain the equilibrium departure rate for early arrivals:

$$\frac{d[s \cdot C_p(t) + C_a(t)]}{dt} = \overline{\alpha} \cdot \left(\frac{r(t)}{K} - 1\right) - \overline{\beta} \cdot \left(1 + \overline{\alpha} \cdot \left(\frac{r(t)}{K} - 1\right)\right) = 0 \tag{A.3}$$

$$\Rightarrow \ r(t) = \frac{K \cdot \overline{\alpha}}{\overline{\alpha} - \overline{\beta}} \ \ \forall \, t \ \in \ [t_s, \tilde{t} \ ) \tag{A.4}$$

Analogous calculations give the following equilibrium departure rate for late arrivals:

$$r(t) = \frac{K \cdot \overline{\alpha}}{\overline{\alpha} + \overline{\gamma}} \ \ \forall \, t \ \in \ [\ \tilde{t}, t_e] \tag{A.5}$$

The rates in Eqs. (A.4) and (A.5) show that the queue builds up linearly from t_s to $\tilde{t}$, and then dissipates linearly until it disappears at t_e.

Using that the first and last flight must experience the same generalized cost in equilibrium, and that the peak duration is f/K, the start and end of the peak period can be derived, together with the equilibrium generalized cost per flight:

$$\overline{\beta} \cdot (t^* - t_s) = \overline{\gamma} \cdot (t_e - t^*) \ \wedge \ t_s - t_e = f/K \tag{A.6}$$

$$\Rightarrow t_s = t^* - \frac{\overline{\gamma}}{\overline{\beta} + \overline{\gamma}} \cdot \frac{f}{K} \ \wedge \ t_e = t^* + \frac{\overline{\beta}}{\overline{\beta} + \overline{\gamma}} \cdot \frac{f}{K} \ \wedge \ s \cdot C_p(t) + C_a(t) = \frac{\overline{\beta} \cdot \overline{\gamma}}{\overline{\beta} + \overline{\gamma}} \cdot \frac{f}{K} \, \forall \, t \, \in \, [t_s, t_e] \tag{A.7}$$

Finally, straightforward calculations yield the departure time for an on-time arrival:

$$\tilde{t} = t^* - \frac{\overline{\beta}}{\overline{\alpha}} \cdot \frac{\overline{\gamma}}{\overline{\beta} + \overline{\gamma}} \cdot \frac{f}{K} \tag{A.8}$$

As can be seen above, the conditions that we need to impose are $\overline{\alpha} > \overline{\beta} > 0$ and $\overline{\gamma} > 0$, so that the variables have the correct sign. As the empirical literature suggests (Morrison and Winston (1989); Lijesen (2006)), the values of time for passengers satisfy these conditions ($\alpha_p > \beta_p > 0$ and $\gamma_p > 0$), and, as a consequence, the results hold also when airlines' do not incur schedule delay costs ($\beta_a = 0$ and $\gamma_a = 0$).

Appendix A.2. Equilibrium fare variation over time

With the equilibrium rates described, we can study how the fare $\rho(t)$ changes over time by using Eqs. (2), (10), and taking the derivative with respect to t:

$$\frac{\partial \rho(t)}{\partial t} = -\frac{\partial C_p(t)}{\partial t} = -\alpha_p \, \frac{\partial T(t)}{\partial t} - \begin{cases} \beta_p \frac{\partial (t^* - t)}{\partial t} \\ \gamma_p \frac{\partial (t - t^*)}{\partial t} \end{cases} = \begin{cases} -\left[\alpha_p \cdot \frac{\overline{\beta}}{\overline{\alpha}} - \beta_p\right] = \beta_p \left[1 - \frac{\alpha_p/\beta_p}{\overline{\alpha}/\overline{\beta}}\right] & \text{if } t \leq t^* \\ \left[\alpha_p \cdot \frac{\overline{\gamma}}{\overline{\alpha}} - \gamma_p\right] = \gamma_p \left[\frac{\alpha_p/\gamma_p}{\overline{\alpha}/\overline{\gamma}} - 1\right] & \text{if } t \geq t^* \end{cases} \tag{A.9}$$

where we use that, in equilibrium, queuing delays, $T(t)$, have a slope of $\overline{\beta}/\overline{\alpha}$ for early arrivals and $-\overline{\gamma}/\overline{\alpha}$ for late arrivals. This reveals that only when the ratios α_p/β_p and α_p/γ_p equal $\overline{\alpha}/\overline{\beta}$ and $\overline{\alpha}/\overline{\gamma}$ respectively, the fare (and generalized cost) is constant over time. This can only occur when the passengers' willingness to accept schedule delays in order to reduce travel times, as represented by the ratios α_p/β_p and α_p/γ_p, equal the airlines' willingness to accept schedule delays in order to reduce travel times (α_a/β_a and α_a/γ_a). On the other hand, when the passengers' ratios α_p/β_p and α_p/γ_p are lower (higher) than the airlines' ratios, the fare will be higher (lower) for passengers traveling closer to t^*.

Appendix B. Fringe's response and leader's first-order conditions

From Eq. (28), we can solve for f and take the derivative with respect to l_c and l_s in order to obtain the fringe's reaction to changes in number of flights by the leader. Solving for f, we obtain:

$$f = \frac{s\left[A - E \cdot s(l_c + l_s)\right] - c - \overline{\delta} \cdot l_c/K}{Bs^2 + \overline{\delta}/K} \tag{B.1}$$

Taking the derivative of Eq. (B.1) with respect to l_c, we find the response of the fringe to a change in the number of flights that the leader schedules in the peak center:

$$\frac{\partial f}{\partial l_c} = -\frac{Es^2}{Bs^2 + \overline{\delta}/K} - \frac{\overline{\delta}/K}{Bs^2 + \overline{\delta}/K} = -\frac{Es^2 + \overline{\delta}/K}{Bs^2 + \overline{\delta}/K} \equiv \phi \tag{B.2}$$

Since $E < B$, it follows that $-1 < \phi < 0$. That is, a frequency change by the leader in the peak center, yields an opposite change in number of flights by the fringe, but that is not equal in magnitude because of the assumed substitution pattern in demand. In the case where outputs are perfect substitutes ($E = B$), $\phi = -1$, which means that any frequency reduction by the leader in the congested period is fully offset by an increase in number of flights by the competitive fringe. When outputs are independent ($E = 0$), the response of the follower still partially offsets a leader's frequency change.

Differentiating Eq. (B.1) with respect to l_s gives the response of the fringe to a change in the number of flights scheduled in the peak shoulders:

$$\frac{\partial f}{\partial l_s} = -\frac{Es^2}{Bs^2 + \bar{\delta}/K} \equiv \lambda > \phi \tag{B.3}$$

When $0 \leq E < B$, the response of the fringe, to an increase of the leader number of flights scheduled in the peak shoulders, satisfies $-1 < \lambda \leq 0$. Note that $\lambda > \phi$ means that the response ϕ is stronger than the response λ, because both are negative.

With these expressions, we can derive the first-order conditions for profit maximization. Taking the derivatives of the profit in Eq. (30), we get the following:

$$\begin{aligned} \frac{\partial \Pi}{\partial l_c} = 0 = \quad & s[A - B \cdot s(l_c + l_s) - E \cdot sf] - c - \tfrac{\bar{\delta}(f + l_c)}{K} \\ & - \left[(B + \phi E) \cdot s^2(l_c + l_s)\right] - \left[\tfrac{\bar{\delta}(l_c + l_s)}{K}(1 + \phi)\right] \end{aligned} \tag{B.4}$$

$$\begin{aligned} \frac{\partial \Pi}{\partial l_s} = 0 = \quad & s[A - B \cdot s(l_c + l_s) - E \cdot sf] - c - \tfrac{\bar{\delta}(f + l_c + l_s)}{K} \\ & - \left[(B + \lambda E) \cdot s^2(l_c + l_s)\right] - \left[\tfrac{\bar{\delta}(l_c + l_s)}{K} \cdot \lambda\right] \end{aligned} \tag{B.5}$$

In both first-order conditions, the last two terms in square brackets on the right-hand side show the market power markup and the reduced incentives to internalize self-imposed congestion, respectively. By subtracting Eqs. (B.4) and (B.5), we can explicitly write the fraction of flights that the leader schedules in the shoulders:

$$\frac{\partial \Pi}{\partial l_c} - \frac{\partial \Pi}{\partial l_s} = \frac{\bar{\delta} l_s}{K} - \left[\phi E \cdot s^2(l_c + l_s)\right] + \left[\lambda E \cdot s^2(l_c + l_s)\right] - \frac{\bar{\delta}(l_c + l_s)}{K}(1 + \phi - \lambda) = 0$$

$$\Rightarrow \frac{l_s}{l_s + l_c} = \frac{E \cdot s^2(\phi - \lambda) + \bar{\delta}/K(1 + (\phi - \lambda))}{\bar{\delta}/K} = 1 - \frac{Es^2 + \bar{\delta}/K}{Bs^2 + \bar{\delta}/K} \tag{B.6}$$

References

Arnott, R., De Palma, A. and Lindsey, R. (1990), 'Economics of a bottleneck', *Journal of Urban Economics* **27**(1), 111–130.

Arnott, R., De Palma, A. and Lindsey, R. (1993), 'A structural model of peak-period congestion: A traffic bottleneck with elastic demand', *The American Economic Review* **83**(1), 161–179.

Arnott, R., De Palma, A. and Lindsey, R. (1994), 'The welfare effects of congestion tolls with heterogeneous commuters', *Journal of Transport Economics and Policy* **28**(2), 139–161.

Basso, L. J. and Zhang, A. (2007), 'Congestible facility rivalry in vertical structures', *Journal of Urban Economics* **61**(2), 218–237.

Basso, L. J. and Zhang, A. (2010), 'Pricing vs. slot policies when airport profits matter', *Transportation Research Part B: Methodological* **44**(3), 381–391.

Brueckner, J. K. (2002), 'Airport congestion when carriers have market power', *The American Economic Review* **92**(5), 1357–1375.

Brueckner, J. K. (2004), 'Network structure and airline scheduling', *The Journal of Industrial Economics* **52**(2), 291–312.

Brueckner, J. K. (2005), 'Internalization of airport congestion: A network analysis', *International Journal of Industrial Organization* **23**(7-8), 599–614.

Brueckner, J. K. (2009), 'Price vs. quantity-based approaches to airport congestion management', *Journal of Public Economics* **93**(5-6), 681–690.

Brueckner, J. K. and Girvin, R. (2008), 'Airport noise regulation, airline service quality, and social welfare', *Transportation Research Part B: Methodological* **42**(1), 19–37.

Brueckner, J. K. and Van Dender, K. (2008), 'Atomistic congestion tolls at concentrated airports? seeking a unified view in the internalization debate', *Journal of Urban Economics* **64**(2), 288–295.

Brueckner, J. K. and Verhoef, E. T. (2010), 'Manipulable congestion tolls', *Journal of Urban Economics* **67**(3), 315–321.

Carlin, A. and Park, R. (1970), 'Marginal cost pricing of airport runway capacity', *The American Economic Review* **60**(3), 310–319.

Carlsson, F. (2003), 'Airport marginal cost pricing: Discussion and an application to swedish airports', *International Journal of Transport Economics* **30**(3), 283–303.

Daniel, J. I. (1995), 'Congestion pricing and capacity of large hub airports: A bottleneck model with stochastic queues', *Econometrica: Journal of the Econometric Society* **63**(2), 327–370.

Daniel, J. I. (2001), 'Distributional consequences of airport congestion pricing', *Journal of Urban Economics* **50**(2), 230–258.

Daniel, J. I. (2009), 'The deterministic bottleneck model with non-atomistic traffic', *Working Papers 09-08, University of Delaware, Department of Economics* .

Daniel, J. I. and Harback, K. T. (2008), '(When) do hub airlines internalize their self-imposed congestion delays?', *Journal of Urban Economics* **63**(2), 583–612.

Daniel, J. I. and Harback, K. T. (2009), 'Pricing the major us hub airports', *Journal of Urban Economics* **66**(1), 33–56.

Dixit, A. (1979), 'A model of duopoly suggesting a theory of entry barriers', *The Bell Journal of Economics* pp. 20–32.

Douglas, G. W. and Miller, J. C. (1974), *Economic regulation of domestic air transport: theory and policy*, Vol. 10, Brookings Institution Washington, DC.

Forbes, S. J. (2008), 'The effect of air traffic delays on airline prices', *International Journal of Industrial Organization* **26**(5), 1218–1232.

Jorge, J. D. and de Rus, G. (2004), 'Cost-benefit analysis of investments in airport infrastructure: a practical approach', *Journal of Air Transport Management* **10**(5), 311–326.

Levine, M. E. (1969), 'Landing fees and the airport congestion problem', *Journal of Law and economics* **12**(1), 79–108.

Lijesen, M. G. (2006), 'A mixed logit based valuation of frequency in civil aviation from sp-data', *Transportation Research Part E: Logistics and Transportation Review* **42**(2), 82–94.

Mohring, H. and Harwitz, M. (1962), *Highway benefits: An analytical framework*, Evanston, IL: Northwestern University Press.

Morrison, S. A. and Winston, C. (1989), 'Enhancing the performance of the deregulated air transportation system', *Brookings Papers on Economic Activity. Microeconomics* **1989**, 61–112.

Oum, T. H., Zhang, A. and Zhang, Y. (1995), 'Airline network rivalry', *The Canadian Journal of Economics/Revue canadienne d'Economique* **28**(4a), 836–857.

Pels, E. and Verhoef, E. T. (2004), 'The economics of airport congestion pricing', *Journal of Urban Economics* **55**(2), 257–277.

Rupp, N. (2009), 'Do carriers internalize congestion costs? empirical evidence on the internalization question', *Journal of Urban Economics* **65**(1), 24–37.

Santos, G. and Robin, M. (2010), 'Determinants of delays at european airports', *Transportation Research Part B: Methodological* **44**(3), 392–403.

Silva, H. E. and Verhoef, E. T. (2013), 'Optimal pricing of flights and passengers at congested airports and the efficiency of atomistic charges', *Journal of Public Economics* **106**, 1–13.

Small, K. A. (1982), 'The scheduling of consumer activities: work trips', *The American Economic Review* **72**(3), 467–479.

van den Berg, V. A. C. (2012), 'Step-tolling with price-sensitive demand: Why more steps in the toll make the consumer better off', *Transportation Research Part A: Policy and Practice* **46**(10), 1608–1622.

van den Berg, V. A. C. and Verhoef, E. T. (2011), 'Winning or losing from dynamic bottleneck congestion pricing?: The distributional effects of road pricing with heterogeneity in values of time and schedule delay', *Journal of Public Economics* **95**(7–8), 983–992.

Verhoef, E. T. (2010), 'Congestion pricing, slot sales and slot trading in aviation', *Transportation Research Part B: Methodological* **44**(3), 320–329.

Vickrey, W. (1973), 'Pricing, metering, and efficiently using urban transportation facilities', *Highway Research Record* **476**, 36–48.

Vickrey, W. S. (1969), 'Congestion theory and transport investment', *The American Economic Review* **59**(2), 251–260.

Zhang, A. and Zhang, Y. (2006), 'Airport capacity and congestion when carriers have market power', *Journal of Urban Economics* **60**(2), 229–247.