INSTITUTO DE ECONOMÍA

DOCUMENTO de TRABAJO

512

2018

Equilibrium in a dynamic model of congestion with large and small users.

Hugo E. Silva, Robin Lindsey y Andre de Palma

www.economia.puc.cl · ISSN (edición impresa) 0716-7334 · ISSN (edición electrónica) 0717-7593

Equilibrium in a dynamic model of congestion with large and small users[☆]

Robin Lindsey[a], André de Palma[b], Hugo E. Silva[c]

[a] *Sauder School of Business, University of British Columbia, Canada.*
[b] *Ecole Normale Supérieure de Cachan (CREST) - University Paris-Saclay, 61 Av. Du Président Wilson, 94230, Cachan, France.*
[c] *Departmento de Ingeniería de Transporte y Logística, Instituto de Economía, Pontificia Universidad Católica de Chile, Santiago, Chile.*

April 4, 2018

Abstract

Individual users often control a significant share of total traffic flows. Examples include airlines, rail and maritime freight shippers, urban goods delivery companies and passenger transportation network companies. These users have an incentive to internalize the congestion delays their own vehicles impose on each other by adjusting the timing of their trips. We investigate simultaneous trip-timing decisions by large users and small users in a dynamic model of congestion. Unlike previous work, we allow for heterogeneity of trip-timing preferences and for the presence of small users such as individual commuters and fringe airlines. We derive the optimal fleet departure schedule for a large user as a best-response to the aggregate departure rate of other users. We show that when the vehicles in a large user's fleet have a sufficiently dispersed distribution of desired arrival times, there may exist a pure-strategy Nash-equilibrium (PSNE) in which the large user schedules vehicles when there is a queue. This resolves the problem of non-existence of a PSNE identified in Silva et al. (2017) for the case of symmetric large users. We also develop some examples to identify under what conditions a PSNE exists. The examples illustrate how self-internalization of congestion by a large user can affect the nature of equilibrium and the travel costs that it and other users incur.

Jel Classifications: C61, C62, D43, D62, R41.

Keywords: departure-time decisions; bottleneck model; congestion; schedule delay costs; large users; user heterogeneity; existence of Nash equilibrium.

[☆]This research was partially funded by FONDECYT project No 11160294, the Complex Engineering Systems Institute (CONICYT – PIA – FB0816) and the Social Sciences and Humanities Research Council of Canada (Grant 435-2014-2050).

Email addresses: robin.lindsey@sauder.ubc.ca (Robin Lindsey), andre.depalma@ens-cachan.fr (André de Palma), husilva@uc.cl (Hugo E. Silva)

1. Introduction

Transportation congestion has been a growing problem for many years, and road traffic congestion is now a blight in most large cities worldwide. Couture et al. (2016) estimate that the deadweight loss from congestion is about US$30 billion per year in large US cities.[1] Hymel (2009) shows that high levels of congestion dampen employment growth, and that congestion pricing could yield substantial returns in restoring growth. Congestion delays are also a problem at airports, on rail lines, at seaports and in the hinterland of major transportation hubs. Ball et al. (2010) estimate that in 2007 air transportation delays in the US imposed a cost of US$25 billion on passengers and airlines.

Research on congestion dates back to Pigou (1920). Yet most economic and engineering models of congestible transportation facilities still assume that users are small in the sense that each one controls a negligible fraction of total traffic (see, e.g., Melo, 2014). This is a realistic assumption for passenger trips in private vehicles. Yet large users are prevalent in all modes of transport. They include major airlines at their hub airports, railways, maritime freight shippers, urban goods delivery companies, large taxi fleets and postal services. In some cases large users account for an appreciable fraction of traffic.[2] Furthermore, major employers such as government departments, large corporations, and transportation service providers can add substantially to traffic on certain roads at peak times.[3] So can large shopping centres, hotels, and major sporting events.[4]

Unlike small users, large users have an incentive to internalize the congestion delays

[1] Methods of estimating the costs of congestion differ, and results vary widely. The Texas Transportation Institute estimated that in 2014, congestion in 471 urban areas of the US caused approximately 6.9 billion hours of travel delay and 3.1 billion gallons of extra fuel consumption with an estimated total cost of US$160 billion (Schrank et al., 2017). It is unclear how institutional and technological innovations such as ridesharing, on-line shopping, electric vehicles, and automated vehicles will affect traffic volumes. The possibility that automated vehicles will increase congestion is raised in National Academies of Sciences, Engineering, and Medicine (2017) and The Economist (2018).

[2] For example, the world market for shipping is relatively concentrated. According to Statista (2017), as of December 31, 2017, the top five shipping operators accounted for 61.5% of the world liner fleet. The top ten accounted for 77.7%, and the top 15 for 85.5%. The top five port operators had a 29.9% global market share (Port Technology, 2014). The aviation industry is another example. The average market share of the largest firm in 59 major US airports during the period 2002-2012 was 42% (Choo, 2014). Similar shares exist in Europe.

[3] For example, Ghosal and Southworth (2017) describe how the Kia Motors Manufacturing plant, a large automobile assembler in West Point, Georgia, affects inbound and outbound transportation flows on highway and rail networks, and at seaports.

[4] Using data from US metropolitan areas with Major League Baseball (MLB) teams, Humphreys and Pyun (2017) estimate that attendance at MLB games increases average daily vehicle-miles traveled by about 6.9%, and traffic congestion by 2%.

their own vehicles impose on each other. This so-called "self-internalization" incentive can affect large users' decisions[5] and raises a number of interesting questions — some of which are discussed further in the conclusions. One is how much a large user gains from self-internalization. Can it backfire and leave the large user worse off after other users respond? Second, do other users gain or lose when one or more large users self-internalize? Does it depend on the size of the large users and when they prefer to schedule their traffic? Are mergers between large users welfare-improving? What about unions of small users that create a large user?

There is now a growing literature on large users and self-internalization – notably on large airlines and airport congestion. Nevertheless, this body of work is limited in two respects. First, as described in more detail below, most studies have used static models. Second, much of the theoretical literature has restricted attention to large users. In most settings, however, small users are also present. Automobile drivers and most other road users are small. Most airports serve not only scheduled commercial service, but also general aviation movements by recreational private aircraft and other non-scheduled users. Low-cost carriers with small market shares serve airports where large legacy airlines control much of the overall traffic.[6]

We contribute to the literature in this paper by developing and analyzing a dynamic model of congestion at a transportation facility with both large users and small users. More specifically, we use the Vickrey bottleneck model to study how large users schedule departure times for their vehicle fleets when small users use the facility too. As we explain in the literature review below, to the best of our knowledge, we are the first to study trip-timing decisions in markets with a mix of large and small users.

Several branches of literature have developed on large users of congestible facilities.[7] They include studies of route-choice decisions on road networks and flight scheduling at congested airports. There is also a literature directed to computer and telecommunications

[5]For example, some seaports alleviate congestion by extending operating hours at truck gates, and using truck reservation systems at their container facilities (Weisbrod and Fitzroy, 2011). Cities and travel companies are also attempting to spread tourist traffic by making off-peak visits more attractive and staggering the arrivals of cruise ships (Sheahan and Bryan, 2018). Airports, especially in Europe, restrict the number of landings and takeoffs during specific periods of time called slot windows (see Daniel (2014) for a discussion of this practice).

[6]Using data from Madrid and Barcelona, Fageda and Fernandez-Villadangos (2009) report that the market share of low-cost carriers is generally low (3-5 carriers with 3-18% of market share). Legacy carriers themselves sometimes operate only a few flights out of airports where another legacy carrier has a hub. For example, at Hartsfield-Jackson Atlanta International (ATL) American Airlines has a 3% market share while Delta's is 73% (Bureau of Transportation Statistics, 2017).

[7]See Silva et al. (2017) for a brief review.

networks on atomic congestion games. However, most studies have adopted static models that disregard the timing decisions of users despite the fact that congestion delays tend to be highly concentrated at peak times (see, e.g., Naroditskiy and Steinberg, 2015). The relatively small body of work that does address the temporal dimension of congestion has taken three approaches to incorporate dynamics. One approach has used dynamic stochastic models designed specifically to describe airport congestion (see, e.g., Daniel, 1995). A second approach, also directed at studying airport congestion, features deterministic congestion and a sequential decision-making structure in which an airline with market power acts as a Stackelberg leader and schedules its flights before other airlines in a competitive fringe (see Daniel, 2001; Silva et al., 2014). As Daniel (2014) discusses, the presence of slot constraints at airports makes the Stackelberg approach relevant. The slots are allocated twice a year with priority for the incumbent airlines; slots allocation for new entrants, which are modeled as followers in this approach, occur only after the incumbents have committed to a slot schedule, and normally come from new airport capacity. In these cases, adopting a sequential decision-making structure seems to be accurate. Nevertheless, at most US airports, the capacity is assigned in a first-come, first-served basis, which makes the simultaneous structure, and Nash as an equilibrium concept, more relevant.

These two approaches lead to outcomes broadly consistent with those of static models. Two results stand out. First, self-internalization of congestion by large users tends to result in less concentration of traffic at peak times, and consequently lower total costs for users in aggregate. Second, the presence of small users limits the ability of large users to reduce congestion. This is because reductions in the amount of traffic scheduled by large users, either at peak times or overall, are partially offset by increases in traffic by small users.

The Stackelberg equilibrium concept adopted in the second approach rests on the assumptions that the leader can schedule its traffic before other agents, and also commit itself to abide by its choices after other agents have made theirs. These assumptions are plausible in some institutional settings (e.g., Stackelberg leadership by legacy airlines at hub airports), but by no means in all settings. The third approach to incorporating trip-timing decisions, which we adopt, instead takes Nash equilibrium as the solution concept so that all users make decisions simultaneously.

Our paper follows up on recent work by Verhoef and Silva (2017) and Silva et al. (2017) who focus on determining under what conditions a Pure Strategy Nash Equilibrium (PSNE) in departure-time decisions exists. These two studies employ different deterministic congestion models that are best suited to describe road traffic congestion. Verhoef and Silva (2017) use the flow congestion model developed by Henderson (1974), and modified by Chu (1995). In this model, vehicles travel at a constant speed throughout their trips with the speed determined by the density of vehicles prevailing when their trip ends. Verhoef and

Silva (2017) show that, if there are two or more large users and no small users, a PSNE always exists. Self-internalization of congestion by the large users results in less concentration of trips at peak times and, not surprisingly, higher efficiency compared to the equilibrium without large users. However, this result is tempered by two well-known drawbacks of the Henderson-Chu model. First, vehicles departing at any given time never interact with vehicles departing at other times.[8] Second, compared to the bottleneck model discussed below, the Henderson-Chu model is less analytically tractable, and for most functional forms it can only be solved numerically.

The second paper to adopt Nash equilibrium, by Silva et al. (2017), uses the Vickrey (1969) bottleneck model in which congestion takes the form of queuing behind a bottleneck with a fixed flow capacity. Silva et al. consider two large users controlling identical vehicles with linear trip-timing preferences. In contrast to Verhoef and Silva (2017), Silva et al. find that under plausible parameter assumptions a PSNE in departure times does not exist. They also prove that a PSNE never exists in which large users queue. These results readily generalize to oligopolistic markets with more than two large users. Silva et al. also show that more than one PSNE may exist in which no queuing occurs, and that ex ante identical users can incur substantially different equilibrium costs. These results are disturbing given the fundamental importance of existence and uniqueness of equilibrium for equilibrium models. The unease is heightened by the facts that the bottleneck model is widely used, and that when all users are small a unique PSNE with a deterministic and finite departure rate exists under relatively unrestrictive assumptions.[9]

[8] In essence, this means that every infinitesimal cohort of vehicles travels independently of other cohorts and is unaffected by the volume of traffic that has departed earlier – contrary to what is observed in practice. The Henderson-Chu model is a special case of the Lighthill-Whitham-Richards hydrodynamic model in which shock waves travel at the same speed as vehicles and therefore never influence other vehicles (see Lindsey and Verhoef, 2007). Henderson (1974) originally assumed that vehicle speed is determined by the density of traffic encountered when a vehicle starts its trip. This formulation has the additional disadvantage that a vehicle departing when density is low may overtake a vehicle that departed earlier when density was higher. As Lindsey and Verhoef (2007) explain, overtaking has no behavioral basis if drivers and vehicles are identical, and it is physically impossible under heavily congested conditions. By contrast, in Chu's (1995) reformulated model overtaking does not occur in equilibrium.

[9] A few experimental economics studies have tested the theoretical predictions of the bottleneck model; see Dixit et al. (2015) for a review. The studies used a variant of the bottleneck model in which vehicles and departure times are both discrete. In all but one study, players controlled a single vehicle. The exception is Schneider and Weimann (2004) who ran two sets of experiments. In the first experiment each player controlled one vehicle, and in the second experiment each player controlled 10 vehicles which were referred to as trucks. Compared to the first experiment, the aggregate departure-time profile in the second experiment was further from the theoretical Nash equilibrium and closer to the system optimum. Schneider and Weimann conclude (p.151) that "players with 10 trucks internalize some of the congestion externality".

In this paper we extend Silva et al. (2017) by investigating the existence and nature of PSNE in the bottleneck model for a wider range of market structures and under more general assumptions about trip-timing preferences. Unlike both Verhoef and Silva (2017) and Silva et al. (2017), we allow for the presence of small users as well as large users. As in the standard bottleneck model, small users each control a single vehicle and seek to minimize their individual trip cost. Each large user operates a vehicle fleet that comprises a positive fraction or measure of total traffic, and seeks to minimize the aggregate trip costs of its fleet.[10] Each vehicle has trip-timing preferences described by a trip-cost function $C(t, a)$, where t denotes departure time and a denotes arrival time. Trip cost functions can differ for small and large users, and they can also differ for vehicles in a large user's fleet.

Our analysis consists of several parts. After introducing the basic model and assumptions in Section 2, in Section 3 we use optimal control theory to derive a large user's optimal fleet departure schedule as a best response to the aggregate departure rate profile of other users. We show that the optimal response can be indeterminate, and the second-order condition for an interior solution is generally violated. Consequently, a candidate PSNE departure schedule may exist in which a large user cannot gain by rescheduling any single vehicle in its fleet, yet it can gain by rescheduling a positive measure of vehicles. These difficulties underlie the non-existence of a PSNE in Silva et al. (2017). We then show in Section 4 that if vehicles in the large user's fleet have sufficiently diverse trip-timing preferences, a PSNE may exist in which some - or even all - of the large user's vehicles do queue. The fact that a PSNE exists given sufficient user heterogeneity parallels the existence of equilibrium in the Hotelling model of location choice given sufficient preference heterogeneity (de Palma et al., 1985).

Next, in Section 5 we revisit the case of symmetric large users that Silva et al. (2017) consider, and derive the minimum degree of preference heterogeneity required to support a PSNE. We show that relative to the PSNE in which large users disregard self-imposed congestion, self-internalization results in substantial efficiency gains from reduced queuing delays even when the number of users is fairly large. Then, in Section 6 we modify the example of symmetric large users by assuming that part of the traffic is controlled by a single large user, and the rest by a continuum of small users. We derive conditions for existence of

Unfortunately, they do not provide information on how the players distributed their vehicles over departure-time slots. Thus, it is not possible to compare their results with the predictions of our model as far as when large users choose to depart.

[10] In game theory, small agents or players are sometimes called "non-atomic" and large agents "atomic". In economics, the corresponding terms are "atomistic" and "non-atomistic". To avoid confusion, we do not use these terms. However, we do refer to the PSNE in which large users do not internalize their self-imposed congestion externalities as an "atomistic" PSNE.

a PSNE and show how the order in which users depart depends on the flexibility implied by the trip-timing preferences of large and small users. We also show that self-internalization of congestion can have no effect on the PSNE at all.

2. The model

The model is a variant of the classical bottleneck model.[11] All users travel from a common origin to a common destination along a single link that has a bottleneck with a fixed flow capacity of s. Without loss of generality, travel times from the origin to the bottleneck and from the bottleneck to the destination are normalized to zero. If there is no queue upstream of the bottleneck, travel time through the bottleneck is also zero and departure time from the origin coincides with arrival time at the destination. Let $r(t)$ denote the aggregate departure rate from the origin at time t, and $R(t)$ denote cumulative departures. If the departure rate exceeds s, a queue develops. Let $\widehat{t}$ be the most recent time at which there was no queue. The number of vehicles in the queue is then $Q(t) = R(t) - R(\widehat{t}) - s\left(t - \widehat{t}\right)$, and travel time through the bottleneck is $q(t) = Q(t)/s$, or

$$q(t) = \widehat{t} - t + s^{-1}\left(R(t) - R(\widehat{t})\right). \tag{1}$$

A user departing at time t arrives at time $a = t + q(t)$.

The cost of a trip is described by a function $C(t, a, k)$, where k denotes a user's index or type.[12] Function $C(t, a, k)$ is assumed to have the following properties:

Assumption 1: $C(t, a, k)$ is differentiable almost everywhere with derivatives $C_t < 0$, $C_a > 0$, $C_{tt} \geq 0$, $C_{tt} + C_{aa} > 0$, $C_{ta} = C_{at} = 0$, $C_{tk} \leq 0$, $C_{ak} \leq 0$, $C_{tkk} = 0$, and $C_{akk} = 0$.

The assumption $C_t < 0$ implies that a user prefers time spent at the origin to time spent in transit. Similarly, assumption $C_a > 0$ implies that a user prefers time spent at the destination to time spent in transit. User types can be defined in various ways. For much of the analysis, type is assumed to denote a user's preferred time to travel if a trip could be made instantaneously (i.e, with $a = t$). For type k, the preferred time is $t_k^* = Arg\min_t C(t, t, k)$. Given Assumption 1, t_k^* is unique. Types are ordered so that if $k > j$, $t_k^* \geq t_j^*$.

As explained in the Appendix, Assumption 1 is satisfied for various specifications of the cost function including the piecewise linear form introduced by Vickrey (1969):

[11] The bottleneck model is reviewed in Arnott et al. (1998), Small and Verhoef (2007), de Palma and Fosgerau (2011), and Small (2015). The exposition in this section draws heavily from Silva et al. (2017). Literal excerpts are not marked as such and are taken to be acknowledged by this footnote.

[12] As explained in the Appendix, the trip cost function can be derived from functions specifying the flow of utility or payoff received at the origin, at the destination, and while in transit.

$$C(t,a,k) = \begin{cases} \alpha_k (a-t) + \beta_k (t_k^* - a) & \text{for } a < t_k^* \\ \alpha_k (a-t) + \gamma_k (a - t_k^*) & \text{for } a > t_k^* \end{cases} . \tag{2}$$

In (2), parameter α_k is the unit cost of travel time, $\beta_k < \alpha_k$ is the unit cost of arriving early, and γ_k is the unit cost of arriving late. The first term in each branch of (2) denotes travel time costs, and the second term denotes schedule delay costs. We refer to this specification of costs as "step preferences".[13]

Because step preferences have a kink at t_k^*, the derivative C_a is discontinuous at $a = t_k^*$. This turns out to affect some of the results in this paper, and makes step preferences an exception to some of the propositions. It is therefore useful to know whether step preferences are reasonably descriptive of reality. Most studies that have used the bottleneck model have adopted step preferences, but this may be driven in part by analytical tractability and convention. Empirical evidence on the shape of the cost function is varied. Small (1982) found that step preferences describe morning commuter behaviour fairly well, but he did find evidence of discrete penalties for arriving late beyond a "margin of safety." Nonconvexities in schedule delay costs have been documented (e.g., Matsumoto, 1988), and there is some empirical evidence that the marginal cost of arriving early can exceed the marginal cost of travel time (Abkowitz, 1981a,b; Hendrickson and Plank, 1984; Tseng and Verhoef, 2008) which violates the assumption $\beta_k < \alpha_k$.

The paper features examples with step preferences where the results depend on the relative magnitudes of parameters α, β, and γ. Estimates in the literature differ, but most studies of automobile trips find that $\beta < \alpha < \gamma$. Small (1982, Table 2, Model 1) estimates ratios of β:α:γ = 1:1.64:3.9. These rates are representative of average estimates in later studies.[14] For benchmark values we adopt β:α:γ = 1:2:4.

In the standard bottleneck model, each user controls a single vehicle of measure zero and decides when it departs. A Pure Strategy Nash Equilibrium (PSNE) is a set of departure

[13] These preferences have also been called "$\alpha - \beta - \gamma$ " preferences.

[14] Estimates of the ratio γ/β vary widely. It is of the order of 8 in Geneva (Switzerland), and 4 in Brussels (Belgium), where tolerance for late arrival is much larger (see de Palma and Rochat, 1997; Khattak and de Palma, 1997). Tseng et al. (2005) obtain a ratio of 3.11 for the Netherlands. Peer et al. (2015) show that estimates derived from travel choices made in the short run can differ significantly from estimates derived from long-run choices when travelers have more flexibility to adjust their schedules. Most studies of trip-timing preferences have considered passenger trips. Many large users transport freight rather than people. Trip-timing preferences for freight transport can be governed by the shipper, the receiver, the transportation service provider, or some combination of agents. There is little empirical evidence for freight transportation on the relative values of α, β, and γ. The values are likely to depend on the type of commodity being transported, the importance of reliability in the supply chain, and other factors. Thus, it is wise to allow for a wide range of possible parameter values.

times for all users such that no user can benefit (i.e., reduce trip cost) by unilaterally changing departure time while taking other users' departure times as given. For brevity, the equilibrium will be called an "atomistic PSNE".

If small users are homogeneous (i.e., they all have the same type), then in a PSNE they depart during periods of queuing when the cost of a trip is constant. Their departure rate will be called their atomistic equilibrium departure rate, or "atomistic rate" for short. The atomistic rate for type k is derived from the condition that $C(t,a,k)$ is constant. Using subscripts to denote derivatives, this implies

$$C_t(t,a,k) + C_a(t,a,k)\left(1 + \frac{dq(t)}{dt}\right) = 0.$$

Given (1), the atomistic rate is

$$\hat{r}(t,a,k) = -\frac{C_t(t,a,k)}{C_a(t,a,k)}s. \tag{3}$$

Since $C_t < 0$ and $C_a > 0$, $\hat{r}(t,a,k) > 0$. Using Assumption 1, it is straightforward to establish the following properties of $\hat{r}(t,a,k)$:[15]

$$\frac{\partial \hat{r}(t,a,k)}{\partial k} \geq 0,\ \frac{\partial^2 \hat{r}(t,a,k)}{\partial k^2} \geq 0,\ Sgn\left(\frac{\partial \hat{r}(t,a,k)}{\partial a}\right) = -C_{aa}. \tag{4}$$

For given values of t and a, the atomistic rate increases with a user's type, and at an increasing rate. In addition, the atomistic rate is increasing with arrival time if $C_{aa} < 0$, and decreasing if $C_{aa} > 0$. With step preferences, $C_{aa} = 0$ except at t_k^*, and the atomistic rate is:

$$\hat{r}(t,a,k) = \begin{cases} \frac{\alpha_k}{\alpha_k - \beta_k}s \text{ for } a < t_k^* \\ \frac{\alpha_k}{\alpha_k + \gamma_k}s \text{ for } a > t_k^* \end{cases}. \tag{5}$$

Silva et al. (2017) consider a variant of the standard bottleneck model in which users are "large". A large user controls a vehicle fleet of positive measure, and internalizes the congestion costs its vehicles impose on each other. A PSNE entails a departure schedule for each user such that no user can reduce the total trip costs of its fleet by unilaterally changing its departure schedule while taking other users' departure schedules as given. We will call the equilibrium the "internalized PSNE". Silva et al. (2017) focus on the case of two large users with step preferences. In the next section we derive the departure schedule of a large user with general trip-timing preferences when other large users and/or small users may be departing too.

[15] See the Appendix.

3. Fleet departure scheduling and equilibrium conditions

3.1. Optimal departure schedule for a large user

This section uses optimal control theory to derive and characterize the optimal departure schedule of a large user with a fleet of vehicles. Call the large user "user A", and let N_A be the measure of vehicles in its fleet. Vehicle k has a cost function $C^A(t, a, k)$. Vehicles are indexed in order of increasing preferred arrival times so that $t_k^* \geq t_j^*$ if $k > j$. It is assumed that, regardless of the queuing profile, it is optimal for user A to schedule vehicles in order of increasing k. The departure schedule for user A can then be written $r_A(t)$ with argument k suppressed. If $R_A(t)$ denotes cumulative departures of A, vehicle $k = R_A(t)$ departs at time t.

User A chooses $r_A(t)$ to minimize the total trip costs of its fleet while taking as given the aggregate departure rate of other users, $r_{-A}(t)$. Trips are assumed to be splittable: they can be scheduled at disjoint times (e.g., some vehicles can travel early in the morning while others travel at midday). Let t_{As} and t_{Ae} denote the first and last departure times chosen by user A. User A's optimization problem can be stated as:

$$\underset{t_{As}, t_{Ae}, r_A(t)}{Min} \int_{t=t_{As}}^{t_{Ae}} r_A(t)\, C^A(t, t + q(t), R_A(t))\, dt, \tag{6}$$

subject to the equations of motion:

$$\frac{dq(t)}{dt^+} = \begin{cases} s^{-1}(r_{-A}(t) + r_A(t)) - 1 \text{ if } q(t) > 0 \text{ or } r_{-A}(t) + r_A(t) > s \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{7}$$

(costate variable $\lambda(t) \geq 0$), and

$$\frac{dR_A(t)}{dt} = r_A(t) \qquad \text{(costate variable } \mu(t)\text{)}, \tag{8}$$

and the following constraints[16]:

$$r_A(t) \geq 0 \quad \text{(multiplier } \xi(t) \geq 0), \tag{9a}$$

$$R_A(t_{As}) = 0, \quad R_A(t_{Ae}) = N_A, \tag{9b}$$

$$q(t_{As}) = \bar{q}(t_{As}) \quad \text{(multiplier } \phi), \tag{9c}$$

$$t_{As},\ t_{Ae} \quad \text{chosen freely.} \tag{9d}$$

Costate variable $\lambda(t)$ for Eq. (7) measures the shadow cost to user A of queuing time. Eq. (8) governs how many vehicles in user A's fleet have left the origin. Costate variable

[16]The nonnegativity constraint on queuing time, $q(t) \geq 0$, is guaranteed by (7).

$\mu(t)$ measures the shadow cost of increasing the number of vehicles in the fleet that have started their trips. Condition (9a) stipulates that the departure rate cannot be negative. Condition (9b) specifies initial and terminal values for cumulative departures. Condition (9c) describes how queuing time is evolving when departures begin. Finally, (9d) indicates that the choice of departure period is unconstrained.

The Hamiltonian for the optimization problem is

$$H(t) = r_A(t) C^A(t, t + q(t), R_A(t)) + \mu(t) \frac{dR_A(t)}{dt} + \lambda(t) \frac{dq(t)}{dt^+}, \tag{10}$$

and the Lagrangian is

$$L(t) = H(t) + r_A(t) \xi(t). \tag{11}$$

Costate variable $\lambda(t)$ for queuing time evolves according to the equation of motion

$$\frac{d\lambda(t)}{dt} = -\frac{\partial H}{\partial q} = -r_A(t) C_a^A(t, t + q(t), R_A(t)) \leq 0. \tag{12}$$

Variable $\lambda(t)$ decreases as successive vehicles in the fleet depart because fewer vehicles remain that can be delayed by queuing.

Costate variable $\mu(t)$ for cumulative departures evolves according to the equation of motion

$$\frac{d\mu(t)}{dt} = -\frac{\partial H}{\partial R_A} = -r_A(t) C_k^A(t, t + q(t), R_A(t)) \geq 0. \tag{13}$$

If vehicles in the fleet are homogeneous, μ is independent of time.

With t_{Ae} chosen freely, transversality conditions at t_{Ae} are:

$$\lambda(t_{Ae}) = 0, \tag{14}$$

$$H(t_{Ae}) = 0. \tag{15}$$

According to condition (14), the shadow cost of queuing time drops to zero when the last vehicle departs. Condition (15) dictates that the net flow of cost is zero when the last vehicle departs. Substituting (14) into (10), and applying (15) yields

$$\mu(t_{Ae}) = -C^A(t_{Ae}, t_{Ae} + q(t_{Ae}), N_A). \tag{16}$$

Condition (16) states that the benefit from dispatching the last vehicle in the fleet is the cost of its trip that has now been incurred, and is no longer a pending liability.

With t_{As} chosen freely, a transversality condition also applies at t_{As}. Following Theorem 7.8.1 in Leonard and Van Long (1992), the transversality condition is:

$$H(t_{As}) - \phi \left. \frac{dq(t)}{dt} \right|_{t_{As}} = 0, \tag{17}$$

where ϕ is a multiplier on the constraint (9c). By continuity, $\phi = \lambda(t_{As})$. Using (10) and (8), condition (17) reduces to

$$r_A(t_{As})\left(C^A(t_{As}, t_{As} + q(t_{As}), 0) + \mu(t_{As})\right) = 0. \tag{18}$$

It remains to determine the optimal path of $r_A(t)$. The optimality conditions governing $r_A(t)$ depend on whether or not there is a queue. Attention is limited here to the case with a queue.[17] If $q(t) > 0$, the optimal departure rate is governed by the conditions

$$\frac{\partial L}{\partial r_A(t)} = C^A(t, t + q(t), R_A(t)) + \xi(t) + \mu(t) + \frac{\lambda(t)}{s} = 0, \tag{19}$$

$$\xi(t) r_A(t) = 0.$$

If $r_A(t)$ is positive and finite during an open time interval containing t, then $\xi(t) = 0$ and (19) can be differentiated with respect to t:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial r_A(t)}\right) = C_t^A(t, t + q(t), R_A(t)) + C_a^A(t, t + q(t), R_A(t))\left(1 + \frac{dq(t)}{dt^+}\right)$$
$$+C_k^A(t, t + q(t), R_A(t)) r_A(t) + \frac{d\mu(t)}{dt} + \frac{1}{s}\frac{d\lambda(t)}{dt} = 0.$$

Using Eqs. (7) and (12), this condition simplifies to

$$C_t^A(t, t + q(t), R_A(t)) + C_a^A(t, t + q(t), R_A(t))\frac{r_{-A}(t)}{s} = 0. \tag{20}$$

The left-hand-side of (20) depends on the aggregate departure rate of other users, $r_{-A}(t)$, but not on $r_A(t)$ itself. In general, derivatives $C_t^A(t, t + q(t), R_A(t))$ and $C_a^A(t, t + q(t), R_A(t))$ depend on the value of $q(t)$, and hence the value of $R(t)$, but not directly on $r_A(t)$. Condition (20) will therefore not, in general, be satisfied regardless of user A's choice of $r_A(t)$. This implies that the optimal departure rate may follow a bang-bang solution between zero flow and a mass departure.[18] This is confirmed by inspecting the Hessian matrix of the Hamiltonian:

$$\begin{bmatrix} \frac{\partial^2 H}{\partial r_A^2(t)} & \frac{\partial^2 H}{\partial r_A(t)\partial q(t)} & \frac{\partial^2 H}{\partial r_A(t)\partial R_A(t)} \\ \frac{\partial^2 H}{\partial r_A(t)\partial q(t)} & \frac{\partial^2 H}{\partial q^2(t)} & \frac{\partial^2 H}{\partial q(t)\partial R_A(t)} \\ \frac{\partial^2 H}{\partial r_A(t)\partial R_A(t)} & \frac{\partial^2 H}{\partial q(t)\partial R_A(t)} & \frac{\partial^2 H}{\partial R_A^2(t)} \end{bmatrix} =$$

$$\begin{bmatrix} 0 & C_a^A(t, t + q(t), R_A(t)) & C_k^A(t, t + q(t), R_A(t)) \\ C_a^A(t, t + q(t), R_A(t)) & r_A(t) C_{aa}^A(t, t + q(t), R_A(t)) & r_A(t) C_{ak}^A(t, t + q(t), R_A(t)) \\ C_k^A(t, t + q(t), R_A(t)) & r_A(t) C_{ak}^A(t, t + q(t), R_A(t)) & r_A(t) C_{kk}^A(t, t + q(t), R_A(t)) \end{bmatrix}.$$

[17] The optimality conditions with no queue, which involve multiple cases, are not very instructive.

[18] See Leonard and Long (1992, Chapter 8).

Since the Hessian is not positive definite, the second-order sufficient conditions for a local minimum are not satisfied. As we will show, if users are homogeneous the necessary condition (20) cannot describe the optimal schedule unless $C_{aa}^{A} = 0$.

In summary, user A will not, in general, depart at a positive and finite rate when a queue exists. To understand why, consider condition (20). Given $C_t^A < 0$ and $C_a^A > 0$, if $r_{-A}(t)$ is "small" the left-hand side of (20) is negative. The net cost of a trip is decreasing over time, and user A is better off scheduling the next vehicle in its fleet later. Contrarily, if $r_{-A}(t)$ is "large", the left-hand side of (20) is positive. Trip cost is increasing, and user A should dispatch a mass of vehicles immediately if it has not already done so. In either case, the optimal departure rate is not positive and finite.

In certain cases, described in the next section, condition (20) will be satisfied. The condition can then be written as a formula for the departure rate of other users:

$$r_{-A}(t) = -\frac{C_t^A\left(t, t+q(t), R_A(t)\right)}{C_a^A\left(t, t+q(t), R_A(t)\right)} s \equiv \widehat{r}_A\left(t, t+q(t), R_A(t)\right). \tag{21}$$

Condition (21) has the same functional form as Eq. (3) for the atomistic rate of small users. Thus, with step preferences, the right-hand side exceeds s for early arrival and is less than s for late arrival. Moreover, the condition depends only on the aggregate departure rate of other users, and not their composition (e.g., whether the other users who are departing are large or small. However, condition (21) is only necessary, not sufficient, to have $r_A(t) > 0$ because the second-order conditions are not satisfied. This leads to**:**

Lemma 1. *Assume that a queue exists at time t. A large user will not depart at a positive and finite rate at time t unless the aggregate departure rate of other users equals the large user's atomistic rate given in Eq. (21).*

Lemma 1 requires qualification in the case of step preferences because the atomistic rate is discontinuous at the preferred arrival time. If vehicles in a large user's fleet differ sufficiently in their individual t_k^*, it is possible to have a PSNE in which the fleet departs at a positive and finite rate with each vehicle arriving exactly on time. The aggregate departure rate of other users falls short of the atomistic rate of each vehicle in the fleet just before it departs, and exceeds it just after it departs. This is illustrated using an example in Section 6.

3.2. Equilibrium conditions with large users

We now explore the implications of Lemma 1 for the existence of a PSNE in which a large user departs when there is a queue and the atomistic rates of all users are continuous. Conditions for a PSNE depend on whether or not small users are present, and the two cases are considered separately below.

3.2.1. Multiple large users and no small users

Suppose there are $m \geq 2$ large users and no small users. User i has an atomistic rate $\hat{r}_i(t, t + q(t), R_i(t))$. For brevity, we write this as $\hat{r}_i(t)$ with arrival time and the index k for vehicles both suppressed. Suppose that a queue exists at time t, and user i departs at rate $r_i(t) > 0$, $i = 1...m$.[19] Necessary conditions for a PSNE to exist are

$$r_{-i}(t) = \hat{r}_i(t), \; i = 1...m. \tag{22}$$

This system of m equations has a solution

$$\begin{aligned} r_i(t) &= \frac{1}{m-1}\textstyle\sum_{j \neq i} \hat{r}_j(t) - \frac{m-2}{m-1}\hat{r}_i(t) \\ &= \frac{1}{m-1}\textstyle\sum_{j} \hat{r}_j(t) - \hat{r}_i(t), \; i = 1...m. \end{aligned} \tag{23}$$

With $m = 2$, the solution is $r_1(t) = \hat{r}_2(t)$, and $r_2(t) = \hat{r}_1(t)$. With $m > 2$, the solution is feasible only if all departure rates are nonnegative. A necessary and sufficient condition for this to hold at time t is

$$\underset{i}{Max}\ \hat{r}_i(t) \leq \frac{1}{m-2}\textstyle\sum_{j \neq i} \hat{r}_j(t). \tag{24}$$

Condition (24) is satisfied if large users have sufficiently similar atomistic rates.

3.2.2. Multiple large users and small users

Assume now that, in addition to $m \geq 1$ large users, there is a group of homogeneous small users comprising a positive measure of total traffic with an atomistic rate $\hat{r}_o(t)$. Suppose that large user i departs at rate $r_i(t) > 0$, $i = 1...m$, and small users depart at an aggregate rate $r_o(t) > 0$. If a queue exists at time t, necessary conditions for a PSNE are

$$r_{-i}(t) = \hat{r}_i(t), \; i = 1...m, \tag{25}$$

$$\textstyle\sum_j r_j(t) + r_o(t) = \hat{r}_o(t). \tag{26}$$

The solution to this system of $m + 1$ equations is

$$r_i(t) = \hat{r}_o(t) - \hat{r}_i(t), \; i = 1...m, \tag{27}$$

$$r_o(t) = \textstyle\sum_j \hat{r}_j(t) - (m-1)\hat{r}_o(t). \tag{28}$$

The solution is feasible only if all departure rates are nonnegative. With $m = 1$, the necessary and sufficient condition is $\hat{r}_1(t) < \hat{r}_0(t)$. With $m > 1$, necessary and sufficient

[19] If a user does not depart at time t, it can be omitted from the set of m "active" users at t.

conditions for nonnegativity are

$$\frac{1}{m-1}\sum_j \hat{r}_j(t) > \hat{r}_o(t), \tag{29}$$

$$\hat{r}_i(t) < \hat{r}_o(t), \; i = 1...m. \tag{30}$$

Condition (30) requires that all large users have lower atomistic rates than the small users. However, condition (29) dictates that the average atomistic rate for large users be close enough to the atomistic rate of small users. Together, (29) and (30) impose relatively tight bounds on the $\hat{r}_i(t)$.

4. Existence of PSNE with queuing by a large user

Silva et al. (2017) consider two identical large users with homogeneous vehicle fleets and step preferences. They show that a PSNE with queuing does not exist. In addition, they show that if $\gamma > \alpha$, a PSNE without queuing does not exist either so that no PSNE exists. In this section we build on their results in two directions. First, we prove that if a large user has a homogeneous vehicle fleet, and $C^A_{aa} \neq 0$ at any time when the large user's vehicles arrive, a PSNE in which the large user queues does not exist for any market structure. Second, we show that if a large user has a heterogeneous vehicle fleet, and the derivative C^A_{ak} is sufficiently large in magnitude, a PSNE in which the large user queues is possible. We illustrate the second result in Section 5.

Consider a large user, "user A", and a candidate PSNE in which queuing time is $\bar{q}(t) > 0$ at time t. (A bar denotes quantities in the candidate PSNE.) User A never departs alone when there is a queue because it can reduce its fleet costs by postponing departures. Thus, if $\bar{r}_A(t)$ is positive and finite, other users must also be departing. The aggregate departure rate of other users must equal user A's atomistic rate as per Eq. (22) or (25):

$$\bar{r}_{-A}(t) = \hat{r}_A\left(t, t + \bar{q}(t), \bar{R}_A(t)\right).$$

In addition, user A must depart at a rate $\bar{r}_A(t)$ consistent with equilibrium for other users as per Eq. (23), or Eqs. (27) and (28). Figure 1 depicts a candidate PSNE on the assumption that $C^A_{aa} > 0$. (The case $C^A_{aa} < 0$ is considered below.) Cumulative departures of other users, $\bar{R}_{-A}(t)$, are shown by the blue curve passing through points y and z. Cumulative total departures, $\bar{R}(t)$, are shown by the black curve passing through points A, D and B. Cumulative departures of user A, $\bar{R}_A(t) = \bar{R}(t) - \bar{R}_{-A}(t)$, are measured by the distance between the two curves.

Suppose that user A deviates from the candidate PSNE during the interval (t_A, t_B) by dispatching its vehicles slightly later so that section ADB of $\bar{R}(t)$ shifts rightwards to $\tilde{R}(t)$

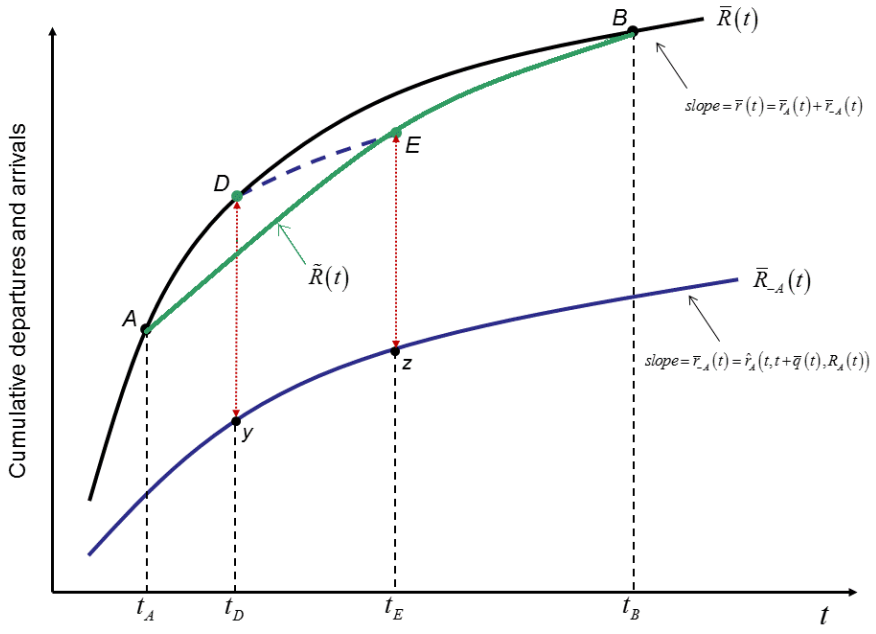

Figure 1: Candidate PSNE with $C^A_{aa} > 0$

shown by the green curve. The rescheduled vehicles still depart in order of increasing k.[20] Vehicle $k = \bar{R}_A(t_D)$ that is initially scheduled to depart at point D and time t_D is therefore rescheduled to point E and time t_E such that distance Ez equals distance Dy. Vehicle k experiences a change in cost of

$$\Delta C^A(k) = C^A(t_E, t_E + \tilde{q}(t_E), k) - C^A(t_D, t_D + \bar{q}(t_D), k),$$

where $\bar{q}(t_D)$ is queuing time at t_D with the candidate equilibrium departure schedule $\bar{R}(t)$, and $\tilde{q}(t_E)$ is queuing time at t_E with the deviated schedule $\tilde{R}(t)$. The path from point D to point E can be traversed along the dashed blue curve running parallel to $\bar{R}_{-A}(t)$ between points y and z. Let $\check{q}(t)$ denote queuing time along this path. The change in cost can then

[20] If vehicles are homogeneous, the order of departure does not matter. The assumption that they depart in the same order is useful for accounting purposes.

be written

$$\Delta C^A(k) = \int_{t=t_D}^{t_E} \left(C_t^A\left(t, t+\check{q}(t), k\right) + C_a^A\left(t, t+\check{q}(t), k\right)\left(1+\frac{d\check{q}(t)}{dt}\right)\right) dt.$$

$$= \int_{t=t_D}^{t_E} \left(C_t^A\left(t, t+\check{q}(t), k\right) + C_a^A\left(t, t+\check{q}(t), k\right)\frac{\widehat{r}_A\left(t, t+\bar{q}(t), \bar{R}_A(t)\right)}{s}\right) dt$$

$$= \frac{1}{s}\int_{t=t_D}^{t_E} C_a^A\left(t, t+\check{q}(t), k\right)\left\{\widehat{r}_A\left(t, t+\bar{q}(t), \bar{R}_A(t)\right) - \widehat{r}_A\left(t, t+\check{q}(t), k\right)\right\} dt$$

$$= \frac{1}{s}\int_{t=t_D}^{t_E} C_a^A\left(t, t+\check{q}(t), k\right)\left\{\begin{array}{c}\left(\widehat{r}_A\left(t, t+\bar{q}(t), \bar{R}_A(t)\right) - \widehat{r}_A\left(t, t+\bar{q}(t), k\right)\right)\\ -\left(\widehat{r}_A\left(t, t+\check{q}(t), k\right) - \widehat{r}_A\left(t, t+\bar{q}(t), k\right)\right)\end{array}\right\} dt. \quad (31)$$

The sign of this expression depends on how $\widehat{r}_A$ varies with arrival time and vehicle index. We begin by showing that, if vehicles are homogeneous, (31) is negative so that $\Delta C^A(k) < 0$ and the candidate is not a PSNE.

4.1. Homogeneous vehicle fleets

If user A has a homogeneous fleet, the first line in braces in (31) is zero. Given $C_{aa}^A > 0$ and $\check{q}(t) < \bar{q}(t)$ for $t \in (t_D, t_E)$, $\hat{r}_A\left(t, t+\check{q}(t), k\right) > \hat{r}_A\left(t, t+\bar{q}(t), k\right)$ and the second line in braces is negative. Hence $\Delta C^A(k) < 0$, and rescheduling the vehicle from D to E reduces its trip cost. Since point D is representative of all points between A and B, all the rescheduled vehicles except those at the endpoints, A and B, experience a reduction in costs. User A therefore gains from the deviation, and the candidate schedule is not a PSNE. In the Appendix we show that if $C_{aa}^A < 0$, user A can benefit by accelerating departures of its fleet. Deviation is therefore beneficial both when $C_{aa}^A > 0$ and when $C_{aa}^A < 0$. This result is formalized in

Lemma 2. *Consider large user A with a homogeneous vehicle fleet. If a queue exists at time t, and $C_{aa}^A(t, t+q(t)) \neq 0$, user A will not depart at a positive and finite rate at time t.*

Lemma 2 shows that although the candidate PSNE is robust to deviations in the departure time of a single vehicle, it is not robust to deviations by a positive measure of the fleet. If $C_{aa}^A > 0$, the departure rate of other users must decrease over time in order for user A to maintain a positive and finite departure rate. By delaying departures, user A enables vehicles in its fleet to benefit from shorter queuing delays. Conversely, if $C_{aa}^A < 0$, the departure rate of other users must increase over time in a PSNE, and user A can benefit by accelerating departures of its fleet.

Lemma 2 contrasts sharply with the results of Verhoef and Silva (2017) who show that, given a set of large users with homogeneous vehicle fleets, a PSNE always exists in the Henderson-Chu model. As noted in the introduction, in the Henderson-Chu model vehicles that arrive (or depart) at different times do not interact with other. In particular, a cohort of vehicles departing at time t is unaffected by the number or density of vehicles that departed before t. Thus, if a large user increases or decreases the departure rate of its fleet at time t, it does not affect the costs incurred by other vehicles in the fleet that are scheduled after t. Equilibrium is determined on a point-by-point basis, and there is no state variable analogous to the queue in the bottleneck model that creates intertemporal dependence in costs.

4.2. Heterogeneous vehicle fleets

Suppose now that user A has a heterogeneous fleet. By (4), $\partial \hat{r}_A(t,a,k)/\partial k \geq 0$ so that the first line in braces in (31) is positive. Expression (31) is then positive if the first line outweighs the second line. We show that this is indeed the case under plausible assumptions. Towards this, we introduce the following two-part assumption:

Assumption 2: (i) The trip cost function depends only on the difference between actual arrival time and desired arrival time, and thus can be written $C^A(t,a,k) = C^A(t, a - t_k^*)$. (ii) t_k^* is distributed according to a density function $f(t_k^*)$ over a range $[t_s^*, t_e^*]$.

The following result is proved in the Appendix:

Theorem 1. *Consider large user A with a heterogeneous vehicle fleet that satisfies Assumption 2. If the density of desired arrival times in user A's fleet never exceeds bottleneck capacity (i.e., $f(t_k^*) \leq s \; \forall \; t_k^* \in [t_s^*, t_e^*]$), a PSNE in which user A queues may exist.*

Theorem 1 identifies necessary conditions such that a large user may queue in a PSNE. In light of Lemma 2 the key requirement is evidently sufficient heterogeneity in the trip-timing preferences of vehicles in the large user's fleet. Condition $f(t_k^*) \leq s$ stipulates that the desired arrival rate of vehicles in the fleet never exceeds bottleneck capacity. Put another way, if user A were the only user of the bottleneck, it could schedule its fleet so that every vehicle arrived precisely on time without queuing delay.

The assumption $f(t_k^*) \leq s$ is plausible for road transport. Freight shippers such as Fedex or UPS operate large vehicle fleets out of airports and central warehouses, and they can make hundreds of daily trips on highways and connecting roads in an urban area. Nevertheless, deliveries are typically made to geographically dispersed customers throughout the day so that the fleet rarely comprises more than a small fraction of total

traffic on a link at any given time. Thus, for any t_k^*, $f(t_k^*)$ is likely to be only a modest fraction of s.

In concluding this section it should be emphasized that Theorem 1 only states that a PSNE in which a large user queues *may* exist. A large user may prefer to avoid queuing by traveling at off-peak times. To determine whether this is the case, it is necessary to consider the trip-timing preferences of all users. We do so in Section 5 for the case of large users studied by Silva et al. (2017). Section 6 examines a variant with both large users and small users.

5. Existence of PSNE and self-internalization: multiple large users

In this section we analyze the existence of PSNE with $m \geq 2$ symmetric large users. We begin with $m = 2$: the case considered by Silva et al. (2017). Consider two symmetric large users, A and B, that each controls $N/2$ vehicles with step preferences. Such a market setting might arise with two airlines that operate all (or most of) the flights at a congested airport. This section revisits Proposition 1 in Silva et al. (2017) which states that a PSNE does not exist with homogeneous vehicles when $\gamma > \alpha$. Their proof entails showing that with $\gamma > \alpha$, a PSNE without queuing does not exist. The proof that a PSNE with queuing does not exist either follows the general reasoning used to prove Lemma 2 above. Here we relax the assumption that vehicles are homogeneous, and suppose that in each vehicle fleet, t_k^* is uniformly distributed with a density $f(t_k^*) = N/(2\Delta)$ over the interval $[t_s^*, t_e^*]$ where $\Delta \equiv t_e^* - t_s^*$. It can be shown that introducing heterogeneity in this way does not upset the proof in Silva et al. (2017) that a PSNE without queuing does not exist. However, a PSNE with queuing does exist if the conditions of Theorem 1 are met. Both conditions of Assumption 2 are satisfied. The remaining condition, $f(t_k^*) \leq s$, is also met if $N/(2\Delta) \leq s$, or $\Delta \geq N/(2s)$.

The candidate PSNE with queuing is shown in Figure 2.[21] The cumulative distribution of desired arrival times for users A and B together is shown by the straight line W with domain $[t_s^*, t_e^*]$ and height N. The two users schedule vehicles at the same rate. During the initial interval (t_s, t_q), both users depart at an aggregate rate of s without creating a queue. Queuing begins at time t_q, and ends at t_e when the last vehicle in each fleet departs. Queuing time reaches a maximum at $\tilde{t}$ for a vehicle arriving at $\hat{t}^* = \frac{\beta}{\beta+\gamma} t_s^* + \frac{\gamma}{\beta+\gamma} t_e^*$. Total departures after time t_q are shown by the piecewise linear curve ALC. Cumulative departures by user B starting at t_q are given by the piecewise linear curve $AP'E$, and cumulative departures by user A are measured by the distance between $AP'E$ and ALC.

[21] Figure 2 is a variant of Figure 2 in Silva et al. (2017). The main difference is that desired arrival times have a nondegenerate distribution rather than being the same for all vehicles.

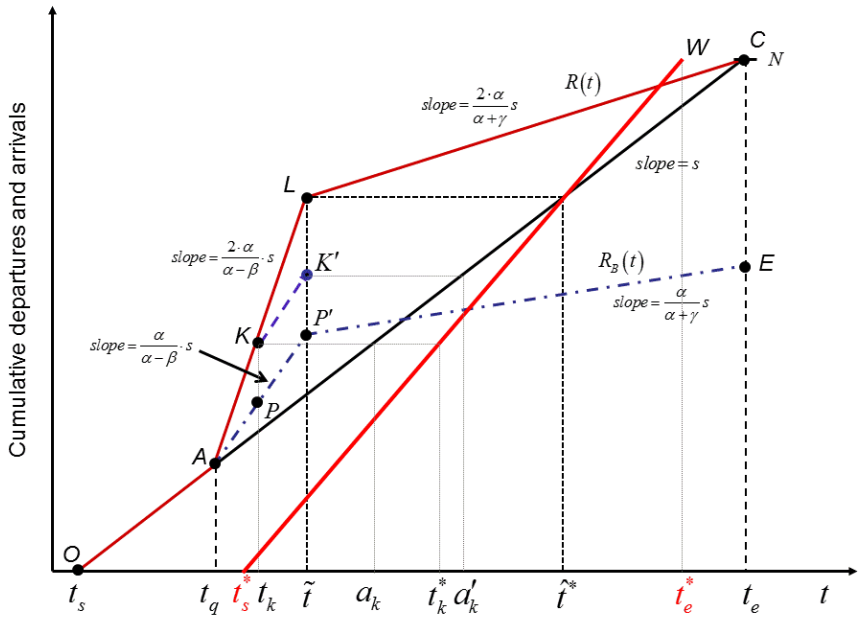

Figure 2: PSNE with two large users

Silva et al. (2017) show that with homogeneous vehicles user A can benefit by deviating from the candidate PSNE and rescheduling part of its fleet later. To see whether this is possible in Figure 2, consider vehicle k in user A's fleet. This vehicle is scheduled to depart at time t_k, when cumulative departures have reached point K, and arrive early at $a_k < t_k^*$. To benefit from rescheduling, vehicle k has to depart after $\tilde{t}$ when user B has decreased its departure rate from $\frac{\alpha}{\alpha-\beta}s$ to $\frac{\alpha}{\alpha+\gamma}s$. Vehicle k therefore has to depart after cumulative departures have reached point K' where distance $K'P'$ equals distance KP. Vehicle k can benefit if it still arrives early. This is possible if all vehicles had a common desired arrival time of $\hat{t}^*$. But, as shown in the Appendix, it is not possible with the distribution of t^* shown, with $\Delta = t_e^* - t_s^* > N/(2s)$, because vehicle k will arrive after time $a_k' > t_k^*$ when it is late. Vehicle k thus cannot take advantage of the drop in user B's departure rate at $\tilde{t}$. The same reasoning applies to all vehicles in user A's fleet. Deviation is therefore unprofitable, and the candidate PSNE is a PSNE.

The existence of a PSNE contrasts with Silva et al. (2017) who show that no PSNE exists if $\gamma > \alpha$ and vehicles are homogeneous. We formalize this result in the following proposition which serves as a counterpart to Proposition 1 in Silva et al. (2017).

Proposition 1. *Consider two identical large users that simultaneously schedule $N/2$ vehi-*

cles each with unit costs α, β, and γ, with $\gamma > \alpha$. Desired arrival times in each fleet are uniformly distributed with a range Δ. If $\Delta > N/(2s)$, a unique PSNE exists that is shown in Figure 2.

The analysis is readily extended to $m > 2$ symmetric users. The necessary and sufficient condition for a PSNE with queuing to exist is $\Delta > N/(ms)$. This condition becomes less stringent the larger is m. The counterpart to Proposition 1 with m users is stated as:

Corollary 1. *Consider m identical large users that simultaneously schedule N/m vehicles each with unit costs α, β, and γ, with $\gamma > \alpha$. Desired arrival times in each fleet are uniformly distributed with a range Δ. If $\Delta > N/(ms)$, a unique PSNE exists.*

If a PSNE exists, total costs in the internalized PSNE, TC^i, are lower than total costs in the atomistic PSNE, TC^n. As shown in the Appendix, the total cost saving from internalization with m users is

$$TC^n - TC^i = \underbrace{\frac{(m-1)\,\beta\,(\alpha+\gamma) + m\alpha\gamma}{2m\,(m-1)\,\beta\gamma + 2m\alpha\gamma}}_{\Psi} \cdot TC^{nH},$$

where $TC^{nH} = \frac{\beta\gamma}{\beta+\gamma}\frac{N^2}{s}$ denotes total costs in the atomistic PSNE with homogeneous vehicles. The composite parameter Ψ depends on parameters α, β, and γ only through the ratios β/α and γ/α. Given the benchmark ratios of $\beta{:}\alpha{:}\gamma$ = 1:2:4, $\Psi = \frac{7m-3}{4m(1+m)}$, which varies with m as shown in Table 1.

m	1	2	3	4	5	10	20	*Large*
Ψ	0.5	0.458	0.375	0.313	0.267	0.152	0.081	$\simeq 7/(4m)$

Table 1: Proportional cost saving from internalization as a function of m

With two users ($m = 2$) the saving is nearly as great as with a single user. Even with 10 users the savings is over 15 percent of the atomistic costs TC^{nH}. These results are similar to those obtained by Verhoef and Silva (2017) with the Henderson-Chu model of congestion and a single desired arrival time, as they also find significant savings from self-internalization. Moreover, with heterogeneity in t^*, total costs in the atomistic PSNE are less than TC^{nH} so that the proportional cost saving from internalization is actually larger than shown in Table 1. The example shows that self-internalization of congestion can boost efficiency appreciably even if no user controls a large fraction of total traffic. This is consistent with Brueckner (2002) who showed, using a Cournot oligopoly model, that internalization of self-imposed delays leads to an equilibrium that is more efficient than the atomistic equilibrium, and correspondingly offers smaller potential efficiency gains from congestion pricing.

6. Existence of PSNE and self-internalization: large and small users

In this section we modify the example in Section 5. We now assume that traffic is controlled by one large user, user A, with a vehicle fleet of measure N_A, and a group of homogeneous small users with a measure N_o. For ease of reference, vehicles in user A's fleet are called "large vehicles" and vehicles driven by small users are called "small vehicles". Large vehicles have the same trip-timing preferences as in Section 5. Their unit costs are denoted by α_A, β_A, and γ_A. Their desired arrival times are uniformly distributed over the interval $[t_s^*, t_e^*]$ with a range of $\Delta \equiv t_e^* - t_s^*$. For future use we define $\delta \equiv s\Delta/N_A$.

The existence and nature of PSNE depend on how the trip-timing preferences of small vehicles compare with those of large vehicles. We adopt a specification that allows the preferences to be either the same, or different in a plausible and interesting way. Small vehicles have step preferences with unit costs of α, β, and γ. The cost of late arrival relative to early arrival is assumed to be the same as for large vehicles so that $\gamma/\beta = \gamma_A/\beta_A$. The distribution of desired arrival times is also the same as for large vehicles.[22]

Small vehicles and large vehicles are allowed to differ in the values of β/α and β_A/α_A. The ratio β_A/α_A measures the cost of schedule delay relative to queuing time delay for large vehicles. It determines their flexibility with respect to arrival time, and hence their willingness to queue to arrive closer to their desired time. If β_A/α_A is small, large vehicles are flexible in the sense that they are willing to reschedule trips in order to avoid queuing delay. Conversely, if β_A/α_A is big, large vehicles are inflexible. Ratio β/α has an analogous interpretation for small vehicles. To economize on writing, we use the composite parameter $\theta \equiv \frac{\beta_A/\alpha_A}{\beta/\alpha}$ to measure the relative flexibility of the two types.

We consider two cases. In Case 1, $\theta \leq 1$ so that large vehicles are (weakly) more flexible than small vehicles. To fix ideas, small vehicles can be thought of as morning commuters with fixed work hours and relatively rigid schedules. Large vehicles are small trucks or vans that can make deliveries within a broad time window during the day. We show below that for a range of parameter values, a PSNE exists in which large vehicles depart at the beginning and end of the travel period without queuing. Small vehicles queue in the middle of the travel period in the same way as if large vehicles were absent.

In Case 2, $\theta > 1$ so that large vehicles are less flexible than small vehicles. This would be the case if large vehicles are part of a just-in-time supply chain, or have to deliver

[22] Within limits, this assumption can be relaxed. Suppose that t^* is uniformly distributed over the interval $[t_{so}^*, t_{eo}^*]$. The existence and nature of PSNE with self-internalization are unaffected if two conditions are satisfied. First, $t_{eo}^* - t_{so}^* \leq N_o/s$. This condition assures that small vehicles queue in the PSNE. Second, $\frac{\beta}{\beta+\gamma} t_{so}^* + \frac{\gamma}{\beta+\gamma} t_{eo}^* = \frac{\beta}{\beta+\gamma} t_s^* + \frac{\gamma}{\beta+\gamma} t_e^*$. This condition assures that small vehicles and large vehicles adopt the same queuing pattern in the atomistic PSNE.

products to receivers within narrow time windows.[23] We show that for a range of parameter values a PSNE exists in which large vehicles depart simultaneously with small vehicles and encounter queuing delays. The PSNE is identical to the atomistic PSNE in which user A disregards the congestion externalities that its vehicles impose on each other. Cases 1 and 2 are analyzed in the following two subsections.

6.1. Case 1: Large vehicles more flexible than small vehicles

In Case 1, large vehicles are more flexible than small vehicles. In the atomistic PSNE, large vehicles depart at the beginning and end of the travel period, and small vehicles travel in the middle. A queue exists throughout the travel period, but it rises and falls more slowly while large vehicles are departing than when small vehicles are departing just before and after the peak.[24] One might expect the same departure order to prevail with self-internalization, but with user A restricting its departure rate to match capacity so that queuing does not occur. The candidate PSNE with this pattern is shown in Figure 3.

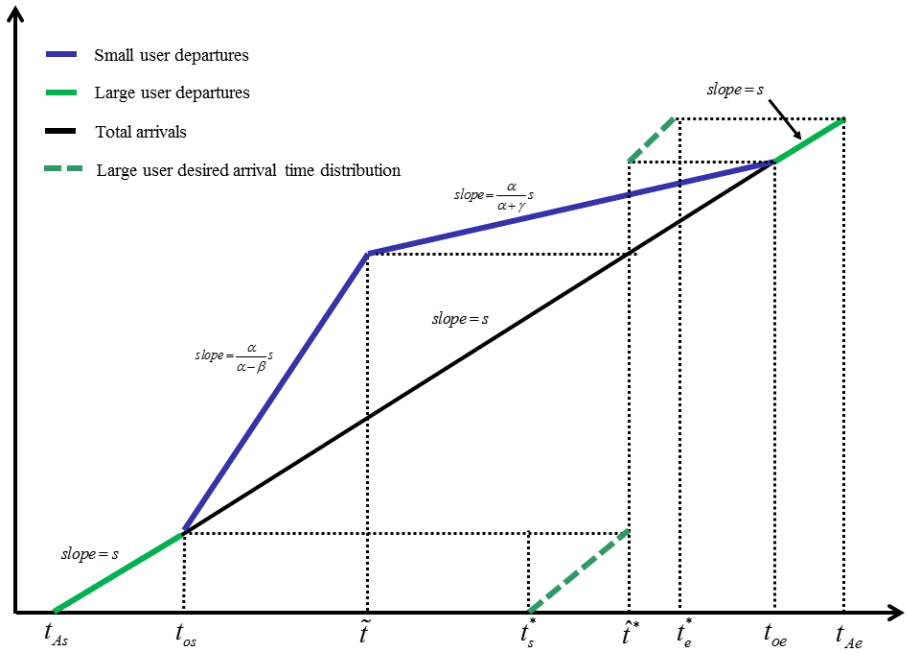

Figure 3: PSNE in which large user does not queue (Case 1)

Large vehicles depart during the intervals (t_{As}, t_{os}) and (t_{oe}, t_{Ae}), where $t_{As} = t^* -$

[23] Another possibility is that large vehicles are commercial aircraft operated by airlines with scheduled service, while small vehicles are private aircraft used mainly for recreational purposes.

[24] This departure pattern was studied by Arnott et al. (1988) and Arnott et al. (1994).

$\frac{\gamma}{\beta+\gamma}\frac{N_o+N_A}{s}$, and $t_{Ae} = t^* + \frac{\beta}{\beta+\gamma}\frac{N_o+N_A}{s}$.[25] Small vehicles depart during the central interval (t_{os}, t_{oe}). The departure schedule for small vehicles and the resulting queue are the same as if user A were absent.

If the candidate departure schedule in Figure 3 is a PSNE, neither small vehicles nor any subset of large vehicles can reduce their travel costs by deviating. The requisite conditions are identified in the two-part assumption:

Assumption 3: (i) $\theta \leq 1$. (ii) $\alpha_A \geq (\beta_A + \gamma_A)(1-\delta)$.

The following proposition identifies necessary and sufficient conditions for the pattern in Figure 3 to be a PSNE.

Proposition 2. *Let Assumption 3 hold. Then the departure pattern in Figure 3 is a PSNE. Small users depart early at rate $\alpha \cdot s/(\alpha - \beta)$ during $\left(t_{os}, \tilde{t}\right)$, and late at rate $\alpha \cdot s/(\alpha + \gamma)$ during $\left(\tilde{t}, t_{oe}\right)$. Large vehicles depart early at rate s during (t_{As}, t_{os}), and late at rate s during (t_{oe}, t_{Ae}). Large vehicles avoid queuing.*

Proof: See the Appendix.

Condition (i) in Assumption 3 is required for Proposition 2 since otherwise large vehicles could reduce their costs by departing during the queuing period (t_{os}, t_{oe}). As explained below, Condition (ii) in Assumption 3 is required so that user A cannot reduce its fleet's costs by rescheduling a mass of vehicles. The condition $\Delta > N_A/s$ (or $\delta \geq 1$) required for Theorem 1 does not apply here because user A does not queue in the PSNE. Nevertheless, Condition (ii) becomes less stringent the larger δ is, and Condition (ii) is guaranteed to hold if $\delta \geq 1$. Thus, similar to the duopoly example in Section 5, heterogeneity is conducive to existence of a PSNE.[26]

The key to the proof of Proposition 2 is to show that user A cannot profitably deviate from the candidate PSNE by rescheduling vehicles departing after t_{oe} to a mass departure at t_{os}. Forcing vehicles into the bottleneck as a mass just as small vehicles are beginning to depart allows user A to reduce the total schedule delay costs incurred by its fleet. Doing so at t_{os} is preferable to later because, with $\theta \leq 1$, large vehicles have a lesser willingness to queue than small vehicles. Queuing delay is nevertheless unavoidable because vehicles that depart later in the mass have to wait their turn. This trade-off is evident in the condition $\alpha_A \geq (\beta_A + \gamma_A)(1-\delta)$. Moreover, the more dispersed desired arrival times are,

[25] Recall that $\gamma_A/\beta_A = \gamma/\beta$.

[26] In the candidate PSNE, large vehicles travel in the tails of the departure period. In the system optimum there is no queuing, and the optimal order of departure depends on the ranking of β_A and β. If $\beta_A < \beta$, large vehicles still travel in the tails, but if $\beta_A > \beta$ they would travel in the middle. Hence the PSNE may be inefficient not only because queuing occurs, but also because total schedule delay costs are excessive.

the lower the fleet's costs in the candidate PSNE, and hence the less user A stands to gain from rescheduling. If $\delta > 1$, rescheduling vehicles actually increases their schedule delay costs because they arrive too quickly relative to their desired arrival times. Rescheduling then cannot possibly be beneficial. Given the benchmark parameter ratios $\beta{:}\alpha{:}\gamma = 1{:}2{:}4$, condition $\alpha_A \geq (\beta_A + \gamma_A)(1 - \delta)$ simplifies to $\delta \geq 3/5$, or $\Delta \geq (3/5)(N_A/s)$. In words: the range of desired arrival times for vehicles in the fleet must be at least 60 percent of the aggregate time required for them to traverse the bottleneck. This condition is plausible, at least for road users.

As noted above, the atomistic PSNE features the same order of departures and arrivals as the internalized PSNE, but with queuing by large vehicles as well as small vehicles. It is easy to show that both large vehicles and small vehicles incur lower travel costs with self-internalization. Thus, self-internalization achieves a Pareto improvement.

Silva et al. (2017) show that a PSNE without queuing exists for a symmetric duopoly and homogeneous users if $\alpha \geq \gamma$. We have effectively replaced one of the duopolists with a continuum of small users. The condition for a PSNE here (with $\delta = 0$) is $\alpha_A \geq \beta_A + \gamma_A$. This is more stringent than for the duopoly with the same unit costs. Hence, counterintuitively, the mixed market with a large user and small users may not have a PSNE even if a PSNE exists for both the less concentrated atomistic market and the more concentrated duopoly. While this nonmonotonic variation in behavior is intriguing, it complicates the analysis of equilibrium with large users.

6.2. Case 2: Large vehicles less flexible than small vehicles

In Case 2, $\theta > 1$ so that large vehicles are less flexible than small vehicles. Large vehicles prefer to travel in the middle of the travel period to reduce their schedule delay costs. However, queuing will be inevitable because small vehicles prefer the same range of arrival times. To meet the requirements of Theorem 1 for an internalized PSNE with queuing, it is necessary to assume that $\Delta > N_A/s$. Given this assumption, the atomistic PSNE is as shown in Figure 4. Large vehicles depart during the interval (t_{As}, t_{Ae}) and arrive at rate N_A/Δ over the interval $[t_s^*, t_e^*]$. Each large vehicle arrives on time. Small vehicles arrive at rate $s - N_A/\Delta$ during this interval, and at rate s during the rest of the interval $[t_{os}, t_{oe}]$. The aggregate departure rate and queuing time are the same as if all vehicles were small.[27]

Given an additional assumption identified in Assumption 4 below, the internalized PSNE turns out to be identical to the atomistic PSNE. All large vehicles thus travel during a

[27] Newell (1987) analyzed a more general version of this arrival pattern in the bottleneck model with small users. See also de Palma and Lindsey (2002).

queuing period, and depart at the same time as in the atomistic PSNE.[28] Thus, in contrast to Case 1, the large user's incentive to internalize self-congestion has no effect on either its fleet or small users.

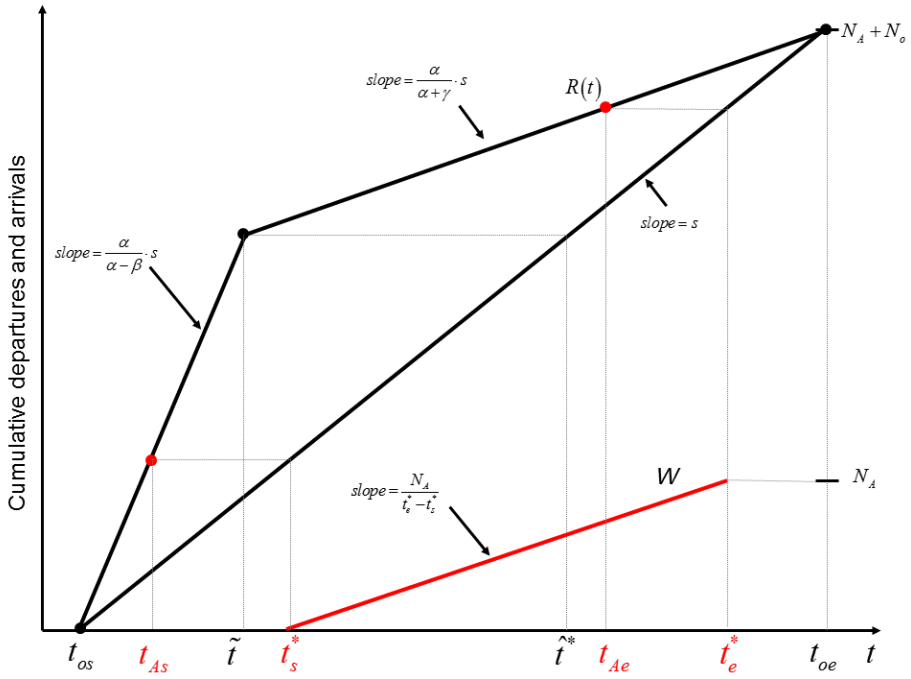

Figure 4: PSNE in which large user does not queue (Case 2).

The candidate departure schedule in Figure 4 is an internalized PSNE if and only if neither small vehicles nor any subset of large vehicles can reduce their travel costs by deviating. The three requisite conditions are identified in Assumption 4:

Assumption 4: (i) $\theta > 1$. (ii) $\Delta > N_A/s$.

$$\text{(iii)} \frac{N_A}{s} < (\theta - 1) \frac{\beta\gamma}{\alpha(\beta+\gamma)} \left(\frac{N_A + N_o}{s} - \Delta \right). \tag{32}$$

Using Assumption 4, the internalized PSNE is stated as:

Proposition 3. *Let Assumption 4 hold. Then the departure pattern in Figure 4 is a PSNE. Large users depart during the queuing period and all arrive on time. Small vehicles arrive*

[28] Recall Condition (30) which requires that in a PSNE with queuing, all large users have lower atomistic rates than the small users. This condition is satisfied in Case 2 because each large vehicle has a lower atomistic rate than small users *after* its preferred arrival time.

at a complementary rate so that the bottleneck is fully utilized. The aggregate departure rate and queuing time are the same as if all vehicles were small.

Proof: See the Appendix.

The roles of Conditions (i) and (ii) in Assumption 4 were explained above. Condition (iii) assures that user A's fleet is small enough that it prefers to schedule all its vehicles on-time during the queuing period, rather than scheduling some vehicles before queuing begins at t_{os}.[29]

7. Conclusions

In this paper we have studied trip-timing decisions by large users in the Vickrey bottleneck model of congestion. We believe that the model is representative of many transportation settings including airlines scheduling flights at airports, rail companies operating on rail networks, and freight shippers using congested roads. We build on previous studies of trip-timing decisions by large users in three ways: (i) we allow for the presence of small users; (ii) we consider general trip-timing preferences; and (iii) we allow for heterogeneity of trip-timing preferences within a large user's fleet as well as between large and small users.

Our paper makes two main contributions. First and foremost, it identifies conditions under which a Nash equilibrium in pure strategies exists in a setting in which large users make trip-timing decisions simultaneously and queue in a dynamic model of congestion with realistic propagation of delays. More specifically, we show that if vehicles in a large user's fleet have sufficiently diverse trip-timing preferences, a PSNE in which the large user queues may exist. We also provide an example in which the conditions for existence of a PSNE become less stringent as the number of large users increases.

Second, we illustrate how self-internalization can affect equilibrium travel costs. In two of the three examples presented, self-internalization reduces costs for all users. In the first example with symmetric large users (Section 5), the cost savings are substantial and can be nearly as large as for a monopolistic user that controls all the traffic. In the second example with one large user and a group of small users, all parties also gain if the large user schedules its fleet during the off-peak period without queuing. However, in the third example in which the large user travels during the peak, the equilibrium is identical to the atomistic PSNE so that no one benefits. The three examples illustrate that the effects

[29] In the candidate PSNE, large vehicles arrive at their individually preferred arrival times because they are less flexible than small vehicles. In the system optimum there is no queuing and, as in Case 1, the optimal order of departure depends on the ranking of β_A and β. If $\beta_A > \beta$, large vehicles would still be scheduled at their individually preferred arrival times, but if $\beta_A < \beta$ they would travel in the tails.

of self-internalization depend on both market structure and the trip-timing preferences of users.

The analysis of this paper can be extended in various directions. One is congestion pricing: either in the form of an optimal fine (i.e., continuously time-varying) toll that eliminates queuing, or a more practically feasible step-tolling scheme. Although the gains from self-internalization can be substantial, there is still scope to improve welfare by implementing congestion pricing. Indeed, this is what Verhoef and Silva (2017) find using the Henderson-Chu model for the case of large users with homogenous trip-timing preferences. A second topic is mergers or other measures to enable users to coordinate their trip-timing decisions gainfully without intervention by an external authority using either tolls or direct traffic control measures. It is not obvious from our preliminary results which users, if any, stand to gain by merging, how a merger would affect other users, and whether there is a case for regulation.

A third extension is to explore more complex market structures and different types of user heterogeneity. Ride-sharing companies or so-called Transportation Network Companies (TNCs) have become a major mode of passenger transportation in some cities and evidence is emerging that they are contributing to an increase in vehicle-km and congestion (Clewlow and Mishra, 2017; The New York Times, 2017). In Manhattan, the number of TNCs exceeds the number of taxis. Transportation services are offered by six types of operators in all: yellow cabs that must be hailed from the street, for-hire vehicles or black cars that must be booked, and four TNC companies: Uber, Lyft, Via, and Juno (Schaller, 2017).[30] The firms differ in their operations and fare structures. Their trip-timing preferences are also dictated by those of their customers. The size of a firm's fleet is not fixed, but varies by time of day and day of week according to when drivers choose to be in service. The simple Vickrey model would have to be modified to incorporate these user characteristics.

A fourth topic that we are studying is whether self-internalization by a large user can make other users worse off, or even leave the large user itself worse off. Such a result is of policy interest because it suggests that the welfare gains from congestion pricing of roads, airports and other facilities in which large users operate could be larger than previously thought.

References

Abkowitz, M. D. (1981a). An analysis of the commuter departure time decision. *Transportation*, 10(3):283–297.

[30] In addition, in 2013 Street Hail Liveries or green taxies began providing service in Northern Manhattan, the Bronx, Brooklyn, Queens, and Staten Island (Taxi and Limousine Commission, 2016).

Abkowitz, M. D. (1981b). Understanding the effect of transit service reliability on work-travel behavior. *Transportation Research Record*, 794:33–41.

Arnott, R., de Palma, A., and Lindsey, R. (1988). Schedule delay and departure time decisions with heterogeneous commuters. *Transportation Research Record*, (1197):56–67.

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.

Arnott, R., de Palma, A., and Lindsey, R. (1998). Recent developments in the bottleneck model. In Button, K. and Verhoef, E., editors, *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility*, pages 79–110. Edward Elgar, Aldershot.

Ball, M., Barnhart, C., Dresner, M., Hansen, M., Neels, K., Odoni, A., Peterson, E., Sherry, L., Trani, A. A., and Zou, B. (2010). Total delay impact study: a comprehensive assessment of the costs and impacts of flight delay in the united states. National Center of Excellence for Aviation Operations Research, Final Report.

Börjesson, M., Eliasson, J., and Franklin, J. P. (2012). Valuations of travel time variability in scheduling versus mean–variance models. *Transportation Research Part B: Methodological*, 46(7):855–873.

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

Choo, Y. Y. (2014). Factors affecting aeronautical charges at major us airports. *Transportation Research Part A: Policy and Practice*, 62:54–62.

Chu, X. (1995). Endogenous trip scheduling: the henderson approach reformulated and compared with the vickrey approach. *Journal of Urban Economics*, 37(3):324–343.

Clewlow, R. R. and Mishra, G. S. (2017). Disruptive transportation: the adoption, utilization, and impacts of ride-hailing in the united states. Technical report, Research Report–UCD-ITS-RR-17.

Couture, V., Duranton, G., Turner, M. A., et al. (2016). Speed. Technical report, Sciences Po.

Daniel, J. I. (1995). Congestion pricing and capacity of large hub airports: A bottleneck model with stochastic queues. *Econometrica*, 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. (2014). The untolled problems with airport slot constraints. *Economics of Transportation*, 3(1):16–28.

de Palma, A. and Fosgerau, M. (2011). Dynamic traffic modeling. In de Palma, A., Lindsey, R., Quinet, E., and Vickerman, R., editors, *Handbook in Transport Economics*, pages 188–212. Edward Elgar, Cheltenham, UK and Northampton, Mass, USA.

de Palma, A., Ginsburgh, V., Papageorgiou, Y. Y., and Thisse, J.-F. (1985). The principle of minimum differentiation holds under sufficient heterogeneity. *Econometrica: Journal of the Econometric Society*, pages 767–781.

de Palma, A. and Lindsey, R. (2002). Comparison of morning and evening commutes in the vickrey bottleneck model. *Transportation Research Record: Journal of the Transportation Research Board*, (1807):26–33.

de Palma, A. and Rochat, D. (1997). Impact of adverse weather conditions on travel decisions: Experience from a behavioral survey in geneva. *International Journal of Transport Economics/Rivista internazionale di economia dei trasporti*, pages 307–325.

Dixit, V. V., Ortmann, A., Rutstrom, E., and Ukkusuri, S. (2015). Understanding transportation systems through the lenses of experimental economics: A review.

Fargier, P.-H. (1983). Effects of the choice of departure time on road traffic congestion: theoretical approach. In *Proceedings of the Eighth International Symposium on Transportation and Traffic Theory*, pages 223–263. University of Toronto Press, Toronto.

Fosgerau, M. and Engelson, L. (2011). The value of travel time variance. *Transportation Research Part B: Methodological*, 45(1):1–8.

Ghosal, V. and Southworth, F. (2017). Advanced manufacturing plant location and its effects on economic development, transportation network, and congestion. Technical report, February 1. GDOT Research Project No. 12-24, https://ssrn.com/abstract=2925939.

Henderson, J. V. (1974). Road congestion: a reconsideration of pricing theory. *Journal of Urban Economics*, 1(3):346–365.

Hendrickson, C., Nagin, D., and Plank, E. (1981). Characteristics of travel time and dynamic user equilibrium for travel-to-work. In *Proceedings of the Eighth International Symposium on Transportation and Traffic Theory, Toronto, Canada.*

Hendrickson, C. and Plank, E. (1984). The flexibility of departure times for work trips. *Transportation Research Part A: General*, 18(1):25–36.

Humphreys, B. R. and Pyun, H. (2017). Professional sporting events and traffic: Evidence from us cities. Technical report, (March 25, 2017). Available at SSRN: https://ssrn.com/abstract2940762.

Hymel, K. (2009). Does traffic congestion reduce employment growth? *Journal of Urban Economics*, 65(2):127–135.

Khattak, A. J. and de Palma, A. (1997). The impact of adverse weather conditions on the propensity to change travel decisions: a survey of brussels commuters. *Transportation Research Part A: Policy and Practice*, 31(3):181–203.

Leonard, D. and Van Long, N. (1992). *Optimal control theory and static optimization in economics*. Cambridge University Press.

Lindsey, R. and Verhoef, E. (2007). Congestion modelling. In Hensher, D. A. and Button, K. J., editors, *Handbook of Transport Modelling: 2nd Edition*, pages 417–441. Emerald Group Publishing Limited.

Melo, E. (2014). Price competition, free entry, and welfare in congested markets. *Games and Economic Behavior*, 83:53–72.

Naroditskiy, V. and Steinberg, R. (2015). Maximizing social welfare in congestion games via redistribution. *Games and Economic Behavior*, 93:24–41.

National Academies of Sciences, Engineering, and Medicine (2017). Strategies to advance automated and connected vehicles. Technical report, Washington, DC: The National Academies Press. (https://doi.org/10.17226/24873).

Newell, G. F. (1987). The morning commute for nonidentical travelers. *Transportation Science*, 21(2):74–88.

Peer, S., Verhoef, E., Knockaert, J., Koster, P., and Tseng, Y.-Y. (2015). Long-run versus short-run perspectives on consumer scheduling: Evidence from a revealed-preference experiment among peak-hour road commuters. *International Economic Review*, 56(1):303–323.

Pigou, A. C. (1920). *The Economics of Welfare*. Macmillan, London.

Port Technology (2014). The world's top 5 terminal operators. December 4, https://www.porttechnology.org/news/the_worlds_top_5_terminal_operators.

Schaller, B. (2017). Empty seats, full streets: Fixing manhattan's traffic problem. Technical report, Schaller Consulting, December 21.

Schneider, K. and Weimann, J. (2004). Against all odds: Nash equilibria in a road pricing experiment. In Schreckenberg, M. and Selten, R., editors, *Human behaviour and traffic networks*, pages 133–153. Springer.

Schrank, D., Eisele, B., Lomax, T., and Bak, J. (2017). 2015 urban mobility scorecard. Technical report, Texas A&M Transportation Institute and INRIX, August.

Sheahan, M. and Bryan, V. (2018). Europe works to cope with overtourism. Technical report, The Globe and Mail. March 10.

Silva, H. E., Lindsey, R., de Palma, A., and van den Berg, V. A. C. (2017). On the existence and uniqueness of equilibrium in the bottleneck model with atomic users. *Transportation Science*, 51(3):863–881.

Silva, H. E., Verhoef, E. T., and van den Berg, V. A. C. (2014). Airlines' strategic interactions and airport pricing in a dynamic bottleneck model of congestion. *Journal of Urban Economics*, 80:13–27.

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

Small, K. A. (2015). The bottleneck model: An assessment and interpretation. *Economics of Transportation*, 4(1):110–117.

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

Statista (2017). Leading ship operator's share of the world liner fleet as of december 31, 2017. December 31, https://www.statista.com/statistics/198206/share-of-leading-container-ship-operators-on-the-world-liner-fleet/.

Taxi and Limousine Commission (2016). 2016 tlc factbook. Technical report, http://www.nyc.gov/html/tlc/downloads/pdf/2016_tlc_factbook.pdf.

The Economist (2018). Free exchange: Jam tomorrow. Technical report, January 20, p.68.

The New York Times (2017). Your uber car creates congestion. should you pay a fee to ride? December 26 (by Winnie Hu), https://www.nytimes.com/2017/12/26/nyregion/uber-car-congestion-pricing-nyc.html?smid$\bar{t}$w-nytimes&smtyp$\bar{c}$ur.

Tseng, Y., Ubbels, B., and Verhoef, E. (2005). Value of time, schedule delay, and reliability–estimation results of a stated choice experiment among dutch commuters facing congestion. In *Department of Spatial Economics, Free University of Amsterdam*.

Tseng, Y.-Y. and Verhoef, E. T. (2008). Value of time by time of day: A stated-preference study. *Transportation Research Part B: Methodological*, 42(7):607–618.

Verhoef, E. T. and Silva, H. E. (2017). Dynamic equilibrium at a congestible facility under market power. *Transportation Research Part B: Methodological*, 105:174–192.

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

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

Weisbrod, G. and Fitzroy, S. (2011). Traffic congestion effects on supply chains: Accounting for behavioral elements in planning and economic impact models. In Renko, S., editor, *Supply Chain Management–New Perspectives*. INTECH Open Access Publisher.

Appendix A. Appendix

Appendix A.1. Properties of the trip cost function (Section 2)

Trip-scheduling preferences can be described by a utility function of departure time and arrival time with the form[31]

$$U(t,a) = \int_{t_h}^{t} u_h(v)\,dv + \int_{a}^{t_w} u_w(v)\,dv, \tag{A.1}$$

where t_h and t_w are such that all trips take place within the interval $[t_h, t_w]$. Function $u_h(\cdot) > 0$ denotes the flow of utility at the origin (e.g., home), and function $u_w(\cdot) > 0$ denotes utility at the destination (e.g., work). It is assumed that $u_h(\cdot)$ and $u_w(\cdot)$ are continuously differentiable almost everywhere with derivatives $u_h' \leq 0$ and $u_w' \geq 0$, and $u_h(t^*) = u_w(t^*)$ for some time t^*. Utility from time spent traveling is normalized to zero. The cost of a trip is the difference between actual utility and utility from an idealized instantaneous trip at time t^*: $C(t,a) \equiv U(t^*,t^*) - U(t,a) \geq 0$.

[31] This formulation of scheduling preferences is due to Vickrey (1969, 1973) and has been used in several studies since; see de Palma and Fosgerau (2011). Defining preferences in terms of utility is appropriate for commuting and certain other types of trips. For trips involving freight transport, the utility function can be interpreted as profit or some other form of payoff or performance metric.

Various specification are possible for the flow-of-utility functions. Vickrey (1969) adopted a piecewise constant form:

$$\begin{aligned} u_h(v) &= u_h \qquad \text{(a constant)}, \\ u_w(v) &= \begin{cases} u_w^E \text{ for } v < t^* \\ u_w^L \text{ for } v > t^* \end{cases}, \end{aligned} \tag{A.2}$$

where $u_h > 0$, $0 < u_w^E < u_h$, and $u_w^L > u_h$. The cost function corresponding to (A.2) is:

$$C(t,a) = \begin{cases} \alpha(a-t) + \beta(t^*-a) \text{ for } a < t^* \\ \alpha(a-t) + \gamma(a-t^*) \text{ for } a > t^* \end{cases}, \tag{A.3}$$

where $\alpha = u_h$, $\beta = u_h - u_w^E$, and $\gamma = u_w^L - u_h$.

Another specification adopted by Fosgerau and Engelson (2011), and called the "slope" model by Börjesson et al. (2012), features linear flow-of-utility functions:

$$u_h(x) = u_{ho} - u_{h1}x, \;\; u_w(x) = u_{wo} + u_{w1}x\,.$$

Preferred travel time is $t^* = \frac{u_{ho}-u_{wo}}{u_{h1}+u_{w1}}$, and the cost function is

$$C(t,a) = \alpha(a-t) + \frac{u_{h1}}{2}(t^*-t)^2 + \frac{u_{w1}}{2}(a-t^*)^2\,, \tag{A.4}$$

where $\alpha = u_{ho} - u_{h1}t^*$. To assure that the model is well-behaved, departure and arrival times are restricted to values such that $u_h(t) > 0$ and $u_w(a) > 0$.

A third specification — used in early studies by Vickrey (1973), Fargier (1983), and Hendrickson et al. (1981) — is a variant of (A.4) with $u_{h1} = 0$:

$$C(t,a) = \alpha(a-t) + \frac{u_{w1}}{2}(a-t^*)^2\,. \tag{A.5}$$

In (A.5), utility at the origin is constant and schedule delay costs depend on arrival time but not departure time. Cost functions (A.3), (A.4), and (A.5) all satisfy Assumption 1 in the text (with t^* in place of k).

Appendix A.2. Atomistic departure rates (Section 2)

The atomistic rate for a user of type k is given by Eq. (3):

$$\hat{r}(t,a,k) = -\frac{C_t(t,a,k)}{C_a(t,a,k)}s.$$

Derivatives of specific interest are (with arguments suppressed to economize on notation)

$$\frac{\partial \hat{r}(t,a,k)}{\partial k} = \frac{C_t C_{ak} - C_a C_{tk}}{C_a^2} \geq 0,$$

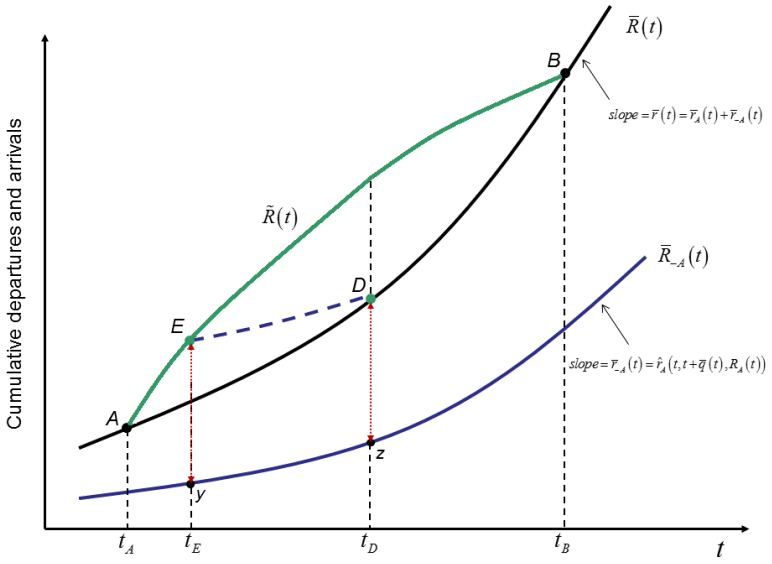

Figure A.5: Candidate PSNE with $C^A_{aa} < 0$

$$\frac{\partial^2 \hat{r}\,(t,a,k)}{\partial k^2} = \frac{C_a\left(C_t C_{akk} - C_a C_{tkk}\right) + 2C_{ak}\left(C_a C_{tk} - C_t C_{ak}\right)}{C_a^3} \geq 0,$$

$$\frac{\partial \hat{r}\,(t,a,k)}{\partial a} = \frac{s}{C_a^2}(C_t C_{aa} - C_a C_{ta}) = \frac{s}{C_a^2} C_t C_{aa} \overset{s}{=} -C_{aa},$$

where $\overset{s}{=}$ means identical in sign.

Appendix A.3. Proof of Lemma 2 with $C^A_{aa} < 0$

Consider Figure A.5, which depicts a candidate PSNE similar to that in Figure 1, but with $C^A_{aa} < 0$ so that curve $\bar{R}_{-A}(t)$ is convex rather than concave. Suppose that user A deviates from the candidate PSNE during the interval (t_A, t_B) by dispatching its vehicles earlier so that section ADB of $\bar{R}(t)$ shifts leftwards to $\tilde{R}(t)$. Vehicle $k = \bar{R}_A(t_D)$ originally scheduled to depart at point D and time t_D is rescheduled earlier to point E and time t_E such that distance Ey equals distance Dz. Vehicle k experiences a change in costs of

$$\Delta C^A(k) = C^A\left(t_E, t_E + \tilde{q}\,(t_E), k\right) - C^A\left(t_D, t_D + \bar{q}\,(t_D), k\right).$$

Let $\breve{q}(t)$ denote queuing time along the path from point D to point E shown by the dashed blue curve that runs parallel to $\bar{R}_{-A}(t)$ between points y and z. The change in cost can be

written

$$\begin{aligned}\Delta C^A(k) &= -\int_{t=t_E}^{t_D}\left(C_t^A\left(t,t+\check{q}(t),k\right)+C_a^A\left(t,t+\check{q}(t),k\right)\left(1+\frac{d\check{q}(t)}{dt}\right)\right)dt\\ &= -\int_{t=t_E}^{t_D}\left(C_t^A\left(t,t+\check{q}(t),k\right)+C_a^A\left(t,t+\check{q}(t),k\right)\frac{\widehat{r}_A\left(t,t+\bar{q}(t),\bar{R}_A(t)\right)}{s}\right)dt\\ &= -\frac{1}{s}\int_{t=t_E}^{t_D}C_a^A\left(t,t+\check{q}(t),k\right)\left\{\widehat{r}_A\left(t,t+\bar{q}(t),\bar{R}_A(t)\right)-\widehat{r}_A\left(t,t+\check{q}(t),k\right)\right\}dt.\end{aligned}$$

Since $\check{q}(t) > \bar{q}(t)$ for $t \in (t_A, t_B)$, with $C_{aa}^A < 0$ and for any j, $\hat{r}_A(t,t+\check{q}(t),j) < \hat{r}_A(t,t+\bar{q}(t),j)$. If user A's fleet is homogeneous, the expression in braces is negative, $\Delta C^A(k) < 0$, and rescheduling the vehicle from D to E reduces its trip cost.

Appendix A.4. Proof of Theorem 1 (Section 4)

Using Eq. (1), the term in braces in (31) can be written

$$Z = \int_{j=k}^{\bar{R}_A(t)}\left(\frac{\partial\widehat{r}_A\left(t,t+\bar{q}(t),j\right)}{\partial j}+\frac{1}{s}\frac{\partial\widehat{r}_A\left(t,t+\bar{q}(t)-\frac{j-k}{s},k\right)}{\partial a}\right)dj. \tag{A.6}$$

A sufficient condition for Z to be positive is that the integrand be positive for all values of j. Given Assumption 2, there is a one-to-one monotonic correspondence between j and t_j^*. The integrand in (A.6), z, can therefore be written

$$z = \frac{\partial\widehat{r}_A\left(t,t+\bar{q}(t),j\right)}{\partial t_j^*}\frac{1}{f\left(t_j^*\right)}-\frac{1}{s}\frac{\partial\widehat{r}_A\left(t,t+\bar{q}(t)-\frac{j-k}{s},k\right)}{\partial t_j^*}. \tag{A.7}$$

Now

$$\begin{aligned}&\frac{\partial\widehat{r}_A\left(t,t+\bar{q}(t),j\right)}{\partial t_j^*}-\frac{\partial\widehat{r}_A\left(t,t+\bar{q}(t)-\frac{j-k}{s},k\right)}{\partial t_j^*}\\ &= \int_{n=k}^{j}\left(\frac{\partial^2\widehat{r}_A\left(t,t+\bar{q}(t)-\frac{n-k}{s},n\right)}{\partial\left(t_n^*\right)^2}\frac{1}{f\left(t_n^*\right)}+\frac{1}{s}\frac{\partial^2\widehat{r}_A\left(t,t+\bar{q}(t)-\frac{n-k}{s},n\right)}{\partial a\partial t_n^*}\right)dn\\ &= \int_{n=k}^{j}\left(\frac{\partial^2\widehat{r}_A\left(t,t+\bar{q}(t)-\frac{n-k}{s},n\right)}{\partial\left(t_n^*\right)^2}\left(\frac{1}{f\left(t_n^*\right)}-\frac{1}{s}\right)\right)dn\end{aligned} \tag{A.8}$$

By (4), the second derivative is positive, and by assumption, $f(t_n^*) \le s$ for all t_n^*. Hence (A.8) is positive. Using this result in (A.7) we have

$$z \ge \frac{\partial\widehat{r}_A\left(t,t+\bar{q}(t)-\frac{j-k}{s},k\right)}{\partial t_j^*}\left(\frac{1}{f\left(t_j^*\right)}-\frac{1}{s}\right) > 0.$$

This establishes that $Z > 0$ in (A.6), and hence that $\Delta C^A(k) > 0$.

Appendix A.5. Section 5

Appendix A.5.1. Proposition 1: Profitability of deviation with 2 users

Since vehicle k arrives early in the candidate PSNE, $k < \frac{\gamma}{\beta+\gamma}N_A$ where $N_A = N/2$. Vehicle k has a desired arrival time of

$$t_k^* = \hat{t}^* - \left(\frac{\gamma}{\beta+\gamma}N_A - k\right)\frac{\Delta}{N_A}.$$

Vehicle k departs at time

$$t_k = \tilde{t} - \left(\frac{\gamma}{\beta+\gamma}N_A - k\right) / \left(\frac{\alpha}{\alpha-\beta}s\right),$$

and arrives at time

$$a_k = \hat{t}^* - \left(\frac{\gamma}{\beta+\gamma}N_A - k\right) / (s/2).$$

As shown by Silva et al. (2017, Eq. (24c)),

$$\tilde{t} = \hat{t}^* - \frac{\beta(\gamma-\alpha)}{2\alpha(\beta+\gamma)}\frac{N}{s}.$$

If user A deviates from the candidate PSNE so that vehicle k departs at $\tilde{t}$ rather than t_k, vehicle k arrives at

$$a_k' = a_k + \left(\tilde{t} - t_k\right)\frac{\alpha}{\alpha-\beta}s \cdot \frac{1}{s} = \hat{t}^* - \frac{1}{s}\left(\frac{\gamma}{\beta+\gamma}N_A - k\right).$$

Vehicle k can benefit from deviation only if $a_k' < t_k^*$: a condition which reduces to $\Delta < N_A/s$. Deviation is not profitable if $\Delta > N_A/s$, or equivalently $f = N_A/\Delta < s$ as per Theorem 1.

Appendix A.5.2. Gain from internalization with m users

With m large users the aggregate equilibrium departure rate in the candidate PSNE during the period of queuing is given by eq. (23):

$$r(t) = \begin{cases} \frac{m}{m-1}\frac{\alpha}{\alpha-\beta}s, & t \in \left(t_q, \tilde{t}\right) \\ \frac{m}{m-1}\frac{\alpha}{\alpha+\gamma}s, & t \in \left(\tilde{t}, t_e\right) \end{cases}.$$

When all vehicles have the same desired arrival time, t^*, the critical times t_s, t_q, $\tilde{t}$, and t_e are determined by the following four equations:

$$t_e - t_s = N/s, \tag{A.9}$$

$$\beta(t^* - t_s) = \gamma(t_e - t^*), \tag{A.10}$$

$$s(t_q - t_s) + \frac{m}{m-1}\frac{\alpha}{\alpha-\beta}s\left(\tilde{t} - t_q\right) + \frac{m}{m-1}\frac{\alpha}{\alpha+\gamma}s\left(t_e - \tilde{t}\right) = N, \tag{A.11}$$

$$s\left(t_q - t_s\right) + \frac{m}{m-1}\frac{\alpha}{\alpha-\beta}s\left(\tilde{t} - t_q\right) = s\left(t^* - t_s\right). \tag{A.12}$$

Eq. (A.9) stipulates that all vehicles complete their trips. Eq. (A.10) states that the first and last vehicles incur the same private cost. Eq. (A.11) stipulates that cumulative departures equal N. Finally, according to eq. (A.12) total departures from t_s to $\tilde{t}$ equals the number of vehicles that arrive early.

Solving (A.9)-(A.12), it is possible to show after considerable algebra that total costs in the candidate PSNE are

$$TC^i = \frac{(m-1)(2m-1)\beta\gamma + m\alpha\gamma - (m-1)\alpha\beta}{2m\gamma\left(\alpha + (m-1)\beta\right)}\frac{\beta\gamma}{\beta+\gamma}\frac{N^2}{s}.$$

Total costs in the atomistic PSNE are

$$TC^n = \frac{\beta\gamma}{\beta+\gamma}\frac{N^2}{s}.$$

When vehicles differ in their desired arrival times, schedule delay costs are reduced by the same amount in the two PSNE. The departure rate is unchanged in the candidate PSNE with internalization. The difference in total costs is thus the same with and without heterogeneity so that, as stated in the text

$$TC^n - TC^i = \frac{(m-1)\beta(\alpha+\gamma) + m\alpha\gamma}{2m(m-1)\beta\gamma + 2m\alpha\gamma}\frac{\beta\gamma}{\beta+\gamma}\frac{N^2}{s}.$$

Appendix A.5.3. Proof of Proposition 2

It is necessary to show that neither user A nor a small user can gain by deviating from the candidate PSNE. In all, seven types of deviations need to be considered-

Deviation 1. A small user cannot gain by deviating.

Small users incur the same cost throughout the candidate departure interval (t_{os}, t_{oe}). Hence, they cannot gain by retiming their trips within this interval. Rescheduling a trip either before t_{os} or after t_{oe} would clearly increase their cost. Thus, no small user can benefit by deviating.

Deviation 2. User A cannot gain by rescheduling vehicles outside the departure period (t_{As}, t_{Ae}).

User A does not queue in the candidate PSNE. Large vehicles therefore do not delay each other. Moreover, the highest costs are borne by the first and last vehicles departing at t_{As} and t_{Ae}, respectively. Rescheduling any vehicles either before t_{As} or after t_{Ae} would increase user A's fleet cost.

Deviation 3. User A cannot gain by rescheduling a single vehicle to another time within the departure period when there is no queue; i.e. to any time $t \in (t_{As}, t_{os}) \cup (t_{oe}, t_{Ae})$.[32]

[32]Much of the following text is drawn, verbatim, from Silva et al. (2017).

During the no-queuing period, the bottleneck is used to capacity. It is therefore necessary to distinguish between the cost that user A saves by removing a vehicle from the departure schedule (which does not affect the costs of other vehicles in the fleet) and the cost user A incurs by adding a vehicle (which creates a queue unless the vehicle is added at t_{Ae}). The respective costs are[33]:

$$C_A^-(t) = \begin{cases} \beta_A \cdot (t^* - t), & t \in [t_{As}, t_{os}] \\ \gamma_A \cdot (t - t^*), & t \in [t_{oe}, t_{Ae}] \end{cases},$$

$$C_A^+(t) = \begin{cases} \beta_A \cdot (t^* - t) \ + \frac{\alpha_A - \beta_A}{s} \cdot \int_t^{t_{os}} r_A(u)\, du \ + \frac{\alpha_A + \gamma_A}{s} \cdot \int_{t_{oe}}^{t_{Ae}} r_A(u)\, du, & t \in [t_{As}, t_{oe}] \\ \gamma_A \cdot (t - t^*) + \frac{\alpha_A + \gamma_A}{s} \cdot \int_t^{t_{Ae}} r_A(u)\, du, & t \in [t_{oe}, t_{Ae}] \end{cases}.$$

A vehicle can be rescheduled in four ways: (i) late to late, (ii) late to early, (iii) early to late, and (iv) early to early. Consider each possibility in turn.

(*i). Rescheduling late to late*: Rescheduling a late vehicle to a later time is never beneficial because the vehicle's trip cost increases, and other vehicles in the fleet do not benefit. Suppose a vehicle is rescheduled earlier from t to t' where $t_{oe} \leq t' < t$. User A's fleet costs change by an amount:

$$\begin{aligned} \Delta C_A &= -C_A^-(t) + C_A^+(t') = -\gamma_A \cdot (t - t') + \frac{\alpha_A + \gamma_A}{s} \cdot \int_{t'}^{t} r_A(u)\, du \\ &= -\gamma_A \cdot (t - t') + \frac{\alpha_A + \gamma_A}{s} \cdot s\,(t - t') = \alpha_A (t - t') > 0. \end{aligned}$$

Since fleet costs increase, the deviation is not gainful.

(*ii). Rescheduling late to early*: The best time to reschedule a vehicle is t_{os} because this minimizes the vehicle's early-arrival cost as well as the queuing delay imposed on the rest of the fleet. But rescheduling the vehicle to t_{os} is no better (or worse) than rescheduling it to t_{oe}, which is not beneficial as per case (i).

(*iii). Rescheduling early to late*: The best option in this case is to reschedule a vehicle from t_{As}. However, the gain is the same as (or worse than) from rescheduling a vehicle from t_{Ae}, and this is not beneficial as per case (i). Rescheduling early to late therefore cannot be beneficial.

(*iv). Rescheduling early to early*: The best option in this case is to reschedule a vehicle from t_{As} to t_{os}. Again, this is not beneficial for the same reason as in case (iii).

Deviation 4. User A cannot gain by rescheduling a single vehicle to a time within the queuing period, (t_{os}, t_{oe}).

[33]The formula for $C_i^+(t)$ can be derived by integrating (12) and applying transversality condition (14).

For any vehicle in user A's fleet that is scheduled to depart early at t, there is another vehicle scheduled to depart late at t' that incurs the same cost (this follows from symmetry of the t_A^* distribution). Removing either vehicle saves the same cost: $C_A^-(t) = C_A^-(t')$. However, removing the early vehicle and inserting it at any time during the queuing period creates a (small) queue that persists until t_{Ae}. Removing the late vehicle creates a queue only until t' because the queue disappears during the departure-time slot opened up by the rescheduled vehicle. Rescheduling a late vehicle is therefore preferred. The best choice is to reschedule the first late-arriving vehicle at t_{oe} so that no later vehicles in the fleet are delayed. Rescheduling a vehicle from $t' > t_{oe}$ would reduce that vehicle's cost by more, but a queue would persist from t_{oe} until t'. The fleet's schedule delay costs would therefore not be reduced, and a greater queuing cost would be incurred as well.

Given $\theta \leq 1$, rescheduling a vehicle from t_{oe} to any time $t \in (t_{os}, t_{oe})$ will (weakly) increase its cost. So rescheduling it not gainful. But if $\theta > 1$, the vehicle will benefit. Hence the candidate can be a PSNE only if $\theta \leq 1$ as per Proposition 2.

Deviation 5. User A cannot gain by rescheduling a positive measure of its fleet (i.e., a mass of vehicles) to times within the departure period when there is no queue.

If user A reschedules a positive measure of vehicles to depart during $(t_{As}, t_{os}) \cup (t_{oe}, t_{Ae})$, queuing will occur during some nondegenerate time interval. By Lemma 1, user A is willing to depart at a positive and finite rate during early arrivals only if $r_{-A}(t) = \hat{r}_A = \alpha_A \cdot s / (\alpha_A - \beta_A) > 0$. Since no other users depart at t, $r_{-A}(t) = 0$ and user A is better off scheduling vehicles later. Similarly, for late arrivals user A is willing to depart at a positive and finite rate only if $r_{-A}(t) = \alpha_A \cdot s / (\alpha_A + \gamma_A)$. Since $r_{-A}(t) = 0$, user A is again better off scheduling vehicles later.

Deviation 6. Any deviation by user A involving multiple mass departures is dominated by a deviation with a single mass departure.

Suppose that user A deviates from the candidate PSNE by scheduling multiple mass departures. All vehicles in the fleet are assumed to depart in order of their index, including vehicles within the same mass. (This assures that fleet costs in the deviation cannot be reduced by reordering vehicles.) We show that such a deviation is dominated by a single mass departure. The proof involves establishing three results: (i) Fleet costs can be reduced by rescheduling any vehicles that are not part of a mass, but suffer queuing delay, to a period without queuing. (ii) Fleet costs can be reduced by rescheduling any vehicles in a mass departure after $\tilde{t}$ to a period without queuing. (iii) Any deviation with multiple mass departures launched before $\tilde{t}$ entails higher fleet costs than a deviation with a single mass departure at t_{os}. These three results show that the candidate PSNE need only be tested against a single mass departure launched at t_{os}.

Result (i): When a queue exists, user A is willing to depart at a positive and finite rate only if condition (21) is satisfied; i.e. $r_{-A}(t) = \hat{r}_A(t)$. For any vehicle that arrives early this requires $r_{-A}(t) = \hat{r}_{AE} = \alpha \cdot s/(\alpha - \theta\beta) > 0$, and for any vehicle that arrives late, $r_{-A}(t) = \hat{r}_{AL} = \alpha \cdot s/(\alpha + \theta\gamma) > 0$. During the departure period (t_{As}, t_{os}), $r_{-A}(t) = 0$, so user A is better off scheduling all vehicles in the mass later. During the departure period $\left(t_{os}, \tilde{t}\right)$, $r_{-A}(t) = \alpha \cdot s/(\alpha - \beta)$. Since $\theta \leq 1$, $r_{-A}(t) \geq \hat{r}_{AE} > \hat{r}_{AL}$ and user A is (weakly) better off scheduling all vehicles in the mass earlier. During the departure period $\left(\tilde{t}, t_{oe},\right)$, $r_{-A}(t) = \alpha \cdot s/(\alpha + \gamma) \leq \hat{r}_{AL} < \hat{r}_{AE}$. User A is (weakly) better off scheduling all vehicles later. Finally, during the departure period $(t_{oe}, t_{Ae},)$, $r_{-A}(t) = 0$ and user A is again better off scheduling vehicles later.

Result (ii): Assume that user A launches the last mass departure after $\tilde{t}$. We show that user A can reduce its fleet costs by rescheduling vehicles in the mass to a later period in which they avoid queuing delay. This is true whether or not each vehicle in the mass is destined to arrive early or late relative to its individual t^*. By induction, it follows that all mass departures launched after $\tilde{t}$ can be gainfully rescheduled. In what follows it is convenient to use the auxiliary variable $\lambda_{t,t'}^{-A} \equiv \int_{t'}^{t} r_{-A}(u)\, du/(s \cdot (t - t'))$ which denotes average departures of small users as a fraction of capacity during the period $[t', t]$.

Suppose user A launches the last mass departure at time t_L with M vehicles. Assume first that at t_L there is a queue with queuing time $q(t_L)$. We show that postponing the mass departure to a later time when a queue still exists reduces user A's fleet costs. By induction, it follows that postponing the mass until the queue disappears is gainful. Let j be the vehicle that departs in position m of the mass, $m \in [0, M]$. Let $D_j[\cdot]$ be the schedule delay cost function of vehicle j, and $c(j,t)$ its trip cost if the mass departs at time t.

If the mass departs at time t_L, vehicle j incurs a cost of

$$c(j, t_L) = \alpha \cdot \left(q(t_L) + \frac{m}{s}\right) + D_j\left[t_L + q(t_L) + \frac{m}{s}\right]. \tag{A.13}$$

If the mass departure is postponed to time $t'_L > t_L$, and a queue still exists at t'_L, vehicle j incurs a cost of

$$c\left(j, t'_L\right) = \alpha \cdot \left(q\left(t'_L\right) + \frac{m}{s}\right) + D_j\left[t'_L + q\left(t'_L\right) + \frac{m}{s}\right]. \tag{A.14}$$

By Result (i), user A does not depart during (t_L, t'_L) because a queue persists during this period. Hence

$$q\left(t'_L\right) = q(t_L) + \int_{t_L}^{t'_L} \frac{r_{-A}(u) - s}{s} du = q(t_L) - \left(t'_L - t_L\right)\left(1 - \lambda_{t_L, t'_L}^{-A}\right). \tag{A.15}$$

Substituting (A.15) into (A.14), and using (A.13), one obtains

$$c\left(j,t'_L\right)-c\left(j,t_L\right)=-\alpha\cdot\left(1-\lambda^{-A}_{t_L,t'_L}\right)\left(t'_L-t_L\right) \tag{A.16}$$
$$+D_j\left[t_L+q\left(t_L\right)+\frac{m}{s}+\lambda^{-A}_{t_L,t'_L}\left(t'_L-t_L\right)\right]-D_j\left[t_L+q\left(t_L\right)+\frac{m}{s}\right].$$

The value of $\lambda^{-A}_{t_L,t'_L}$ depends on the timing of t_L and t'_L. If $t'_L\leq t_{oe}$, small users depart at rate $\frac{\alpha}{\alpha+\gamma}\cdot s$ throughout the interval (t_L,t'_L) so that $\lambda^{-A}_{t_L,t'_L}=\frac{\alpha}{\alpha+\gamma}$. If $t'_L>t_{oe}$, $\lambda^{-A}_{t_L,t'_L}<\frac{\alpha}{\alpha+\gamma}$. Hence, $\lambda^{-A}_{t_L,t'_L}\leq\frac{\alpha}{\alpha+\gamma}$ for all values of t'_L and the first line of (A.16) is negative. For the second line there are three possibilities to consider according to when vehicle j arrives: (a) early both before and after the mass is postponed, (b) early before postponement and late after, and (c) late both before and after postponement. In case (a), the second line of (A.16) is negative, in case (c) it is positive, and in case (b) the sign is a priori ambiguous. To show that (A.16) is negative it suffices to show this for case (c). The second line is an increasing function of $\lambda^{-A}_{t_L,t'_L}$. Using $\lambda^{-A}_{t_L,t'_L}\leq\frac{\alpha}{\alpha+\gamma}$ and $D_j\left[x\right]=\gamma x$ for $x>0$, (A.16) yields

$$c\left(j,t'_L\right)-c\left(j,t_L\right)\leq-\alpha\cdot\frac{\gamma}{\alpha+\gamma}\left(t'_L-t_L\right)+\theta\gamma\cdot\left(\frac{\alpha}{\alpha+\gamma}\left(t'_L-t_L\right)\right)<0.$$

This proves that postponing the mass departure (weakly) reduces the cost for every vehicle in the mass. We conclude that if there is a queue when the last mass departs, user A can (weakly) reduce its fleet costs by postponing the mass departure to the time when the queue just disappears (user A's later vehicles are not affected by postponing the mass).

Now assume there is no queue at t_L when the last mass is launched, which is possible only if $t_L>t_{oe}$. Small users do not depart after t_{oe}, and by Result (i) none of A's vehicles outside the mass depart until the queue from the mass has disappeared. The queue produced by the mass departure thus disappears at time t_L+M/s. We show that user A can reduce its fleet costs by rescheduling vehicles in the mass to depart at rate s over the time interval $[t_L,t_L+M/s]$. Every vehicle arrives at the same time as in the mass, but avoids queuing delay.

To see this, let j be the index of the vehicle that departs in position m in the mass. In the mass departure, vehicle j incurs a cost of

$$c\left(j,t_L\right)=D_j\left[t_L+\frac{m}{s}\right]+\alpha\cdot\frac{m}{s}.$$

In the deviation where vehicle j delays departure until $t'_L=t_L+m/s$, it incurs a cost of

$$c\left(j,t'_L\right)=D_j\left[t_L+\frac{m}{s}\right].$$

Its cost changes by

$$c\left(j,t'_L\right)-c\left(j,t_L\right)=-\alpha\cdot\frac{m}{s}<0.$$

Every vehicle enjoys a reduction in queuing time cost with no change in schedule delay cost. Hence, in any deviation from the candidate PSNE, fleet costs can be reduced by eliminating the last mass departure. By induction, any mass departure launched after $\tilde{t}$ can be rescheduled without increasing fleet costs.

Next, we show that any deviation entailing multiple mass departures before $\tilde{t}$ is dominated by scheduling a single mass departure at t_{os}.

Result (iii): Suppose that more than one mass departure is scheduled before $\tilde{t}$. Assume the first mass is launched at time t_E with M vehicles, and the second mass is launched at time t'_E with M' vehicles. There are three cases to consider depending on the timing of t_E and t'_E.

Case 1: $t_E < t'_E \leq t_{os}$. Both masses are scheduled before small users start to depart.

Since $r_{-A}(t) = 0$ for $t < t_E$, by Result (i), there is no queue at t_E. If the queue from the first mass disappears before t'_E, as in the proof of Result (ii), user A can reduce its fleet costs simply by rescheduling vehicles in the first mass to depart at a rate of s during $t \in (t_E, t'_E)$. Since user A does not depart in the original deviation until the first queue has dissipated, the rescheduled vehicles in the alternative deviation avoid queuing and arrive at the same time – thereby reducing their queuing delay costs without affecting their schedule delay costs. If the queue from the first mass does not disappear before t'_E, user A can still reduce its fleet costs by rescheduling $s \cdot (t'_E - t_E)$ vehicles at a rate s during (t_E, t'_E), and letting the remaining $M - s \cdot (t'_E - t_E)$ vehicles join the head of the second mass at t'_E. The first set of vehicles in the first mass avoids queuing and incur the same schedule delay costs. The remaining vehicles in the first mass also incur lower queuing costs since they no longer queue between t_E and t'_E. Vehicles in the second mass that departs at t'_E still depart and arrive at the same time because the same number of vehicles depart before them, and the bottleneck operates at capacity throughout.

Case 2: $t_E < t_{os} < t'_E$. The second mass is scheduled after small users start to depart.

If the queue from the first mass disappears before t_{os}, the reasoning for Case 1 applies. If the queue from the first mass does not disappear before t_{os}, the queue will not dissipate until after small users have stopped departing at t_{oe}. However, user A can still reduce its fleet costs by rescheduling some of the M vehicles in the first mass to t_{os}, and rescheduling the remainder to the head of the second mass at t'_E.

Case 3: $t_{os} \leq t_E < t'_E$. The first mass departs when, or after, small users begin to depart.

In this case, user A can reduce its fleet costs by rescheduling the second mass to depart immediately after the first mass. To show this, let $q(t)$, $t \geq t_E$, denote queuing time after the first mass of M vehicles departs. Let j be the index of the vehicle that departs in position m of the second mass, where $m \in [0, M']$. Vehicle j arrives at time $a'_j =$

$t'_E + q\left(t'_E\right) + m/s$ and incurs a cost of

$$c\left(j, t'_E\right) = \alpha\left(q\left(t'_E\right) + \frac{m}{s}\right) + D_j\left[a'_j\right].$$

If the second mass is instead dispatched immediately after the first mass at t_E, vehicle j arrives at time $a_j = t_E + q\left(t_E\right) + m/s$ and incurs a cost of

$$c\left(j, t_E\right) = \alpha\left(q\left(t_E\right) + \frac{m}{s}\right) + D_j\left[a_j\right].$$

The cost saving is

$$c\left(j, t'_E\right) - c\left(j, t_E\right) = \alpha\left(q\left(t'_E\right) - q\left(t_E\right)\right) + D_j\left[a'_j\right] - D_j\left[a_j\right]. \tag{A.17}$$

Now

$$a'_j = a_j + t'_E - t_E + q\left(t'_E\right) - q\left(t_E\right), \tag{A.18}$$

and

$$q\left(t'_E\right) - q\left(t_E\right) = \frac{\beta}{\alpha - \beta}\left(t'_E - t_E\right). \tag{A.19}$$

Substituting (A.18) and (A.19) into (A.17) yields

$$\begin{aligned} c\left(j, t'_E\right) - c\left(j, t_E\right) &= \frac{\alpha\beta}{\alpha - \beta}\left(t'_E - t_E\right) + D_j\left[a_j + \frac{\alpha}{\alpha - \beta}\left(t'_E - t_E\right)\right] - D_j\left[a_j\right] \\ &= \beta\Delta q_{-A} + D_j\left[a_j + \Delta q_{-A}\right] - D_j\left[a_j\right] \geq 0, \end{aligned} \tag{A.20}$$

where $\Delta q_{-A} \equiv \frac{\alpha}{\alpha - \beta}\left(t'_E - t_E\right)$ is the gross contribution of small users to queuing time during the period $\left(t_E, t'_E\right)$. The weak inequality in (A.20) holds as an equality if vehicle j arrives early when the second mass departs at t'_E. The inequality is strict if vehicle j arrives late. Since this conclusion holds for all vehicles in the second mass, user A can reduce its costs by merging the later mass with the earlier mass.

By induction, all but one of any mass departures launched before $\tilde{t}$ can be eliminated in a way that decreases user A's fleet costs. Using similar logic, it is straightforward to show that user A can do no better than to schedule the single mass at t_{os} rather than later. In summary, results (i)–(iii) show that, of all deviations from the candidate PSNE entailing mass departures, a deviation with a single mass departure launched at t_{os} is the most viable.

Deviation 7. User A cannot gain by rescheduling a positive measure of its fleet to times during the queuing period $\left(t_{os}, t_{oe}\right)$.

To prove that Deviation 7 is not gainful, we must determine whether total fleet costs can be reduced by deviating from the candidate PSNE. Since user A has weaker preferences for on-time arrival than small users, user A prefers not to schedule departures in the interior of $\left(t_{os}, t_{oe}\right)$. User A's best deviation is to schedule a mass departure at t_{os}. Let N_{Am} be

the measure of vehicles in the mass. If N_{Am} is small, the best choice is to reschedule the first vehicles departing late during the interval $(t_{oe}, t_{oe} + N_{Am}/s)$. (As explained in proving that Deviation 4 is not beneficial, this strategy avoids queuing delay for large vehicles that are not part of the mass.) The first of the rescheduled vehicles has a preferred arrival time of $\hat{t}^*$. In the candidate PSNE, this vehicle incurs a cost

$$C^A\left(t_{oe}, t_{oe}, \hat{t}^*\right) = \gamma_A\left(t_{oe} - \hat{t}^*\right). \tag{A.21}$$

The last of the rescheduled vehicles has a preferred arrival time of $\hat{t}^* + \delta N_{Am}/s$. It incurs a cost

$$C^A\left(t_{oe} + \frac{N_{Am}}{s}, t_{oe} + \frac{N_{Am}}{s}, \hat{t}^* + \frac{\delta}{s}N_{Am}\right) = \gamma_A\left(t_{oe} - \hat{t}^* + (1-\delta)\frac{N_{Am}}{s}\right). \tag{A.22}$$

The average cost of the rescheduled vehicles is the unweighted mean of eqs. (A.21) and (A.22). Total costs for the N_{Am} vehicles before they are displaced are therefore

$$TC^c_{dev} = \left[\gamma_A\left(t_{oe} - \hat{t}^*\right) + \frac{\gamma_A(1-\delta)N_{Am}}{2s}\right]N_{Am}, \tag{A.23}$$

where superscript c denotes the candidate PSNE.

The first of the rescheduled vehicles departs at t_{os} and incurs a cost

$$C^A\left(t_{os}, t_{os}, \hat{t}^*\right) = \beta_A\left(\hat{t}^* - t_{os}\right). \tag{A.24}$$

The last of the rescheduled vehicles incurs a cost of

$$\begin{aligned} &C_A\left(t_{os}, t_{os} + \frac{N_{Am}}{s}, \hat{t}^* + \delta\frac{N_{Am}}{s}\right) \\ &= \beta_A\left(\hat{t}^* + \delta\frac{N_{Am}}{s} - (t_{os} + \frac{N_{Am}}{s})\right) + \alpha_A\frac{N_{Am}}{s} \\ &= \beta_A\left(\hat{t}^* - t_{os} + \delta\frac{N_{Am}}{s}\right) + (\alpha_A - \beta_A)\frac{N_{Am}}{s}. \end{aligned} \tag{A.25}$$

The average cost of the rescheduled vehicles is the unweighted mean of eqs. (A.24) and (A.25). Total costs for the rescheduled vehicles are therefore

$$TC^d_{dev} = \left[\beta_A\left(\hat{t}^* - t_{os}\right) + \frac{\alpha_A - \beta_A(1-\delta)}{2}\frac{N_{Am}}{s}\right]N_{Am}, \tag{A.26}$$

where superscript d denotes the deviation. Given (A.23) and (A.26), the change in total costs is

$$TC^d_{dev} - TC^c_{dev} = \left[\alpha_A - (\beta_A + \gamma_A)(1-\delta)\right]\frac{N^2_{Am}}{2s}. \tag{A.27}$$

The deviation is unprofitable if $TC^d_{dev} \geq TC^c_{dev}$; that is, if

$$\alpha_A \geq (\beta_A + \gamma_A)(1-\delta). \tag{A.28}$$

When condition (A.28) is met, user A cannot profit by rescheduling some vehicles from the early-departure interval (t_{As}, t_{os}) in addition to all large vehicles from the late-departure interval (t_{oe}, t_{Ae}). To see why, note that the net benefit from rescheduling the vehicle at t_{As} is the same as the net benefit from rescheduling the vehicle at t_{Ae}. The benefit from rescheduling vehicles after t_{os} is lower.

Appendix A.5.4. Proof of Proposition 3

The aggregate departure rate is given by Eq. (5)

$$r(t) = \begin{cases} \frac{\alpha}{\alpha-\beta} s \text{ for } t_{os} < t < \tilde{t} \\ \frac{\alpha}{\alpha+\gamma} s \text{ for } \tilde{t} < t < t_{oe} \end{cases}.$$

Large vehicles depart at rate

$$r_A(t) = \begin{cases} 0 & \text{for } t < t_{As} \\ \frac{\alpha}{\alpha-\beta}\frac{N_A}{\Delta} & \text{for } t_{As} < t < \tilde{t} \\ \frac{\alpha}{\alpha+\gamma}\frac{N_A}{\Delta} & \text{for } \tilde{t} < t < t_{Ae} \\ 0 & \text{for } t > t_{Ae}. \end{cases}$$

Critical travel times are

$$t_{os} = t^* - \frac{\gamma}{\beta+\gamma}\frac{N_A+N_o}{s},$$

$$t_{As} = t_s^* - \frac{\beta\gamma}{\alpha(\beta+\gamma)}\left(\frac{N_A+N_o}{s} - \Delta\right),$$

$$\tilde{t} = t^* - \frac{\beta\gamma}{\alpha(\beta+\gamma)}\frac{N_A+N_o}{s},$$

$$t^* = \frac{\beta}{\beta+\gamma}t_s^* + \frac{\gamma}{\beta+\gamma}t_e^*,$$

$$t_{Ae} = t_e^* - \frac{\beta\gamma}{\alpha(\beta+\gamma)}\left(\frac{N_A+N_o}{s} - \Delta\right),$$

$$t_{oe} = t^* + \frac{\beta}{\beta+\gamma}\frac{N_A+N_o}{s}.$$

Clearly, user A cannot reduce the cost for any single vehicle in its fleet by rescheduling it to another time. It is necessary to check that user A cannot reduce its fleet cost by rescheduling a positive measure of vehicles. The first and last large vehicles to depart incur the same travel cost of

$$C^A(t_s^*) = C^A(t_e^*) = \alpha_A \frac{\beta}{\alpha}\frac{\gamma}{\beta+\gamma}\left(\frac{N_A+N_o}{s} - \Delta\right). \tag{A.29}$$

The last large vehicle imposes no delay on others in the fleet, whereas the first large vehicle imposes a delay of $1/s$ on all the others. The first vehicle can be rescheduled to just

before the travel period at a lower cost than the other vehicles. Thus, if deviation from the candidate PSNE is profitable, it must be profitable to reschedule the vehicle departing at t_{As} to t_{os}. It is straightforward to show that user A can retime departures of the remaining large vehicles so that they continue to arrive on time and incur no schedule delay cost. The net gain to the other large vehicles is therefore $\alpha_A N_A/s$. The first vehicle incurs a cost of (A.29) in the candidate PSNE, and a cost of $(\beta_A\beta/\alpha)(t_s^* - t_{os})$ if it rescheduled. The net change in costs for the fleet is

$$\begin{aligned}\Delta TC^A &= \left(\beta_A - \alpha_A\frac{\beta}{\alpha}\right)\frac{\gamma}{\beta+\gamma}\left(\frac{N_A+N_o}{s} - \Delta\right) - \alpha_A\frac{N_A}{s} \\ &= (\theta-1)\,\alpha_A\frac{\beta\gamma}{\alpha(\beta+\gamma)}\left(\frac{N_A+N_o}{s} - \Delta\right) - \alpha_A\frac{N_A}{s}.\end{aligned}$$

Deviation is not profitable if this difference is positive, which is assured by condition (32).