The Economics of Crowding in Rail Transit

André de Palma[a,b], Robin Lindsey[c], Guillaume Monchambert[d,*]

[a] *CREST, ENS Cachan, CNRS, Université Paris-Saclay, 94235 Cachan, France*
[b] *CECO, Ecole polytechnique, Université Paris-Saclay, 91128 Palaiseau, France*
[c] *Sauder School of Business, University of British Columbia, Vancouver, V6T 1Z2 British Columbia, Canada*
[d] *University of Lyon, Université Lyon 2, LAET, F69007, Lyon, France*

Abstract

We model trip-timing decisions of rail transit users who trade off crowding costs and disutility from traveling early or late. With no fare or a uniform fare, ridership is too concentrated on timely trains. Marginal-cost-pricing calls for time-dependent fares that smooth train loads and generate more revenue than an optimal uniform fare. The welfare gains from time-dependent fares are unlikely to increase as ridership grows. However, imposing time-dependent fares raises the benefits of expanding capacity by either adding trains or increasing train capacity. We illustrate these results by calibrating the model to the Paris RER A transit system.

Keywords: rail transit; crowding; pricing; optimal capacity

JEL Codes: R41; R48; D62

*Corresponding author. Present address: LAET-ISH, 14 avenue Berthelot, 69363 Lyon, France.
Email addresses: andre.depalma@ens-cachan.fr (André de Palma), robin.lindsey@sauder.ubc.ca (Robin Lindsey), g.monchambert@univ-lyon2.fr (Guillaume Monchambert)

1. Introduction

Since the pioneering work of Pigou (1920) and Knight (1924), economists have made major strides in studying automobile traffic congestion. They have developed models that describe individuals' trip-making decisions and that capture the evolution of congestion over space and time. They have analyzed first-best and second-best congestion pricing, characterized optimal road capacity, and examined the linkages between optimal capacity and the way in which usage is priced.[1] By comparison, economists have devoted much less attention to congestion delays and crowding in public transportation, and they have not explored the time dimension of transit usage in as much depth. Yet, as described below, peak-period travel delays and crowding on urban transit systems are becoming major problems.

The goal of this paper is to analyze transit crowding congestion using an approach that broadly parallels what has been done for automobile traffic congestion. We develop a simple but general model of trip-timing decisions on a crowded rail transit line with a dedicated right-of-way. We derive equilibrium usage patterns with no fare, an optimal flat fare, and an optimal time-of-day-varying (henceforth TOD) fare, and optimal transit capacity for each fare regime. To focus the analysis we address two specific questions: one concerning optimal fares and the other optimal capacity. First, how does the welfare gain from optimal TOD fares depend on the severity of crowding? Is it the case that, as in road traffic congestion models, the gains from congestion pricing increase more than proportionally with traffic volume? Second, how does congestion pricing using TOD fares affect optimal system capacity? Is it true that, as is widely believed, congestion pricing is a substitute for investment?

These questions are important because of the large and growing costs of travel delays and crowding in transit systems around the world. A recent roundtable report (OECD, 2014) identifies crowding as a major component of the cost of travel. Transit crowding imposes disutility on riders in several ways.[2] It increases waiting time and in-vehicle travel time, and reduces travel time reliability. It causes stress and feelings of exhaustion (Mohd Mahudin et al., 2012). Disutility from in-vehicle time increases with the number of users (Wardman and Whelan, 2011; Haywood and Koning, 2015). In a comprehensive analysis, Wardman and Whelan (2011) find that the monetary

[1] See Small and Verhoef (2007) for a review.
[2] See Tirachini et al. (2013).

valuation of the disutility from public transport travel time is, on average, multiplied by a factor of 2.32 if a rider has to stand. Discomfort also occurs while entering and exiting transit vehicles, accessing stations on walkways and escalators, and so on.

Several recent studies have documented the aggregate cost of crowding. For example, Prud'homme et al. (2012) estimate that the eight percent increase in passenger densities in the Paris subway between 2002 and 2007 imposed a welfare loss in 2007 of at least €75 million. Veitch et al. (2013) estimate the annual cost of crowding in Melbourne's metropolitan trains in 2011 at €208 million.

The costs of crowding are likely to increase as transit usage grows faster than capacity. As urbanization proceeds in both developed and developing countries, the number of city residents who rely on transit is rising. Younger people are delaying acquisition of a driver's license, and choosing to live in areas where transit service provides most of their travel needs. Though the automobile still dominates in the US and Canada, transit ridership is rising there too.[3] Cities, meanwhile, struggle to obtain adequate funding for capacity expansion and operations.

City planners are now recognizing that crowding should be considered in cost-benefit analysis of transit projects as well as travel-demand management policies (Parry and Small, 2009). For example, bus service was improved in London prior to introduction of the Congestion Charge in 2003. Similarly, bus, metro, and rail service were expanded in Stockholm before the Congestion Tax trial in 2006. Nevertheless, crowding and other dimensions of transit quality are still often undervalued in project evaluation relative to more easily measured metrics such as in-vehicle speed (OECD, 2014).

Expanding capacity is a natural way to alleviate transit crowding, but it is expensive and time-consuming, and residents and businesses located near transit routes often oppose it. An alternative is to use transit fares as a rationing mechanism. More than sixty years ago, Vickrey (1955) undertook a thorough study of New York City's subway fare structure. He remarked on the severe crowding at peak times, and the high cost and time lags incurred to expand capacity. He advocated a fare system based on marginal-cost-pricing principles with due regard for collection costs and revenue generation constraints, and pointed out the disadvantages of uniform fares and fares strictly proportional to

[3]Transit ridership in the US has been growing since 1995 (www.apta.com/mediacenter/pressreleases/2015/Pages/150309_Ridership.aspx). In Canada, the share of morning commute trips taken by public transit in each of the ten largest cities increased from 2006 to 2011 (http://www12.statcan.gc.ca/nhs-enm/2011/as-sa/99-012-x/2011003/tbl/tbl1b-eng.cfm; http://www12.statcan.gc.ca/nhs-enm/2011/as-sa/99-012-x/2011003/tbl/tbl1a-eng.cfm).
Information from the 2016 Canadian census is scheduled to be released in November 2017.

distance. In Vickrey (1963) he also argued for peak-load fares, and noted the common economic principles underlying congestion pricing of transit systems and congestion pricing of roads.

A number of transit agencies do vary fares by time-of-day,[4] but uniform fares are still levied in many large urban areas including Paris, Tokyo, and Toronto. Opinions differ as to whether peak-period pricing is cost-effective. One argument against it is that travelers such as morning commuters lack the flexibility to reschedule their trips. Retiming morning arrival times may also be inhibited by departure time constraints in the evening (Daniels and Mulley, 2013). However, several studies have concluded that congestion can be reduced appreciably by a suitable combination of fare surcharges and discounts. Using a simulation model, Whelan and Johnson (2004) determine that combining a fare increase in the peak with a fare reduction in the off-peak generates significant reductions in overcrowding with only marginal changes to total demand and operator revenue. Douglas et al. (2011) examine the potential of TOD fare variation to spread morning peak train usage. They find that surcharges have a bigger effect than discounts of the same magnitude, and that discounts are more effective at peak spreading on early trains than late trains.

Several surveys (e.g., of London and Melbourne) have found that many travelers are willing to change travel time by 15 minutes, and in some cases more, if they are compensated in some way (e.g., by fare reductions, faster trains, or less crowding). Off-peak discounts have been implemented in some cities, and they are popular with travelers.[5]

Beginning with Mohring (1972), most economic studies of transit pricing, investment, and subsidy policy have employed static models that do not account for travelers' time-of-use decisions and the large daily variations in ridership and crowding typical of major transit systems. Dynamic models have been used to study traffic congestion for many years, but they are not applicable to public transport because of differences in the nature of supply as well as the form that congestion

[4]These include the London subway (http://content.tfl.gov.uk/tube-dlr-lo-adult-fares.pdf), the Long Island Rail Road (http://web.mta.info/lirr/about/TicketInfo/), the Washington, D.C. metro (www.wmata.com/fares/), Seattle (http://kingcounty.gov/depts/transportation/metro/fares-orca/what-to-pay.aspx), Sydney (http://www.transportnsw.info/en/tickets/tickets-opal-fares/train.page#opaltrainfares), and Melbourne (https://www.ptv.vic.gov.au/tickets/fares/regional-fares/).

[5]In Singapore, commuting to the downtown core is free before 7:45 am, and a discount of up to 50 cents applies when arriving between 7:45 am and 8:00 am (www.lta.gov.sg/apps/news/page.aspx?c=2&id=c3983784-2949-4f8d-9be7-d095e6663632). See Lovrić et al. (2016) for a recent assessment. Similarly, in Melbourne, weekday trips on the electrified train network before 7am are free for users with the myki travel card (http://ptv.vic.gov.au/tickets/myki/myki-money/). In 2014, Hong Kong's Mass Transit Railway system introduced an Early Bird Discount Promotion to encourage users to travel before the peak. The effects are described in Halvorsen et al. (2016).

takes.[6] Transit service is provided in batch form according to a timetable so that travelers have to choose among a discrete set of departure-time alternatives rather than being able to depart whenever they want. Moreover, transit congestion often involves crowding rather than travel delay. Users pay fares that may or may not depend on time of day and distance traveled. Transit capacity also has several dimensions whereas the capacity of a road can usually be described by a single variable denoting its flow capacity in vehicles per hour.

Kraus and Yoshida (2002) take a major step toward overcoming these limitations by using a variant of the Vickrey (1969) bottleneck model to study transit congestion. Kraus and Yoshida consider a rail service between a single origin and destination. Service capacity is defined by the number of train runs, the number of trains, the capacity of each train, and the time headway between trains. The number of people who board a train is limited by its capacity, and congestion takes the form of queuing delay. Service discipline is first-come-first-served, and if the fare does not depend on time of day, in equilibrium users traveling at the peak have to wait for several trains to pass before they can board. The social optimum with no queuing can be decentralized by levying a time-varying (i.e., train-dependent) fare. Kraus and Yoshida assume that travelers cannot arrive late. Yoshida (2008) extends their model to allow late arrivals, and also considers random-access boarding discipline.

Train capacity in the Kraus and Yoshida (2002) and Yoshida (2008) model is "hard" in the sense that it has no effect on users' costs until the capacity constraint is reached, but the number of passengers who board a train cannot exceed capacity at any cost. In practice, congestion on most transit systems does not develop as abruptly as this, but rather increases steadily with passenger loads as crowding develops on walkways, escalators, and platforms, as well as in transit vehicles themselves.

A few studies have taken steps towards modeling crowding on transit systems. Huang et al. (2005) assume that all travelers waiting for a train can get on, but the discomfort incurred while aboard increases with the passenger load. Tian et al. (2007) consider a many-to-one network in which riders board trains at multiple stations and derive the equilibrium departure pattern from each originating station. Tian et al. (2009) show that the socially optimal usage pattern can be

[6] Gonzales and Daganzo (2012) study the car vs public transit competition for a single bottleneck. They assume a generalized cost of a transit trip constant, which does not depend on the number of transit users or on the road congestion.

supported using time-dependent fares. These studies do not consider the welfare gains from TOD fares or optimal service capacity, and Tian et al. (2009) assume that the number of trains is large enough that not all of them are used. de Palma et al. (2015) focus on the functional form of the crowding cost function for seated and standing passengers. Their work is complementary to our paper. They propose some parameterization of in-vehicle congestion function. They provide a continuous description of congestion under three regimes: all users have a seat; some passengers sit and others are standing; and finally very high in-vehicle congestion.

Our analysis builds on previous work in two main ways. First, we explore in depth the welfare gain from TOD transit fares, and how it depends on the functional form of the crowding cost function and capacity parameters. Second, we derive the optimal number of trains and train capacity for three fare regimes: no-fare, optimal uniform-fare, and optimal TOD fares, and compare optimal service supply across regimes. Our analysis parallels much of that in Kraus and Yoshida (2002). We highlight the distinction between crowding and queuing congestion, and note the similarities and differences between our results and those of Kraus and Yoshida (2002). Our analysis ventures beyond theirs in considering general trip-timing preferences, a regime with zero fare in which transit usage is underpriced, the excess burden of public funds, and the welfare gain from TOD fares. We also develop a numerical example based on the Paris RER A transit line.

To preview results, the answers to the two questions posed in the second paragraph are as follows. First, when capacity is fixed the welfare gain from implementing optimal TOD fares may not increase with the total number of users or, therefore, with the severity of crowding. Indeed, if the cost of crowding aboard a train grows at an increasing rate with the passenger load, the welfare gain actually decreases with the total number of users. As we elaborate later in the paper, this contrasts with intuition as well as results of automobile traffic congestion models. Second, even if the total number of users is fixed, the optimal number of trains and train capacity can be higher with optimal TOD fares than when fares are uniform for all trains. Thus, while congestion pricing improves utilization of a given transit service, it can actually increase the benefits of expanding capacity.

Section 2 describes the model. Section 3 derives the equilibrium trip-timing decisions of users with a zero fare, and Section 4 derives the optimal uniform fare. Section 5 analyzes the socially optimal distribution of users across trains, the TOD fare schedule that supports the optimum, and the welfare gain from the TOD fare. Section 6 considers the long run in which the number of

trains and train capacity are endogenous, and compares optimal capacities in the three fare regimes. Section 7 examines how the results are affected by two market distortions: the excess burden in raising public funds, and unpriced traffic congestion when users can choose between transit and driving. A numerical example based on the Paris RER A line is presented in Section 8. Section 9 concludes. Major proofs are relegated to appendixes.

2. A model of crowding on a rail transit line

In this section we introduce a general model of rail transit crowding which we call the "PTC" model. A rail line with a dedicated right-of-way connects two stations without intermediate stops.[7] There are m trains, indexed in order of departure, which run on a timetable.[8] Trains have the same capacity which is determined by the number of cars, the number of seats per car, and the amount of standing area. Service is perfectly reliable.[9] Train k leaves the origin station on schedule at time t_k, $k = 1, ..., m$. Travel time aboard a train is independent of both departure time and train occupancy, and without loss of generality it is normalized to zero.

Each morning a fixed number, N, of identical users take the line to work. They know the timetable and the crowding level on each train, and choose which train to take. By assumption, they cannot increase their chances of securing a good seat by arriving at the origin station early. Users therefore arrive just before a train comes in, and never have to wait.[10] When a train arrives at the station, all users on the platform board it. Such behavior is observed in many public transport networks around the world. See, for example Huang et al. (2005) for Beijing.[11]

Users of a train incur an expected crowding disutility, $g(n)$, where n is the number of users taking the same train.[12] Crowding disutility is zero on an empty train (i.e., $g(0) = 0$), strictly increasing

[7] Given the separate right-of-way, the model is not applicable to regular bus service that shares the road with other vehicles. With some modifications, the model could be applied to a Bus Rapid Transit (BRT) service with dedicated traffic lanes.

[8] A notational glossary is provided in Appendix D.

[9] Some implications of unreliable service are discussed in the conclusion.

[10] Similar results would obtain if users do not know the timetable. Each user would then incur an expected waiting time equal to half the headway.

[11] The model therefore excludes congestion in the form of either waiting time or longer in-vehicle travel time. In reality, any train has a finite capacity to accommodate passengers. However, as long as users anticipate the crowding cost they will incur, in equilibrium the number of passengers who attempt to board a train will never exceed its capacity. It is therefore unnecessary to include absolute train capacity constraints in the model.

[12] Function $g(n)$ is an average over possible states: securing a good seat, getting a bad seat, having to stand in the middle of the corridor, standing close to the door, etc.. The function can include disutility that is correlated with the number of people who ride a train such as the time required to buy tickets, crowding on platforms, jostling to board and alight from trains, fear of pickpockets and so on.

with n (i.e., $g'(.) > 0$), and twice continuously differentiable. Several of the results we derive depend on the shape of $g(n)$. Most empirical studies find that crowding costs are approximately linear. Whelan and Crockett (2009) conduct a stated-preference study of crowding in the UK. They find that crowding costs are approximately linear above a threshold load factor or passenger density. Linearity is also supported by Wardman and Whelan (2011) in a comprehensive analysis of crowding cost estimates for the UK, and Haywood and Koning (2015) who estimate time multiplier coefficients in the Paris subway.[13] In light of this evidence, for much of the paper we assume that $g(n)$ is linear. With linearity, the model can be readily extended to allow elastic demand, and the optimal number of trains and train capacity can be characterized as well. Nevertheless, it is clear that $g(n)$ becomes steep at very high passenger densities, and approaches a vertical line when physical limits to crowding are reached. For the short-run analysis we begin by considering general functional forms for $g(n)$, and measure its curvature by the elasticity of $g'(n)$ with respect to n:

$$\varepsilon(n) \equiv g''(n)\, n / g'(n).$$

Because trains are costly to procure and operate, it is natural to assume that all m trains are used. Letting n_k denote the number of users on train k we thus assume that for all $k = 1, ..., m$, $n_k > 0$ which implies that $g(n_k) > 0$: users incur a crowding disutility on every train.

Since travel time is normalized to zero, an individual is either at home or at work. Following Vickrey (1969) and Small (1982), time at home yields an instantaneous time-varying utility $u_h(t)$, and time at work an instantaneous time-varying utility $u_w(t)$. Let (t_B, t_E) denote the time interval (beginning to end) during which all travel takes place. It is assumed that during this interval, $u_h(t)$ is weakly decreasing, $u_w(t)$ is weakly increasing, and the functions intersect at time t^* which is the *desired arrival time* (i.e., $u_h(t^*) = u_w(t^*)$). A user taking train k gains a total utility of $U(t_k) = \int_{t_B}^{t_k} u_h(t)\, dt + \int_{t_k}^{t_E} u_w(t)\, dt - g(n_k)$. If a train with unlimited capacity left at t^*, the user could travel from home to work at t^* without suffering crowding disutility. As a consequence, his utility would be maximal and equal to $U^{\max} = \int_{t_B}^{t^*} u_h(t)\, dt + \int_{t^*}^{t_E} u_w(t)\, dt$. We define the *user travel cost*, c_k, as the difference between this hypothetical maximal utility and the actual utility of taking train k:

$$c(t_k) \equiv U^{\max} - U(t_k) = g(n_k) + \delta(t_k),$$

[13] In a recent empirical study of transit crowding costs, Tirachini et al. (2016) take linearity of the crowding cost function as given.

where $\delta(t_k)$ is the *schedule delay cost* such that $\delta(t_k) = \int_{t_k}^{t^*} (u_h(t) - u_w(t))\, dt$. Note that maximizing $U(t_k)$ is equivalent to minimizing $c(t_k)$. The schedule delay cost is the disutility an individual accumulates from not being where their utility is the highest (i.e., at home until t^*, and at work after t^*). Function $\delta(t)$ is weakly decreasing for $t < t^*$ and weakly increasing for $t > t^*$. Trains that arrive close to t^* have small values of $\delta(t)$, and will sometimes be called *timely trains*. As shown in the next section, timely trains are more heavily used than other trains.

For much of the analysis it is assumed that $\delta(t)$ has a piecewise linear form: $\delta(t) = \beta(t^* - t)$ if $t < t^*$, and $\delta(t) = \gamma(t - t^*)$ if $t \geq t^*$, where β and γ are marginal disutilites from arriving early and late, respectively. This specification, called "step preferences", is used in most studies of road traffic congestion.

In the general case, a user taking train k with n_k users incurs a combined schedule delay and crowding disutility $c(t_k) = \delta(t_k) + g(n_k)$, $k = 1, ..., m$. To economize on writing, $\delta(t_k)$ is written δ_k, and $c(t_k)$ is written c_k, unless time dependence is required for clarity.

3. Equilibrium departure times with a zero fare

In this section we characterize user equilibrium when there is a uniform fare (i.e., independent of k). This regime serves as a benchmark against which to compare TOD fares. With the total number of users, N, fixed, a uniform and positive fare would not affect either the division of users between trains or crowding costs. We first describe user equilibrium given the general crowding cost function, and then consider the linear version.

3.1. General crowding costs

Let superscript "e" denote the no-fare or user equilibrium (UE), and c^e the equilibrium trip cost which is to be determined. In UE, users distribute themselves between trains so that the user cost on every train is c^e. The equilibrium is a pure-strategy Nash equilibrium with departure times as the strategy variables. Hence, $\delta_k + g(n_k^e) = c^e$, $k = 1, ..., m$. Given $g'(.) > 0$, the inverse function $g^{-1}(.)$ exists, with $g^{-1}(0) = 0$ and $\left(g^{-1}\right)'(.) > 0$. The UE can therefore be solved for the equilibrium number of users on each train, n_k^e, as a function of c^e:

$$n_k^e = g^{-1}(c^e - \delta_k), \quad k = 1, ..., m. \tag{1}$$

Since every user has to take some train, $\sum_{k=1}^{m} n_k^e = N$, or $\sum_{k=1}^{m} g^{-1}(c^e - \delta_k) - N = 0$. This equation implicitly determines a unique value of c^e. Figure 1 depicts a UE for seven trains ($m = 7$).

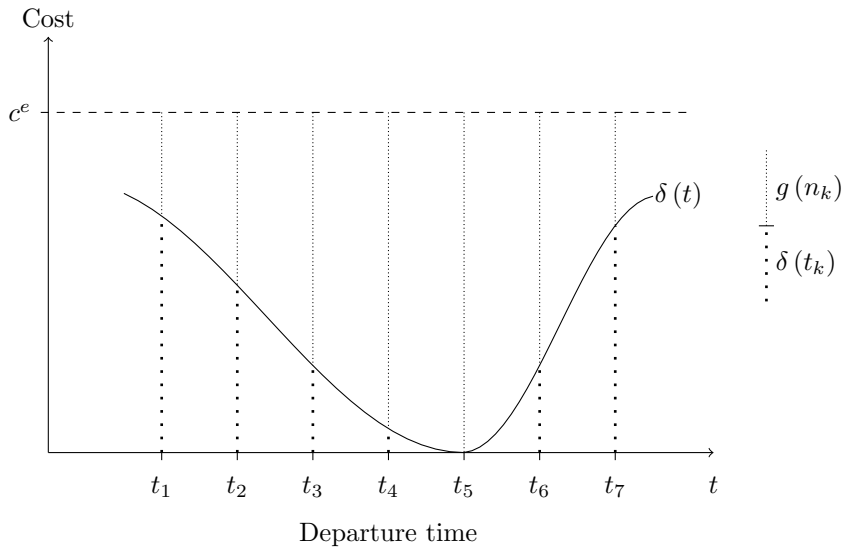

Figure 1: Schedule delay $\delta(t_k)$, crowding cost $g(n_k)$ and equilibrium cost c^e for seven trains, $t_5 = t^*$

Train $k = 5$ arrives on time and carries the most users.

Comparative statics properties of UE with respect to N are described in:[14]

Proposition 1. *In equilibrium with no fare, user cost is an increasing function of N. It is convex, linear, or concave if $g(.)$ is convex, linear, or concave, respectively.*

As expected, user cost is an increasing function of total patronage, N. Less obvious is that the curvature of $c^e(N)$ depends on the crowding cost function rather than the schedule delay cost function. This is because the train timetable is fixed in the short run, and the period of usage cannot be extended earlier or later in response to growing demand. Furthermore, since each train's arrival time is fixed, the schedule delay cost incurred when taking a given train does not depend on N. The only way to accommodate additional demand is to carry more passengers on each train. Equilibrium user cost therefore increases at an increasing (resp. decreasing) rate with N if the marginal cost of crowding aboard a train increases (resp. decreases) with ridership. This property of the model implies that, using estimates of the crowding cost function, a rail transit operator can predict the effects of rising patronage on user cost without knowing the schedule delay cost

[14]Proof: $\partial c^e/\partial N = 1/\sum_{k=1}^{m} \left(g'(u_k)\right)^{-1} > 0$, where $u_k \equiv g^{-1}(c^e(N) - \delta_k)$. The second derivative, $\partial^2 c^e/\partial N^2$, has the same sign as $\sum_{k=1}^{m} g''(u_k) \left(g'(u_k)\right)^{-3}$ which in turn has the same sign as $g''(.)$.

function.

3.2. Linear crowding costs

Given the empirical evidence, noted above, supporting linearity of crowding costs we examine further the solution when the crowding cost function $g(n)$ is linear. Assume that $g(n) = \lambda n/s$, where $s > 0$ is a measure of train capacity, and $\lambda > 0$. The private cost of taking train k is then $c_k = \delta_k + \lambda n_k/s$, $k = 1, ..., m$. Define $\overline{\delta} \equiv \frac{1}{m}\sum_{k=1}^{m}\delta_k$ as the unweighted average scheduling cost for trains. Eq. (1) can be solved explicitly to obtain:

$$n_k^e = \frac{N}{m} + \frac{s}{\lambda}\left(\overline{\delta} - \delta_k\right), \tag{2}$$

$$c^e = \overline{\delta} + \frac{\lambda N}{ms}. \tag{3}$$

For given values of m and s, equilibrium usage of each train and equilibrium trip cost are linear increasing functions of ridership, N. The difference in loads between two successive trains is proportional to parameter s, and inversely proportional to λ.[15]

4. The optimal uniform fare

User equilibrium in the PTC model is inefficient because users impose external crowding costs on each other. To assess the role of fares in internalizing the externality we briefly consider the general crowding cost function, and then turn to the linear variant.

4.1. General crowding costs

The marginal social cost of a trip, MSC, is derived by differentiating the equilibrium total cost function, $TC^e = c^e \times N$, with respect to N: $MSC^e \equiv \partial TC^e/\partial N = c^e + (\partial c^e/\partial N)\, N$. The average marginal external cost of a trip is therefore $MEC^e \equiv MSC^e - c^e = (\partial c^e/\partial N)\, N$. With elastic demand (introduced in Section 5), transit is overused with a zero fare.[16] If the fare system is restricted to uniform fares, the fare should be set equal to the average marginal external cost:

$$\tau^u = \frac{\partial c^e}{\partial N} N, \tag{4}$$

where superscript "u" denotes the optimal uniform fare.

[15] All trains are used provided N is sufficiently large; see on-line Appendix E.

[16] In Section 7.2 we show that this is not necessarily true if transit is an alternative to driving and traffic congestion is severe.

4.2. Linear crowding costs

With linear crowding costs, the user cost function is given in (3) and the marginal social cost of a trip is

$$MSC^e = \overline{\delta} + 2\frac{\lambda N}{ms}. \tag{5}$$

The optimal uniform fare in (4) works out to:

$$\tau^u = \frac{\lambda N}{ms},$$

and fare revenue is

$$R^u = \tau^u N = \frac{\lambda N^2}{ms}. \tag{6}$$

Total schedule delay costs, SDC, total crowding costs, TCC, and total travel costs net of the fare, TC, are given by

$$SDC^e = \overline{\delta}N - 4RV^o,\ \ TCC^e = \frac{\lambda N^2}{ms} + 4RV^o,\ \ TC^e = \overline{\delta}N + \frac{\lambda N^2}{ms}, \tag{7}$$

where

$$RV^o \equiv \frac{s}{4\lambda}\left(\sum_{k=1}^{m}\delta_k^2 - \frac{1}{m}\left[\sum_{k=1}^{m}\delta_k\right]^2\right). \tag{8}$$

By the Cauchy-Schwarz inequality, $RV^o > 0$. As shown in Section 5, RV^o corresponds to the welfare gain from the socially-optimal time-varying fare. Equation (7) reveals that total schedule delay costs are lower than if users were equally distributed across trains (in which case $SDC^e = \overline{\delta}N$). Total crowding costs are higher by the same amount. This is because users crowd onto timely trains that arrive closer to t^*.

The optimal uniform fare does not support the social optimum because the marginal external cost of crowding varies with train occupancy which is larger on timely trains. As explained in the next section, the social optimum can be achieved by levying time-dependent (i.e., train-specific) fares.

5. Socially optimal departure time pattern

In this section we derive the socially optimal departure time pattern and TOD fare structure that supports it. We then characterize the welfare gain from the TOD fare. Finally, we rank usage, private costs, consumers' surplus, and social surplus for the no-fare, uniform-fare, and socially optimal fare regimes.

5.1. General crowding costs

In the social optimum (SO), users are distributed between trains to equalize the marginal social costs of trips, rather than user's private costs as in the UE. The marginal social cost of using train k is $MSC_k = \partial (c_k n_k) / \partial n_k = \delta_k + v(n_k),\ k = 1, ..., m$, where $v(n_k) \equiv g(n_k) + g'(n_k) n_k$ is the marginal social crowding cost on train k. Let superscript "o" denote the SO. Total costs in the SO are $TC^o = \sum_{k=1}^{m} c_k n_k$, and the marginal social cost of a trip is $MSC^o = \partial TC^o / \partial N$. At the optimum, users are distributed across trains so that $MSC_k = MSC^o$ for every train:

$$\delta_k + v(n_k^o) = MSC^o,\ k = 1, ..., m. \tag{9}$$

Since $g'(.) > 0$ for $n > 0$, the marginal social crowding cost is always positive. In practice, it may not increase monotonically at all levels of ridership.[17] To facilitate analysis, however, we assume that $v'(.) > 0$:

Assumption 1. *The marginal social crowding cost, $v(n)$, is an increasing function of n.*

Assumption 1 is equivalent to assuming that $\varepsilon(n) > -2$. It is satisfied for all increasing and convex $g(.)$ functions and in particular for all power functions $g(n) \propto n^r$, $r > 0$. Given Assumption 1, the inverse function $v^{-1}(.)$ exists and it is increasing. Eq. (9) yields

$$n_k^o = v^{-1}(MSC^o - \delta_k). \tag{10}$$

Since all users must take some train in the SO, $\sum_{k=1}^{m} n_k^o = N$. Given Eq. (10), $\sum_{k=1}^{m} v^{-1}(MSC^o - \delta_k) - N = 0$, which implicitly determines a unique value of MSC^o. A counterpart to Prop. 1 then follows:

Proposition 2. *In the social optimum, the marginal social cost of a trip is an increasing function of N. It is convex, linear, or concave if $v(.)$ is convex, linear, or concave respectively.*

Comparing Prop. 2 with Prop. 1 it is clear that $v(.)$ plays the same role in shaping the SO as $g(.)$ does for the UE.[18]

[17] For example, $v(.)$ may increase as seats fill up, and then flatten out once all seats are occupied because standing passengers impose little inconvenience on others as long as the corridor and doors of rail cars remain clear. For details see de Palma et al. (2015).

[18] Note that $v''(n) = 3g''(n) + ng'''(n)$. The marginal social cost of a trip can therefore be a convex function of N even if the user cost function is concave in N, and vice versa.

We now examine how riders are distributed over trains. Intuition suggests that, as discussed in de Palma et al. (2015), passenger loads are spread more evenly in the SO than the UE because smoother loads should reduce the total costs of crowding. In fact, this is not invariably true but depends on how the marginal external crowding cost varies with usage. For any train, the derivative of the marginal external crowding cost is $d\left(g'\left(n\right)n\right)/dn = g'\left(n\right)\left(1+\varepsilon\left(n\right)\right)$. The marginal external crowding cost increases with usage if $\varepsilon\left(n\right) > -1$, and decreases with usage if $\varepsilon\left(n\right) < -1$. The load patterns in the SO and UE are compared in:

Proposition 3. *If $\varepsilon\left(n\right) > -1$ (respectively, $\varepsilon\left(n\right) < -1$) the socially optimal distribution of users across trains is a mean-preserving spread (respectively contraction) of the user equilibrium distribution of users across trains.*

The SO load pattern is a mean-preserving spread of the UE load pattern if the SO load pattern has more weight in the tails than the UE load pattern.[19] If the marginal external crowding cost increases monotonically with passenger load, then $\varepsilon\left(n\right) > -1$.[20] If so, the marginal social costs of trips on two trains with unequal loads differ by more than their user costs. Consequently, the SO balance between crowding costs and schedule delay costs calls for a smaller range of train loads than in the UE. Conversely, if $\varepsilon\left(n\right) < -1$, which is possible only if $g(.)$ is sufficiently concave,[21] then passenger loads are more peaked in the SO than the UE.

In summary, the difference between the SO and UE train loads depends on the curvature of the crowding cost function. If $g(.)$ is linear or convex, $\varepsilon\left(n\right) \geq 0$ and ridership in the UE is too concentrated on timely trains and should be spread out.

Regardless of whether the SO is more or less peaked than the UE, the SO usage pattern can be decentralized by charging a fare on train k equal to the marginal external cost of usage.[22] We will call the fare pattern the *SO-fare*. Given $MSC_k - c_k = g'\left(n_k\right)n_k$, the SO-fare is

$$\tau_k^o = g'\left(n_k^o\right)n_k^o,\ k = 1, ..., m. \tag{11}$$

[19] The relevant definition of MPS for discrete distributions is found in Rothschild and Stiglitz (1970).

[20] Similar to Assumption 1, which is weaker, $\varepsilon\left(n\right) > -1$ is satisfied for all convex crowding cost functions, and crowding cost functions that belong to the class of power functions: $g\left(n\right) \propto n^r$, $r > 0$.

[21] For example, inequality $\varepsilon\left(n\right) < -1$ holds for the function $g\left(n\right) = c_0 + c_1 \ln\left(n\right) - kn$ for $c_0 > k$ and over the range $n \in \left[1, c_1/k\right)$.

[22] The fare is set according to Pigouvian principles. Revenue generation incentives are considered in Section 7. Tian et al. (2009) show that the SO usage pattern can be supported using train-dependent fares under the assumption that the number of trains is large enough that not all of them are used.

With this fare structure in place, users of train k incur a private cost equal to the social cost of a trip: $p_k^o = c_k^o + \tau_k^o = MSC^o,\ k = 1, ..., m$. The SO is more efficient than the UE because users are better distributed between trains. However, inclusive of the SO-fare users incur a higher private cost in the SO. To see this, note that at least one train is more crowded in the SO than the UE. Compared to the UE, in the SO a rider of that train incurs the same schedule delay cost but a higher crowding cost and a positive fare. Since all users incur the same private cost in the UE, and all users incur the same private cost in the SO, private costs are higher in the SO.

Unless fare revenues are used to improve service in some way, charging fares to price crowding costs in the PTC model leaves users worse off. We prove in on-line Appendix F that $\partial R^i/\partial N = \partial MSC^i/\partial N \cdot N,\ i = u, o$. Thus, in both fare regimes fare revenue increases with total usage (N) as long as the marginal social cost of a trip increases with N.

Next, we examine how the welfare gain from implementing the SO-fare varies with usage. Let G^{eo} denote the welfare gain in shifting from the UE to the SO:

$$G^{eo} \equiv TC^e - TC^o.$$

We have seen that the rate at which the cost of crowding increases with load depends on the curvature of the crowding cost function, $g(\cdot)$. It turns out that properties of the crowding cost function also govern how G^{eo} depends on N.

Proposition 4. *(a) If the crowding cost is linear, then the welfare gain G^{eo} is independent of N. (b) If $\frac{d\varepsilon(n)}{dn} \geq 0$, $\varepsilon(n) > -1$, and $v''(n) < 0$, then the welfare gain G^{eo} increases with N. (c) If $\frac{d\varepsilon(n)}{dn} \leq 0$, $\varepsilon(n) > -1$, and $v''(n) > 0$, then the welfare gain G^{eo} decreases with N.*

Note that $v''(n) > 0$ if the marginal social cost of crowding is a strictly convex function of load, and $v''(n) < 0$ if it is a strictly concave function. Proposition 4 identifies conditions under which G^{eo} increases, decreases, or is independent of total ridership. Since the conditions are not collectively exhaustive, Prop. 4 does not establish the direction of change for all cases. Nevertheless, the conditions span a broad set of functions. For example, suppose the crowding cost function satisfies $g(n) \propto n^r$. Then part (a) of Prop. 4 holds if $r = 1$, part (b) holds if $0 < r < 1$, and part (c) holds if $r > 1$.

According to Prop. 4, if $g(\cdot)$ is convex the welfare gain from TOD fares actually decreases as ridership increases. Intuition might suggest otherwise. If the severity of crowding increases

with ridership, one would expect the welfare gain from congestion pricing to increase per user. In addition, more users would benefit from congestion relief. Such is the case in the Vickrey (1969) bottleneck model of queuing congestion used by Kraus and Yoshida (2002). In the bottleneck model the variable cost of a trip with no toll, a uniform toll, or a fine (i.e., queue-eliminating) toll is proportional to N. The total cost of N trips is therefore proportional to N^2, and the welfare gain from implementing the fine toll is proportional to N^2 (Arnott et al., 1993). The welfare gain G^{eo} thus increases with the square of ridership.

To understand why the result is so different with transit crowding congestion, note that the welfare gain derives from reallocating users between trains as in Prop. 3. If $g(.)$ is convex, users are reallocated more evenly, but the UE and SO train loads become more similar as N rises. The *amount* of user reallocation decreases, and the total welfare gain therefore falls. As $g(n)$ becomes increasingly convex, it approaches an inverse-L curve in shape. The cost of crowding remains very low until trains are nearly full, and then rises rapidly. Reallocating users between trains then provides little benefit.[23]

To obtain further insights, we now turn to the linear crowding cost function. As indicated in Prop. 4, the linear function is a knife-edge case in which the welfare gain from TOD fares is independent of total ridership.

5.2. Linear crowding costs

With the linear crowding cost function, the marginal social cost of crowding is $v(n) = 2\lambda n/s$ and the social cost of taking train k is $MSC_k = \delta_k + 2\lambda n_k/s$, $k = 1, ..., m$.

Eqs. (10) and (11) give:

$$n_k^o = \frac{N}{m} + \frac{s}{2\lambda}\left(\overline{\delta} - \delta_k\right), \tag{12}$$

$$\tau_k^o = \frac{\lambda}{s} n_k^o. \tag{13}$$

[23]Another way to view Prop. 4 is in terms of the marginal social cost of usage, which is MSC^e in the UE and MSC^o in the SO. If $MSC^o < MSC^e$, an additional user raises total costs by less in the SO than the UE, and G^{eo} rises. Conversely, if $MSC^o > MSC^e$, total costs rise more in the SO and G^{eo} falls. Thus, if $g(.)$ is convex an additional user is, paradoxically, more costly to accommodate in the SO than in the UE even though users are distributed optimally between trains in the SO. If $g(.)$ is linear, $MSC^o = MSC^e$ and the difference in total costs between UE and SO is independent of N. In effect, the benefits of internalizing the crowding cost externality are exhausted once total usage is large enough for all trains to be used. We illustrate this case diagrammatically in the next subsection.

The marginal social cost of a trip is the same for all trains and equal to

$$MSC^o = \overline{\delta} + 2\frac{\lambda N}{ms}$$

which is the same as Eq. (5) for the MSC^e in UE. Since MSC^o is an increasing function of N, by Prop. 3 train loads are more evenly distributed in the social optimum than the uniform-fare equilibrium. Indeed, given (2) and (12) the difference in loads between successive trains is only half as large.[24] Compared to the optimal uniform fare in Eq. (4), the SO-fare is lower on the earliest and latest trains with $\delta_k > \overline{\delta}$, and higher on timely trains with $\delta_k < \overline{\delta}$.

Given Eqs. (12) and (13), total revenue from the SO-fare is

$$R^o = \frac{\lambda}{ms}N^2 + RV^o, \tag{14}$$

where RV^o is given in Eq. (8), and repeated here for ease of reference:

$$RV^o \equiv \frac{s}{4\lambda}\left(\sum_{k=1}^{m}\delta_k^2 - m\overline{\delta}^2\right). \tag{15}$$

The first term in (14) matches revenue from the optimal uniform fare, R^u, in Eq. (6). The second term, RV^o, is extra revenue (when $m > 1$) due to variation of the fare, and it will be called *variable revenue*. Total schedule delay costs, total crowding costs, and total travel costs net of the fare are given by

$$SDC^o = SDC^e + 2RV^o,\ \ TCC^o = TCC^e - 3RV^o,\ \ TC^o = TC^e - RV^o. \tag{16}$$

Total schedule delay costs are higher in the SO than the UE, but crowding costs are smaller and total costs are lower by an amount equal to variable revenue. With linear crowding costs, the welfare gain from imposing the SO-fare is therefore equal to variable revenue: $G^{eo} = RV^o$. This means that welfare gains can be estimated by comparing total crowding costs in the no-fare equilibrium with the costs that would accrue if users were equally distributed over trains. Information on scheduling costs is not needed.[25]

Consistent with Prop. 4, G^{eo} is independent of total usage, N. To see why, consider a simple case with two trains. The cost of using train k is $c_k = \delta_k + gn_k$ where $g \equiv \lambda/s$ measures the rate

[24] Again, all trains are used provided N is sufficiently large. As explained in on-line Appendix E, this is assured if all trains are used in the no-fare or uniform-fare equilibrium.

[25] Likewise, in the bottleneck model used by Kraus and Yoshida (2002) the welfare gains from the optimal fare equal fare revenues. However, TOD pricing is more effective because it eliminates queuing congestion without increasing schedule delay costs.

at which crowding costs increase with train load. Figure 2 depicts the UE and SO using a diagram with two vertical axes separated by N. Usage of train 1 is measured to the right from the left-hand axis, and usage of train 2 to the left from the right-hand axis. By assumption, $\delta_2 > \delta_1$ so that train 1 is overused in the UE. The welfare gained in shifting users from train 1 to train 2 is shown by the triangular shaded area. The height of the triangle is $\delta_2 - \delta_1$, and the width of the triangle is $(\delta_2 - \delta_1)/(2g)$. The area of the triangle is therefore $(\delta_2 - \delta_1)^2/(4g)$. It does not depend on N because neither dimension of the triangle depends on N. The height of the triangle equals the difference in marginal external costs of using the two trains in the UE. This is determined by the difference in their attractiveness, $\delta_2 - \delta_1$, not N. The width of the triangle is the optimal number of users to redistribute between trains, which is proportional to $\delta_2 - \delta_1$ and inversely proportional to g. This, too, is independent of N. With multiple trains, the welfare gain from redistributing users between trains is the sum of analogous triangular areas.

The numerical example in Section 8 features linear schedule delay costs and a constant headway, h, between trains. Given these assumptions, and $G^{eo} = RV^o$, it is straightforward to show that for large values of m,

$$G^{eo} \simeq \frac{s}{48\lambda}\left(\frac{\beta\gamma}{\beta+\gamma}\right)^2 h^2 m\left(m^2-1\right). \tag{17}$$

Eq. (17) reveals how the welfare gain from the SO-fare varies with parameters. First, G^{eo} varies with the square of the unit costs of schedule delay β and γ together. This is consistent with the quadratic dependence of G^{eo} on the schedule delay costs, δ_1 and δ_2, in the example with $m = 2$. For the same reason, G^{eo} varies with the square of the headway, h. G^{eo} varies inversely with the ratio $g = \lambda/s$ because the scope to alleviate crowding by redistributing riders between trains decreases if trains become crowded more quickly.

Finally, G^{eo} varies approximately with the cube of the number of trains, m. This highly nonlinear dependence is due to two multiplicative factors. First, with h given, the average schedule delay cost of trains is proportional to m. The average difference in schedule delay costs is therefore proportional to m, and the welfare gain from redistributing passengers between two trains varies with m^2. Second, the number of trains between which passenger loads can gainfully be redistributed is approximately proportional to m. Hence, the overall welfare gain varies approximately with m^3.

In the introduction to the paper we noted that the distribution of passengers between trains is governed by the trade-off users face between scheduling costs and crowding costs. It is therefore surprising that the parameters measuring the strength of these two costs affect the welfare gain

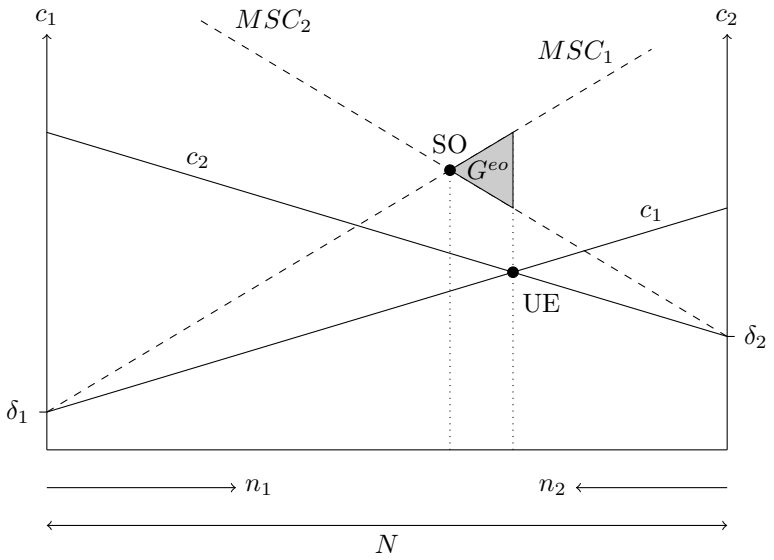

Figure 2: User equilibrium (UE), social optimum (SO) and welfare gain (G^{eo}) with two trains

from congestion pricing in very different ways. According to Eq. (17), doubling the unit costs of schedule delay, β and γ, increases the welfare gain four-fold. By contrast, doubling the crowding cost parameter, λ, reduces the gain by half. In assessing the potential benefits from implementing congestion pricing, it is therefore important to predict how the parameter values will evolve over time. If parameters β, γ, and λ all grow at a rate r, G^{eo} grows at rate r too. By contrast, if work hours become more flexible in the future, β and γ could stagnate while λ continues to rise. Other things equal, G^{eo} would then decline.

To enhance tractability and to obtain more definitive results, for the rest of the paper it is assumed that the crowding cost function is linear.

5.3. Elastic demand

So far it has been assumed that transit ridership is exogenous. In practice, travelers can often use other transport modes and they may choose to forego travel if it is too costly. To admit these possibilities we now assume that transit demand is a smooth and decreasing function of the private cost:

$$N = N(p), \quad \frac{\partial N}{\partial p} < 0. \tag{18}$$

Consumers' surplus from trips is $CS(p) = \int_p^\infty N(u)\,du$, and social surplus (gross of capacity costs) is the sum of consumers' surplus and fare revenue: $SS(p,\tau) = CS(p) + R$. Any market failures

arising other than in transit usage are ignored.

Let superscript n denote the no-fare regime and a hat ($\hat{}$) denote an equilibrium value with elastic demand. Define $\widehat{G}^{ij} \equiv \widehat{SS}^{j} - \widehat{SS}^{i}$ as the welfare gain in shifting from regime i to regime j when demand is elastic. The equilibrium values in the no-fare, optimal uniform-fare, and SO-fare regimes are compared in:

Proposition 5. *With linear crowding costs, elastic demand, and given values of m and s, equilibrium private costs are the same in the SO-fare (o) and optimal uniform-fare (u) regimes, and lower in the no-fare regime (n): $\widehat{p}^{o} = \widehat{p}^{u} > \widehat{p}^{n}$.*

Equilibrium usage satisfies $\widehat{N}^{o} = \widehat{N}^{u} < \widehat{N}^{n}$, consumers' surplus satisfies $\widehat{CS}^{o} = \widehat{CS}^{u} < \widehat{CS}^{n}$, and social surplus satisfies $\widehat{SS}^{o} > \widehat{SS}^{u} > \widehat{SS}^{n}$. Consequently, $\widehat{G}^{no} > \widehat{G}^{nu} > 0$, and $\widehat{G}^{no} > \widehat{G}^{uo} > 0$.

Proposition 5 is proved in on-line Appendix G. The rankings of social surplus and welfare gains across the three regimes are intuitive. As noted in Section 5.2, the marginal social cost of a trip is the same in the uniform-fare and SO-fare regimes. Optimal usage and consumers' surplus in the two regimes are therefore the same too. Thus, introducing differentiated fares creates a welfare gain without making users worse off. Looked at another way, it allows the operator to boost farebox revenues without reducing ridership.

6. Optimal transit service

We now turn attention to the long run when the transit authority can choose the number of trains, m, train capacity, s, and the train timetable. For tractability, we assume that schedule delay costs are linear. We also assume that the headway between trains is a given constant, h. This is reasonable if the headway is set to the *safe headway*: the shortest technologically feasible interval consistent with safe operations.[26] The optimal timetable is derived for given values of m and s in on-line Appendix H. In this section we begin by deriving properties of the optimal m and s for a general capacity cost function. Then we adopt a specific function and derive formulas for the optimal m and s.

[26]Since all users wish to arrive at the same time, t^*, it is optimal to choose the shortest headway possible. In practice it may not be possible to maintain a minimum headway throughout the travel period. As Kraus and Yoshida (2002) and Kraus (2003) show, if the number of train sets is small compared to the time required for a train to make a round trip, train sets may be run in a series of clusters. Trains in the same cluster are separated by the minimum headway, but there is a longer gap between the last train in one cluster and the first train in the next one.

6.1. General capacity cost function

Let $K(m,s)$ denote the cost of providing service,[27] where m is treated as a continuous variable. Function $K(m,s)$ is assumed to be a strictly increasing and differentiable function of m and s. Let superscript i denote the pricing regime, $i = n, u, o$. Social surplus net of capacity costs is

$$SS^i = \int_0^N p(n)\,dn - \left(\bar{\delta}N + \frac{\lambda N^2}{ms} - RV^i(m,s) + K(m,s)\right),$$

where $RV^n = RV^u = 0$, and RV^o is a function of m and s, but does not depend on usage. The transit authority chooses m and s to maximize SS^i. To economize on notation, let K_m and K_s denote the derivatives of $K(m,s)$ with respect to m and s respectively.

Proposition 6. *Let $i = n, u, o$ index the pricing regime. Then, first-order conditions for a maximum of SS^i are*

$$\textit{For } s \quad : \qquad \frac{\lambda N^2}{ms^2} \cdot D^i = K_s - RV_s^i, \tag{19a}$$

$$\textit{For } m \quad : \qquad \left(\frac{\lambda N}{m^2 s} - \frac{\partial \bar{\delta}}{\partial m}\right) N \cdot D^i = K_m - RV_m^i, \tag{19b}$$

where $D^n = \frac{p_N N}{p_N N - \frac{\lambda N}{ms}} < 1$ and $D^u = D^o = 1$.

Proposition 6 is proved in on-line Appendix I. The LHS of (19a) is the marginal benefit of expanding train capacity. The first term of the product is the marginal benefit from expanding train capacity if usage remained fixed. The cost of crowding would decrease by $\lambda N/\left(ms^2\right)$ for each of the N users. In the no-fare regime, $D^n < 1$, and the actual reduction in crowding cost is smaller than this because improved service attracts new users who value trips less than the marginal social cost they impose. This is the induced demand effect that has been well studied in the road traffic congestion literature (e.g., Duranton and Turner, 2011). Indeed, in the limit of perfectly elastic demand (i.e., $p_N \to \infty$), the potential benefit from expanding m or s is completely dissipated. The RHS of (19a) is the marginal cost of expanding train capacity. Since $RV_s^o > 0$, in regime o the marginal financial cost of expanding capacity is effectively reduced by the generation of additional variable revenue. For the other two regimes there is no such benefit.

Equation (19b) has a similar interpretation to Eq. (19a). The LHS of (19b) is the marginal benefit of adding a train, and the RHS is the marginal cost. Term $\lambda N/\left(m^2 s\right)$ is the marginal

[27] Service cost includes capital, operations, and maintenance. It is assumed to be independent of usage. Adding N as an argument of the service cost function would not affect results of interest.

benefit per user from less crowding, and term $\partial\bar{\delta}/\partial m$ is the marginal disbenefit from greater average schedule delay costs. In the no-fare regime, the marginal net benefit is diluted by the same factor, D^n, as in Eq. (19a).

In contrast to the no-fare regime, the marginal benefit from expanding capacity in the optimal uniform-fare regime is not diluted by additional usage because usage is priced efficiently. This might suggest that the optimal values of s and m, s^u_* and m^u_*, are larger than their counterparts with a zero fare, s^n_* and m^n_*. However, at least for given values of s and m, usage is higher in the no-fare regime as per Prop. 5. This leaves the rankings of s^u_* and s^n_*, and m^u_* and m^n_*, ambiguous in general.

As just noted, the generation of variable revenue from the SO-fare, RV^o, effectively reduces the financial cost of expanding either s or m. Hence, conditional on the values of N and m, optimal train capacity is larger in the social optimum than in the optimal uniform-fare regime: $s^o_*(m,N) > s^u_*(m,N)$. Similarly, conditional on N and s, the optimal number of trains is larger in the social optimum than in the optimal uniform-fare regime: $m^o_*(s,N) > m^u_*(s,N)$. These rankings may seem surprising given that total system costs are lower in the social optimum than the uniform-fare regime. Inequality $m^o_*(s,N) > m^u_*(s,N)$ is explained by the fact that ridership is distributed more evenly across trains in the social optimum. More users take the earliest and latest trains in the social optimum, which makes adding extra trains more beneficial. To understand the inequality $s^o_*(m,N) > s^u_*(m,N)$, recall from Eq. (17) that in the uniform-fare regime the deadweight loss from imbalanced ridership between trains *increases* with s. Expanding capacity is therefore more valuable in the SO.

Despite the inequalities $s^o_*(m,N) > s^u_*(m,N)$ and $m^o_*(s,N) > m^u_*(s,N)$, there is no guarantee that the unconditionally optimal values (s^o_*, m^o_*) in the social optimum are both larger than their counterparts (s^u_*, m^u_*) in the uniform-fare regime. One reason is that $s^u_*(m,N)$ is a decreasing function of m, and $m^u_*(s,N)$ is a decreasing function of s, and one function can shift much more than the other. The other reason is that usage generally differs in the two regimes; i.e., $N^o_* \neq N^u_*$.

6.2. A specific capacity function

As we do here, Kraus and Yoshida (2002) derive optimal capacity for a rail service. They distinguish in their model between the number of train runs and the number of train sets (a train set can make more than one run). They also account for the time required for a train set to make a round trip. Our model excludes these variables, and we adopt a simpler service cost function of

the form

$$K(m,s) = (\nu_0 + \nu_1 s)\, m + \nu_2 s, \tag{20}$$

where ν_0, ν_1, and ν_2 are all non-negative parameters. The term $\nu_0 + \nu_1 s$ in (20) is the incremental capital, operating, and maintenance costs of running an additional train. It increases linearly with the train capacity. If $\nu_0 > 0$, there are scale economies with respect to train size. The second term in (20), $\nu_2 s$, accounts for costs that depend on train capacity but not the number of trains. Kraus and Yoshida (2002) interpret this term as capital costs for terminals.[28] In this subsection we focus on the optimal uniform fare and SO-fare regimes because the slope of the demand function does not affect the optimal values of m or s, and properties of the solution can be derived while treating N parametrically.

Given the service cost function (20), first-order conditions (19a) and (19b) for s and m in the uniform fare and SO-fare regimes are

$$\text{For } s \quad : \quad \frac{\lambda N^2}{m s^2} = \nu_1 m + \nu_2 - RV_s^i,\ i = u, o, \tag{21a}$$

$$\text{For } m \quad : \quad \left(\frac{\lambda N}{m^2 s} - \frac{\partial \bar{\delta}}{\partial m}\right) N = \nu_0 + \nu_1 s - RV_m^i,\ i = u, o. \tag{21b}$$

As noted above, there is no guarantee that capacity is larger in the social optimum than the uniform-fare regime in the sense that $s_*^o(N) > s_*^u(N)$ and $m_*^o(N) > m_*^u(N)$. Nevertheless, with capacity function (20) it is possible to establish some results on how capacity depends on the fare regime. These results apply not only to the optimal uniform-fare and social optimum regimes, but also to fare regimes with intermediate efficiencies. To formalize this idea, let $f \in [0,1]$ be an index of fare-regime efficiency, with regime f yielding variable revenue of $f \cdot RV^o$. For the optimal uniform-fare regime, $f = 0$, and for the social optimum regime, $f = 1$. Regimes with intermediate values of $f \in (0,1)$ impose an optimal average level of fare, and fare variations by time of day that partly internalize differences between trains in external crowding costs. Given two regimes with indexes f_1 and $f_2 > f_1$, we will say that the second regime is *more efficient* than the first.

As formalized in the following proposition, definitive capacity rankings can be derived when transit demand is inelastic:[29]

[28] They note that the linear specification is applicable if terminal cost is proportional to terminal area, and terminal area is proportional to train capacity.

[29] The rankings of m, s, and ms in Proposition 7 hold under ceratin conditions on the general capacity function $K(m,s)$ that include the specific capacity function (20) as a special case. The more general conditions do not have a ready interpretation, and since the derivations are tedious the more general results are not stated here.

Proposition 7. *Assume that transit demand is price inelastic. Then a more efficient fare regime has a larger number of trains (m), a smaller train capacity (s), and a larger fleet capacity (ms).*

For the optimal uniform-fare and social optimum regimes, Proposition 7 implies that, for any value of N, $m_*^o > m_*^u$, $s_*^o < s_*^u$, and $m_*^o s_*^o > m_*^u s_*^u$. As noted in subsection 6.1, inequality $m_*^o > m_*^u$ is consistent with the fact that additional trains (which are less convenient than existing trains) are more valuable in the social optimum than the uniform-fare equilibrium because the additional trains are more heavily used. However, the intuitive argument that train capacity is also larger in the social optimum does not go through. This highlights the importance of treating separately different dimensions of transit service capacity, rather than only considering some aggregate measure of capacity such as total seats.[30]

When transit demand is price elastic, the capacity rankings are conclusive only for the number of trains:[31]

Proposition 8. *Assume that transit demand is price elastic. Then a more efficient fare regime has a larger number of trains (m). Train capacity (s), fleet capacity (ms), and total usage (N) can be larger or smaller with a more efficient fare regime.*

Together, Propositions 7 and 8 reveal that whether train capacity should be increased or decreased in response to a change of fare regime can depend on demand elasticity. In the numerical example of Section 8 it turns out that when demand is relatively inelastic, train capacity is smaller with the SO-fare than the optimal uniform-fare, but the ranking is reversed when the elasticity is large enough. Since transit demand elasticities depend on various factors including fare levels, modal split, severity of traffic congestion and so on, the appropriate direction of capacity adjustment can differ between cities. Contrary to what is often claimed, introducing a more efficient pricing regime such as TOD fares does not necessarily reduce the need for capacity expansion.

To conclude this section we briefly investigate Mohring's square-root rule and the effects of capacity cost function parameter values on cost recovery from fare revenues. In the optimal uniform-fare regime, Eqs. (21a) and (21b) can be solved jointly to obtain explicit formulas for the unconditionally

[30] Proposition 7 is a counterpart to Lemma 3 in Kraus and Yoshida (2002). Consistent with Prop. 7, they find that the number of train runs is longer, and train capacity smaller, in the SO than the uniform-fare regime.

[31] Prop. 8 is a counterpart to Prop. 3 in Kraus and Yoshida (2002). Their conclusions are similar except that in their model total usage is unambiguously larger in the SO than with a uniform fare.

optimal values, s_*^u and m_*^u. The characteristics of the solution depend on the relative magnitudes of parameters ν_0 and ν_1. If $\nu_1 = 0$, the cost of a train is independent of its capacity and there are scale economies with respect to train service. In the other limiting case with $\nu_0 = 0$, there are no scale economies. In either case, Mohring's square-root rule does not apply because the headway is held fixed. Details are provided in on-line Appendix J.

The degree of cost recovery from fare revenue is readily derived for both the optimal uniform-fare and SO-fare regimes. From Eq. (6), fare revenue in the uniform-fare regime is $R^u = \lambda N^2/(m_u s_u)$. Given first-order condition (21a) this implies $R_*^u = (\nu_1 m_*^u + \nu_2)\, s_*^u$. From Eq. (14), fare revenue in the SO-fare regime is $R^o = \lambda N^2/(m_o s_o) + RV^o$. Given (21a), this yields $R_*^o = (\nu_1 m_*^o + \nu_2)\, s_*^o + RV^o - s_*^o RV_s^o$. From Eq. (15) for RV^o, the formula simplifies to $R_*^o = (\nu_1 m_*^o + \nu_2)\, s_*^o$. Hence, for both fare regimes the cost recovery ratio, ρ, is $\rho^i = \frac{R_*^i}{K(m_*^i, s_*^i)} = \frac{(\nu_1 m_*^i + \nu_2) s_*^i}{\nu_0 m_*^i + (\nu_1 m_*^i + \nu_2) s_*^i}$, $i = u, o$. With no scale economies with respect to train size (i.e., $\nu_0 = 0$), fare revenue fully covers capacity costs. Otherwise, costs are only partially recovered and the service runs a deficit.

7. Market distortions

The fare and capacity choices derived thus far are based on the assumption that first-best conditions apply. Any alternative travel modes to transit are priced at marginal social cost, and no distortions exist elsewhere in the economy. Fares can then be based on Pigouvian principles, and the benefits and costs of capacity can be assessed on Marshallian partial-equilibrium principles. In practice first-best conditions rarely, if ever, hold. In this section we briefly consider two instances in which the conditions fail: one in which public funds are scarce, and the other in which driving is an alternative to transit and unpriced traffic congestion exists.

7.1. Marginal cost of public funds

Transit authorities are often short of funds, and may be obliged to cover a portion of their costs from fare-box revenues. Governments may also be willing to sacrifice some allocative efficiency in return for stronger finances. These factors can be captured by assigning to fare revenues and capacity costs a weight of $1 + \phi$, $\phi \geq 0$, where $1 + \phi$ is the marginal cost of public funds (MCPF). The MCPF measures "the efficiency cost of raising one unit of tax revenue, given that the tax revenue is spent on a public good that does not affect the consumption of taxed commodities"[32]

[32] As an anonymous referee pointed out, in the case of fare revenues $1 + \phi$ could be viewed as a marginal benefit, or premium, from raising funds, rather than a cost. Here and in the numerical example of Section 8 we assume that

(Proost et al., 2007, p. 66).

We focus here on how the MCPF affects the optimal choice of fares. With a MCPF of $1+\phi$, weighted social surplus in the uniform-fare regime becomes

$$SS_a^u = \int_p^\infty N(u)\,du + (1+\phi)\,\tau^u N(p),$$

where $p = \bar{\delta} + \frac{\lambda N}{ms} + \tau^u$. If $\phi > 0$, a one-dollar loss of consumers' surplus from fare payment is outweighted by a one-dollar gain in revenue for the authority. Let η, $(\eta < 0)$, denote the local price elasticity of demand. It is straightforward to show that the optimal uniform fare works out to

$$\tau^u = \frac{\lambda N}{ms} - \frac{\phi}{\phi + \eta(1+\phi)}\left(\bar{\delta} + 2\frac{\lambda N}{ms}\right), \tag{22}$$

where $\phi + \eta(1+\phi) < 0$ to assure that $\tau^u < \infty$. The fare is set above the Pigouvian level of $\frac{\lambda N}{ms}$ by an amount that increases with ϕ.

Adding a premium to revenue generation does not affect the principle that users should be allocated efficiently between trains. The SO-fare thus varies between trains in the same way as before, but a constant, τ^o, is added to the fare for each train. Weighted social surplus thus becomes

$$SS_a^o = \int_p^\infty N(u)\,du + (1+\phi)\left(\tau^o N(p) + \frac{\lambda N}{ms} + RV^o\right),$$

where $p = \bar{\delta} + \frac{2\lambda N}{ms} + \tau^o$. The optimal flat component of the fare works out to

$$\tau^o = -\frac{\phi}{\phi + \eta(1+\phi)}\left(\bar{\delta} + 2\frac{\lambda N}{ms}\right). \tag{23}$$

Eq. (23) matches Eq. (22) for the uniform fare except that the component $\frac{\lambda N}{ms}$ is absent.

7.2. Modal choice and unpriced traffic congestion

So far it has been assumed that any alternatives to transit are priced at marginal social cost so that the welfare analysis can be restricted to transit. Yet automobile travel, the main alternative to transit in most cities, is generally underpriced. Here we examine how unpriced traffic congestion affects optimal transit policy. We first consider fares, and then briefly capacity. To keep the analysis simple, we treat transit and driving as perfect (demand) substitutes and limit attention to interior

the same value of ϕ applies to fare revenues and capacity costs. To simplify terminology, we stick to convention and refer to marginal costs. The MCPF varies widely in both theory and practice. It can be less than one for taxes that enhance economic efficiency. It is large for distortionary taxes, and it can be undefined for taxes with a narrow base for which an increase in the tax rate causes revenues to drop.

equilibria in which both modes are used. To economize on writing, we also use general notation for the functions.[33]

Let N_A denote the number of trips taken by automobile, N_R the number of trips taken by transit, and $C_i(N_i)$ the user cost function for mode i, $i = A, R$. Let N denote the total number of trips, and $p(N)$ the inverse demand curve corresponding to demand function (18). In the no-fare regime, N_A, N_R, and N can be solved using the identity $N = N_A + N_R$ and the equilibrium conditions

$$p(N) = C_A(N_A) = C_R(N_R).$$

In the uniform-fare regime the equilibrium conditions are

$$p(N) = C_A(N_A) = C_R(N_R) + \tau^u.$$

It is straightforward to show that the second-best optimal uniform fare is

$$\tau^u = C'_R(N_R) N_R - C'_A(N_A) N_A \frac{p_N(N)}{p_N(N) - C'_A(N_A)}, \tag{24}$$

where the prime ($'$) symbol denotes a derivative. Eq. (24) is none other than Eq. (2) in Verhoef et al. (1996) who study second-best pricing of a congestible facility when there is an unpriced alternative that is also congested. The second-best uniform fare is equal to the optimal uniform fare with no mode choice, $C'_R(N_R) N_R$, minus a fraction of the optimal uniform toll for the road, $C'_A(N_A) N_A$. The fraction is larger the less elastic is travel demand (i.e., the larger is $\|p_N(N)\|$) and hence the greater the potential for transit fare reductions to curtail traffic congestion rather than simply inducing more travel. If traffic congestion is severe enough, the second-best uniform fare can be negative.

Consider now the SO-fare regime. For any given number of transit users it is clearly optimal to price transit efficiently. The SO-fare therefore varies between trains in the same way as with inelastic travel demand, but the fare schedule is shifted down by the same amount as the uniform fare in Eq. (24). The propositions established in Section 5 therefore continue to hold with N_R in place of N.[34]

[33] The analysis in this subsection does not depend on linearity of the crowding cost function.

[34] The equivalent result holds in Kraus and Yoshida's (2002) model. The propositions do not generally hold if users are heterogeneous. To see this, suppose that individuals who prefer to drive at peak times also prefer to take transit at peak times. Since these individuals contribute the most to traffic congestion, it is desirable to reduce peak-period fares so that some will switch to transit. As Glaister (1974) showed, it is possible for second-best peak fares to be below off-peak fares, and even negative.

Finally, consider optimal capacity investment. In the optimal uniform-fare and SO-fare regimes, transit usage is efficiently priced and the investment rules given in Eqs. (19a) and (19b) remain optimal. However, usage is not optimal in the no-fare regime and the investment rules need to be modified. Depending on the relative numbers of drivers and transit users, the severity of traffic congestion, and the price elasticity of demand, the modified investment rules can call for more or less transit capacity than when traffic congestion is absent. For details see on-line Appendix K.

8. Optimal pricing and capacity for the Paris RER A line

The numerical example draws on recent empirical estimates of crowding costs. It is calibrated to describe service on the Paris RER A line during the morning peak although the example should be considered only illustrative for the line.[35] RER A crosses the Île-de-France region (which includes Paris) from West to East. It is one of Europe's busiest rail transit lines with over 1.1 million users per weekday, and is considered to be at capacity.[36] The line is 109 km long and is served by 46 stations. Base-case parameter values are: $\beta = 7.4$ [€/(hr·user)], $\gamma = 17.2$ [€/(hr·user)], $\lambda = 4.4$ [€/user], and $h = 2.5$ [min/train]. The demand function (18) is assumed to have a constant-elasticity form $N = N_0 p^{\eta}$ with $\eta = -1/3$.[37] Parameter N_0 and parameters ν_0, ν_1, and ν_2 of the capacity cost function are chosen to yield equilibrium values for the optimal uniform-fare equilibrium of $N^u = 32,600$, $m_*^u = 24$, $s_*^u = 1,733$, and a cost recovery rate of 5/6. The resulting values are: $N_0 = 69,003$ [users], $\nu_0 = 936.7$ [€/train], $\nu_1 = 0.1344$ [€/user], and $\nu_2 = 61.63$ [€· train/user]. In adopting this procedure we are not assuming that the existing fare scheme actually follows first-best pricing principles. Indeed, in 2013 the cost recovery rate for the Île-de-France region was only about 40 percent.[38] Rather, we are using the optimal uniform-fare regime for calibration because it is an intermediate regime, and also the most qualitatively descriptive of actual practice. The relative efficiency of the optimal uniform-fare regime can be measured by taking the no-fare and social optimum regimes as polar benchmarks and using the index $Eff_u = \frac{\widehat{SS}^u - \widehat{SS}^n}{\widehat{SS}^o - \widehat{SS}^n}$. Results for

[35] Parameter values are explained in Appendix C.

[36] Page 36 in http://www.stif.org/IMG/pdf/Deliberation_no2012-0163_relative_au_schema_directeur_du_RER_A.pdf.

[37] An elasticity of $-1/3$ is in the mid-range of empirical estimates (Oum et al., 2008, p.249). Consumers' surplus is infinite with $\eta > -1$. To enable comparisons of consumers' surplus between regimes, the area to the left of the demand curve is computed only for $p \leq$ €100.

[38] See www.stif.org/transports-aujourd-hui/tarication-francilienne/les-recettes-tarifaires-source-de.html.

Table 1: Comparison of no-fare, optimal uniform fare, and SO-fare (i.e., social optimum) regimes: base-case parameter values

	Fare regime		
	No-fare (n)	Optimal uniform fare (u)	Social optimum (o)
m	25.26	24	26.70
s	1,762	1,733	1,710
N	37,173	32,600	32,907
p	6.40	9.48	9.22
$Rev/user$	0	3.45	3.39
TCC	161,558	133,499	111,520
SDC	76,210	63,244	80,376
TC	237,768	196,743	191,896
K	138,270	134,889	136,528
R	0	112,407	111,520
ρ	0	0.833	0.817
CS	1,873,288	1,766,213	1,774,816
SS	1,735,018	1,743,732	1,749,807
$Total\ gain$		8,714	14,789
$Gain/user$		0.27	0.45
$Rel.eff$	0	0.59	1

the three fare regimes are reported in Table 1.[39]

8.1. No fare

With no fare, the equilibrium private cost (which equals the equilibrium user cost) is €6.40. There are $N^n = 37,173$ users who are accommodated in $m^n_* = 25.26$ trains with nominal capacities of $s^n_* = 1,762$. Total crowding costs (TCC^n) and total schedule delay costs (SDC^n) account for respectively 68 percent and 32 percent of total user costs (TC^n). Capital costs (K^n) are about 58 percent as large as total user costs (TC^n). Given no fare, the degree of cost recovery is zero.

8.2. Optimal uniform fare

The optimal uniform fare is $\tau^u =$€3.45. It boosts the equilibrium private cost to $p^u =$ €9.48 which is €3.08 above the no-fare equilibrium price. Ridership drops to $N^u = 32,600$: about 12 percent below the no-fare level. Both the number of trains and train capacity are lower than with no fare although capacity costs are reduced by only 2.4 percent. By design, fare revenue of $R^u =$€112,407 covers a fraction $\rho^u = 0.833$ of capacity costs. Consumers' surplus is lower than

[39]Throughout the numerical example, m is treated as a continuous variable. The results change very little if m is restricted to integer values (see on-line Appendix M).

with no fare, but social surplus is higher by €8,714 or about €0.27 per rider in the uniform-fare equilibrium. With the base-case parameter values, $Eff_u \simeq 0.59$ so that the optimal uniform fare yields nearly 3/5 of the efficiency gain from the SO-fare.

8.3. Social optimum

The social optimum calls for more trains than either the no-fare or the optimal uniform-fare regime. This is consistent with the rankings for the SO and optimal-uniform fare regimes in Prop. 8. However, train capacity is slightly lower than in the other two regimes.[40] Ridership and consumers' surplus are slightly higher than with a uniform fare. Price, revenue per user, and cost recovery are slightly lower. Crowding costs are significantly lower than in the other regimes (58 percent of total user costs), while, schedule delay costs are correspondingly higher because the SO-fare spreads usage more evenly over trains. Capacity costs are intermediate between the other regimes. Social surplus is higher than with no fare by about €0.45 per rider.[41]

8.4. Short-run versus long-run welfare gain from pricing

In Table 1, capacity is chosen optimally for each fare regime. Because rail transit capacity can take years to adjust, it is of interest to compare fare regimes in the "short run" when capacity is fixed. If pricing is assumed to become more efficient over time, there are three cases to consider: regime u with capacity fixed at (m_*^n, s_*^n), regime o with capacity fixed at (m_*^n, s_*^n), and regime o with capacity fixed at (m_*^u, s_*^u). Let G_x^{xy} denote the welfare gain in shifting from regime x to regime y when capacity remains fixed at its optimal level for regime x. With the base-case parameter values, $G_n^{nu} =$ €8,336, $G_u^{uo} =$ €5,273, and $G_n^{no} =$ €14,589. By comparison, from Table 1 the long-run welfare gains when capacity is adjusted optimally are $G^{nu} =$ €8,714, $G^{uo} =$ €6,076, and $G^{no} =$ €14,788. The long-run gains are higher by 4.5 percent, 15.2 percent, and 1.4 percent respectively. The difference between short-run and long-run gains is appreciable only for G^{uo}. This is mainly

[40] According to Prop. 8, with price-elastic demand optimal capacity rankings are theoretically ambiguous. One reason why, in the example, capacity is lowest in the social optimum is that the number of trains is largest. Capacity is highest for the no-fare regime because the greater benefits when there are more users turns out to outweigh the reduction in benefits due to latent demand. Nevertheless, the differences between regimes in optimal capacity are fairly small.

[41] As shown in on-line Appendix M, the welfare gains from the optimal uniform and SO fares are sensitive to the elasticity of demand. With $\eta = 0$, the uniform fare has no effect and the gain from the SO fare drops to €0.185 per rider. By contrast, if the elasticity is doubled to $\eta = -2/3$, the per-capita gains from the optimal uniform fare and SO fare rise to €0.51 and €0.70 respectively.

Table 2: Effects of increasing parameters β and γ (or parameter h) by 10 percent

	Fare regime		
	No-fare (n)	Opt. unif. fare (u)	Soc. opt. (o)
m	-4.98%	-4.98%	-4.44%
s	$+1.69\%$	$+1.70\%$	$+1.64\%$
N	-1.07%	-0.99%	-0.96%
Welf. gain		$+0.7\%$	$+6.3\%$

because regimes u and o differ the most in terms of optimal number of trains.[42]

In all three fare regimes the equilibrium price is an increasing function of parameters β, γ, λ, and h. Equilibrium usage thus decreases if these parameters increase in value. Varying parameters β, γ, λ, or h also induces changes in m and s, and to determine the size of the effects it is necessary to solve for the new equilibria.

As a first experiment, the unit schedule delay cost parameters β and γ were both increased by 10 percent. The results are shown in Table 2. In each fare regime the number of trains drops by nearly 5 percent because users incur higher schedule delay costs, which reduces demand. Train capacity increases by about 1.7 percent, but equilibrium prices still rise and usage drops slightly. Welfare gain G^{nu} increases by 0.7 percent, and welfare gain G^{uo} increases by 6.3 percent. An increase in headway, h, has exactly the same effect as an equal percentage increase in β and γ.

As a second experiment, the crowding cost parameter λ was increased by 10 percent. The results are shown in Table 3. In all fare regimes the number of trains rises by about 3 percent while train capacity increases by just over 2 percent. Usage drops by about 1 percent. Welfare gains G^{nu} and G^{uo} both increase slightly.[43]

Tables 2 and 3 depict long-run effects of changes in parameter values. These effects can differ significantly from the short-run effects when capacity is given. Consider, for example, welfare gain G^{uo}. In the short run with s and m fixed, G^{uo} is given by Eq. (17). With a 10 percent increase in β and γ, G^{uo} rises by a factor of $(1.1)^2$, or 21 percent. This is more than triple the long-run increase of 6.3 percent shown in Table 2. A 10 percent increase in λ causes G^{uo} to fall in the short

[42]Note that by Prop. 5, usage with the SO-fare and capacity fixed at (m^u_*, s^u_*) is the same as usage with the optimal uniform fare. Thus, in the short run, regimes u and o differ only in how passengers are distributed between trains.

[43]It is plausible that with economic growth, parameters β, γ, and λ will all increase — and perhaps at similar rates as considered in subsection 5.2. As a third experiment, the three parameters were all increased by 10 percent. The overall effects were close to the sum of the effects shown in Tables 2 and 3 except that m decreased by a slightly smaller percentage. To economize on space, the results are not reported in a third table.

Table 3: Effects of increasing parameter λ by 10 percent

	Fare regime		
	No-fare (n)	Opt. unif. fare (u)	Soc. opt. (o)
m	+3.02%	+3.02%	+2.92%
s	+2.14%	+2.14%	+2.15%
N	−1.06%	−1.08%	−1.08%
Welf. gain		+2.4%	+1.4%

run by a factor $(1.1)^{-1}$ or about 9 percent. Yet Table 3 shows that the long-run gain actually rises by 1.4 percent.

The large differences between the short-run and long-run effects highlight the importance of the planning time horizon. For example, recent empirical research has led to improved estimates of the costs of public transport crowding (OECD, 2014). A rise in the estimated unit cost of crowding (i.e., parameter λ) might dissuade a planner with a short-run perspective from implementing train-dependent fares. By contrast, a planner with a long-run perspective could be spurred to go ahead. This illustrates the well-known lesson that pricing and capacity investment decisions are interdependent, and should be considered jointly (Lindsey, 2012).

8.5. Marginal cost of public funds

To examine the effect of the marginal cost of public funds (MCPF) on fares and capacity, the calculations were repeated for values of parameter ϕ ranging from 0 up to 0.3.[44] The same value of ϕ was applied to fare revenues and capacity costs. Fares for the optimal uniform-fare and SO regimes were computed using Eqs. (22) and (23). Values for the numbers of trains and train capacities were computed using variants of Eqs. (19a) and (19b). As shown in Figure 3, optimal fares increase rapidly with the MCPF, and reach over €20 with $\phi = 0.3$. These high fares are explained by the relatively low price elasticity of demand of $\eta = -1/3$ which permits significant revenue to be generated with only a modest loss of allocative efficiency.[45] The relative efficiency of the optimal uniform-fare regime in generating weighted social surplus increases from 0.59 with $\phi = 0$ up to 0.97 with $\phi = 0.3$. The steep increase is attributable to the fact that, with homogeneous users, a

[44]The prescribed value for project evaluation in France is 0.3 (Comité directeur des transports, 2004). This falls within the range of theoretical and empirical values identified in the literature. See, for example, Snow and Warren (1996) and Gahvari (2006). Small and Verhoef (2007, §4.2.5 and pp. 177−8) provide a discussion and some empirical estimates of MCPF from a transportation perspective.

[45]If the elasticity is doubled to $\eta = -2/3$, fares only rise to about €8 with $\phi = 0.3$. Some estimates of the long-run elasticity of public transit demand are in this range. See Schimek (2015).

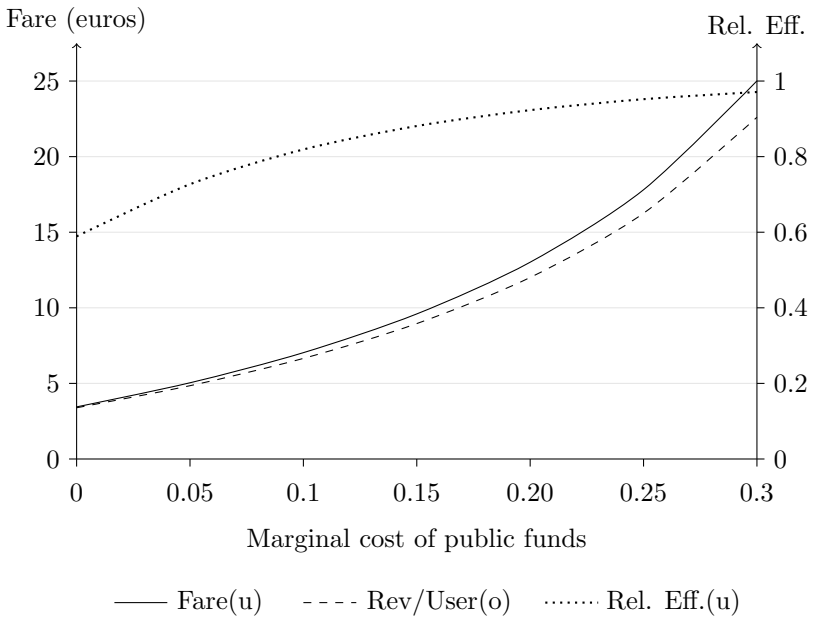

Figure 3: Fare revenues and relative efficiency as functions of MCPF

uniform fare is fully efficient at generating revenues, and as the weight on revenues increases the performance gap between the optimal uniform-fare and SO-fare narrows.

Numbers of trains and train capacities are presented in Figure 4 as fractions of their base-case values with $\phi = 0$. Train capacities decrease in all three fare regimes. The largest reduction occurs in the uniform-fare regime, and the smallest in the no-fare regime which is unaffected on the demand side by revenue generation incentives. Numbers of trains are affected proportionally less than train capacities, and the number actually increases slightly in the social optimum.[46]

9. Conclusion

We have analyzed the time pattern of usage and crowding on a rail transit line using a model (the *PTC* model) of trip-timing preferences. Users face a trade-off between riding a crowded train that arrives at a convenient time, and riding a less crowded train that arrives earlier or later than

[46] One reason for this is that capacity costs become more burdensome as the MCPF increases. Since train-capacity-specific costs (i.e., term $\nu_2 s$ in Eq. (20)) account for over three-quarters of total capital costs in the example, it is more effective to trim costs by reducing train capacity than decreasing the number of trains. However, generating fare revenues also becomes more important as MCPF rises. To partly offset the loss of service quality and ridership due to the reduction in s, it can be optimal to increase m.

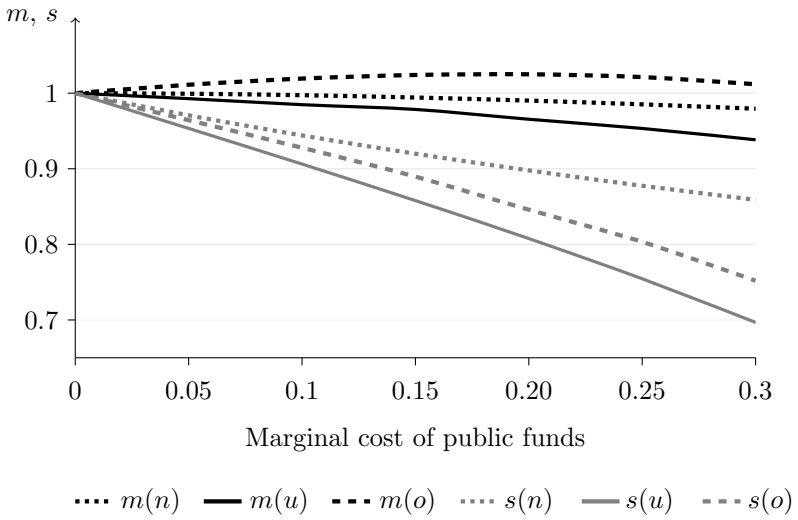

Figure 4: Optimal numbers of trains and train capacities relative to base case as functions of MCPF

desired. We solve user equilibrium for three fare regimes: no fare, an optimal uniform fare that controls the total number of users, and an optimal time-dependent fare that controls the distribution of users between trains as well. We also solve for the optimal long-run number and capacities of trains.

Timely trains are more crowded in all fare regimes. Under plausible assumptions, passenger loads are more evenly distributed in the social optimum than in the other regimes. Because crowding is assumed to occur at all levels of train occupancy, it is impossible to eliminate crowding costs even if fares can be varied freely. Consequently, imposing Pigouvian fares makes users worse off, at least before accounting for how fare revenue is used.

We show that if the crowding cost function is convex, the short-run welfare gain from introducing optimal time-dependent fares decreases with total ridership. The marginal social cost of accommodating an additional passenger is then higher in the social optimum than with a uniform fare even though passengers are distributed optimally across trains in the social optimum. This finding contrasts with both conventional wisdom and models of road traffic congestion.

Solving for optimal transit supply in the PTC model is complicated by the fact that (even treating the headway as uniform and fixed) capacity has two dimensions: the number of trains and the capacity of each train. We treat a special case with linear crowding and schedule delay cost functions, and a uniform headway between trains. The ranking of optimal capacity in the

no-fare and optimal uniform-fare regimes is ambiguous in general. More users take transit in the no-fare regime, but the benefit from expanding capacity is diluted by latent demand. Comparison of the uniform-fare and SO regimes is more clear-cut. With inelastic demand, the SO has a larger number of trains, a lower train capacity, and a larger fleet capacity. With elastic demand, the same result holds for number of trains but the other rankings are ambiguous. The prospect that capacity investments yield higher benefits in the social optimum again contrasts with conventional wisdom that capacity investments and efficient pricing are substitutes for relieving congestion.

For illustration, we calibrate the model to describe the Paris RER A line during morning-peak conditions. We find that total schedule delay costs account for roughly one third of total user costs, with total crowding costs comprising the rest. Ignoring schedule delay costs, as is implicitly done with static models, thus leads to underestimation of the total costs of congestion by about one third. With the base-case parameter values, the welfare gain from implementing efficient pricing is €0.27 per user for the optimal uniform fare, and €0.45 for the optimal time-dependent fare. While these amounts may seem modest, the system-wide gain could be large. The RER A line carries more than 300 million users per year, and on average more than 1.5 million individuals used public transport in the Île-de-France region during the morning peak (7am-9am) in 2010.[47] Given 250 working days per year, a welfare gain of about €0.50 per trip, and doubling the number of trips to account (roughly) for evening travel, the annual total welfare gain from optimal pricing amounts to nearly €400 million per year.[48]

The analysis in this paper could be extended in various directions. One is to allow travelers to differ in their trip-timing preferences and disutility from crowding. Doing so would permit consideration of the equity implications of alternative fare regimes and capacity expansion policies. Another extension is to consider rewards as a way to redistribute passengers across trains. An alternative to penalizing peak-period users with high fares is to reward off-peak users with low, or possibly even negative, fares.[49] Pricing usage below marginal social cost is inefficient when it induces excessive travel, but the induced deadweight loss may be an acceptable price to pay if discounting fares helps overcome opposition to time-of-day pricing. A third extension, which we

[47] See p.11 in www.lvmt.fr/IMG/pdf/RAPA_Chaire_Stif_2013-2014_v1.pdf.

[48] This figure is comparable to the social benefit from congestion pricing of roads in Paris. By modeling the Île-de-France region as a monocentric city with congestion and endogenous road capacity, De Lara et al. (2013) estimated an annual social saving of €606 million from a cordon toll.

[49] As noted in the introduction, some cities have implemented such schemes.

examined briefly, is to add driving on congested roads as an alternative to public transit. Reducing fares below first-best levels is one way to reduce traffic congestion. The efficacy of discounting fares naturally depends on cross-price elasticities of demand between driving and public transit which vary widely between and within cities.[50]

Another important consideration is service reliability. Service can be disrupted for many reasons: special events that create surges in ridership, bad weather, terrorist threats or actual attacks, mechanical failures, sick passengers, people or objects falling on the tracks, etc.. Delays at one station can cascade or snowball downstream as well as onto connecting lines. Service is deteriorating in some large cities. During the first quarter of 2016, 16.5% of RER A users arrived more than five minutes later than prescribed by the timetable.[51] From 2012 to 2016, delays more than doubled on the New York City subway system, and on weekdays only two-thirds of trains reach their final station within five minutes of the schedule.[52] Travelers can adapt to travel time uncertainty by departing earlier (or possibly later) or switching to a different transport mode (Monchambert and de Palma, 2014). System operators can improve reliability by padding schedules, increasing the number of reserve vehicles, and increasing boarding rates. Technological advances can reduce the frequency of mechanical failures and intrusions onto track (Transport for London, 2014). For example, modular designs facilitate the removal and replacement of components more quickly, and fully-automatic train operation helps to eliminate human error. Modeling these elements will be challenging, and likely require numerical as well as analytical methods.

Acknowledgments

We are grateful to the editor and two anonymous referees for very helpful suggestions on the content and organization of the paper. For useful comments we would also like to thank Martin Koning, Clifford Winston and participants at: the Annual Conference of the Spatial Economic Research Center, London School of Economics, May 2014; the Annual Conference of the International Transportation Economics Association (ITEA) in Toulouse, June 2014; the Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, July 2014; the XVI Conference of the Italian Association of Transport Economics and Logistics

[50] A recent study by Anderson (2014) suggests that the elasticities can be large. Using data from a 2003 transit strike in Los Angeles he finds that cessation of transit service had a large effect on traffic speeds on heavily congested roads.

[51] Source: http://www.stif.org/IMG/pdf/bqst_1er-trimestre-2016_ok.pdf, in French.

[52] Source: https://nyti.ms/2l4ALiI

(SIET) in Florence, October 2014; the Department of Spatial Economics, Free University of Amsterdam, February 2015; the 50th Annual Conference of the Canadian Transportation Research Forum (CTRF), Montreal, May 2015; the Department of Economics, University of Laval, October 2015; and Département Economie et Gestion, Ecole Normale Supérieure de Cachan, April 2016. Financial support from the Social Sciences and Humanities Research Council of Canada (Grant 435-2014-2050), from the French National Agency of Research ANR (Project Elitisme), from ENS Cachan, France, and from CES, KU Leuven, Belgium is gratefully acknowledged.

References

Anderson, M. L., 2014. Subways, strikes, and slowdowns: The impacts of public transit on traffic congestion. American Economic Review 104 (9), 2763–96.

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

Comité directeur des transports, 2004. Instruction cadre relative aux méthodes d' évaluation économique des grandes infrastructures de transport. Tech. rep., Ministère de l' équipement, Paris, retrieved from http://cpdp.debatpublic.fr/cpdp-lnpn/site/DOCS/AUTRES_DOCUMENTS/INSTRUCTION_CADRE_ROBIEN__MEEDD.PDF.

Daniels, R., Mulley, C., 2013. The paradox of public transport peak spreading: Universities and travel demand management. International Journal of Sustainable Transportation 7 (2), 143–165.

De Lara, M., de Palma, A., Kilani, M., Piperno, S., 2013. Congestion pricing and long term urban form: Application to Paris region. Regional Science and Urban Economics 43 (2), 282–295.

de Palma, A., Kilani, M., Proost, S., 2015. Discomfort in mass transit and its implication for scheduling and pricing. Transportation Research Part B: Methodological 71 (0), 1–18.

Douglas, N. J., Henn, L., Sloan, K., 2011. Modelling the ability of fare to spread am peak passenger loads using rooftops. In: Australasian Transport Research Forum (ATRF), 34th, 2011, Adelaide, South Australia, Australia. Retrieved from http://atrf.info/papers/2011/2011_Douglas_Henn_Sloan.pdf.

Duranton, G., Turner, M. A., 2011. The fundamental law of road congestion: Evidence from us cities. American Economic Review 101 (6), 2616–52.

Gahvari, F., 2006. On the marginal cost of public funds and the optimal provision of public goods. Journal of Public Economics 90 (6), 1251–1262.

Glaister, S., 1974. Generalised consumer surplus and public transport pricing. The Economic Journal 84 (336), 849–867.

Gonzales, E. J., Daganzo, C. F., 2012. Morning commute with competing modes and distributed demand: user equilibrium, system optimum, and pricing. Transportation Research Part B: Methodological 46 (10), 1519–1534.

Halvorsen, A., Koutsopoulos, H. N., Lau, S., Au, T., Zhao, J., 2016. Reducing subway crowding: Analysis of an off-peak discount experiment in Hong Kong. Transportation Research Record: Journal of the Transportation Research Board (2544), 38–46.

Haywood, L., Koning, M., 2015. The distribution of crowding costs in public transport: New evidence from paris. Transportation Research Part A: Policy and Practice 77 (0), 182–201.

Huang, H.-J., Tian, Q., Gao, Z.-Y., 2005. An equilibrium model in urban transit riding and fare polices. Algorithmic Applications in Management 3521, 112–121.

Knight, F. H., 1924. Some fallacies in the interpretation of social cost. The Quarterly Journal of Economics 38 (4), 582–606.

Kraus, M., 2003. A new look at the two-mode problem. Journal of Urban Economics 54 (3), 511–530.

Kraus, M., Yoshida, Y., 2002. The commuter's time-of-use decision and optimal pricing and service in urban mass transit. Journal of Urban Economics 51 (1), 170–195.

Lindsey, R., 2012. Road pricing and investment. Economics of Transportation 1 (1), 49–63.

Lovrić, M., Raveau, S., Adnan, M., Pereira, F. C., Basak, K., Loganathan, H., Ben-Akiva, M., 2016. Evaluating off-peak pricing strategies in public transportation with an activity-based approach. Transportation Research Record: Journal of the Transportation Research Board 2544, 10–19.

Mohd Mahudin, N. D., Cox, T., Griffiths, A., 2012. Measuring rail passenger crowding: Scale development and psychometric properties. Transportation Research Part F: Traffic Psychology and Behaviour 15 (1), 38–51.

Mohring, H., 1972. Optimization and scale economies in urban bus transportation. American Economic Review 62 (4), 591–604.

Monchambert, G., de Palma, A., 2014. Public transport reliability and commuter strategy. Journal of Urban Economics 81, 14–29.

OECD, 2014. Valuing convenience in public transport. Tech. rep., ITF Round Tables, retrieved from http://www.internationaltransportforum.org/jtrc/DiscussionPapers/DP201402.pdf.

Parry, I. W., Small, K. A., 2009. Should urban transit subsidies be reduced? American Economic Review 99 (3), 700–724.

Pigou, A. C., 1920. The Economics of Welfare. London: Macmillan.

Proost, S., De Borger, B., Koskenoja, P., 2007. Public finance aspects of transport charging and investments. In: de Palma, A., Lindsey, R., Proost, S. (Eds.), Investment and the use of tax and toll revenues in the transport sector. Vol. 19 of Research in transportation economics. Amsterdam, The Netherlands: Elsevier, pp. 59–80.

Prud'homme, R., Koning, M., Lenormand, L., Fehr, A., 2012. Public transport congestion costs: The case of the Paris subway. Transport Policy 21, 101–109.

Rothschild, M., Stiglitz, J. E., 1970. Increasing risk: I. a definition. Journal of Economic Theory 2 (3), 225–243.

Schimek, P., 2015. Dynamic estimates of fare elasticity for us public transit. Transportation Research Record (2538), 96–101.

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

Small, K. A., Verhoef, E. T., 2007. The economics of urban transportation. Routledge.

Snow, A., Warren, R. S., 1996. The marginal welfare cost of public funds: theory and estimates. Journal of Public Economics 61 (2), 289–305.

Tian, Q., Huang, H.-J., Yang, H., 2007. Equilibrium properties of the morning peak-period commuting in a many-to-one mass transit system. Transportation Research Part B: Methodological 41 (6), 616–631.

Tian, Q., Yang, H., Huang, H., Mou, H., 2009. Dynamic congestion pricing in urban transit system. In: International Conference on Transportation Engineering 2009. ASCE, pp. 1542–1547.

Tirachini, A., Hensher, D. A., Rose, J. M., 2013. Crowding in public transport systems: Effects on users, operation and implications for the estimation of demand. Transportation Research Part A: Policy and Practice 53, 36–52.

Tirachini, A., Sun, L., Erath, A., Chakirov, A., 2016. Valuation of sitting and standing in metro trains using revealed preferences. Transport Policy 47, 94–104.

Transport for London, 2014. New tube for london feasibility report. Tech. rep., retrieved from http://content.tfl.gov.uk/ntfl-feasibility-report.pdf.

Veitch, T., Partridge, J., Walker, L., 2013. Estimating the costs of over-crowding on melbourne's rail system. In: Australasian Transport Research Forum (ATRF), 36th, 2013, Brisbane, Queensland, Australia. Retrieved from http://atrf.info/papers/2013/2013_veitch_partridge_walker.pdf.

Verhoef, E., Nijkamp, P., Rietveld, P., 1996. Second-best congestion pricing: the case of an untolled alternative. Journal of Urban Economics 40 (3), 279–302.

Vickrey, W. S., 1955. A proposal for revising new york's subway fare structure. Journal of the Operations Research Society of America 3 (1), 38–68.

Vickrey, W. S., 1963. Pricing in urban and suburban transport. American Economic Review 53 (2), 452–465.

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

Wardman, M., Chintakayala, P., de Jong, G., Ferrer, D., 2012. European wide meta-analysis of values of travel time. Tech. rep., Significance, retrieved from http://www.significance.nl/papers/2012-European%20wide%20meta-analysis%20of%20values%20of%20travel%20time.pdf.

Wardman, M., Whelan, G., 2011. Twenty years of rail crowding valuation studies: Evidence and lessons from british experience. Transport Reviews 31 (3), 379–398.

Whelan, G., Johnson, D., 2004. Modelling the impact of alternative fare structures on train overcrowding. International Journal of Transport Management 2 (1), 51–58.

Whelan, G. A., Crockett, J., 2009. An investigation of the willingness to pay to reduce rail overcrowding. In: International Choice Modelling Conference 2009.

Yoshida, Y., 2008. Commuter arrivals and optimal service in mass transit: Does queuing behavior at transit stops matter? Regional Science and Urban Economics 38 (3), 228–251.

Appendix A. Proof of Proposition 3

Let j index trains in order of decreasing schedule delay cost so that $\delta_1 > \delta_2 > ... > \delta_m$. (Because trains arrive early and late, the index does not correspond to the temporal sequence in which trains are run.) Since in the UE $n_j^e = g^{-1}\left[c^e - \delta_j\right]$ and $g'(.) > 0$, n_j^e increases with j: $n_1^e < n_2^e < ... < n_m^e$.

We show that $n_j^e \lessgtr n_j^o \iff n_j^e g'\left(n_j^e\right) \lessgtr MSC^o - c^e$. Given $n_j^o = v^{-1}[MSC^o - \delta(t_j)]$, it follows that

$$\begin{aligned}
n_j^e &\lessgtr n_j^o \\
\iff g^{-1}\left[c^e - \delta_j\right] &\lessgtr v^{-1}\left[MSC^o - \delta_j\right] \\
\iff v\left\{g^{-1}\left[c^e - \delta_j\right]\right\} &\lessgtr MSC^o - \delta_j \\
\iff c^e - \delta_j + g^{-1}\left[c^e - \delta_j\right] \times g'\left\{g^{-1}\left[c^e - \delta_j\right]\right\} &\lessgtr MSC^o - \delta_j \\
\iff g^{-1}\left[c^e - \delta_j\right] \times g'\left\{g^{-1}\left[c^e - \delta_j\right]\right\} &\lessgtr MSC^o - c^e \\
\iff n_j^e g'\left(n_j^e\right) &\lessgtr MSC^o - c^e.
\end{aligned}$$

Variables n_j^e and n_j^o have the same ranking as $n_j^e g'\left(n_j^e\right)$, the marginal external cost of crowding in the UE, and $MSC^o - c^e$, which is constant. Because total patronage, N, is fixed, some trains are more more heavily loaded in the UE, and the others are more heavily loaded in the SO. Consequently, if $ng'(n)$ is a strictly increasing function of n (i.e., $\varepsilon(n) > -1$), there exists a unique train $\hat{\jmath}$ such that $n_j^e < n_j^o$ when $j < \hat{\jmath}$, $n_{\hat{\jmath}}^e \geq n_{\hat{\jmath}}^o$, and $n_j^e > n_j^o$ when $j > \hat{\jmath}$. Conversely, if $ng'(n)$ is a strictly decreasing function of n (i.e., $\varepsilon(n) < -1$), there exists a unique train $\hat{\jmath}$ such that $n_j^e > n_j^o$ when $j < \hat{\jmath}$, $n_{\hat{\jmath}}^e \leq n_{\hat{\jmath}}^o$, and $n_j^e < n_j^o$ when $j > \hat{\jmath}$.

Appendix B. Proof of Proposition 4

We prove the case for which the welfare gain G^{eo} decreases with N. The proof for the case in which G^{eo} increases follows the same steps, and is omitted. As mentioned in the text, G^{eo} decreases if the marginal social cost of an additional user is higher in the SO than the UE. Thus, it suffices to show that $MSC^o > MSC^e$.

As Appendix A, let k index trains in order of decreasing schedule delay cost so that in the no-fare equilibrium, $n_1^e < n_2^e < ... < n_m^e$. Equilibrium cost with no fare, c^e, is determined implicitly by Eq. (1):

$$\sum_{k=1}^{m} g^{-1}\left[c^e - \delta_k\right] - N = 0. \tag{B.1}$$

This equation can be written

$$\sum_{k=1}^{m} f\left[g\left(n_k^e\right) + g'\left(n_k^e\right) n_k^e\right] = N, \tag{B.2}$$

where $f(n) \equiv v^{-1}(n)$. Since $f(v(n)) = n$,

$$f'(n) = \frac{1}{v'(n)} = \frac{1}{2g'(n) + g''(n)\, n}. \tag{B.3}$$

The marginal social cost of a trip in the no-fare equilibrium is $MSC^e = \frac{\partial(c^e N)}{\partial N} = c^e + \frac{\partial c^e}{\partial N} N$. Using Eq. (B.1) to derive $\frac{\partial c^e}{\partial N}$ one obtains

$$MSC^e = c^e + \frac{N}{\sum_{k=1}^{m} \frac{1}{g'\left(n_k^e\right)}}. \tag{B.4}$$

From Eq. (10) the marginal social cost of a trip in the social optimum is defined implicitly by:

$$\sum_{k=1}^{m} f\left[MSC^o - \delta_k\right] = N. \tag{B.5}$$

By Assumption 1, the LHS of Eq. (B.5) is a strictly increasing function of MSC^o. Suppose we substitute eqn. (B.4) for MSC^e in place of MSC^o in Eq. (B.5). If the resulting LHS is less than N, then $MSC^o > MSC^e$ and the proof is complete. To economize on notation, let g_k denote $g\left(n_k^e\right)$, g_k' denote $g'\left(n_k^e\right)$, and n_k denote n_k^e. After a few substitutions one can write

$$\sum_{k=1}^{m} f\left[MSC^e - \delta_k\right] = \sum_{k=1}^{m} f\left[g_k + \frac{N}{m} \frac{m}{\sum_{k=1}^{m} \frac{1}{g_k'}}\right].$$

Define

$$mec_k \equiv g_k + g_k' n_k, \tag{B.6}$$

and

$$\widetilde{mec}_k \equiv g_k + \frac{m}{\sum_{k=1}^{m} \frac{1}{g_k'}} \frac{N}{m}. \tag{B.7}$$

Given Eq. (B.2), we need to prove that the following expression is negative:

$$\Delta F \equiv \sum_{k=1}^{m} \underbrace{f\left[g_k + \frac{m}{\sum_{k=1}^{m} \frac{1}{g_k'}} \frac{N}{m}\right]}_{\tilde{n}_k} - \sum_{k=1}^{m} \underbrace{f\left[mec_k\right]}_{n_k}.$$

Given Assumption 1, $\tilde{n}_k > n_k$ for small k, and $\tilde{n}_k < n_k$ for large k. The rankings of $\tilde{n}_k$ and n_k, and of $\tilde{c}_k$ and c_k, are shown in Figure B.5.

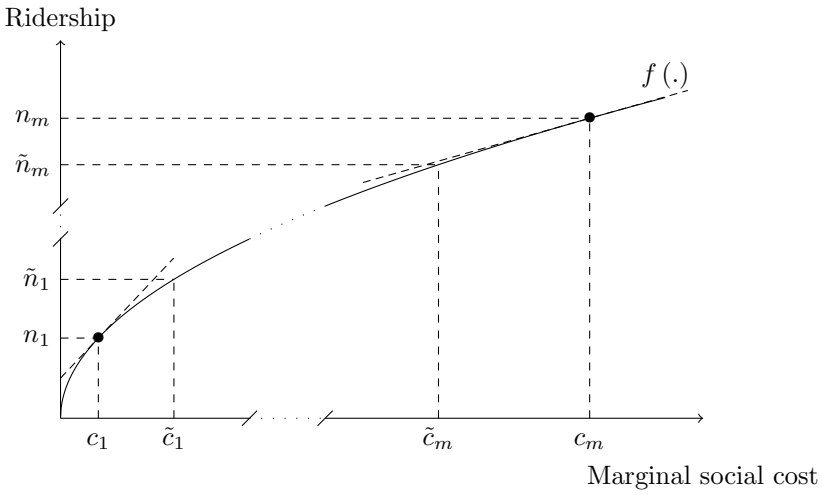

Figure B.5: Ridership and marginal social cost

Function $f\,()$ is concave if we assume that the marginal social cost of crowding is a strictly convex function of load (i.e., $v''(n) > 0$). Clearly, for all trains $\tilde{n}_k - n_k < (\tilde{c}_k - c_k) f'[c_k]\,, \quad k = 1...m$. Using Eqs. (B.6), (B.7) and (B.3) this implies

$$\begin{aligned}\Delta F &= \sum_{k=1}^{m} \tilde{n}_k - \sum_{k=1}^{m} n_k < \sum_{k=1}^{m} (\tilde{c}_k - c_k) f'[c_k] \\ &= \sum_{k=1}^{m} \left(\frac{m}{\sum_{k=1}^{m} \frac{1}{g'_k}} \frac{N}{m} - g'_k n_k \right) \frac{1}{2g'_k + g''_k n_k}.\end{aligned}$$

Now, $\sum_{k=1}^{m} \frac{1}{g'_k} = \frac{\sum_{j=1}^{m} \prod_{i \neq j} g'_i}{\prod_{i=1}^{k} g'_i}$. Hence

$$\begin{aligned}\Delta F &= \sum_{k=1}^{m} \left(N \frac{\prod_{i=1}^{m} g'_i}{\sum_{j=1}^{m} \prod_{i \neq j} g'_i} - g'_k n_k \right) \frac{1}{2g'_k + n_k g''_k} \\ &= \sum_{k=1}^{m} \left(N \frac{\prod_{i \neq k} g'_i}{\sum_{j=1}^{m} \prod_{i \neq j} g'_i} - n_k \right) \frac{g'_k}{2g'_k + n_k g''_k} \\ &= \sum_{k=1}^{m} \left(\frac{\left(\sum_{l \neq k} n_l \right) \prod_{i \neq k} g'_i}{\sum_{j=1}^{m} \prod_{i \neq j} g'_i} + \left(\frac{\prod_{i \neq k} g'_i}{\sum_{j=1}^{m} \prod_{i \neq j} g'_i} - 1 \right) n_k \right) \frac{g'_k}{2g'_k + n_k g''_k} \\ &= \sum_{k=1}^{m} \left(\underbrace{\left(\sum_{l \neq k} n_l \right)}_{(1)} \underbrace{\frac{\prod_{i \neq k} g'_i}{\sum_{j=1}^{m} \prod_{i \neq j} g'_i}}_{(2)} - \underbrace{\frac{\sum_{j \neq k} \prod_{i \neq j} g'_i}{\sum_{j=1}^{m} \prod_{i \neq j} g'_i}}_{(3)} \underbrace{n_k}_{(4)} \right) \frac{g'_k}{2g'_k + n_k g''_k} \qquad \text{(B.8)}\end{aligned}$$

In the second line of eqn. (B.8),

$$\begin{aligned} & \sum_{k=1}^{m}\left(N\frac{\prod_{i\neq k}g_i'}{\sum_{j=1}^{m}\prod_{i\neq j}g_k'}-n_k\right) \\ = \; & N\sum_{k=1}^{m}\left(\frac{\prod_{i\neq k}g_i'}{\sum_{j=1}^{m}\prod_{i\neq j}g_i'}\right)-\sum_{k=1}^{m}n_k \\ = \; & N-N=0. \end{aligned}$$

Terms (1) and (2) in the last line of Eq. (B.8) are decreasing functions of k. Terms (3) and (4) are increasing functions of k. Hence Eq. (B.8) is negative if $\frac{g_k'}{2g_k'+n_k g_k''}$ is a non-decreasing function of k, or equivalently if $\varepsilon(n)=\frac{g_k''n_k}{g_k'}$ is a non-increasing function of k which is guaranteed if we assume that $\varepsilon(n)$ is a nonincreasing function of load (i.e., $\frac{d\varepsilon(n)}{dn}\leq 0$).

Appendix C. Parameter values for numerical example

The numerical example requires base-case parameter values for β, γ, λ and h, and target values for N, m and s. The operating period was set to one hour, and target values were chosen for the optimal uniform-fare regime. This regime is intermediate in efficiency between the no-fare and SO-fare regimes, and it is arguably the most descriptive of public transit service in Paris where fares are positive and constant throughout the day.

Consider first the supply-side parameters m, h and s. According to the document "Schéma Directeur du RER A" written in June 2012 by the STIF (Syndicat des Transport d'Île-de-France), 30 trains per hour are supposed to operate during the morning peak in the East-West direction on the RER A line. However, the frequency actually achieved over the 4-year period February 2008 to February 2012 was only 24.4 trains per hour.[53] The target value for number of trains was thus set to $m=24$, and the headway was set to $h=\frac{60}{24}=2.5$ mins.

Two types of bi-level train sets are operated during the morning peak:[54]

- MI2N train sets with 904 seats and standing room for 1,636 users (4 users/m^2) for a total capacity of 2,540 riders
- MI09 train sets with 948 seats and standing room for 1,683 users (4 users/m^2) for a total capacity of 2,614 riders

[53] See p.36 in www.stif.org/IMG/pdf/Deliberation_no2012-0163_relative_au_schema_directeur_du_RER_A.pdf.
[54] See p.54 in www.stif.org/IMG/pdf/Deliberation_no2012-0163_relative_au_schema_directeur_du_RER_A.pdf.

This suggests a value for capacity of about $s = 2,600$. However, in the model users are assumed to travel from a single origin to a single destination whereas the RER A line serves many stations. La Défense is the most popular destination, but a substantial fraction of users pass through it. Only part of train capacity is thus effectively devoted to users who exit at La Défense. After experimentation with alternative values of s, and other parameters described below, we settled on a capacity equal to two-thirds of nominal train capacity so that $s = \frac{2}{3} \cdot 2,600 = 1,733$.

Consider now the demand-side parameters. According to a January 2011 document "Étude La Défense Analyse des Trafics" prepared by the DRIEA (Direction Régionale et Interdépartementale de l'Équipement et de l'Aménagement), in 2009, 32,600 users arrived at La Défense by RER A between 8:25am and 9:25am.[55] This count includes users traveling in both East-West and West-East directions, but it excludes users who are passing through. Including travel in both directions results in overestimation of traffic in one direction, whereas excluding users who pass through La Défense results in underestimation this traffic. Lacking an indication as to which bias dominates, we set $N = 32,600$.

Wardman et al. (2012) conduct a meta-analysis of estimates of β, γ, and the value of travel time; call it α. They report point estimates of $\beta = 0.74 \cdot \alpha$ and $\gamma = 1.72 \cdot \alpha$ (see Table 19, p.25). For commuters in France, $\alpha =$€15/hr (see Table 15, p.21) which is consistent with the government-recommended value. This suggests setting $\beta = 0.74 \cdot 15 =$ €11.1/hr, and $\gamma = 1.72 \cdot 15 =$ €25.8/hr. However, in the model it is assumed that users have the same desired arrival time, t^*. In reality, trip-timing preferences vary. The assumption of a common t^* leads to overestimation of schedule delay costs. In addition, with $\beta =$ €11.1/hr and $\gamma =$ €25.8/hr., condition (E.1) that all trains are used was violated given plausible values for other parameters. After experimentation with alternative values of β, γ, and s (noted above) we scaled down β and γ by one-third to $\beta =$ €7.4/hr and $\gamma =$ €17.2/hr.

Empirical studies of public transit crowding often report crowding costs as time multipliers. This is consistent with evidence that disutility from crowding is proportional to amount of time spent in crowded conditions. The crowding cost parameter can then be written

$$\lambda = \alpha \cdot tt \cdot (tm - 1)\,, \tag{C.1}$$

[55] See Figure 2 on page 8 in http://cpdp.debatpublic.fr/cpdp-grandparis/site/DEBATPUBLIC_GRANDPARIS_ORG/_SCRIPT/NTSP_DOCUMENT_FILE_DOWNLOADCB59.PDF.

where tt is travel time and tm is the time multiplier.

According to the survey "Étude mobilité transports à la Défense - Profils, usages et modes de déplacements des salariés et habitants du quartier d'affaires" by the EPAD (Établissement Public de la Région Pour l'Aménagement de la Défense), in 2006, the average travel time incurred by public transport riders who used only one transport mode to reach La Défense was 40 mins.[56] This is consistent with a study by the Enquête Global Transport in 2010, which found an average travel time for commuters of 41 mins.[57] We thus set $tt = 40$ mins or $2/3$ hrs.

Haywood and Koning (2015) have estimated time multipliers for Paris. They obtain a linear approximation of the time multiplier (see Eq. (10), p.194) of $tm = 1 + 0.11 \cdot d$, where d is the density of passengers per square metre. Substituting the estimates of α, tt, and tm into Eq. (C.1) one obtains $\lambda = 15 \cdot 2/3 \cdot 0.11 \cdot d$. With a density of 4 users/m^2 for standing room on the train sets used on the RER A line (see above), this yields $\lambda = 4.4$.

Appendix D. Glossary

Appendix D.1. Latin characters

c : user cost of a trip [€/user]

CS : total consumers' surplus [€]

e : superscript for uniform-fare regime

$g(n)$: expected crowding cost function [€/user]

G^{xy} : welfare gain in shifting from pricing regime x to y

h : time interval between successive trains [hr/train]

k : index of train

K : capacity cost function [€]

m : number of trains used [trains]

MEC : marginal external cost of a trip [€/user]

MSC : marginal social cost of a trip [€/user]

n : number of users on a train [users/train]

n_k : number of users taking train k [users/train]

[56] See p.11 in http://www.ladefense-seine-arche.fr/fileadmin/site_internet/user_upload/8-ENLIEN/etudes/etude-mobilite-transports.pdf.

[57] See p.3 in http://www.driea.ile-de-france.developpement-durable.gouv.fr/IMG/pdf/Fiche_Actifs__cle0cecb9.pdf.

N : total number of users [users]

o : subscript for socially-optimal fare regime

p : private trip cost including fare [€/user]

R : total fare revenue [€]

RV : variable fare revenue from socially optimal fare schedule [€]

s : measure of train capacity [users/train]

SDC : total schedule delay costs [€]

SS : social surplus [€]

t : departure time from origin station [clock time]

t^* : desired arrival time at destination [clock time]

TC : total user costs [€]

TCC : total crowding costs [€]

u : superscript for optimal uniform-fare regime

$v(n)$: marginal social crowding cost function [€/user]

Appendix D.2. Greek characters

β : cost per minute of arriving early [€/(hr·user)]

γ : cost per minute of arriving late [€/(hr·user)]

δ : schedule delay cost function [€/user]

ε : elasticity of $g'(n)$

η : elasticity of demand

λ : crowding cost parameter [€/user]

ν_0 : capacity cost function coefficient on m [€/train]

ν_1 : capacity cost function coefficient on $m \cdot s$ [€/user]

ν_2 : capacity cost function coefficient on s [€·train/user]

τ : fare [€/user]

ϕ : $1 + \phi$: marginal cost of public funds

On-line appendices

Appendix E. Conditions for positive usage of all trains

Appendix E.1. Uniform fare

Because the first (or last) train carries the fewest passengers, the equilibrium usage pattern solution satisfies all the non-negativity constraints $n_k^e > 0$ if $n_1^e > 0$ and $n_m^e > 0$. Given Eq. (2) the requisite condition is:

$$N > \frac{ms}{\lambda}\left(\max\left[\delta_1, \delta_m\right] - \overline{\delta}\right). \tag{E.1}$$

Since service is costly to provide, condition (E.1) is satisfied when m and s are chosen optimally as in Section 6.

Appendix E.2. Social optimum

Given Eq. (12) for n_k^o, the non-negativity constraint on usage of all trains is satisfied if

$$N > \frac{ms}{2\lambda}\left(\max\left[\delta_1, \delta_m\right] - \overline{\delta}\right).$$

This condition is satisfied if condition (E.1) is satisfied for the no-fare equilibrium.

Appendix E.3. Optimal capacity

It is not obvious from Eqs. (19a) and (19b) whether condition (E.1) is satisfied at the long-run optimum. However, if m is restricted to integer values (as is the case in reality), it is possible to show that condition (E.1) is indeed satisfied.

Appendix F. Proof of $\partial R^i/\partial N = \partial MSC^i/\partial N \cdot N$, $i = u, o$

Total fare revenue from the optimal uniform fare is $R^u = \tau^u N$. Hence $\frac{\partial R^u}{\partial N} = \tau^u + \frac{\partial \tau^u}{\partial N} N$.
Now $MSC^u = \frac{\partial TC^u}{\partial N} = \frac{\partial (c^u N)}{\partial N} = c^u + \frac{\partial c^u}{\partial N} N = c^u + \tau^u$.
Thus $\frac{\partial MSC^u}{\partial N} N = \left(\frac{\partial c^u}{\partial N} + \frac{\partial \tau^u}{\partial N}\right) N = \tau^u + \frac{\partial \tau^u}{\partial N} N = \frac{\partial R^u}{\partial N}$.
Total fare revenue from the SO-fare is $R^o = \sum_{k=1}^{m} \tau_k^o n_k^o$. Hence

$$\frac{\partial R^o}{\partial N} = \sum_{k=1}^{m} \left(\tau_k^o + \frac{\partial \tau_k^o}{\partial n_k^o} n_k^o\right) \frac{\partial n_k^o}{\partial N}.$$

The marginal social cost of a trip is the same for all trains that are used: $MSC^o = c_k^o + \tau_k^o$. Hence:

$$\begin{aligned}
\frac{\partial MSC^o}{\partial N} &= \left(\frac{\partial c_k^o}{\partial n_k^o} + \frac{\partial \tau_k^o}{\partial n_k^o}\right)\frac{\partial n_k^o}{\partial N}, \\
\frac{\partial MSC^o}{\partial N} n_k^o &= \left(\frac{\partial c_k^o}{\partial n_k^o} n_k^o + \frac{\partial \tau_k^o}{\partial n_k^o} n_k^o\right)\frac{\partial n_k^o}{\partial N} = \left(\tau_k^o + \frac{\partial \tau_k^o}{\partial n_k^o} n_k^o\right)\frac{\partial n_k^o}{\partial N}, \\
\frac{\partial MSC^o}{\partial N} N &= \sum_{k=1}^{m} \frac{\partial MSC^o}{\partial N} n_k^o = \sum_{k=1}^{m} \left(\tau_k^o + \frac{\partial \tau_k^o}{\partial n_k^o} n_k^o\right)\frac{\partial n_k^o}{\partial N} = \frac{\partial R^o}{\partial N}.
\end{aligned}$$

Appendix G. Proof of Proposition 5

We first consider uniform fares (which include no fare and the optimal uniform fare as special cases), and then the SO-fare.

Appendix G.1. Uniform-fare regimes

With a uniform fare, the equilibrium private cost of a trip, p^e, equals the user cost plus the fare:

$$p^e = \bar{\delta} + \frac{\lambda N}{ms} + \tau. \tag{G.1}$$

Eq. (G.1) serves as a supply function for trips. Solving (G.1) and the demand function (18) yields the equilibrium private cost and number of trips, $\widehat{p}^e$ and $\widehat{N}^e$. If the fare is zero, the equilibrium price is

$$\widehat{p}^n = \bar{\delta} + \frac{\lambda \widehat{N}^n}{ms}. \tag{G.2}$$

Social surplus equals consumers' surplus: $\widehat{SS}^n = \widehat{CS}^n = \int_{\widehat{p}^n}^{\infty} N(u)\, du$. The optimal uniform fare is given by Eq. (4): $\widehat{\tau}^u = \frac{\lambda \widehat{N}^u}{ms}$, and fare revenue is $\widehat{R}^u = \tau^u \widehat{N}^u = \frac{\lambda \left(\widehat{N}^u\right)^2}{ms}$. The efficient price of a trip equals marginal social cost:

$$\widehat{p}^u = \widehat{MSC}^n = \widehat{c}^u + \widehat{\tau}^u = \bar{\delta} + \frac{2\lambda \widehat{N}^u}{ms}. \tag{G.3}$$

Social surplus is equal to $\widehat{SS}^u = \int_{\widehat{p}^u}^{\infty} N(u)\, du + \tau^u \widehat{N}^u$. Finally, the welfare gain in switching from no fare to the optimal uniform fare is $G^{nu} = \widehat{SS}^u - \widehat{SS}^n$.

Appendix G.2. Social optimum

The social optimum can be supported by imposing time-dependent fares $\tau_k^o = \lambda n_k^o / s$. Given (7) and (16), total travel costs are $\widehat{TC}^o = \bar{\delta} \widehat{N}^o + \frac{\lambda \left(\widehat{N}^o\right)^2}{ms} - RV^o$. Since variable revenue in Eq. (15)

does not depend on the number of trips, the marginal social cost of a trip is $\widehat{MSC}^o = \bar{\delta} + \frac{2\lambda\widehat{N}^o}{ms}$. Similar to the optimal uniform-fare regime, the efficient price of a trip equals marginal social cost:

$$\widehat{p}^o = \widehat{MSC}^o\left(\widehat{N}^o\right) = \bar{\delta} + \frac{2\lambda\widehat{N}^o}{ms}. \tag{G.4}$$

Eqs. (G.3) and (G.4) reveal that the optimal price is the same function of usage in regimes u and o. This is consistent with the observation that, if the crowding cost function is linear, the marginal social cost of trips is the same in the SO and UE. Social surplus is equal to

$$\widehat{SS}^o = \int_{\widehat{p}^o}^{\infty} N(u)\,du + R^o\left(\widehat{N}^o\right) = \int_{\widehat{p}^o}^{\infty} N(u)\,du + \frac{\lambda\left(\widehat{N}^o\right)^2}{ms} + RV^o.$$

The welfare gain in switching from no fare to the SO-fare is $G^{no} = \widehat{SS}^o - \widehat{SS}^n$, and the welfare gain in switching from the optimal uniform fare to the SO-fare is $G^{uo} = \widehat{SS}^o - \widehat{SS}^u$.

Appendix G.3. Comparison of the regimes

Private costs in regimes n, u and o are given by Eqs. (G.2), (G.3), and (G.4) respectively. For given values of m, s, and N, it is clear that private costs are the same in regimes u and o, and lower in regime n. With elastic demand this implies that equilibrium usage is the same in regimes u and o, and higher in regime n. Correspondingly, the equilibrium private cost and consumers' surplus are the same in regimes u and o, and higher in regime n. Social surplus is highest in regime o, lowest in regime n, and intermediate in regime u.

Appendix H. Optimal timetable

Appendix H.1. Optimal timetable

The optimal timetable is derived by minimizing users' total costs. For given m and s, the optimal timetables for the UE and SO generally differ because their load patterns differ. However, the timetables coincide given linear schedule delay costs and a constant headway.

With a constant headway between trains, the timetable is fully described by the arrival time of the last train, t_m. Let $\mathbb{1}_x$ be the indicator function with $\mathbb{1}_x = 1$ if x is true, and $\mathbb{1}_x = 0$ otherwise. The optimal timetable is described in:

Proposition 9. *With the optimal timetable, the last train leaves at time*

$$t_m^o = t^* + h\left(m - \varphi_m - \mathbb{1}_{\frac{\gamma m}{\beta+\gamma} > \varphi_m}\right),$$

where $\varphi_m \equiv \left\lfloor \frac{\gamma m}{\beta+\gamma} + \frac{1}{2} \right\rfloor$. *Train* k, *with* $k = \varphi_m + \mathbb{1}_{\frac{\gamma m}{\beta+\gamma} > \varphi_m}$, *arrives on time at* t^*. *The unweighted average schedule delay cost is* $\overline{\delta} \simeq \frac{\beta\gamma}{\beta+\gamma}\frac{mh}{2}$.

According to Prop. 9, the higher the cost of late arrival (γ) relative to early arrival (β) the earlier service begins. The fraction of trains that arrive before t^* is approximately $\gamma/(\beta+\gamma)$. This formula is approximate because the number of trains is integer-valued. For the same reason, the formula for average schedule delay cost, $\overline{\delta}$, is approximate too.

We first derive the optimal timetable for the UE, and then show that this timetable is also optimal for the SO.

Appendix H.2. Optimal timetable for user equilibrium

Total costs in the UE are $TC^e = \overline{\delta}N + \frac{\lambda N^2}{ms}$. The timetable should therefore be chosen to minimize average schedule delay cost, $\overline{\delta}$. The timetable can be defined by the arrival time of the last train, t_m. It is clearly not optimal to set $t_m < t^*$, and have all trains arrive early, since $\overline{\delta}$ could be reduced by setting $t_m = t^*$. Similarly, it is not optimal to set $t_m > t^* + (m-1)h$, and have all trains arrive late, since $\overline{\delta}$ could be reduced by setting $t_m = t^* + (m-1)h$. Thus, one train must arrive during the interval $(t^* - h, t^*]$. Call it train $\hat{k}$. Train $\hat{k}$ is the last train to arrive at or before t^*. Average schedule delay cost is

$$\begin{aligned}
\overline{\delta} &= \frac{1}{m}\left(\sum_{k=1}^{\hat{k}} \beta\left(t^* - t_k\right) + \sum_{k=\hat{k}+1}^{m} \gamma\left(t_k - t^*\right)\right) \\
&= \frac{1}{m}\left(\sum_{k=1}^{\hat{k}} \beta\left(t^* - t_{\hat{k}} + h\left(\hat{k} - k\right)\right) + \sum_{k=\hat{k}+1}^{m} \gamma\left(t_{\hat{k}} - t^* + h\left(k - \hat{k}\right)\right)\right) \\
&= \frac{1}{m}\left(\left(t^* - t_{\hat{k}}\right)\left[(\beta+\gamma)\hat{k} - \gamma m\right] + (\beta+\gamma)h\frac{\hat{k}\left(\hat{k}-1\right)}{2} + \gamma h\frac{m\left(m+1-2\hat{k}\right)}{2}\right). \quad \text{(H.1)}
\end{aligned}$$

The first component of the RHS of Eq. (H.1), $\left(t^* - t_{\hat{k}}\right)$, is the time between the arrival time of train $\hat{k}$ and t^*. If $t_{\hat{k}} < t^*$ we can differentiate Eq. (H.1):

$$\frac{\partial\overline{\delta}}{\partial\left(t^* - t_{\hat{k}}\right)} = \frac{(\beta+\gamma)\hat{k}}{m} - \gamma.$$

If $\hat{k} > \gamma m/(\beta+\gamma)$, then $\partial\overline{\delta}/\partial\left(t^* - t_{\hat{k}}\right) > 0$ and $\overline{\delta}$ is minimized by setting $t^* - t_{\hat{k}}$ to its minimal value: $t^* - t_{\hat{k}} = 0$. Conversely, if $\hat{k} < \gamma m/(\beta+\gamma)$, then $\partial\overline{\delta}/\partial\left(t^* - t_{\hat{k}}\right) < 0$ and $\overline{\delta}$ is minimized by

setting $t^* - t_{\hat{k}} = h$. Hence it is optimal to schedule one train at t^*. Call it train k^*. Replacing $\hat{k}$ in Eq. (H.1) with k^* one obtains

$$\overline{\delta} = (\beta + \gamma)\, h \frac{k^* (k^* - 1)}{2m} + \gamma h \frac{m + 1 - 2k^*}{2}.$$

Treating k^* as a continuous variable for the moment, the first-order condition for minimizing $\overline{\delta}$ with respect to k^* is $k^{*o} = \frac{\gamma m}{\beta+\gamma} + \frac{1}{2}$. Since k^* is an integer, we have to compare $\overline{\delta}$ when $k^* = \lfloor k^{*o} \rfloor$ and when $k^* = \lfloor k^{*o} \rfloor + 1$. We find

$$\overline{\delta}_{k^* = \lfloor k^{*o} \rfloor} - \overline{\delta}_{k^* = \lfloor k^{*o} \rfloor + 1} \lesseqgtr 0 \Longleftrightarrow \frac{\gamma m}{\beta + \gamma} \lesseqgtr \left\lfloor \frac{\gamma m}{\beta + \gamma} + \frac{1}{2} \right\rfloor.$$

Hence,

$$\begin{aligned} k^* &= \left\lfloor \frac{\gamma m}{\beta + \gamma} + \frac{1}{2} \right\rfloor + 1 \times \mathbb{1}_{\frac{\gamma m}{\beta+\gamma} > \left\lfloor \frac{\gamma m}{\beta+\gamma} + \frac{1}{2} \right\rfloor} \\ t_m &= t^* + h \left(m - \left\lfloor \frac{\gamma m}{\beta + \gamma} + \frac{1}{2} \right\rfloor - 1 \times \mathbb{1}_{\frac{\gamma m}{\beta+\gamma} > \left\lfloor \frac{\gamma m}{\beta+\gamma} + \frac{1}{2} \right\rfloor} \right), \end{aligned}$$

where $\mathbb{1}$ is an indicator function with $\mathbb{1}_x = 1$ if x is true, and $\mathbb{1}_x = 0$ otherwise. In summary, if $\gamma m / (\beta + \gamma) > \lfloor \gamma m / (\beta + \gamma) + 1/2 \rfloor$, then

$$\begin{aligned} k^* &= \lfloor \gamma m / (\beta + \gamma) + 1/2 \rfloor + 1, \text{ and} \\ t_m &= t^* + h(m - 1 - \lfloor \gamma m / (\beta + \gamma) + 1/2 \rfloor). \end{aligned}$$

Conversely, if $\gamma m / (\beta + \gamma) < \left\lfloor \frac{\gamma m}{\beta+\gamma} + \frac{1}{2} \right\rfloor$, then

$$\begin{aligned} k^* &= \lfloor \gamma m / (\beta + \gamma) + 1/2 \rfloor, \text{ and} \\ t_m &= t^* + h \left(m - \lfloor \gamma m / (\beta + \gamma) + 1/2 \rfloor \right). \end{aligned}$$

Appendix H.3. Optimal timetable for social optimum

Total costs in the social optimum are given by Eqs. (16) and (15):

$$TC^o = \overline{\delta} N + \frac{\lambda N^2}{ms} - \frac{s}{4\lambda} \left(\Delta - m \overline{\delta}^2 \right),$$

where

$$\Delta - m\overline{\delta}^2 = \sum_{k=1}^{m} \delta_k^2 - \frac{1}{m} \left[\sum_{k=1}^{m} \delta_k \right]^2, \tag{H.2}$$

and

$$\delta_k = \beta \left[t^* - t_m + h\,(m - k) \right]^+ + \gamma \left[t_m - t^* - h\,(m - k) \right]^+. \tag{H.3}$$

TC^o differs from TC^e in including the third term on the right-hand side. As above, let $\hat{k}$ be the last train to arrive at or before t^*. Differentiating (H.2) with respect to t_m, and using (H.3), it is possible to show after considerable algebra that

$$\frac{\partial\left(\Delta - m\overline{\delta}^2\right)}{\partial t_m} = \frac{\gamma}{\beta+\gamma}m + 1 - \hat{k}.$$

The term $\Delta - m\overline{\delta}^2$ therefore reaches an extreme point for the same $\hat{k}$ as does $\overline{\delta}$. Hence TC^o reaches a minimum for the same timetable as TC^e.

Appendix I. Proof of Proposition 6

First-order conditions for a maximum of SS^i are[58]

$$\frac{\partial SS^i}{\partial s} = p\left(N\right)\frac{\partial N}{\partial s} - \left(-\frac{\lambda N^2}{ms^2} + \left(\bar{\delta} + \frac{2\lambda N}{ms}\right)\frac{\partial N}{\partial s} + K_s\right) + RV_s^i = 0, \tag{I.1}$$

$$\frac{\partial SS^i}{\partial m} = p\left(N\right)\frac{\partial N}{\partial m} - \left(\frac{\partial \bar{\delta}}{\partial m}N - \frac{\lambda N^2}{m^2 s} + \left(\bar{\delta} + \frac{2\lambda N}{ms}\right)\frac{\partial N}{\partial m} + K_m\right) + RV_m^i = 0. \tag{I.2}$$

The private cost of usage is given by Eq. (G.1) which can be written

$$p\left(N\right) - \left(\bar{\delta} + \frac{2\lambda N}{ms}\right) = \tau - \frac{\lambda N}{ms}. \tag{I.3}$$

The fare, τ, depends on the pricing regime. To maintain generality we assume for the moment that τ can depend on N, m, and s. Substituting (I.3) into (I.1) and (I.2) yields:

$$\frac{\lambda N^2}{ms^2} + \left(\tau - \frac{\lambda N}{ms}\right)\frac{\partial N}{\partial s} - K_s + RV_s^i = 0, \tag{I.4}$$

$$\frac{\lambda N^2}{m^2 s} - \frac{\partial \bar{\delta}}{\partial m}N + \left(\tau - \frac{\lambda N}{ms}\right)\frac{\partial N}{\partial m} - K_m + RV_m^i = 0. \tag{I.5}$$

The demand derivatives are obtained by totally differentiating (G.1):

$$\frac{\partial N}{\partial s} = \frac{-\frac{\lambda N}{ms^2} + \frac{d\tau}{ds}}{p_N - \frac{\lambda}{ms} - \frac{d\tau}{dN}} > 0, \tag{I.6}$$

$$\frac{\partial N}{\partial m} = \frac{\frac{\partial \bar{\delta}}{\partial m} - \frac{\lambda N}{m^2 s} + \frac{d\tau}{dm}}{p_N - \frac{\lambda}{ms} - \frac{d\tau}{dN}} > 0. \tag{I.7}$$

[58] Given $\bar{\delta} \cong \beta\gamma/\left(\beta+\gamma\right)hm/2$ as per Prop. (9), $\partial\bar{\delta}/\partial m \cong \beta\gamma/\left(\beta+\gamma\right)h/2$ which is a constant.

Substituting (I.6) and (I.7) into (I.4) and (I.5), the first-order conditions become

$$\begin{aligned}
\frac{\lambda N^2}{ms^2}\cdot\frac{p_N N-\tau-\frac{d\tau}{dN}N}{p_N N-\frac{\lambda N}{ms}-\frac{d\tau}{dN}N}+\frac{\left(\tau-\frac{\lambda N}{ms}\right)\frac{d\tau}{ds}N}{p_N N-\frac{\lambda N}{ms}-\frac{d\tau}{dN}N} &= K_s-RV_s^i,\\
\left(\frac{\lambda N}{m^2 s}-\frac{\partial\bar{\delta}}{\partial m}\right)N\cdot\frac{p_N N-\tau-\frac{d\tau}{dN}N}{p_N N-\frac{\lambda N}{ms}-\frac{d\tau}{dN}N}+\frac{\left(\tau-\frac{\lambda N}{ms}\right)\frac{d\tau}{dm}N}{p_N N-\frac{\lambda N}{ms}-\frac{d\tau}{dN}N} &= K_m-RV_m^i.
\end{aligned}$$

Appendix J. Optimal capacity in the uniform-fare regime

In the uniform-fare regime, Eqs. (21a) and (21b) can be solved jointly to obtain quartic equations for the unconditionally optimal values, s_*^u and m_*^u:

$$\begin{aligned}
\nu_1(s_*^u)^4+Z(s_*^u)^3-\lambda Z^2N^2/\nu_2^2 &= 0,\\
\nu_1(m_*^u)^4+\nu_2(m_*^u)^3-\lambda\nu_2^2N^2/Z^2 &= 0,
\end{aligned}$$

where $Z\equiv N\partial\bar{\delta}/\partial m+\nu_0$. If $\nu_1=0$, an explicit solution obtains:

$$\begin{aligned}
s_*^u(N) &= \left(\frac{\partial\bar{\delta}}{\partial m}\frac{\lambda}{\nu_2^2}N^3+\frac{\lambda\nu_0}{\nu_2^2}N^2\right)^{1/3},\\
m_*^u(N) &= \frac{\lambda}{\nu_2\left[s_*^u(N)\right]^2}N^2.
\end{aligned}$$

According to Mohring's (1972) square-root rule, both optimal service frequency and the number of passengers carried per train (or bus) increase with $\sqrt{N}$. In the PTC model, service frequency is constant because headway is fixed. s_*^u rises[59] with N at a rate faster than $N^{2/3}$, and m_*^u grows at a rate slower than $N^{2/3}$, but since it does increase with N the duration of the travel period increases. As N increases, adding trains becomes less attractive because they are scheduled increasingly early or late. The number of trains, m_*^u, approaches a constant value and s_*^u increases approximately linearly with N.[60] With $\nu_1=0$, it is possible to show that equilibrium user cost is a U-shaped function of N with a minimum at $N=\nu_0\left(\partial\bar{\delta}/\partial m\right)^{-1}$. However, both the equilibrium price, p, and the average system cost, $c^u\left(m_*^u,s_*^u\right)+K\left(m_*^u,s_*^u\right)/N$, decline monotonically with N. This is attributable to the fact that, with $\nu_1=0$, the user cost function has constant returns to scale while the service cost function has increasing returns.

[59] In Kraus and Yoshida's (2002) model, the effect of N on s is ambiguous. Nevertheless, they remark (p.178) that s is likely to increase with N.

[60] Eventually a physical limit to train capacity would be reached due to constraints on platform size or tractive power.

The limiting case $\nu_0 = 0$ applies if there are no scale economies with respect to train size:

$$\begin{aligned} \nu_1 + \frac{\partial \bar{\delta}}{\partial m}\frac{N}{s_*^u} - \frac{\lambda}{\nu_2^2}\left(\frac{\partial \bar{\delta}}{\partial m}\right)^2\left(\frac{N}{s_*^u}\right)^4 &= 0, \\ \nu_1 \left(m_*^u\right)^4 + \nu_2 \left(m_*^u\right)^3 - \lambda\nu_2^2\left(\frac{\partial \bar{\delta}}{\partial m}\right)^{-2} &= 0. \end{aligned}$$

The first equation in the system solves for a unique value of N/s_*^u which implies that train capacity is chosen proportional to ridership. The second equation solves for a unique value of m_*^u which implies that the number of trains is independent of ridership. These properties imply that equilibrium user cost, c^u, price, p^u, and average system cost are all constant. Hence, unlike in Mohring's model (in which $\nu_1 = \nu_2 = 0$, $\nu_0 > 0$) there are no scale economies with respect to traffic density.

Appendix K. Optimal transit capacity with no fare and traffic congestion

With no fare and no traffic congestion, the second-best optimal values of s and m are given by Eqs. (19a) and (19b) in Proposition 6 with $i = n$:

$$\begin{aligned} \frac{\lambda N^2}{ms^2}\cdot\frac{p_N}{p_N - \frac{\lambda}{ms}} &= K_s, \\ \left(\frac{\lambda N}{m^2 s} - \frac{\partial \bar{\delta}}{\partial m}\right) N \cdot \frac{p_N}{p_N - \frac{\lambda}{ms}} &= K_m. \end{aligned}$$

When transit demand is price sensitive, the potential benefit from expanding transit capacity is partly undermined by latent demand.

It is straightforward but tedious to show that, with unpriced traffic congestion, the corresponding conditions are

$$\frac{\lambda N_R^2}{ms^2}\cdot\frac{p_N}{p_N - \frac{\lambda}{ms}}\left[\frac{1 + \frac{N_A}{N_R}}{1 + \frac{p_N \frac{\lambda}{ms}}{C_A'\left(p_N - \frac{\lambda}{ms}\right)}}\right] = K_s, \tag{K.2a}$$

$$\left(\frac{\lambda N_R}{m^2 s} - \frac{\partial \bar{\delta}}{\partial m}\right) N_R \cdot \frac{p_N}{p_N - \frac{\lambda}{ms}}\left[\frac{1 + \frac{N_A}{N_R}}{1 + \frac{p_N \frac{\lambda}{ms}}{C_A'\left(p_N - \frac{\lambda}{ms}\right)}}\right] = K_m. \tag{K.2b}$$

The LHS of each equation is multiplied by a scaling factor shown in square brackets. Because the scaling factor is the same, traffic congestion does not alter the relative benefits of expanding train capacity versus increasing the number of trains. The scaling factor can be greater than or less than

1. Expanding transit capacity does induce more travelers to take transit, but the benefit is diluted by the fact that the reduction in traffic congestion induces some travelers to return to their cars. Eqs. (K.2a) and (K.2b) reveal that the presence of traffic congestion is more likely to increase the net benefit from capacity expansion when N_A and C'_A are large. This is because the benefits from congestion relief are then high.

Appendix L. Proof of Propositions 7 and 8

In fare regime f, the average fare paid is $\tau = \frac{\lambda N}{ms}$. Variable revenue is $RV^f = f \cdot RV^o$ with $RV^o = \frac{s}{4\lambda}\left(\sum_{k=1}^{m}\delta_k^2 - \frac{1}{m}\left[\sum_{k=1}^{m}\delta_k\right]^2\right)$, or $RV^o = \frac{s}{4\lambda}V$ for short, where V is a function of m but does not depend on s or N. In parts of the proof it is necessary to impose an upper bound on RV^o. For this purpose we first prove the following

Lemma: $RV^o < \frac{1}{12}\frac{\lambda N^2}{ms} = \frac{1}{12}\tau N$.

Proof: Condition (E.1) that all trains are used implies

$$\frac{\lambda N}{ms} > \max\left[\delta_1, \delta_m\right] - \overline{\delta}.$$

For large values of m, Prop. 9 yields $\delta_1 \simeq \delta_m \simeq \frac{\beta\gamma}{\beta+\gamma}mh$, and $\overline{\delta} \simeq \frac{\beta\gamma}{\beta+\gamma}\frac{mh}{2}$. Hence

$$\frac{\lambda N}{ms} > \frac{\beta\gamma}{\beta+\gamma}\frac{mh}{2}.$$

This inequality can be rearranged to obtain

$$\frac{\lambda N^2}{ms} > \frac{s}{4\lambda}\left(\frac{\beta\gamma}{\beta+\gamma}\right)^2 h^2 m^3 = 12RV^o. \quad \blacksquare \tag{L.3}$$

The Lemma establishes that variable revenue from the SO-fare cannot exceed 1/12 of the revenue collected from the average fare.

Total demand (N) and the optimal values of m and s are given by Eqs. (I.3), (I.4) and (I.5) in Appendix I. Applying $\tau = \frac{\lambda N}{ms}$, these equations become:

$$p\left(N\right) - \left(\overline{\delta} + \frac{2\lambda N}{ms}\right) = 0, \tag{L.4}$$

$$\frac{\lambda N^2}{ms^2} - K_s + RV_s^f = 0, \tag{L.5}$$

$$\frac{\lambda N^2}{m^2 s} - \phi N - K_m + RV_m^f = 0, \tag{L.6}$$

where $K_s = \nu_1 m + \nu_2$, $K_m = \nu_0 + \nu_1 s$, and $\phi \equiv d\bar{\delta}/dm$ is a constant. Totally differentiating (L.4), (L.5), and (L.6) with respect to f yields a system of three equations:

$$\begin{bmatrix} \frac{2\lambda N}{m^2 s} & \frac{2\lambda N}{ms^2} & p_N - \frac{2\lambda N}{ms} \\ -\frac{\lambda N^2}{m^2 s^2} - \nu_1 + RV^f_{ms} & -\frac{2\lambda N^2}{ms^3} & \frac{2\lambda N}{ms^2} \\ -\frac{2\lambda N^2}{m^3 s} + RV^f_{mm} & -\frac{\lambda N^2}{m^2 s^2} - \nu_1 + RV^f_{ms} & \frac{2\lambda N}{m^2 s} - \phi \end{bmatrix} \begin{bmatrix} \frac{dm}{df} \\ \frac{ds}{df} \\ \frac{dN}{df} \end{bmatrix} = \begin{bmatrix} 0 \\ -RV^f_{sf} \\ -RV^f_{mf} \end{bmatrix}. \quad \text{(L.7)}$$

The determinant of the coefficient matrix is positive. Each derivative is considered in turn.

Appendix L.1. Derivative of m

Applying Cramer's Rule to (L.7) one obtains:

$$\frac{dm}{df} \overset{s}{=} \left(p_N - \frac{2\lambda}{ms}\right) \frac{1}{ms^2} \left(\underbrace{-5\frac{\lambda N^2}{ms} + ms\nu_1 - 3fRV^o}_{A} \right) - 8\frac{\lambda^2 N^2}{m^3 s^4}, \quad \text{(L.8)}$$

where $\overset{s}{=}$ means has the same sign as. Expression (L.8) is positive if it is positive in the limiting case of perfectly inelastic demand ($p_N \to -\infty$) as well as perfectly elastic demand ($p_N = 0$). In the limiting case $p_N \to -\infty$, $dm/df > 0$ if term A is negative. From Eq. (L.5)

$$\frac{\lambda N^2}{ms} = s\left(\nu_1 m + \nu_2\right) - fRV^o, \quad \text{(L.9)}$$

and term A can be written

$$A = -4\nu_1 ms - 5\nu_2 s + 2fRV^o. \quad \text{(L.10)}$$

According to the Lemma, $RV^o < \frac{1}{12}\frac{\lambda N^2}{ms}$. Given (L.5), this implies $RV^o < \frac{1}{12}s\left(\nu_1 m + \nu_2\right)$. Substituting this inequality into (L.10) yields

$$A < -4\nu_1 ms - 5\nu_2 s + \frac{1}{6}fs\left(\nu_1 m + \nu_2\right) < 0.$$

In the opposite limiting case with $p_N = 0$,

$$\begin{aligned} \frac{dm}{df} &\overset{s}{=} -\frac{2\lambda}{m^2 s^3}\left(-5\frac{\lambda N^2}{ms} + ms\nu_1 - 3fRV^o\right) - 8\frac{\lambda^2 N^2}{m^3 s^4} \\ &\overset{s}{=} \frac{\lambda N^2}{ms} - ms\nu_1 + 3fRV^o \\ &\overset{s}{=} s\nu_2 + 2fRV^o > 0. \end{aligned}$$

where the last line follows from (L.9). This proves that $dm/df > 0$.

Appendix L.2. Derivative of s

Applying Cramer's Rule to (L.7) again one obtains:

$$\frac{ds}{df} \overset{s}{=} \left(p_N - \frac{2\lambda}{ms}\right)\frac{1}{m}\left(\underbrace{\frac{\lambda N^2}{ms} + ms\nu_1 - 3fRV^o}_{B}\right) - 4\left(\phi - \frac{2\lambda N}{m^2 s}\right)\frac{\lambda N}{ms}. \tag{L.11}$$

Expression (L.11) is negative if it is negative in both limiting cases of perfectly inelastic demand and perfectly elastic demand. In the limiting case $p_N \to -\infty$, $ds/df < 0$ if term B is positive. This can be confirmed following the same steps as for (L.8). With $p_N = 0$,

$$\frac{ds}{df} \overset{s}{=} -\frac{2\lambda}{m^2 s}\left(\frac{\lambda N^2}{ms} + ms\nu_1 - 3fRV^o\right) - 4\left(\phi - \frac{2\lambda N}{m^2 s}\right)\frac{\lambda N}{ms}.$$

Using (L.9), this simplifies to

$$\frac{ds}{df} \overset{s}{=} 3s\nu_2 - 2\phi N.$$

which can be positive or negative. Hence, with sufficiently inelastic demand $ds/df < 0$, but the sign is otherwise indefinite.

Appendix L.3. Derivative of ms

The derivative of fleet capacity is

$$\frac{d(ms)}{df} = s\frac{dm}{df} + m\frac{ds}{df}.$$

Applying formulas for the derivatives dm/df and ds/df one obtains

$$\frac{d(ms)}{df} \overset{s}{=} \frac{2\lambda}{ms}\left(-4ms\nu_1 + 6fRV^o\right) - p_N\left(4s\nu_2 + 2fRV^o\right). \tag{L.12}$$

In the limiting case $p_N \to -\infty$, (L.12) is clearly positive. With $p_N = 0$, the sign is indefinite.

Appendix L.4. Derivative of N

For total usage

$$\frac{dN}{df} \overset{s}{=} \left(\phi - \frac{2\lambda N}{m^2 s}\right)\left(\frac{\lambda N^2}{ms} - 3fRV^o - 6\frac{\lambda N^2}{ms}\right) - \frac{2\lambda N}{m^2 s}\left(\frac{\lambda N^2}{ms} + 3ms\nu_1 - 3fRV^o\right).$$

Applying first-order conditions (L.5) and (L.6), and combining terms, one obtains

$$\begin{aligned}\frac{dN}{df} \overset{s}{=}\ & (5\nu_0 m + 2\nu_1 ms + 3\nu_2 s)\underbrace{s(\nu_1 m + \nu_2)}_{C} \\ & - (2\nu_0 m + 8\nu_1 ms + 12\nu_2 s)\, fRV^o.\end{aligned} \tag{L.13}$$

Expression (L.13) is positive with $f = 0$. With $f > 0$, the second line is negative and the sign of (L.13) depends on the relative magnitudes of two terms. From the Lemma, RV^o cannot exceed 1/12 of the revenue collected from the average fare. Term C in line 1 comprises the portion of capacity costs that vary with train capacity. If there are no scale economies with respect to train size (i.e., $\nu_0 = 0$), term C accounts for total capacity costs. In addition, fare revenue fully covers capacity costs. Expression (L.13) then becomes

$$\frac{dN}{df} \overset{s}{=} (2\nu_1 ms + 3\nu_2 s) K(m,s) - (8\nu_1 ms + 12\nu_2 s) f \frac{1}{12} K(m,s)$$

which is unambiguously positive. In summary, total usage increases with the efficiency of the fare regime unless, possibly, there are very substantial scale economies with respect to train capacity.

Appendix M. Sensitivity analysis

Appendix M.1. Integer-valued number of trains

The number of trains, m, has been treated as a continuous variable although it is discrete in reality. An integer constraint can be imposed by fixing m, and then choosing s for regimes n, u and o. To assess how the integer constraint affects results, m was first set to the largest integer smaller than the real-valued solution, and then the next integer larger. Thus, for the no-fare regime m was first set to $\lfloor m_*^n \rfloor$, and then $\lfloor m_*^n \rfloor + 1$. Since m_*^u was calibrated to be an integer value, this was unnecessary for regime u. The integer value yielding the higher social surplus was then selected. The results changed very little, and social surplus was virtually unchanged. Integer constraints also had little effect for a range of other parameter values.

Appendix M.2. Demand elasticity

If the price elasticity of demand is reduced to $\eta = 0$, ridership is the same in the three fare regimes. With $p_N = -\infty$, the first-order conditions (19a) and (19b) for s and m are the same for regimes n and u so that $s_*^u = s_*^n$, and $m_*^u = m_*^n$. Imposing the uniform fare yields no welfare gain at all, and merely transfers money from users to the transit authority. The SO-fare does yield a welfare gain although (with ridership fixed at $32,600$) it is only €0.185 per rider compared to €0.45 per rider in the base case.

To examine the effects of a higher price elasticity, η was doubled in magnitude to $-2/3$. To maintain equilibrium ridership at $32,600$ in the optimal uniform-fare regime, parameter N_0 was

Table M.1: Comparison of no-fare, optimal uniform fare, and SO-fare (i.e., social optimum) regimes: $\eta = -2/3$

	Fare regime		
	No-fare (n)	Optimal uniform fare (u)	Social optimum (o)
m	26.34	24	26.75
s	1,764	1,733	1,725
N	41,006	32,600	33,220
p	6.72	9.48	9.22
$Rev/user$	0	3.45	3.39
TCC	187,604	133,499	112,503
SDC	88,044	63,244	81,248
TC	275,648	196,743	193,751
K	139,632	134,889	137,558
R	0	112,407	112,503
ρ	0	0.833	0.818
CS	1,206,851	1,106,343	1,115,033
SS	1,067,219	1,083,862	1,089,978
$Totalgain$		16,643	22,759
$Gain/user$	0	0.51	0.70
$Rel.eff$	0	0.73	1

increased to 146,056. The results are shown in Table M.1. With the higher price elasticity, consumers' surplus and social surplus in each regime are lower than with the base-case parameters. Regime u is otherwise unaffected. However, the welfare gain per rider nearly doubles from €0.27 to €0.51. The welfare gain per rider in the social optimum increases from €0.45 to €0.70, but by a smaller percentage so that the relative efficiency of regime u increases.