LEM
LILLE ÉCONOMIE MANAGEMENT

Document de travail du LEM / Discussion paper LEM
2022-05

Social Optimum in the Basic Bathtub Model

Richard Arnott
Department of Economics, University of California, Riverside CA 92506

Moez Kilani
Univ. Lille, CNRS, IESEG School of Management, UMR 9221 - LEM - Lille Économie Management, F-59000 Lille, France | University of Littoral Opal Coast, F-59140, Dunkerque

https://lem.univ-lille.fr/

Social Optimum in the Basic Bathtub Model

Richard Arnott[1] and Moez Kilani*[2,3]

[1]Department of Economics, University of California, Riverside CA 92506

[2]University of Lille, CNRS, IESEG School of Management, UMR 9221 - LEM - Lille Économie Management, F-59000 Lille, France

[3]University of Littoral Opal Coast, F-59140, Dunkerque

January, 2022

Abstract

The basic bathtub model extends Vickrey's bottleneck model to admit hypercongestion (traffic jam situations). A fixed number of identical commuters travel a fixed distance over a dense network of identical city streets between home and work in the early morning rush hour under dynamic MFD congestion. This paper investigates social optima in the basic bathtub model, and contrasts them with the corresponding competitive equilibria (reported in Arnott and Buli (2018) and Buli (2019)). The model gives rise to delay-differential equations, which considerably complicate analysis of the solution properties and design of computational solution algorithms. The paper considers the cases of smooth and strictly concave travel utility functions and of α-β-γ tastes. For each it develops a customized solution algorithm, which it applies to several examples, and for α-β-γ tastes it derives analytical properties as well. Departures may occur continuously, in departure masses, or a mix of the two. As well, hypercongestion may occur in the social optimum. The paper explores how these qualitative solution properties are related to tastes.

JEL codes: C60, D60, R40

Keywords: traffic congestion, hypercongestion, rush hour traffic dynamics, delay differential equations, bathtub model, optimum

*Corresponding author.

Contents

1 Introduction

Hypercongestion occurs when an increase in traffic density is associated with a decrease in traffic flow; intuitively, it corresponds to traffic jam situations. Starting from the seminal paper by Geroliminis and Daganzo (2008), over the last decade evidence has been accumulating on the empirical importance in major metropolitan areas of hypercongestion during rush hours. The workhorse economic model of rush hour traffic dynamics, Vickrey's bottleneck model (1969), rules out hypercongestion by assumption. Urban transportation economists and scientists have long recognized the potential importance of developing models of rush hour traffic dynamics with endogenous trip timing that account for hypercongestion. One has been the basic bathtub model (with endogenous trip timing)[1] in which a fixed number of identical commuters per unit area each travels a fixed distance from home to work over a dense network of city streets during the morning rush hour, and traffic velocity at a point in time is negatively related to the contemporaneous traffic density. While the model is conceptually simple, physically sound, and intuitively appealing, unfortunately it is analytically intractable; in particular, it gives rise to delay differential equations, which are at the research frontier in applied mathematics. The literature has taken three different approaches to deal with this intractability: approximation (Small and Chu 2003, Geroliminis and Levinson 2009, Arnott 2013), special cases (Arnott et al. 2016, Fosgerau 2015, Lamotte and Geroliminis 2017), and computation (Arnott and Buli 2018, AB hereafter). Progress is being made in understanding the model's properties, albeit rather slowly and inelegantly.

This paper contributes to the line of literature that explores the properties of the basic bathtub model "proper" – without approximation. It builds on two earlier papers. Arnott et al. (2016) undertook a preliminary examination of the special case with α-β tastes (no late

[1] We entitled this paper "Social optimum in the basic bathtub model" since it is a companion to Arnott and Buli (2018), which has the title, "Solving for equilibrium in the basic bathtub model." After Arnott and Buli (2018) was published, Wen-Long Jin brought to Arnott's attention an unpublished manuscript, dated 1994, in the Vickrey Archives, now published as Vickrey (2019), which defined the term "bathtub model" for the first time: "Here a maze of congested streets is treated as an undifferentiated movement area in which movement takes place at a speed which is a function of the density of cars in the area". Vickrey's implicit definition is broader than Arnott's previous usage of the term, which added a trip-timing equilibrium condition. We recommend that future research employ Vickrey's definition.

arrivals), congestion technology described by Greenshields' Relation (that traffic velocity is a negative linear function of traffic density), and a fixed population. Under the assumption that all departures occur in contiguous departure masses (so that a departure mass departs from home immediately after the previous departure mass arrives at work) it derived a closed-form solution for the unique equilibrium, but made only limited progress in solving the social optimum problem. In the equilibrium, traffic may be hypercongested at the peak of the rush hour.

AB developed an algorithm to solve numerically for equilibrium in the basic bathtub model[2] when utility is a smooth and strictly concave function of departure time and trip duration, the congestion technology is described by Greenshields' Relation, and the population is fixed. Two results are particularly noteworthy. First, while, density and velocity are continuous functions of time over the rush hour, the rate of entry into the traffic stream (from home) and the rate of exit from the traffic stream (to work) exhibit discontinuities. Second, there may be two equilibria, one (which is said to exhibit aggregate hypercongestion) Pareto dominated by the other.

This paper explores the corresponding social optima. Section 3 is a companion to AB. It develops an algorithm to solve numerically for the social optimum with a similar model variant that AB employ to examine equilibrium, and applies it to the same examples that AB employs. For reasons that shall be explained, it is less successful in deriving qualitative properties of the general solution for the social optimum than was AB is deriving qualitative properties of the general solution for the equilibrium. Section 4 is a companion to Arnott et al. (2016). It develops an algorithm to solve numerically for the social optimum with the same model variant that Arnott et al. (2016) employed to examine equilibrium, and applies it to solve similar examples. In contrast to the situation with a smooth and strictly concave utility function, with the assumed α-β-γ tastes many qualitative solution properties of the social optimum are derived analytically.

Aiming to integrate the results for the two cases of smooth and strictly concave utility

[2]Buli (2019) draws on the algorithm to investigate existence and uniqueness of equilibrium in the basic bathtub model, as well as to derive some comparative static properties of equilibrium.

functions and of α-β-γ tastes, the paper poses three general questions about properties of the social optimum in the basic bathtub model, without providing complete answers to any of them:

1. Under what conditions can hypercongestion occur in the social optimum? Conversely, what restrictions ensure that hypercongestion does not occur in the social optimum?

2. At least for the examples considered, with smooth and strictly concave utility functions traffic density is continuous over the rush hour. In contrast, with α-β-γ tastes, traffic density as a function of time exhibits discontinuities, corresponding to all departures occurring in contiguous departure masses. What is the root cause of the qualitative difference?

3. In both cases, the entry and exit rate functions of time exhibit discontinuities. Why?

Section 2 sets the stage, providing a formal statement of the social optimum problem. Sections 3 treats the case of smooth and strictly concave utility functions. Section 3.1 briefly reviews the solution method developed by AB to solve for equilibrium with smooth and strictly concave utility functions and presents a sample solution. Section 3.2 develops a numerical solution strategy to solve numerically for the social optimum with smooth and strictly concave utility functions. Section 3.3 applies it to several numerical examples, and section 3.4 compares the numerical solutions for the optimum with those of the corresponding equilibrium for a particular numerical example. Section 4 treats the case of α-β-γ tastes. Section 4.1 provides an analytical solution to the social optimum, which includes a sufficient condition for hypercongestion not to occur and a discussion of why departure masses may be optimal; section 4.2 develops a customized algorithm to solve for the social optimum; and section 4.3 presents a numerical example. Section 5 briefly reviews the paper's contributions, and discusses outstanding issues and directions for future research.

2 The Social Optimum Problem

The notation is given Table 1. In the basic bathtub model, a fixed number of identical commuters per unit area, $\underline{N}$, travel the same exogenous distance, $\underline{L}$, from home to work in the early morning rush hour over a dense network of homogeneous city streets. The technology of traffic congestion is described by a function relating traffic velocity, v, at a point in time to traffic density, k at that point in time: $v = v(k)$. When combined with the fundamental identity of traffic flow theory that flow equals density times velocity, the velocity function is assumed to be such that it yields a macroscopic fundamental diagram (MFD) relating traffic flow to traffic density with an inverted U-shape, in accordance with Godfrey (1969) and subsequent empirical analysis. Maximum or capacity flow occurs at capacity density. With traffic density below capacity density, traffic flow increases with traffic density; in this region, traffic is congested. With traffic density above capacity density, traffic flow decreases with traffic density; in this region of density, traffic is hypercongested. For reasons that will be explained shortly, "distance into the rush hour", m, is used as the running variable rather than t, clock time.

The first commuter to leave home/depart in the morning rush hour does so at $m = 0$. Since the distance of her trip is L, her trip runs from distance $m = 0$ to distance $m = L$, so that she arrives at work at $m = L$. The commuter who departs at $m = L$ arrives at $m = 2L$, and so on. M is the distance at which the last commuter arrives at work, so that $M - L$ is the distance at which the last commuter departs for work. We term this form of congestion "dynamic MFD congestion".

A commuter's travel utility is a function of her departure time and the duration of her trip in terms of time. Letting $t(m)$ denote the departure time of the commuter who departs from home at distance m, and $T(m)$ denote her trip duration, the travel utility of a commuter who departs at distance m is $U(t(m), T(m))$. In social surplus analysis, it is assumed that "\$ = \$ = \$" – that a dollar is valued equally by society from whomever it is taken or to whomever it is given.[3] Where $c(m)$ is the money expenditure on a commuter at distance m, social surplus

[3]Most of the welfare analysis done in transportation economics is social surplus analysis. There is an

a	:	arrival time	c	:	consumption
D	:	index of departure mass	$e(m)$	:	entry rate as a function of departure distance/location
$\widehat{e}(t)$	:	entry rate as a function of departure time	$E(m)$	:	cumulative entries as a function of distance/location
g	:	parameter in the generalized Greenshields' Relation	h	:	index in the approximation grid; see Figure 3
I	:	maximum possible number of departure masses	i, j	:	indices
k_j	:	jam density	$k(m)$	:	traffic density at distance/location m
$\widehat{k}(t)$	:	traffic density at clock time t	L	:	distance of a trip
$\mathscr{L}$	:	the Lagrangian in the optimization problem	m	:	departure distance/location
M	:	distance of the morning rush hour	$\mathcal{M}$	:	set of points in the approximation grid
N, $\underline{N}$	:	Population, exogenous population	q	:	flow ($\equiv v\,k$)
SW	:	social welfare	t	:	depending on context clock time and departure time
$\underline{t}$, $\overline{t}$	:	time of first departure, time of last departure	t^*, $t^\#$	:	desired arrival time, pivotal arrival time
$\widetilde{t}$	:	time of last arrival	$t(m)$	:	departure time as a function of departure distance/location
T_f	:	trip duration at free-flow velocity	$T(m)$	:	trip duration as a function of departure location
$\widehat{T}(t)$	:	trip duration as a function of departure time	u, $\underline{u}$	:	utility level, exogenous utility level
u_E, u_X	:	entry subutility, exit subutility	$U(t,T)$	:	travel utility as a function of departure time and trip duration
v_f	:	free-flow velocity	$v(k)$	:	velocity as a function of traffic density
$x(m)$	:	exit rate as a function of distance/location	$\widehat{x}(t)$	:	exit rate as a function of clock time
$X(m)$	:	cumulative exits as a function of distance/location	α	:	unit cost of travel time
β	:	unit cost of time early	γ	:	unit cost of time late
λ	:	Lagrange multiplier on population constraint	$\tau(m)$	:	travel time per unit distance ($= 1/v(k(m))$)

Table 1: Notational glossary.

.

analysis assumes that the total utility of a commuter at m is $U(t(m), T(m)) + c(m)$, so that $U(t(m), T(m))$ is a dollar measure of the travel utility of a commuter at m, which depends on the cardinalization of the utility function. This entails the assumptions that a commuter's total utility is additively separable between travel utility and consumption utility and that the marginal utility of consumption is unity, as well as a particular cardinalization of travel utility. Where $e(m)$ is the entry rate function of commuters per unit area who start their journey at m (which is the entry rate per unit area at m), aggregate total utility, A, is

$$A = \int_0^{M-L} e(m) \left[U(t(m), T(m)) + c(m) \right] \mathrm{d}m \tag{1}$$

The social planner chooses $e(m)$ to maximize aggregate total utility.[4] It is furthermore assumed that the resources available per unit area for consumption are independent of the allocation of commuters over the rush hour, so that $\int_0^{M-L} e(m)c(m)\,\mathrm{d}m$ is a constant. Thus, the entry rate function that maximizes (1) also maximizes aggregate travel utility, which is here termed social welfare, SW:

$$\mathrm{SW} = \int_0^{M-L} e(m)\, U\left(t(m), T(m)\right) \mathrm{d}m \tag{2}$$

Different restrictions are placed on the form of the travel utility function in different sections of the paper.

In the social optimum, the planner chooses the entry rate function to maximize social welfare subject to the following constraints:

alternative and more general "welfarist" approach that does not make the restrictive assumption that \$ = \$ = \$. In this alternative approach, the social planner is assumed to maximize the social welfare function(al) $W(\{u(t(m), T(m), c(m))\})$, where $u(t(m), T(m), c(m))$ is the total utility of a commuter who departs at m and $\{\cdot\}$ denotes a function. This specification allows for the marginal social welfare of consumption/income of a commuter at m to depend on the commuter's total utility at m – informally, to place more weight on a dollar given to a commuter who is worse off. Arnott and Jinushi (2021) contrast the two approaches in an extended numerical example that uses the model of this paper but with two discrete commuters.

[4] At this stage, there are no specific restrictions on $e(m)$: it is nonnegative and sums up to the total size of the population (i.e. Eq. (3)). For integrals in (2) and (3) to be defined, we assume that $e(m)U(t(m), T(m))$ and $e(m)$ are Lebesgue integrable, respectively. In Sections 3.2 and 4.2 we describe and use algorithms that restrict entries to mass points, i.e. we approximate $e(m)$ by a sum of measures.

1. Population constraint

$$\int_0^{M-L} e(m)\,\mathrm{d}m = N \tag{3}$$

2. Congestion technology

   Where $\tau(m)$ is the travel time per unit distance at m,

$$\tau(m) = v(k(m))^{-1}. \tag{4}$$

3. Relationship between clock time and distance

$$t(m, \underline{t}) = \underline{t} + \int_0^m \tau(u)\,\mathrm{d}u \quad \text{so that} \quad t'(m) = \tau(m), \tag{5}$$

   where $\underline{t}$ is the time of the first departure.

4. Trip duration[5]

$$T(m) = t(m + L) - t(m) \quad \text{so that} \quad T'(m) = \tau(m + L) - \tau(m). \tag{6}$$

Let $x(m)$ denote the arrival rate at work at distance m, $X(m) = \int_0^m x(u)\,\mathrm{d}u$ denote the cumulative number of arrivals at work by distance m, $E(m) = \int_0^m e(u)\,\mathrm{d}u$ denote the cumulative number of departures from home by distance m, and $k(m)$ denote traffic density at distance m.

5. Conservation of commuters

$$k(m) = E(m) - X(m) \quad \text{so that} \quad k'(m) = e(m) - x(m). \tag{7}$$

[5]The literature on the macroscopic fundamental diagram (MFD) distinguishes between trip-based models and accumulation-based models. Trip-based models employ (6). Accumulation-based models, in contrast, make the Little's Law approximating assumption that the exit flow from traffic equals the density of traffic divided by the average time a vehicle would spend in traffic at the current speed, which is $L/v(m)$ (so that $x(m) = k(m) \div L/v(m)$).

6. FIFO condition

$$X(m) = E(m - L), \text{ so that } x(m) = e(m - L). \tag{8}$$

7. Boundary conditions and non-negativity constraints:[6]

$$\begin{cases} E(0) = 0 \\ E(M - L) = N \end{cases} \xrightarrow{\text{by (8)}} \begin{cases} X(L) = 0 \\ X(M) = N. \end{cases} \tag{9}$$

Thus, the social optimum problem is to choose $e(m)$, the entry rate function, so as to maximize social welfare (given by (2)) subject to constraints (3) through (9).

The delay that gives rise to the delay-differential equation structure of the maximization problem is the delay in distance between the distance at which a commuter exits the traffic stream and enters the traffic stream. With distance as the running variable, this delay is exogenous. Combining the conservation of cars with the FIFO condition gives $k(m) = E(m) - X(m) = E(m) - E(m - L)$. The density of cars on the road at distance m equals the number of commuters who have entered the road up to distance m minus the number of commuters who have exited from it by distance m, which is also the number of commuters who have entered the road up to distance $m - L$. Differentiating this equation gives $k'(m) = e(m) - e(m - L)$. The change in the density of commuters between distances m and $m + \mathrm{d}m$ equals the number of commuters who enter the traffic stream over that distance interval minus the number who exit from it, which is the number of commuters who entered the traffic stream between distance $m - L$ and $m - L + \mathrm{d}m$.

In social surplus analysis, a commuter's travel utility is measured in dollar terms. Her travel utility may then be interpreted as the negative of her trip cost. Noting this, the sum of travel utilities over the population equals minus the sum of their trip costs over the

[6]AB proves that, conditional on the departure set being connected and the density function being continuous, the conditions of the equilibrium problem imply the non-negativity of entry rates, and hence of the exit, density, and cumulative arrival and departure functions. We conjecture that an analogous result holds for the social optimum problem.

population. Thus, maximizing social welfare, the sum of travel utilities, is equivalent to minimizing the sum of trip costs.

3 Smooth and Strictly Concave Utility Functions

Section 3.1 briefly reviews the solution method developed by AB to solve for equilibrium with smooth and strictly concave utility functions, presents a sample solution, and discusses whether any of the insights it generates into the mathematical structure of the problem carry over to the social optimum problem. Section 3.2 develops a strategy to solve numerically for the social optimum with smooth and strictly concave utility functions. Section 3.3 applies it to several numerical examples, and section 3.4 compares the numerical solutions for the optimum with those of the corresponding equilibrium for a particular example.

3.1 The Algorithm of Arnott and Buli (2018) to Solve for Equilibrium

This paper uses "distance into the rush hour", m, as the running variable. AB in contrast uses clock time, t, as the running variable. In what follows, where there is possible ambiguity, we shall write functions of m without a hat and functions of t with a hat; thus, for example, $e(m)$ denotes the entry rate as function of distance and $\widehat{e}(t)$ the entry rate as a function of clock time.

The trip timing equilibrium condition is that utility is constant over the departure set and is no higher everywhere outside the departure set. With clock time as the running variable, over the departure set therefore,[7] $U(t, \widehat{T}(t)) = \underline{u}$, where $\underline{u}$ is the equilibrium utility level. With the assumed restrictions on the form of the utility function, this function can be inverted to give the trip duration function $\widehat{T}(t; \underline{u})$, which gives the trip duration with departure at time t consistent with utility $\underline{u}$. With clock time as the running variable, the

[7]The equilibrium solution depends on ordinal properties of the utility function, which are preserved under a monotonic transformation. In contrast, as noted earlier, the solution to the social optimum problem entails a particular cardinalization of the utility function.

distance condition is that $\int_t^{t+T} v(\widehat{k}(u))\,\mathrm{d}u = L$; a commuter's trip distance equals the integral of velocity from her departure at t to her arrival at $t+T$. Inserting the trip duration function into the distance condition gives $\int_t^{t+T(t;\underline{u})} v(\widehat{k}(u))\,\mathrm{d}u = L$, which combines the trip timing equilibrium condition and the distance condition. Given the forms of the utility function and the velocity function, the numerical problem is to solve for an equilibrium utility level, the departure set, and an entry rate function over the departure set that are consistent with the trip timing equilibrium condition, the distance condition for the exogenous trip distance $\underline{L}$, the FIFO condition, the population constraint for the exogenous population $\underline{N}$, and non-negativity constraints.

AB proceeds under the assumptions that the departure set is connected and that the velocity function is continuous.[8] The algorithm it develops has three loops. Taking as given the time of the first departure, $\underline{t}$, and the utility level, $\underline{u}$, the innermost loop solves for the entry rate function consistent with the trip timing equilibrium condition with $\underline{u}$, the FIFO condition, and the condition that all commuters have the same trip distance. In general, the solution implies an L that is different from $\underline{L}$ and an N that is different from $\underline{N}$. In the middle and outer loops, $\underline{u}$ and $\underline{t}$ are adjusted so as to satisfy the exogenous trip distance and population using a gradient method. The innermost loop is the non-standard component of the solution algorithm.

The innermost loop takes as fixed trip duration as a function of t and $\underline{u}$, via $\widehat{T}(t;\underline{u})$, and also the start of the rush hour, $\underline{t}$. From this it derives the time of the last departure, the traffic density function and hence the velocity function over the rush hour, and from these the entry rate function over the departure interval and the exit rate function over the arrival interval. To reduce the paper's length, here we only sketch the logic.

The innermost loop starts with the condition $\int_t^{t+\widehat{T}(t;\underline{u})} v(\widehat{k}(u))\,\mathrm{d}u = L$. When this condition holds throughout the departure interval, all commuters receive the same utility level, $\underline{u}$,

[8] Buli (2019) provides incomplete proofs of these assumptions. It also draws on the mathematical insights obtained from the solution method in AB to derive analytical properties of equilibrium with smooth and strictly concave utility functions.

and travel the same distance. First, differentiate this condition with respect to t, yielding

$$v(\widehat{k}(t+\widehat{T}(t;\underline{u})))(1+\dot{\widehat{T}}(t;\underline{u}))-v(\widehat{k}(t))=0. \tag{10}$$

AB refer to this as the *velocity condition*. Consider two commuters, the first of whom departs at t, the second of whom departs $\mathrm{d}t$ after the first. They travel most of their trips together, but during the initial interval $[t, t+\mathrm{d}t]$, the first commuter travels a distance $v(\widehat{k}(t))\,\mathrm{d}t$ that the second commuter does not. To satisfy the equal trip distance requirement, the second commuter must make up this trip distance at the end of her trip. To receive the utility level $\underline{u}$, the first commuter has a trip duration of $\widehat{T}(t;\underline{u})$ while the second commuter has trip duration $\widehat{T}(t+\mathrm{d}t;\underline{u}) = \widehat{T}(t;\underline{u}) + \dot{\widehat{T}}(t;\underline{u})\,\mathrm{d}t$. Thus, the second commuter travels for a period of time $\mathrm{d}t + \dot{\widehat{T}}(t;\underline{u})\,\mathrm{d}t = (1+\dot{\widehat{T}}(t;\underline{u}))\,\mathrm{d}t$ after the first commuter has completed her trip. Over this time interval she travels a distance $v(\widehat{k}(t+\widehat{T}(t;\underline{u})))(1+\dot{\widehat{T}}(t;\underline{u}))\,\mathrm{d}t$. For the two commuters to travel the same distance, it must therefore be the case that $v(\widehat{k}(t+\widehat{T}(t;\underline{u})))(1+\dot{\widehat{T}}(t;\underline{u}))\,\mathrm{d}t = v(\widehat{k}(t))\,\mathrm{d}t$.

The algorithm is formulated in terms of cycles. The first cycle begins when the first commuter departs at $\underline{t}$ and ends when she arrives at work, at $\underline{t}+\widehat{T}(\underline{t};\underline{u})$; the second cycle begins when the first commuter arrives at work, and ends when the commuter who departs at that time arrives at work; and so on. The beginning of each cycle is termed a breakpoint. In equilibrium there are I full cycles and one partial cycle at the end of the rush hour when the last commuter departs. At the start of the rush hour, velocity equals free-flow velocity. Eq. (10) then gives the equilibrium velocity at the start of the second cycle, and by recursion the equilibrium velocity at the start of all active cycles can be determined. Since commuters prefer more central departure times, to satisfy the trip timing condition velocity falls over the early morning rush hour as congestion intensifies and then rises in the late morning rush hour. Eventually a breakpoint is reached, say breakpoint J, at which the velocity required to satisfy (10) rises above free-flow velocity. From this the equilibrium number of full cycles conditional on $\underline{u}$ can be determined.

Differentiating (10) with respect to t gives

$$v'(\widehat{k}(t+\widehat{T}(t;\underline{u})))\dot{\widehat{k}}(t+\widehat{T}(t;\underline{u}))(1+\dot{\widehat{T}}(t;\underline{u}))^2 + v(\widehat{k}(t+\widehat{T}(t;\underline{u})))\,\ddot{\widehat{T}}(t,\underline{u}) - v'(\widehat{k}(t))\,\dot{\widehat{k}}(t) = 0, \quad (11)$$

which AB refer to as the *acceleration condition*. This equation implies that there is a discontinuous increase in the entry rate at $\underline{t}$.[9] The discontinuous increase in the entry rate at $\underline{t}$ generates a discontinuous increase in the exit rate at $\underline{t}+\widehat{T}(\underline{t};\underline{u})$, and to keep traffic density continuous, this in turn requires a discontinuous increase in the entry rate at $\underline{t}+\widehat{T}(\underline{t};\underline{u})$; and so on. Thus, there are discontinuous increases in the entry rates at all breakpoints.

To simplify the exposition, the congestion technology is assumed to be Greenshields' Relation: $v(\widehat{k}(t)) = v_f(1-\widehat{k}(t)/k_j)$ and units are chosen such that $v_f = 1$ and $k_j = 1$, so that Greenshields' Relation becomes $v(\widehat{k}(t)) = 1-\widehat{k}(t)$, and $v'(\widehat{k}(t)) = -1$. Since cycles are interlinked, the notation is modified. Let $t_I(t)$ be the time in cycle I corresponding to time t in the first cycle (so that a commuter who enters at time $t_I(t)$ exits at time $t_{I+1}(t) = t_I(t)+\widehat{T}\left(t_I(t);\underline{u}\right)$). Employing these changes, the above equation reduces to

$$-\dot{\widehat{k}}(t_{I+1}(t))\left(1+\dot{\widehat{T}}(t_I(t);\underline{u})\right)^2 + v\left(\widehat{k}(t_{I+1}(t))\right)\ddot{\widehat{T}}\left(t_I(t);\underline{u}\right) + \dot{\widehat{k}}(t_I(t)) = 0,$$

which is a delay-differential equation with a variable delay.

Letting $\widehat{e}(t_I(t))$ and $\widehat{x}(t_I(t))$ denote the entry and exit rates at $t_I(t)$, $\dot{\widehat{k}}(t_I(t)) = \widehat{e}(t_I(t)) - \widehat{x}(t_I(t))$. Furthermore, since $\widehat{x}(t_I(t)) = \widehat{e}(t_{I-1}(t))$, $\dot{\widehat{k}}(k_I(t)) = \widehat{e}(t_I(t)) - \widehat{e}(t_{I-1}(t))$. Using this

[9] The acceleration condition (11) reveals mathematically why the entry rate function is discontinuous at $t=\underline{t}$. Evaluate the limit of (11) as t falls to $\underline{t}$. Suppose that the entry rate function is not discontinuous at $\underline{t}$. Then the limit of $\dot{\widehat{k}}(t)$ as t falls to $\underline{t}$ goes to zero. Since $v>0$, $v'<0$, and $\ddot{\widehat{T}}<0$, in the limit as t falls to $\underline{t}$, (11) can be satisfied only if the limit of $\dot{\widehat{k}}(T(t;\underline{u}))$ as t falls to $\underline{t}$ is negative. In words, the first car to depart arrives when traffic density is decreasing. Since the entry rate at $\underline{t}$ is zero, so too must be the exit rate at $\underline{t}+T(\underline{t};\underline{u})$. But this implies that the entry rate at $\underline{t}+T(\underline{t};\underline{u})$ is negative, which is not physically possible.

relationship and substituting out $v(\widehat{k}(t_{I+1}(t)))$ using (10), this equation becomes

$$-\left[\widehat{e}(t_{I+1}(t))-\widehat{e}(t_I(t))\right]\left(1+\dot{\widehat{T}}\left(t_I(t);\underline{u}\right)\right)^2+$$
$$v\left(\widehat{k}(t_I(t))\right)\ddot{\widehat{T}}\left(t_I(t);\underline{u}\right)\left(1+\dot{\widehat{T}}\left(t_I(t);\underline{u}\right)\right)^{-1}+\left[\widehat{e}(t_I(t))-\widehat{e}(t_{I-1}(t))\right]=0 \quad (12)$$

This is a linear equation relating the entry rate at $t_I(t)$ to the entry rates at $t_{I-1}(t)$ and $t_{I+1}(t)$ as well as the exogenous magnitudes $\left(1+\dot{\widehat{T}}\left(t_I(t);\underline{u}\right)\right)^2$, $\ddot{\widehat{T}}\left(t_I(t);\underline{u}\right)$, and $\left(1+\dot{\widehat{T}}(t_I(t);\underline{u})\right)^{-1}$. Thus, the entry rate at a particular point in time is linked to the entry rates at the corresponding point in time in both the previous and subsequent entry cycles.

Now apply (12) at the time of the first departure, $t_1(\underline{t})$, on the assumption there is only a partial entry cycle with utility level $\underline{u}$. $\widehat{e}(t_2(\underline{t}))=0$ since there are no entries in the second cycle, and $\widehat{e}(t_0(\underline{t}))=0$ since there are no entries before the first departure. Furthermore, $v\left(\widehat{k}(t_1(\underline{t}))\right)=1$ since velocity equals free-flow velocity at the time of the first departure. Thus, (12) can be solved for the equilibrium entry rate at the time of the first departure, $t_1(\underline{t})$. Now time step forward in increments of $\mathrm{d}t$. At $\underline{t}+\mathrm{d}t$, $\widehat{k}(t_1(\underline{t}+\mathrm{d}t))$ equals the cumulative number of cars that have entered by time $t_1(\underline{t}+\mathrm{d}t)$, which is $\widehat{e}(t_1(\underline{t}))\,\mathrm{d}t$, minus the cumulative number of cars that have exited, which is zero since the first car to enter does not exit until $t_2(\underline{t})$. Applying Greenshields' Relation, velocity at $t_1(t+\mathrm{d}t)$ can be calculated, from which, using (10), the corresponding equilibrium exit velocity of that commuter can be determined. Continue time stepping forward. When the last commuter to depart arrives, traffic density is zero and free-flow velocity is 1. Thus, the exit velocity of the last commuter to depart is 1. This condition is used to determine the time of the last departure. This procedure generates the complete solution.

Now extend the analysis to the case where there is a single full cycle and a partial entry cycle. Then $\widehat{e}(t_1(\underline{t}))$ and $\widehat{e}(t_2(\underline{t}))$ are positive, while $\widehat{e}(t_0(\underline{t}))$ and $\widehat{e}(t_3(\underline{t}))$ are both zero. Thus, (12) generates a pair of equations (the first corresponding to $I=1$, the second to $I=2$) in the two unknowns $\widehat{e}(t_1(\underline{t}))$ and $\widehat{e}(t_2(\underline{t}))$, from which the complete solution can be generated by time stepping forward from $t_1(\underline{t})$ until the time of the last departure. The procedure may

be extended to an arbitrary number of full entry cycles followed by a partial entry cycle.

We have noted that there are discontinuities in the entry rates at what we have termed breakpoints. More properly, these should be called primary breakpoints. The problem exhibits a form of symmetry. There are two boundary conditions. The first is that velocity equals free-flow velocity at the time of the first departure, the second that velocity equals free-flow velocity at the time of the last arrival. The problem can also be solved by running backwards in time from the time of the last arrival. Corresponding to this symmetry, there are also cycles that run backwards in time, corresponding to which there are secondary breakpoints. The first secondary breakpoint is the time of the last arrival; the second secondary breakpoint is the time at which the last arrival departs, and so. Again corresponding to this symmetry, when time is run forward there is a discontinuous decrease in the exit rate at all secondary breakpoints. More precisely: At all primary breakpoints in the departure interval (the interval of times over which there are positive departures), there is a discontinuous increase in the entry rate; at all primary breakpoints in the arrival interval, there is a discontinuous increase in the exit rate; at all secondary breakpoints in the departure interval, there is a discontinuous decrease in the entry rate; and at all secondary breakpoints in the arrival interval, there is a discontinuous decrease in the exit rate.

Of particular note is that the solution algorithm is built around the exogenous trip duration function, $\widehat{T}(t;\underline{u})$, which is obtained from the trip timing equilibrium condition. AB chose to use time as the running variable since the trip duration function is much easier to express in terms of time than in terms of distance. This argument for using time as the running variable does not apply to the social optimum since the trip timing condition does not apply. Using distance as the running variable rather than time has the advantage that the distance lag between a commuter's departure and her arrival is exogenous. When distance is used as the running variable, the primary breakpoints occur at $m = 0, L, 2L, \ldots$. And where M denotes the distance of the last arrival, the secondary breakpoints are at $M, M - L, M - 2L, \ldots$.

Figure 1 displays equilibrium for a particular numerical example.[10] Panel A displays

[10]Units are commuters, miles, and hours, with $t = 0$ corresponding to the first feasible departure time and

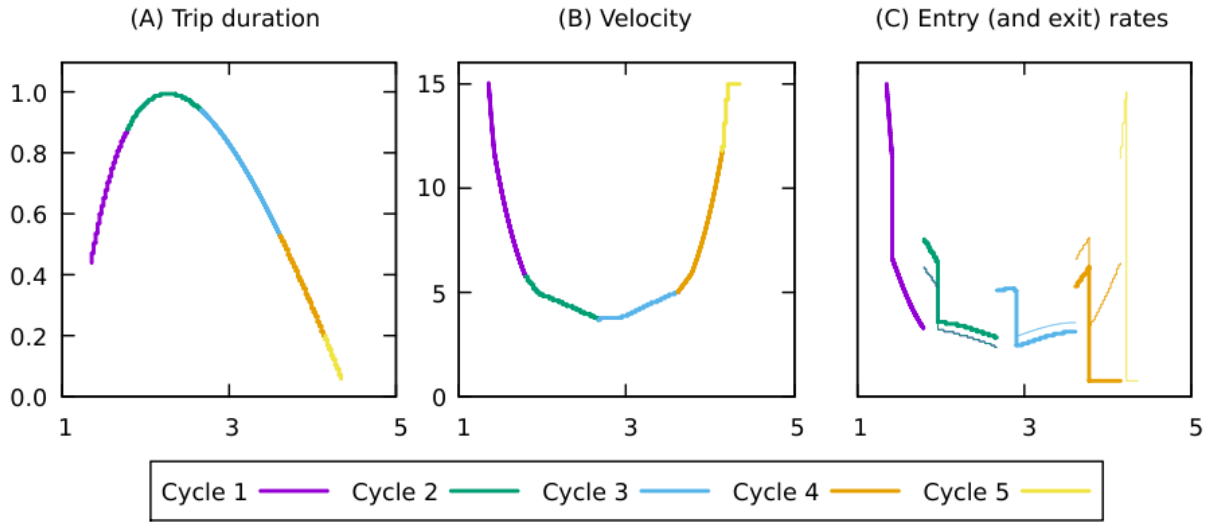


Figure 1: An equilibrium example from AB. On panel C, exit rates are given in thin lines. Units are hours and miles.

the trip duration function, $\widehat{T}(t)$, panel B the velocity function, $v(\widehat{k}(t)) = \widehat{v}(t)$, and panel C the entry and exit rate functions, $\widehat{e}(t)$ and $\widehat{x}(t)$. As expected, velocity falls over the early morning rush hour and then rises over the late morning rush hour, and trip duration has the opposite pattern. As discussed above, notice the discontinuous increases in the entry and exit rates at the primary breakpoints, as well as the discontinuous decreases in the entry and exit rates at the secondary breakpoints (which, as drawn, occur about one-quarter of the way through each cycle).

The rest of this subsection discusses whether the insights developed in AB and Buli (2019) into the mathematical structure of the equilibrium problem with smooth and strictly concave utility functions can be usefully applied to develop a customized algorithm to solve the social optimum problem and to derive analytical properties of the social optimum with smooth and strictly concave utility functions.

$t = t^{\#}$ to the latest feasible arrival time. The utility function has the form $U = r_0\, log(r_1\, t) + s_0\, \log(s_1\, (t^{\#} - (t + \widehat{T}(t))))$, and its parameters are $r_0 = 15$, $r_1 = 0.5$, $s_0 = 18$, $s_1 = 1$, and $t^{\#} = 6$. The congestion function is Greenshields' Relation: $v = v_f(1 - k/k_j)$, where v_f is free-flow velocity and k_j is jam density, and its parameters are $v_f = 15$ and $k_j = 10^6$. The trip distance is $\underline{L} = 6$ and the exogenous population is $N = 2.246 \times 10^6$.

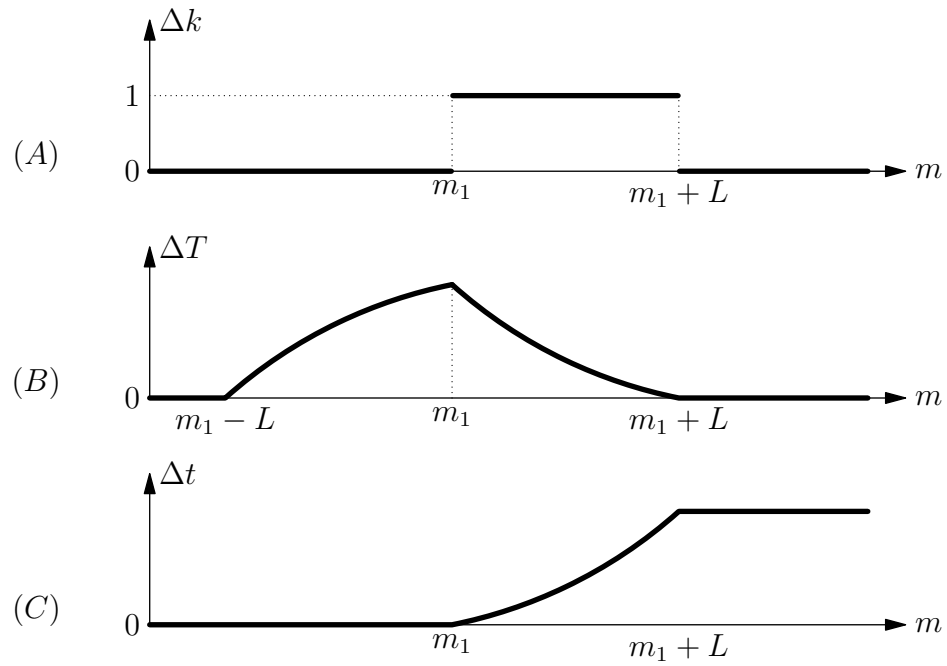


Figure 2: Impacts of an added commuter at $m = m_1$ on traffic density, trip duration and entry time.

Solving for an optimum differs generically from solving for an equilibrium. However, it often facilitates computation and generates additional insight to set up an optimization problem as an equilibrium problem or vice versa. One of the features of the solution to the social welfare optimization described in the previous section is that the increase in social welfare from a marginal increase in the number of commuters must be the same, at whatever distance m in the departure set the marginal commuter is inserted into the traffic stream. If it were otherwise, it would be possible to increase social welfare with a fixed population by changing some commuter's departure time. This observation suggests constructing a solution algorithm by replacing the trip timing condition of the equilibrium problem with the property of the optimum problem that the increase in social welfare from an adding a commuter must be the same at all locations in the departure set and no higher at locations outside the departure set. Via the Envelope Theorem, the increase in social welfare from adding a commuter at distance m is the same whether the entry rate function is held fixed at its pre-perturbation level or treated as endogenous.

Figure 2 graphically depicts the effects of adding a commuter to the traffic stream at distance m_1, without changing the entry rate at other distances. Panel A shows that she causes traffic density to increase by one commuter between the time she enters the traffic stream and the time she exits it at $m_1 + L$, but has no effect on traffic density outside the interval. Panel B shows that the increase in traffic density over the interval $[m_1, m_1 + L)$ increases the trip duration of all commuters who enter between $(m_1 - L, m_1 + L]$ and by more the longer the distance interval a commuter shares the streets with the added commuter. Since the entry rate at all other distances is held constant, the added commuter also causes commuters entering at distances greater that m_1 to delay (in terms of clock time) their entries onto the city streets, as shown in Panel C. Calculating the increase in social welfare from adding a commuter at distance m_1, $\mathrm{SW}(m_1)$, entails applying this perturbation to the social welfare function.

The procedure can be simplified by altering the perturbation. Instead of adding a commuter at distance m_1, move the commuter's departure distance a distance $\mathrm{d}m$ earlier or later. At the social optimum, this marginal change in the commuter's departure distance should have zero effect on social welfare. This perturbation has simpler effects but they nonetheless extend over intervals. For example, moving the departure time of a commuter from m_1 to $m_1 + \mathrm{d}m$: (i) decreases traffic density by one commuter between m_1 and $m_1 + \mathrm{d}m$, which reduces the trip duration of all commuters on the streets at that time, who are those who depart between $m_1 - L$ and m_1; (ii) increases traffic density by one commuter between $m_1 + L$ and $m_1 + L + \mathrm{d}m$, which adds to the trip duration of all commuters on the road at $m_1 + L$, who are those who depart between m_1 and $m_1 + L$. The perturbation also affects the departure time of all those who depart after m_1. Thus, even this very simple perturbation has a complicated effect on social welfare.

These thought experiments are discouraging. First, they suggest that reformulating the social optimum problem as an equilibrium problem using the property of the social optimum that marginal social welfare is the same at all distances in the departure interval is unlikely to be fruitful. As a result, the algorithm developed in the next subsection to solve

numerically for the social optimum does not exploit this property. Second, they suggest that deriving analytical properties[11] of the social optimum with smooth and strictly concave utility functions is likely to be very difficult. For example, how to proceed in identifying primitive conditions under which hypercongestion can occur at the social optimum with smooth and strictly concave utility functions is unclear.

In contrast, considerably more success was achieved in the case of α-β-γ tastes, the results for which are reported in Section 4.

It was noted earlier that social welfare, the sum of travel utilities over the population, equals the negative of the sum of trip costs over the population, so that maximizing the sum of travel utilities over the population generates the same allocation as minimizing the sum of trip costs. This interpretation of the social optimum problem permits the analysis to be cast in the terminology used in partial equilibrium analysis to describe externalities. Under this interpretation, the amount by which the addition of a commuter at distance m increases total trip cost is the marginal social cost of a trip at m. Accordingly, at the social optimum, the marginal social cost of a trip is the same for all departures in the departure interval. In the partial equilibrium theory of externalities, the marginal social cost of a trip can be decomposed into the marginal private cost (or user cost) of the trip and the marginal external (congestion) cost. Under this interpretation, the social optimum problem is so complicated because the addition of a commuter at distance m_1 generates such a complex pattern of externalities. The externalities operate through two channels, departure time and trip duration. Return to Figure 2. The addition of a commuter at m_1 changes the trip durations of all commuters who depart between $m_1 - L$ and $m_1 + L$, in the way displayed in Panel B of Figure 2. Consider a commuter at location m' in this interval. Through the trip duration channel, the addition of the commuter at m_1 increases her trip cost by $\Delta T(m'; m_1) U_T(t(m'), T(m'))$, where $\Delta T(m'; m_1)$ denotes the change in trip duration at m' induced by the addition of a commuter at m_1. Similarly, through the departure time

[11] The following conjectures concerning properties of the social optimum with smooth and strictly concave utility functions remain unproved: (*i*) The departure set is connected. (*ii*) The entry function is continuous except at primary and second breakpoints. (*iii*) The density and velocity functions are continuous. (*iv*) Utility as a function of departure location has an inverted U-shape.

channel, her trip cost increases by $\Delta t(m'; m_1)U_t(t(m'), T(m'))$, where $\Delta t(m'; m_1)$, the change in departure time at m' induced by the addition of a commuter at m_1, is displayed in Panel C of Figure 2. The marginal external cost of a trip at m_1 is therefore the integral over all distances in the departure interval of the external trip duration costs and external departure time costs.

3.2 The Algorithm to Solve for the Social Optimum

Because the properties of the social optimum with smooth and strictly concave utility functions conjectured in fn. [11] remain unproved, the algorithm presented here is more "off the rack" than the algorithm developed by AB to solve numerically for equilibrium with smooth and strictly concave utility functions. In one respect, however, the algorithm does draw on one of these unproved conjectures. Drawing on the conjecture that discontinuities in the departure and arrival rates occur only at primary and secondary breakpoints, it structures its discretization of time around the primary breakpoints and secondary breakpoints, which were defined above.[12]

The algorithm we employ uses discretization.[13] A basic step is the construction of an approximation grid for variable m. We first describe this step and then give a description of the algorithm. Entries and exits occur only on the grid. Without loss of generality, we assume $m_1 = 0$ (this origin can be shifted when the utility function is not defined in zero, e.g. for the logarithmic utility formulation). When a point m_i belongs to the grid, it could be an entry point, so point $m_i + L$, which is the corresponding exit point, should be in the grid too. Point m_i could also be an exit point, and in this case point $m_i - L$, which is the

[12] A referee suggested that this discretization may give the false impression that there are discontinuities in the entry and exit rates at the breakpoints when, if the conjecture is not true, there are not. This point will be addressed in the discussion of Figure 6.

[13] Restricting entries to occur at discrete points is somewhat a strong assumption on $e(m)$. This choice is mainly motivated by computational objectives and as we explain in the numerical examples it yielded satisfactory and robust results. Other schemes are possible, like constant entry rates over given intervals (we initially tried this option but it yielded slow convergence) or polynomial approximation. Since we lack a full analytical characterization of the optimum solution, using approximation leaves open an issue that we leave for future research: prove (theoretically) that the solution of the approximate problem converges to the solution of the original problem.

corresponding entry point, should belong to the grid too. Thus, when m_i belongs to the grid, points $m_i - L$ and $m_i + L$ should also belong to the grid.[14] The grid includes all primary and secondary breakpoints because they satisfy this requirement and discontinuities in entry and exit rates are expected to occur there. Indeed, as illustrated by the bottom line in Figure 3, the simplest grid is limited to primary and secondary breakpoints. As is standard with discretization a finer grid is expected to yield a better approximation. In our case, other points can be added and any construction is acceptable as long as it meets the entry and exit requirement described above. The top line in Figure 3 provides an example of a finer approximation where four evenly spaced points were added between each pair of successive primary breakpoints. From this graphical illustration, it is clear that there exists a positive integer h such that when an entry occurs at point m_i, for $i = 1, \ldots, n$, the corresponding exit occurs at point m_{i+h}: the exogenous delay in the continuous problem is now expressed as a lag in the indices of the approximation grid. This feature is particularly useful to the formulation of the optimization problem below.

Once the approximation grid is constructed, we obtain a set of points, $\mathcal{M} = \{m_1, \ldots, m_n, \ldots, m_{n+h}\}$, where all entries and exits occur. The first entry occurs at m_1 and the last entry at m_n. The first exit occurs at point m_{1+h} and the last exit at point m_{n+h}.[16] The computation of the utility for each agent is straightforward. Traffic density between any points m_i and m_{i+1} is constant. Given entries e_i for $i = 1, \ldots, n$, the densities are given by $k_i = \sum_{j=\max(1,i-h+1)}^{\min(i,n)} e_j$ for $i = 1, \ldots, n+h-1$. Travel speed in the same interval is $v_i = v(k_i)$, and the time to run from m_i to m_{i+1}, for $i = 1, \ldots, n+h-1$, is $\Delta t_i = (m_{i+1} - m_i)/v_i$. We then compute clock time t_i at each approximation point as $t_{i+1} = t_i + \Delta t_i$ with the starting point $t_1 = \underline{t}$. Each agent enters at a given time t_i, for $i = 1, \ldots, n$, and exits at t_{i+h}. His trip duration is $T_i = t_{i+h} - t_i$, and his utility level is $U_i = u(t_i, T_i)$. Social welfare is computed by summing over all the population, i.e. $\text{SW} = \sum_{i=1}^{n} e_i\, u_i$. The objective is then to maximize SW subject to population constraint $\sum_{i=1}^{n} e_i = N$ and nonegativity of decision variables, i.e.

[14] But, when $m_i < L$ (resp. $m_i > M - L$) only $m_i + L$ (resp. $m_i - L$) should be in the grid.

[15] Notice that other procedures can be used to add new points.

[16] Thus, n is the number of points where entries (or exits) occur, and h is the number of entry (resp. exit) points without exits (resp. entries).

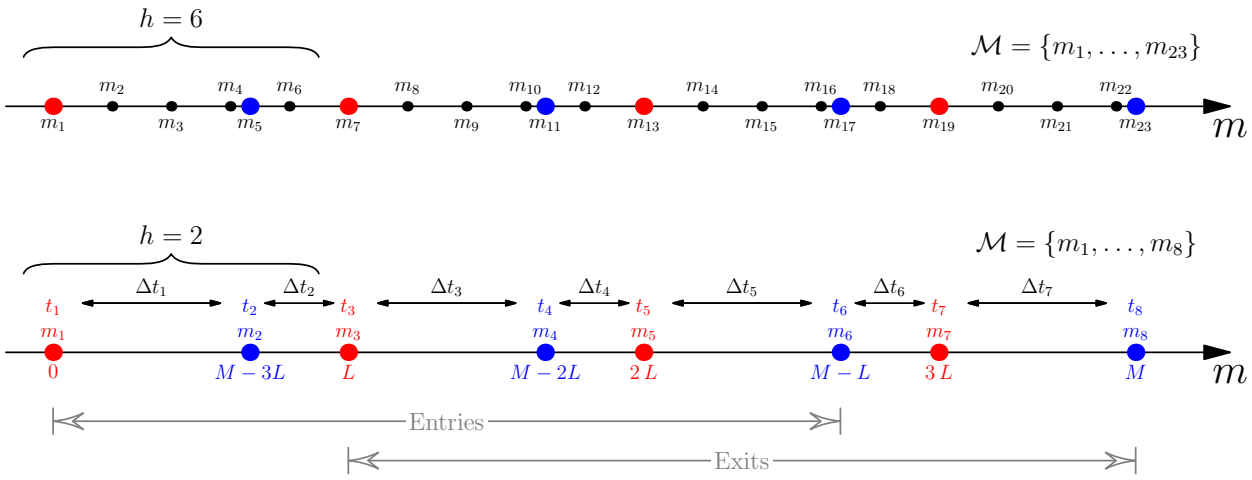


Figure 3: Approximation grid and complementary notation. The grid in the bottom line has only primary (red) and secondary (blue) breakpoints. In the top line, a finer grid is obtained by adding four evenly spaced points within each of the intervals $(0, L)$, $(L, 2L)$ and $(2L, 3L)$; then m_{14}, m_{15} and m_{16} are shifted to the right by distance L to obtain m_{20}, m_{21} and m_{22}.[15] It is clear how to obtain the value of h so that a mass entering at point m_i exits at point m_{i+h}.

$e_i \geq 0$ for $i = 1, \ldots, n$.

The algorithm itself has three loops with an objective that can be stated as

$$\max_{\underline{t},\, M} \left\{ \max_{\mathcal{M}} \left\{ \max_{m_1,\ldots,m_n} \quad \text{SW} \right\} \right\}.$$

The innermost loop uses an optimizer to solve for the social optimum, conditional on the number of entry points over the rush hour, as well as the start of the rush hour, $\underline{t}$, and the distance of the rush hour, M. The optimization problem is the same as that described in Section 2, but with distance being divided into intervals rather than being treated as continuous. The central loop successively refines the estimate of the optimum by increasing the number of entry points, but still holding constant the start and the distance of the rush hour. The loop is terminated when a satisfactory degree of convergence is reached. The outermost loop first determines the gradient of the sum of utilities with respect to $\underline{t}$ and M jointly and then updates the guess of the socially optimum $\underline{t}$ and M. The loop is terminated

when a satisfactory degree of convergence is reached. As is standard with such algorithms, the computational efficiency depends on the step size that is employed in the central loop (with a larger step size decreasing stability), and on the how the tolerance is adjusted in the process of convergence to equilibrium. The implementation of this algorithm involves the coding of the objective function, and depending on the solver to be used, the gradient and Hessian of the Lagrangian among other materials.[17] The detailed description of these steps is given in Appendix B.

3.3 Social Optima for Several Numerical Examples

The examples here are based on utility functions employed in AB, a logarithmic utility function and an exponential utility function. The logarithmic utility function is

$$u(t,T) = r_0 \log\left(r_1\, t\right) + s_0 \log\left(s_1\, (t^{\#} - t - T)\right), \tag{13}$$

where r_0, r_1, s_0 and s_1 are positive parameters. This form of the utility function is consistent with the Vickrey (1973) utility maximization formulation, rather than the Vickrey (1969) cost minimization formulation. The first term in (13) is the entry subutility, which increases in t as the commuters prefer to delay their departure from home. The second term is the exit subutility and reflects the preference of early arrival at work. Utility maximization is obtained as a trade-off between these two parts. Since the terms in the logarithm should be positive, the first entry must occur after time zero and exit time is necessarily smaller than $t^{\#}$. The second expression for a smooth and strictly concave utility function has an exponential form and is given by

$$u(t,T) = \frac{A_0}{a_1}\left(1 - e^{-a_1\, t}\right) + \frac{B_0}{b_1}\left(1 - e^{-b_1\,(t^{\#} - t - T)}\right), \tag{14}$$

[17] All the source code used in this paper is available at `https://gogs.univ-littoral.fr/mkilani/optBathtub`.

where A_0, a_1, B_0 and b_1 are positive parameters. As in the former case the utility is the sum of entry and exit subutilities, but the exponential formulation in (14) is more flexible insofar as entries can occur before time zero and last exits can occur after $t^{\#}$. Several properties of these formulations are discussed in AB, which showed that traffic equilibrium can entail hypercongestion under both formulations. We use their parameter values for the utility functions and trip characteristics and consider several formulations relating velocity to traffic density. As we report below, hypercongestion is absent in the examples we have considered. For the numerical examples where we are interested in the comparison between the optimum and the equilibrium (AB paper), we have used Greenshields' Relation, i.e. $v(k) = v_f(1-k/k_j)$. But, since this expression is frequently viewed as unrealistic, we have run some computations with the generalized Greenshields' Relation, i.e. $v(k) = v_f(1-(k/k_j)^g)$, with $g > 0$, and we have also used the relationship $v(k) = v_f(1-k/k_j)/(1+3k/k_j)$, which yields maximum traffic flow at one third of jam capacity, and velocity at capacity flow equal to one third free-flow velocity.

For the logarithmic utility function we report in Figure 4 the output of several cases corresponding to the three formulations of velocity mentioned above and two population sizes, $N = 16.6$ and $N = 22.5$. During the first cycle of the rush hour, the cumulative entry curve coincides with the traffic density curve since there are no exits. For each one of the three formulations of velocity, we see that when the number of commuters increases, the distance of the rush hour increases, and the traffic density curves move upwards; the first departure occurs earlier and the last arrival occurs later. The maximum value reached by traffic density increases with the population size, but, in all cases we check that it remains smaller than the value of traffic density where traffic flow is at the maximum. For Greenshields' Relation this value is half jam capacity, i.e. $k_j/2 = 5$, for the generalized Greenshields' Relation it is $k_j/(1+g)^{1/g} < k_j/2$, for $0 < g < 1$, and for $v(k) = v_f(1-k/k_j)/(1+3k/k_j)$ it is $k_j/3$. This confirms the absence of hypercongestion for the six cases reported here and all the other experiments we have conducted with utility functions (13) and (14). Notice that dashed curves (corresponding to the generalized Greenshields' Relation) are

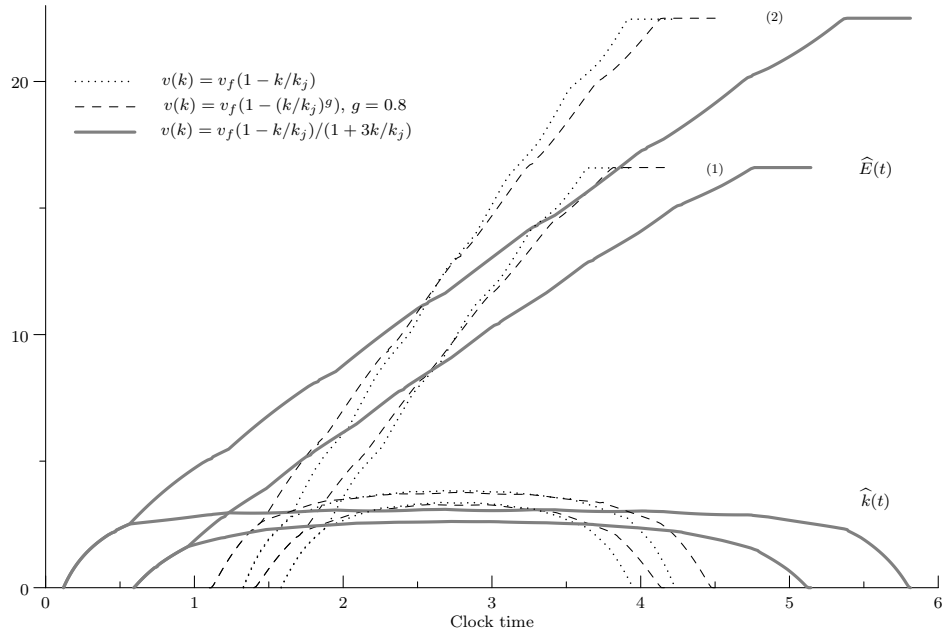


Figure 4: Cumulative entries and traffic density for six cases with the logarithmic utility function: three formulations of velocity with respect to traffic density and two population sizes, a moderate one, $N = 16.6$ and corresponding to curves (1), and a larger one, $N = 22.5$ and corresponding to curves (2). Parameter values used in all examples in this section: $k_j = 10$, $v_f = 15$, $L = 4$ and $t^{\#} = 6$. For utility function (13) we use $r_0 = 15$, $r_1 = 1/2$, $s_0 = 18$ and $s_1 = 1$.

slight deviations from the dotted curves (Greenshields' Relation). This can be seen as a sensitivity analysis with respect to the congestion technology; when the threshold for the occurrence of hypercongestion is smaller, the rush hour is longer and the maximum traffic density is smaller. With a technology where the maximum traffic flow occurs at a smaller value, as in the case of the third relation considered here, and for the same population, the rush hour is longer with higher congestion, but still no hypercongestion occurs. Apart from these quantitative observations, notice that the optimum solution exhibits a similar qualitative structure for all the reported cases.

To give a closer picture of a specific, though representative, case we provide a more detailed presentation for an optimum corresponding to the exponential utility function given in (14). As we have noticed above, with this expression the rush hour can expand before time $t = 0$ and after time $t = t^{\#}$. A large population size is considered to illustrate this point:

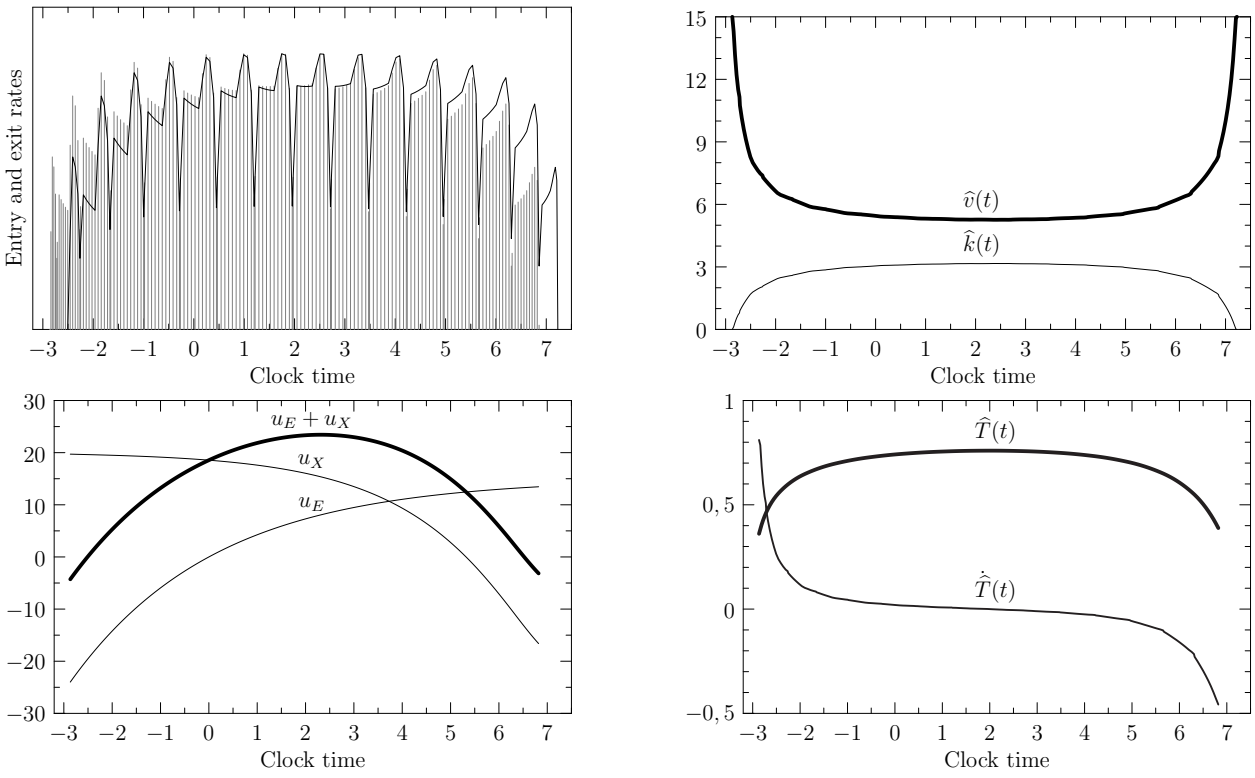


Figure 5: Output for the exponential utility function and a large population $N = 40.0$, yielding an average utility $\overline{u} = 14.9474$. For the model parameters see Figure 4, and for parameters in the utility function (14) we use $A_0 = 5$, $a_1 = 1/3$, $B_0 = 10$ and $b_1 = 1/2$. Travel speed is given by $v(k) = v_f(1 - k/k_j)/(1 + 3k/k_j)$.

$N = 40$. The other model parameters are those used in the previous examples (Figure 4) and the parameters of the utility function are $A_0 = 5$, $a_1 = 1/3$, $B_0 = 10$ and $b_1 = 1/2$. The main variables corresponding to the optimum are illustrated in Figure 5.

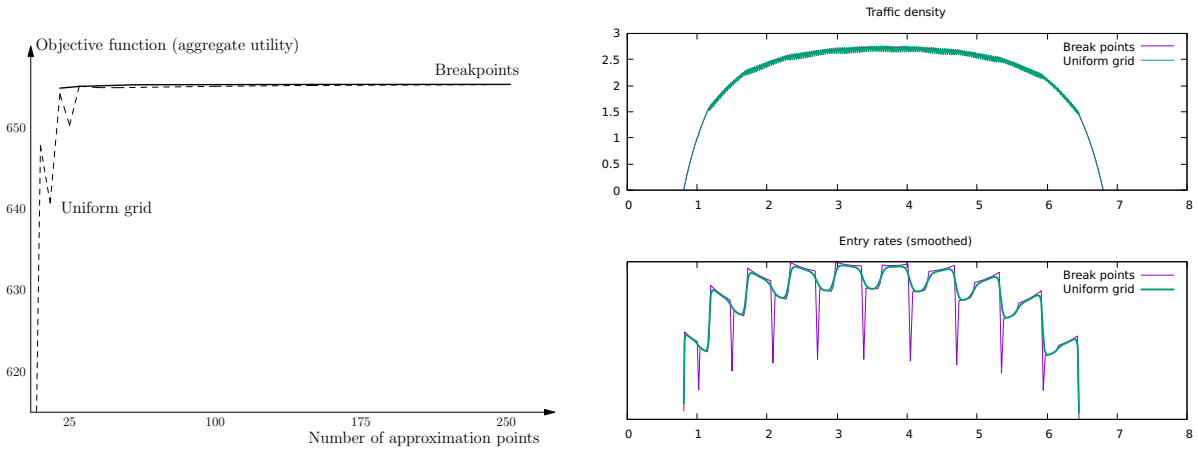
Entry and exit rates are given in the upper left panel. The rush hour starts at $\underline{t} = -2.84 < 0$ and ends at $\tilde{t} = 7.22 > 6.0 = t^{\#}$, so the rush hour is 10.06 hours long with sixteen entry cycles. The upper right panel shows velocity and traffic density as functions of clock time. The minimum travel speed is equal to 5.26 mph for a corresponding maximum value of traffic density equal to $3.16 < 3.33 = k_j/3$, i.e. traffic is highly congested but not hypercongested. Traffic density is inverted U-shaped and traffic speed is U-shaped. Entries into the bathtub, even though they feature discontinuities, produce smooth evolution of traffic density, velocity and traffic flows. It is important to notice that the discontinuities

in entry and exit rates occur at the breakpoints we have used in the approximation grid. Indeed, and as we discuss below, these points turn out to be very important to obtain fast and monotonic convergence to the optimum solution. The functions $T(t)$ and $\dot{T}(t)$ are given in the lower right hand panel. During a large part of the rush hour, trip duration does not vary much since the derivative of $\dot{T}(t)$ is small in absolute value. Early and late departures benefit from lower traffic densities and incur smaller travel times. Groups departing in the middle of the rush hour (the peak) have higher travel time but get higher utility because they have smaller penalties from departing too early or arriving too late (the shoulders). This can be seen in the lower left panel which reports utility levels. The first term in the utility function, (14), is denoted u_E (for "entry subutility") and the second term is denoted u_X (for "exit subutility"). The average utility level is 14.947 and the difference between the highest and lowest utility levels is equal to 27.719. The commuters departing at the beginning or the end of the rush hour are particularly penalized. The marginal social cost is the sum of the user cost and the marginal external cost.[18] At the optimum, the marginal social cost is equal for all departure times and the marginal external cost is higher for the peak. It follows that the user cost is smaller at the peak than in the shoulders.

The solution procedure we are using here is based on the approximation grid described in Figure 3, which is based on breakpoints given in the lower panel of that figure. Then, one may ask how a simpler approximation grid, based on equidistant approximation points, would perform by comparison to our scheme. Also, it is useful to evaluate the convergence rate of a given solution procedure with respect to the number of approximation points. To discuss these issues we have considered an example with the logarithmic utility function and the same parameter values used before and $N = 30$. For each approximation scheme, the optimum solution is computed for several numbers of approximation points. Our results show a clear benefit from using the breakpoints to construct the approximation grid.

The output reported in Figure 6 is useful to illustrate this point. On the left panel (Fig. 6a), we report the values of the objective functions with respect to the number of approximation points. Both methods get close to the optimum with less than one hundred

[18] In the next section we show how to relate utility levels to user costs.



(a) Convergence to the optimum

(b) Traffic density and entry rates

Figure 6: Comparison with a simple grid of equidistant approximation points.

approximation points, but clearly the method based on the breakpoints is much faster. The error in this case is relatively smaller even when only the breakpoints are used (like the grid in the bottom of Figure 3). The convergence with the uniform grid is not monotonic. Indeed, a uniform grid with approximation points that are close to the breakpoints provides a better solution than a uniform grid with a few more approximation points that are more distant from the breakpoints. Still, with more than 50 approximation points, the uniform grid provides a good approximation of the objective function (social welfare). However, increasing the number of approximation points with the uniform grid will fail to converge to the true solution as is shown on the right panel (Fig. 6b). The optimum solution involves a discontinuity in entry rates at the breakpoints (bottom panel in Fig. 6b). The uniform grid fails to capture this dynamic, instead exhibiting persisting oscillations in traffic density, velocity, and flow during the rush hour (top panel in Fig. 6b). These features were observed for several other examples and our experiments show a clear benefit of using the algorithm described in section 3.2.

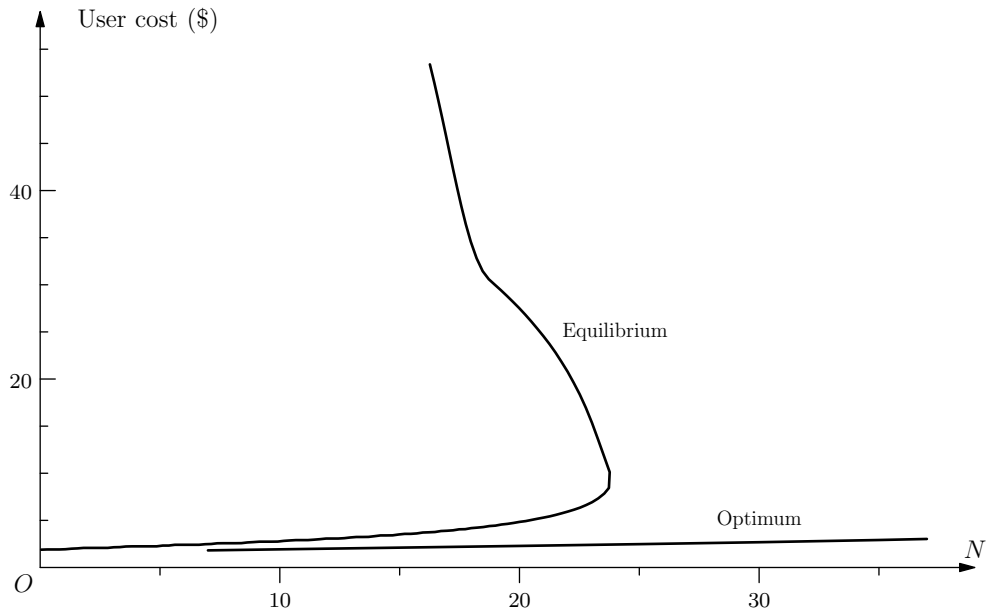


Figure 7: Average user cost for the equilibrium and the optimum as a function of the population size: the case of the logarithmic utility function. Parameter values are those given in Figure 4.

3.4 Comparison of the Optimum and the Equilibrium

For the optimum solutions reported above, entry rates exhibit discontinuities at primary and secondary breakpoints. Comparable discontinuities in entry rates are observed in equilibrium (cf. AB). Examples of equilibrium with hypecongestion usually show a large entry rate at the beginning of the rush hour and then many fewer entries during the rest of the period. We did not find any optimum configuration where entries are comparably high in the beginning of the rush hour. With a small population, the equilibrium may not be hypercongested but, from the examples in AB, entry rates remain higher at the beginning of the rush hour.

For the comparison between the equilibrium and the optimum, it may be interesting to notice that utility levels and user costs can be related.[19] Define $\nu = \max_t U(t, 0)$ to be maximal travel utility. It is the travel utility of a commuter with a zero trip duration and departure at the utility-maximizing time conditional on zero trip duration. Then define the user cost of a trip with trip duration T and departure time t as $\nu - U(t, T)$. It therefore

[19]For a more detailed presentation see AB.

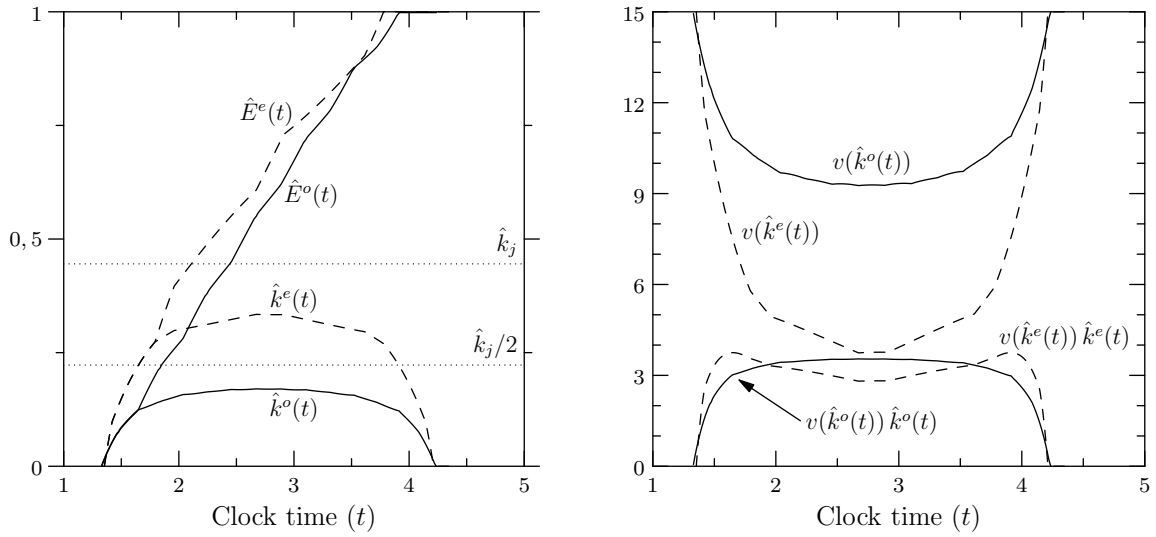


Figure 8: Comparison of the social optimum and the equilibrium (aggregate congestion) a population size $N = 22.46$. The logarithmic utility function is used and the other parameter values are those given in Figure 4. Traffic density and cumulative entries are normalized. Superscripts "e"and "o", respectively, refer to equilibrium and optimum.

equals maximal utility minus actual utility. Using these definitions, we can then draw the curve relating user cost to population in equilibrium and the curve relating average user cost to population in the social optimum.

One of the main findings in AB for the equilibrium solution is that the user cost curve can be backward bending. When this occurs, up to a critical level of population there are two equilibria, while above this population level no equilibrium exists. This is illustrated in Figure 7 for the logarithmic utility function. Equilibria on the increasing part of the curve, which are stable, are referred to as equilibria with "aggregate congestion". Equilibria lying on the decreasing part of the curve, which are unstable and correspond to a relatively higher user cost, are referred to as equilibria with "aggregate hypercongestion". In Figure 7 we add the average user cost corresponding to the optimum. As expected, this curve is below the equilibrium user cost curve, increasing but at a much slower rate.

We now compare the dynamics for two specific cases. We use here Greenshields' Relation since AB produced equilibria with this formulation. The logarithmic utility function is considered in both cases for which we characterize the optimum. The optimum is compared

to the equilibrium with aggregate congestion in the first case, and then compared to the equilibrium with aggregate hypercongestion in the second case. Figure 8 shows the structure of the optimum and the equilibrium for the first case. The left panel provides a comparison of the traffic densities and cumulative entries, and the right panel plots the velocity and traffic flow as a function of clock time t. The rush hour is slightly longer in the optimum. It starts at $\underline{t} = 1.32$ (instead of 1.35), the last entry occurs at $\bar{t} = 3.91907$ (instead of 3.78) and the last arrival occurs at $\tilde{t} = 4.2376$ (instead of 4.21). The main difference between the equilibrium and the optimum is that, at the beginning of the rush hour, the entry rate is considerably higher in the equilibrium than in the optimum. In the right panel, and by comparing the traffic velocity and flow curves, we can conclude that traffic is hypercongested when both curves are increasing or both are decreasing.[20] For the equilibrium, this is the case during a large part of the rush hour, yielding longer travel times and lower utility levels. Notice that rush hour durations are almost equal for the equilibrium and the optimum. The rush hour duration is equal to the population size divided by the average inflow (or average outflow). Since trip distance is fixed, this implies that average traffic flow is more or less the same in the equilibrium and the optimum, which is consistent with different average velocities only if the optimum traffic flow is on average congested, while the equilibrium traffic flow is on average hypercongested.

Figure 9 reports the same variables for the second case. The population size is decreased to $N = 16.56$, and we consider the equilibrium with aggregate hypercongestion. For the optimum, the rush hour is significantly shorter, traffic density has the usual inverted U-shape and remains below $k_j/2$, i.e. no hypercongestion. By comparison to the first case, the average utility increases from 22.4 to 23.1. The equilibrium, however, entails severe hypercongestion and the difference between equilibrium and optimum is strong and spreads over the long rush hour period, which starts very soon, at $t = 0.0787$ and ends shortly before $t^{\#}$. Travel speed quickly decreases below 1 mph, driving traffic flow down too.

[20] Taking the derivative of $v(\widehat{k}(t))\,\widehat{k}(t)$ with respect to t, we obtain $(v'\,k + v)\,k'$ after omitting arguments. If traffic flow and traffic speed are both decreasing, then $(v'\,k + v) < 0$ (since $k' > 0$), which correspond to hypercongested traffic. When both traffic density and traffic speed are increasing a similar argument proves that traffic is hypercongested.

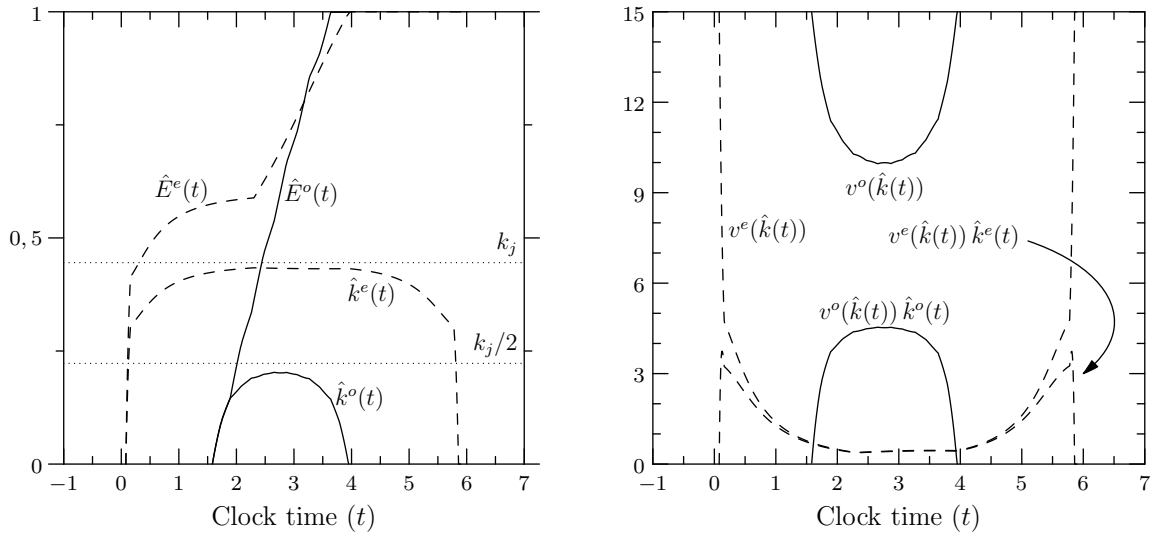


Figure 9: Comparison of the social optimum and the equilibrium (aggregate hypercongestion) with a population size $N = 16.56$. The logarithmic utility function is used and the other parameter values are those given in Figure 4. Traffic density and cumulative entries are normalized. Superscripts "e"and "o", respectively, refer to equilibrium and optimum.

4 Social Optimum with α-β-γ tastes

With α-β-γ tastes, the utility function takes the form:

$$\widehat{U}(t,T(t)) = \begin{cases} -\alpha T(t) - \beta\left(t^* - (t+T(t))\right), & \text{for early arrivals} \\ -\alpha T(t) - \gamma\left(-t^* + (t+T(t))\right), & \text{for late arrivals.} \end{cases} \tag{15}$$

This section examines properties of the social optimum in the model of this paper with α-β-γ tastes. It takes over where (Arnott et al. 2016, AKN hereafter) left off. AKN focused on equilibrium for a special case: identical commuters, α-β tastes (no late arrivals) and MFD congestion with Greenshields' Relation. It provided a closed-form solution for equilibrium and examined solution properties. Section 6 of that paper provided a preliminary examination of the social optimum for the special case. Since it was unsuccessful in finding a closed-form solution with more than two departure masses, it restricted analysis to the case where the population is sufficiently small that the social optimum has only one or two

departure masses.

This section goes beyond AKN's analysis of the social optimum in three major respects. First, it imposes no restrictions on the population; second, it works with a general functional form relating velocity and density; and third, it focuses on qualitative properties of the social optimum.

AKN considered only restricted equilibria, in which the pattern of departures is a succession of departure masses, with a mass departing as soon as the previous departure mass arrives. The paper conjectured that equilibria can take only this form but did not prove it. Analogously, in this section we consider only restricted social optima, which have the same qualitative departure pattern. In particular: (*i*) The travel interval is connected; and (*ii*) All entries occur in entry masses at primary breakpoints, which implies that all exit masses occur at primary breakpoints. All the analysis is predicated on the conjecture that restricted social optima are the global optima. In what follows, we refer to restricted social optima simply as social optima.

4.1 Properties of the Social Optimum

In the social optimum, the marginal social cost of trips are equalized. How should the marginal social cost of a trip be calculated? Think of the social optimum in the basic bottleneck model, in which the entry rate equals the flow capacity of the bottleneck over the connected departure interval, with the timing of the rush hour such that the time early cost of the first commuter to depart equals the time late cost of the last commuter to depart. If one were to measure the marginal social cost of a commuter at time t as the increase in total social costs from adding a commuter at that time, *holding fixed the departure times of all other commuters*, evaluated at the social optimum, one would obtain the wrong answer. Adding a commuter at the start of the rush hour would generate a queue of length one for the duration of the rush hour, whereas adding a commuter at the end of the rush hour would generate no queue. Since the time early cost of the first commuter equals the time late cost of the last commuter, one would incorrectly calculate the marginal social cost at the start

of the rush hour to be higher than the marginal social cost at the end of the rush hour. The error arises because the technology of congestion in the bottleneck model is not smooth, so that the Envelope Theorem does not apply.

A similar problem arises when calculating the marginal social cost of a trip in the basic bathtub model with α-β-γ tastes. We shall start off by arguing that the marginal social cost of a trip is correctly calculated by holding fixed the size of each commuter's departure mass and adjusting the timing of the departure masses such that a departure mass arrives at the desired arrival time. Having established this, the analysis proceeds routinely.

Theorem 1. *With α-β-γ tastes, if there is a social optimum for which the desired arrival time is not a primary breakpoint, there is another social optimum with the same total travel cost for which the desired arrival time is a primary breakpoint.*

Proof. Consider a social optimum for which the desired arrival time is not a primary breakpoint. Let N_e denote the number of commuters in this social optimum who arrive strictly early and N_l denote the number of commuters who arrive strictly late. Since there is no departure mass that arrives on time, $N_e + N_l = \underline{N}$, where $\underline{N}$ is the exogenous population. Delaying departures of all commuters by $\mathrm{d}t$ decreases the total time early cost of all early commuters by $\beta N_e \, \mathrm{d}t$, increases the total time late cost of all late commuters by γN_l, and has no effect on total travel time cost. Similarly, bringing forward departures of all commuters by $\mathrm{d}t$ increases the total time early costs by $\beta N_e \, \mathrm{d}t$, decreases the total time late cost of all late commuters by $\gamma N_l \, \mathrm{d}t$, and has no effect on total travel time cost. Thus, a necessary condition for a social optimum for which the desired arrival time is not a primary breakpoint is that $\beta N_e = \gamma N_l$. But when this condition holds, delaying departures of all commuters by a sufficient amount that the arrival of the latest early departure mass coincides with the desired arrival time has no effect on total travel cost. □

In what follows we restrict our attention to social optima for which the desired arrival time is a primary breakpoint.

Let $i = -I, \ldots, -1, 0, +1, \ldots, I$ denote departure masses, from the earliest to the latest, with some departure masses at the tails possibly being empty/inactive, and with $i = 0$

corresponding to the departure mass that exits immediately before the desired arrival time. The number of commuters in mass i is k_i; the trip duration of commuters in mass i is $T(k_i)$; the time early for commuters in mass i, $i = -I, \ldots, -1$ is $\sum_{j=i+1}^{0} T(k_j)$; the time early for commuters in mass 0 is zero; and the time late for commuters in mass i, $i = 1, \ldots, I$, is $\sum_{j=1}^{i} T(k_j)$. Total travel costs, TTC, are therefore

$$\text{TTC} = \alpha \sum_{i=-I}^{I} k_i T(k_i) + \beta \sum_{i=-I}^{-1} \left(k_i \sum_{j=i+1}^{0} T(k_j) \right) + \gamma \sum_{i=1}^{I} \left(k_i \sum_{j=1}^{i} T(k_j) \right). \tag{16}$$

Travel in mass i is congested if $\mathrm{d}\big(k_i\, v(k_i)\big) / \mathrm{d}k_i > 0 \Leftrightarrow v(k_i) + k_i\, v'(k_i) > 0$ and hypercongested if the inequalities are reversed. Since the trip duration for mass i is $T(k_i) = \underline{L}/v(k_i)$, traffic in mass i is congested if $T(k_i) - k_i\, T'(k_i) > 0$, or equivalently when the elasticity of trip duration with respect to density is less than one. Traffic in mass i is hypercongested when the last inequality is reversed. The size of the departure masses are chosen to minimize TTC subject to the population constraint that $\sum_{i=-I}^{I} k_i = N$ and non-negativity constraints on the k_i, $k_i \geq 0$. Traffic density should also be (strictly) lower than jam capacity, $k_i < k_j$, but this constraint can be ignored as long as we focus on an interior solution with positive traffic flow and velocity. Thus, the minimization problem is

$$\begin{cases} \min\limits_{k_{-I},\ldots,k_I} & \alpha \sum\limits_{i=-I}^{I} k_i T(k_i) + \beta \sum\limits_{i=-I}^{-1} \left(k_i \sum\limits_{j=i+1}^{0} T(k_j) \right) + \gamma \sum\limits_{i=1}^{I} k_i \left(\sum\limits_{j=1}^{i} T(k_j) \right) \\ \text{s.t.} & \sum\limits_{i=-I}^{I} k_i \geq \underline{N} \qquad \text{and} \qquad k_i \geq 0 \text{ for } i = -I, \ldots, 0, \ldots, I. \end{cases} \tag{17}$$

The corresponding Lagrangian is $\mathscr{L} = \alpha \sum_{i=-I}^{I} k_i T(k_i) + \beta \sum_{i=-I}^{-1} \left(k_i \sum_{j=i+1}^{0} T(k_j) \right) + \gamma \sum_{i=1}^{I} k_i \left(\sum_{j=1}^{i} T(k_j) \right) + \lambda \left(\underline{N} - \sum_{i=-I}^{I} k_i \right)$ where λ is the Lagrange multiplier cooresponding the population constraint. The necessary conditions state that (i) for each k_i, $i = -I, \ldots, I$, we have $\partial \mathscr{L} / \partial k_i \geq 0$, $k_i \geq 0$ and the complementary slackness (CS) condition, i.e. $k_i \cdot \partial \mathscr{L} / \partial k_i = 0$; and (ii) that $\underline{N} - \sum_{i=-I}^{I} k_i \leq 0$, $\lambda \geq 0$ and $\lambda \left(\underline{N} - \sum_{i=-I}^{I} k_i \right) = 0$. By differentiating the Lagrangian, we state the necessary condition for the three groups of

masses (early arrivals, on time arrival and late arrivals, respectively). For $i = -I, \ldots, -1$ we have

$$\alpha[k_i T'(k_i) + T(k_i)] + \beta \sum_{j=i+1}^{0} T(k_j) + \beta\, T'(k_i) \sum_{j=-I}^{i-1} k_j \geq \lambda, \tag{18a}$$

with $k_i \geq 0$ and CS condition. For $i = 0$, we have

$$\alpha[k_i T'(k_i) + T(k_i)] + \beta\, T'(k_i) \sum_{j=-I}^{-1} k_j \geq \lambda \tag{18b}$$

with $k_i \geq 0$ and CS condition. For $i = 1, \ldots, I$, we have

$$\alpha[k_i T'(k_i) + T(k_i)] + \gamma \sum_{j=1}^{i} T(k_j) + \gamma\, T'(k_i) \sum_{j=i}^{I} k_j \geq \lambda \tag{18c}$$

with $k_i \geq 0$ and CS condition. Equation (18a) identifies three elements of the marginal social cost of a commuter in early departure masses: the direct cost associated with her travel, $\alpha T(k_i) + \beta \sum_{j=i+1}^{0} T(k_j)$, the congestion externality cost she imposes on commuters in the same departure mass, and the schedule delay externality cost she imposes on commuters in earlier departure masses by causing them to depart earlier. Equation (18b) is the same as (18a), except that commuters in this mass arrive on time. Equation (18c) is the same as (18a), except that it applies to late departure masses, for which the schedule delay externality cost derives from the added commuter causing those to depart in her departure mass and later departure masses to arrive later.

Restricting the analysis to social optima for which the desired arrival time is a primary breakpoint makes the analysis easier. The first-order conditions for early and on time arrivals, (18a) and (18b), do not contain any terms related to the size of late departure masses, and the first-order conditions for late arrivals, (18c), do not contain any terms related to the size of early and on time departure masses. This points to a natural separability in the travel cost minimization problem between early/on time arrivals and late arrivals. In our analysis, we shall exploit this separability in two ways:

1. Obtain properties of the social optimum by: (i) Deriving relationships between the sizes of the early and on time departure masses based on their marginal social costs being the same; (ii) doing the same for late departure masses; and (iii) linking early/on time arrivals and late arrivals by imposing the condition that the marginal social cost of travel in the on time arrival mass is that same as that for the first late departure mass.

2. Imagine that the social optimum problem has been solved, including the division of the population between early/on time arrivals and late arrivals, as given. Taking the population of early and on time arrivals as fixed at their optimal levels, minimize the total travel costs of this sub-population, and then proceed analogously for late arrivals.

We first examine properties of the earliest active departure mass when $\alpha > \beta$. We start with early arrivals; Properties of late arrivals are examined in Appendix A.1. Let $-D$ denote the index of the earliest "active" departure mass, so that there are no departures for $i = -I, \ldots, -D-1$. At the optimum, the marginal social cost of departure in mass $-D-1$ must be greater than or equal to the marginal social cost of departure in mass $-D$, which implies that

$$\alpha[k_{-D-1}T'(k_{-D-1}) + T(k_{-D-1})] + \beta \sum_{-D}^{0} T(k_j) + \beta \sum_{j=-I}^{-D-2} k_j T'(k_{-D-1})$$
$$\geq \alpha[k_{-D}T'(k_{-D}) + T(k_{-D})] + \beta \sum_{-D+1}^{0} T(k_j) + \beta \sum_{j=-I}^{-D-1} k_j T'(k_{-D})$$

Now, $k_i = 0$ for $i = -I, \ldots, -D-1$. Also, $T(k_{-D-1}) = T_f$, where T_f is trip duration at free-flow travel speed. Using these results, the above inequality reduces to

$$\alpha T_f + \beta \sum_{j=-D}^{0} T(k_j) \geq \alpha[k_{-D}T'(k_{-D}) + T(k_{-D})] + \beta \sum_{j=-D+1}^{0} T(k_j).$$

Eliminating common terms, this inequality reduces to

$$0 \geq \alpha[k_{-D}T'(k_{-D}) + T(k_{-D}) - T_f] - \beta T(k_{-D}) = \\ \alpha[k_{-D}T'(k_{-D}) - T_f] + (\alpha - \beta)T(k_{-D}). \quad (19)$$

Adding a commuter to mass $-D-1$ rather than $-D$ increases total schedule delay cost by $\beta T(k_{-D})$ but decreases total social travel time costs by $\alpha[k_{-D}T'(k_{-D}) + T(k_{-D}) - T_f]$. At the optimum, the former must weakly exceed the latter.

Lemma 1. *With α-β-γ tastes, if $\alpha > \beta$, then in the social optimum the earliest active departure mass is congested rather than hypercongested.*

Proof. Since $\alpha > \beta$, (19) implies that $\alpha\big(k_{-D}T'(k_{-D}) - T_f\big) < 0$. From the definition of hypercongestion, if the earliest departure mass were hypercongested, then $k_{-D}T'(k_{-D}) > T(k_{-D}) > T_f$, establishing a contradiction. □

We now consider the relationship between congestion in contiguous early departure masses. The marginal social costs of departures in masses $-D$ to 0 are equal. For successive early, and active, departure masses, it follows that

$$\alpha\Big[k_iT'(k_i) + T(k_i)\Big] + \beta\sum_{j=i+1}^{0} T(k_j) + \beta\sum_{j=-I}^{i-1} k_jT'(k_i) \\ = \alpha\Big[k_{i+1}T'(k_{i+1}) + T(k_{i+1})\Big] + \beta\sum_{j=i+2}^{0} T(k_j) + \beta\sum_{j=-I}^{i} k_jT'(k_{i+1}) \quad (20)$$

Eliminating common terms yields

$$\alpha\Big[k_iT'(k_i) + T(k_i)\Big] + \beta T(k_{i+1}) + \beta\sum_{j=-I}^{i-1} k_jT'(k_i) \\ = \alpha\Big[k_{i+1}T'(k_{i+1}) + T(k_{i+1})\Big] + \beta\sum_{j=-I}^{i} k_jT'(k_{i+1}) \quad (21)$$

After some rearrangement, (21) reduces to

$$\Big[\alpha\big(T(k_{i+1}) - T(k_i)\big) - \beta T(k_{i+1})\Big] + \alpha\left[k_{i+1}T'(k_{i+1}) - k_i T'(k_i)\right] + \beta\big(T'(k_{i+1}) - T'(k_i)\big)\sum_{j=-I}^{i-1} k_j + \beta T'(k_{i+1})k_i = 0. \quad (22)$$

Consider transferring a commuter from mass i to mass $i+1$. This causes her private travel cost to change by $\alpha(T(k_{i+1}) - T(k_i)) - \beta T(k_{i+1})$, the travel time externality cost she imposes on other commuters by $\alpha(k_{i+1}T'(k_{i+1}) - k_i T'(k_i))$, the schedule delay externality cost she imposes on commuters in earlier ($j = -I$ to $i-1$) departure masses by $\beta\left(T'(k_{i+1}) - T'(k_i)\right)\sum_{j=-I}^{i-1} k_j$, and increases the schedule delay externality cost she imposes on commuters in departure mass i by $\beta T'(k_{i+1})k_i$.

For future reference, we rearrange (22) to give

$$(\alpha - \beta)\Big[\big(T(k_{i+1}) + k_{i+1}T'(k_{i+1})\big) - \big(T(k_i) + k_i T'(k_i)\big)\Big] + \beta\left[k_{i+1}T'(k_{i+1}) - T(k_i)\right] + \beta\left[\big(T'(k_{i+1}) - T'(k_i)\big)\sum_{j=-I}^{i} k_j\right] = 0. \quad (23)$$

Lemma 2. *With α-β-γ tastes, if $\alpha > \beta$, then all early departure masses are congested rather than hypercongested*

Proof. Suppose that mass i is congested and mass $i+1$ is hypercongested. Then each of the terms within square brackets on the LHS of (23) would be strictly positive, establishing a contradiction. Since mass $-D$ is congested rather than hypercongested, this result establishes that mass $-D+1$ must be congested rather than hypercongested, and by recursion that all early departure masses must be congested rather than hypercongested. □

Lemma 3. *With α-β-γ tastes, if $\alpha > \beta$, then in the social optimum later early departure masses are more congested.*

Proof. Suppose the contrary. Then it would have to be the case for some i that mass $i+1$ is less congested than mass i. Then each of the terms within square brackets (23) would be

strictly negative, establishing a contradiction (the middle term would be negative since, by Lemma 2, $T(k_i) > k_i\,T'(k_i)$ and $k_i\,T'(k_i) > k_{i+1}T'(k_{i+1})$). □

The results of Lemmas 1, 2 and 3 are brought together in

Proposition 1. *With α-β-γ tastes, if $\alpha > \beta$, then in the social optimum all early departure masses are congested rather than hypercongested, and, among early departure masses, later masses are more congested.*

The case $\alpha < \beta$ is more complex. To gain some insight into it, we consider an extreme and simple example; it is extreme in that $\alpha = 0$, and it is simple in that trip distance equals unity and the normalized Greenshields' Relation holds, so that $T(k) = (1-k)^{-1}$.

Recall that the full social optimum can be derived by solving separately for the early morning and late morning rush hours, taking as fixed the optimal division of the population between early/on time and late arrivals. Let n denote the optimal early/on time population. Consider how the optimum evolves at n increases. Up to $n = 1$, all commuters travel in the departure mass that arrives on time: $k_0 = n$. As n increases from zero, a commuter's trip duration increases from free-flow trip duration up to infinite trip duration, but this does not affect total travel time costs since the value of travel time equals zero. Total schedule delay costs are zero since all commuters arrive on time, so that marginal social cost is zero for $n \in (0, 1)$. As n rises above 1, since it is not possible to accommodate the entire early/on time population in a single departure mass, a second early departure mass is formed. Since commuters in mass 0 arrive on time, total time early costs are minimized by minimizing the total time early costs of commuters in mass -1. Reducing the number of commuters in mass 0 increases the number of commuters in mass -1 but also reduces the time early cost each experiences. Each commuter in mass -1 experiences time early equal to the travel time in mass zero, which is $(1-k_0)^{-1} = (1-(n-k_{-1}))^{-1}$. Thus, the total time early cost of commuters in mass -1 is $\beta k_{-1}(1-n+k_{-1})^{-1}$. The derivative of this expression with respect to k_{-1} is $-\beta(n-1)(1-n+k_{-1})^{-2} < 0$. Thus, the total time early cost of commuters in mass -1 is minimized by making k_{-1} being as large as possible, consistent with the constraints imposed by the congestion technology that departure masses cannot be negative and cannot

be greater than 1. Total time early cost of commuters in mass -1 is therefore minimized by setting $k_{-1} = 1$. Thus, at $n = 1^+$, all the population switches from mass 0 to mass -1, and as the population increases the increase goes to mass 0. As the population continues to rise, a critical population is reached at which it becomes optimal to form a third early departure mass, with mass -2 having a unit population, and the residual population being divided across masses 0 and -1 such that their marginal social costs are equalized. And so on. The phenomenon whereby the earliest departure mass has traffic density equal to jam density is termed extreme hypercongestion.

We highlight this result in

Proposition 2. *With α-β-γ tastes and with Greenshields' Relation, when $\alpha = 0$ and $n > 1$, the earliest departure mass exhibits extreme hypercongestion.*

Proposition 2 covers only an extreme case. When $\alpha > 0$, extreme hypercongestion does not occur in the social optimum since travel time costs are given weight in total trip costs. Nevertheless, hypercongestion does occur in the social optimum for a range of parameter values that are not unreasonable.[21]

We have also established

Proposition 3. *With α-β-γ tastes, in the social optimum all late departure masses are congested rather than hypercongested and, among late departure masses, earlier departure masses are more congested.*

The proof is provided in Appendix A.1, and follows a logic similar to that used to prove Proposition 1.

Proposition 4. *In the social optimum with α-β-γ tastes, with $\alpha > \beta$, the central departure mass is the most congested.*

The proof is provided in Appendix A.2, and follows a logic similar to that used to prove the other Propositions.

Drawing together the results of Propositions 1, 3 and 4 gives

[21] We omit analysis of the intermediate situation in which $\alpha \in (0, \beta)$ because of its complexity.

Proposition 5. *In the social optimum with α-β-γ tastes, with $\alpha > \beta$, all the departure masses are congested rather than hypercongested. In both the early and late morning rush hours, more central departure masses are more congested, with the central departure mass being the most congested.*

All of these results are predicated on the conjecture that, with α-β-γ tastes, the social optimum takes the form of a succession of contiguous departure masses, with the central departure mass arriving exactly on time. Now is a good time to consider the validity of this conjecture. With α-β-γ tastes, there is a kink point in the indifference curves at the desired arrival time, t^*, which generates a social benefit from having a mass of commuters arriving at the desired arrival time. This benefit does not arise if departures are continuous over the rush hour. But because the congestion technology is convex, there is also a social cost associated with having departures occur in contiguous departure masses rather than continuously. Is it not possible that the relative size of the social cost and social benefit depends on parameter values?

To address this question, the solution algorithm of Section 3, which does not assume that the social optimum takes the form of contiguous departure masses, was applied to solve numerically for the social optimum with α-β-γ tastes for a wide variety of parameter sets. In all the examples, the optimum entailed contiguous departure masses. Thus, it remains an open issue as to whether, with α-β-γ tastes, the restricted social optimum is always the global optimum.

4.2 A Numerical Method to Solve for the Social Optimum

As indicated earlier, Arnott et al. (2016) made some progress in solving analytically for the social optimum in the bathtub model with α-β-γ tastes and under Greenshields' Relation. They conjectured but did not succeed in proving that departures occur only in contiguous masses, each occurring at a primary breakpoint, with the exit of one departure mass followed immediately by the entry of another departure mass, and with the common desired arrival time being one of the breakpoints. If this conjecture is correct, then a considerably simpler

numerical method can be developed for this case than for the case of a smooth and strictly concave utility function.

The way we proceeded was to apply two solution methods. The first is the method of the previous section. The second is a method that builds on the conjecture. We first describe the logic underlying its construction, and then the algorithm.

Consider a velocity relationship with a jam density k_j. The optimum number of departure masses must therefore strictly exceed the integer portion of N/k_j. As well, we know from Proposition A2 that, if $\alpha > \beta$, hypercongestion does not occur in the social optimum, so that the density of cars on the street system never exceeds a given proportion ρ of jam capacity. The optimum number of masses must therefore strictly exceed $N/(\rho\, k_j)$. If there is only one departure mass, it must depart early and arrive exactly on time. Now consider adding a second departure mass. The added departure mass may either depart early and arrive at the time that the "central" departure mass departs, or depart immediately after the central mass arrives and arrive late. Solve for the optimum for both configurations, and choose the one that minimizes total costs. Proceed in an analogous way as more departure masses are added.

The issues then arise as to how an exogenous population of commuters should be allocated across a given configuration of departure masses so as to minimize total costs, and to when an additional departure mass should be added. We turn first to the former issue. The exogenous population should be allocated across a given configuration of departure masses so as to equalize the marginal social cost of an adding a commuter to each departure mass. How the marginal social cost of adding a commuter to each departure mass is calculated was shown in section 4.1.

Determining the optimal number of departure masses is an integer problem. Considering that the optimal number of departure masses is unlikely to be large, if the aim is to solve for the optimum for a particular N, the simplest numerical method is likely to be just straightforward comparison of total travel cost for the various configurations of early and late departure masses, with the minimum number of departure masses exceeding $N/(\rho\, k_j)$. An

obvious conjecture that we have not proved is that: Starting with the optimal configuration conditional on the number of departure masses equaling $N/(\rho\, k_j) + 1$, and adding departure masses one by one, if adding a departure mass on the early side of the morning rush hour increases total travel costs and if adding a departure mass on the late side of the morning rush hour increases total travel costs, then the optimum configuration has been found.

The algorithm is constructed in accordance with the above logic. Basically, one may consider two loops, an inner one and an outer one. The inner loop solves for the optimal allocation of commuters over departure masses conditional on the number of early and late departure masses. The outer loop solves for the optimal number and configuration of departure masses. But, it is possible to remove the outer loop and consider, instead, a relatively large value of I (the number of early and late arrival masses). By allocating optimally the population to equalize the marginal social costs among all used masses, i.e. minimizing the aggregate social cost, there will be no commuters in some masses that are on the edges (masses on both edges will not be active). Thus if the inner loop works efficiently, the outer loop can be disregarded. For the inner loop we may proceed directly by solving the nonlinear programming problem in (17). Doing so, however, does not exploit the properties of the problem discussed above. It is preferable to solve the system of nonlinear equations given in (18) instead. This requires particular care of the nonegativity constraints on decision variables k_i.

We proceed as follows. Let the total number of commuters $\underline{N}$ be given. Start with a very small number of commuters N. The optimal solution will be $k_0 = N$ and $k_i = 0$ for the all the other masses. The marginal social cost of mass $i = 0$ in this case is $\alpha[NT'(N) + T(N)]$ which should be smaller than the marginal social cost of mass -1 and mass 1. As the value of N increases, a population threshold, N_1, is reached at which it becomes optimal to add a second departure mass, either departure mass -1 or departure mass 1. Suppose that it becomes optimal to add departure mass -1. From (18a), the marginal social cost of mass -1 (when $k_{-1} = 0$) is $\alpha T(0) + \beta T(N)$. From (18c), the marginal social cost of mass 1 (when $k_1 = 0$) is $\alpha T(0) + \gamma T(0)$, which does not depend on N. Let N_1 be the value of N where

the marginal social costs of masses 0 and -1 or 0 and 1 are equal. For $N > N_1$ the optimum solution involves two masses at least, and N slightly higher than N_1 there are exactly two active masses. The optimum allocation is then found by solving a system of two nonlinear equations with two unknowns. For numerical requirement $k_0 = N$ and $k_{-1} = 0$ can be used as a starting point for the computation.

By increasing again the value of N we can compute the optimum allocation of commuters. At each step we can check that the solution remains optimal by comparing the marginal social cost of the two active masses -1 and 0 with those corresponding to masses -2 and mass 1. Nonactive masses should have higher marginal social costs; otherwise it is optimal to allocate commuters there. This shows how the next threshold N_2 can be found. The procedure can be continued as long as N is smaller than $\underline{N}$. For the numerical implementation there are several algorithms to solve the system of nonlinear equations. Since we have a good guess of a starting point, as mentioned above, and since the optimum problem usually yields a well behaved solution we have used a Newton algorithm which insures quadratic convergence. Our algorithm worked efficiently for the several examples we have tested.

This algorithm has an additional advantage: it can be quickly tuned to solve numerically for the equilibrium. The main required change is to replace the marginal social cost with the marginal private cost. But, since the equilibrium solutions involve higher congestion we had to increase N by a smaller amount compared to the optimum. Otherwise, the Newton procedure may not converge.

4.3 A Discussion of a Numerical Example

The computation of the optimum with α-β-γ preferences follows our description in Section 4.2. For the numerical illustration, travel time cost is set to $\alpha = 10$ \$/h, early arrival cost is set to $\beta = 8$ \$/h and late arrival cost to $\gamma = 15$ \$/h. Jam density is $k_j = 6$, free-flow velocity is $v_f = 15$ mph and travel distance is $L = 5$. These values are comparable to those used in Arnott et al. (2016). For velocity, we use $v_f(k) = v_f(1 - k/k_j)/(1 + 3k/k_j)$ which yields maximum traffic flow at one third of jam capacity.

Figure 10 illustrates the traffic densities and cumulative entries for the optimum and the equilibrium, both computed on the basis of a population size $N = 18$. The optimum is obtained with fourteen departure masses, eight arriving early and five arriving late. Since each mass enters immediately after the one before exits, the mass sizes can be observed and compared easily in the left panel that displays traffic densities. For the optimum, the size of the masses is increasing until the central one, the one that arrives on time, and then decreases at a higher speed since commuters prefer early to late arrivals. From the cumulative entries curve (right panel), we observe that about 60% of the commuters arrive early, 30% arrive late and the remaining 10% are in the central mass and arrive on time. The first mass departs four hours ($t = -5.084$) before the preferred arrival time (set to 0 in this example) and the latest arrival occurs at 2.802, so the rush hour is about six hours long. The travel speed, which is decreasing with traffic density, falls to 5.56 mph for the central mass.

The equilibrium exhibits a clearly distinct dynamic. The rush hour is about three times longer, starts at more than ten hours before the desired arrival time and ends more five hours after. Traffic density quickly increases to get close to jam capacity, significantly decreasing travel speed to less than 0.5 mph for the central mass. There are only six masses, two arriving early and three arriving late. This contrasts with the social optimum, in which there are more departure masses and a higher proportion of the population arrives early. At equilibrium, only 21% of the population arrive early and 50% arrive late. The central mass is overused with a very low travel speed and a long travel time. Thus, masses that depart earlier incur a high schedule penalty that can be reduced by departing later.

Changing the parameter values will change the magnitudes of the mass sizes, but the output will remain qualitatively similar. For instance, if for the same example the values of α and β are switched the number of masses decreases to twelve. This impact is intuitively clear since smaller α reduces the user cost of travel time relative to the penalty of arriving earlier.

For the optimum, the average user cost in this example is equal to \$ 25.13, composed of travel time cost (\$ 6.07), early arrival cost (\$ 12.54) and late arrival cost (\$ 6.52). The

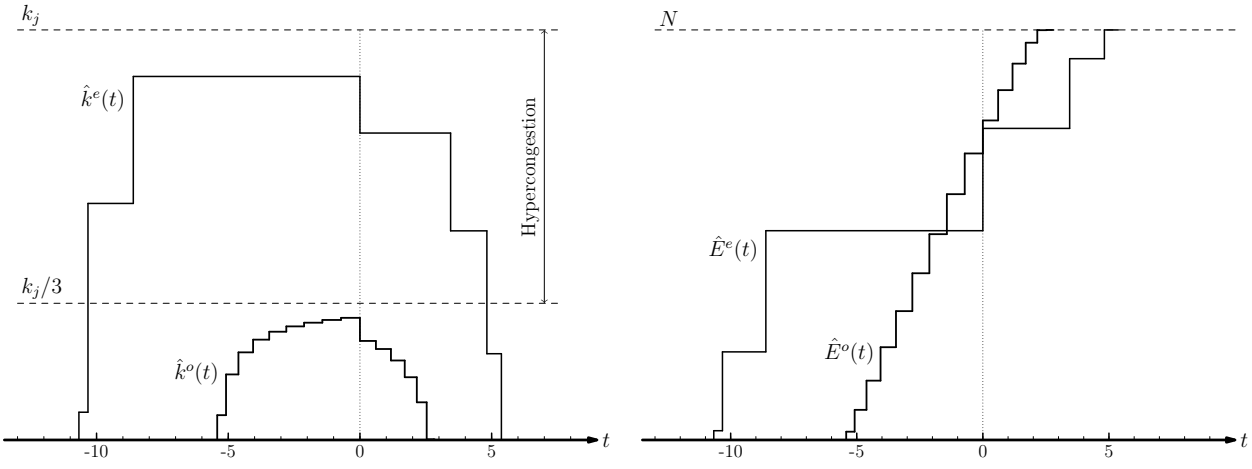


Figure 10: Traffic density (left panel) and cumulative entries (right panel) for the social optimum and the equilibrium with $N = 18$. Parameter values are $k_j = 6$, $v_f = 15$, $L = 4$ and $t^* = 0$; $\alpha = 10$, $\beta = 8$ and $\gamma = 15$. Velocity is given by $v(k) = v_f(1 - k/k_j)/(1 + 3k/k_j)$.

marginal social cost at the optimum is equal to \$ 44.76; and the difference from the average user cost is the average external cost. The proportion of early arrival cost may seem too high, but at the optimum the external costs also matter. In the case under consideration, early arrivals produce moderate external costs in comparison to late arrivals, and also in comparison to trip durations. At the optimum, the aggregate cost is minimized when there are more masses (and commuters) arriving early than masses arriving late. At equilibrium, equal user costs yield more commuters arriving on time. This increases external costs through an increase in early arrival times, for masses departing earlier, and an increase in the trip duration, for commuters in the same mass. To compensate for the larger schedule delay cost from arriving too early, more commuters choose masses arriving late. In this example, the user cost at equilibrium is \$ 86.05, composed of travel time cost (\$ 40.13), early arrival cost (\$ 15.10) and late arrival cost (\$ 30.81).

It is interesting to compare the equilibrium and the optimum for several population sizes. Average user cost and marginal social cost for the equilibrium and the optimum, as functions of the population size, are given in Figure 11. To construct these curves we start from a small value of the user cost and then find the corresponding population size. In the next step, we increase the user cost by a small value and find the corresponding population size.

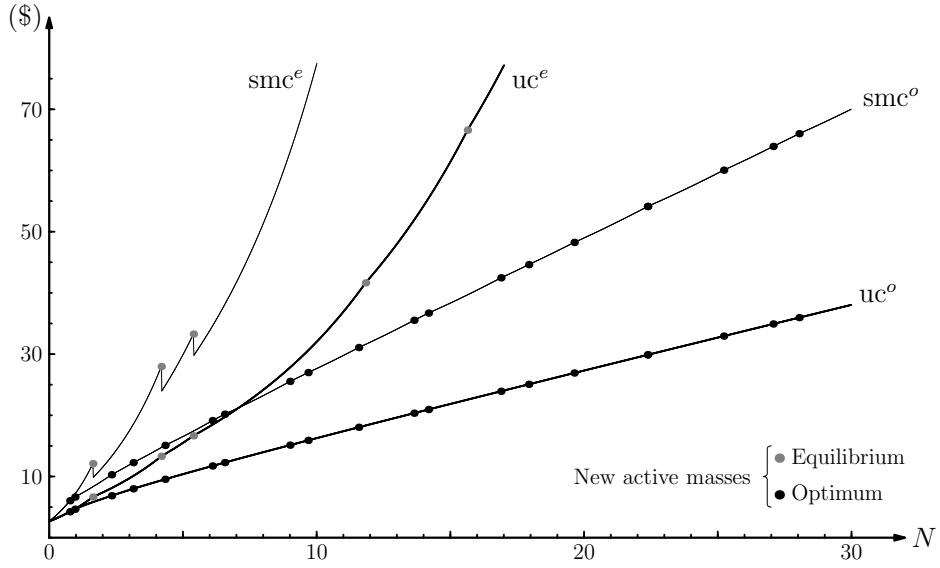


Figure 11: User cost (thick lines) and marginal social costs (thin lines) for the equilibrium (superscript "e") and the optimum (superscript "o"). The dots are located where new active masses are created. Parameter values are $k_j = 6$, $v_f = 15$, $L = 4$ and $t^* = 0$; $\alpha = 10$, $\beta = 8$ and $\gamma = 15$ (the same as in Fig. 10).

This iteration is repeated until a predefined limit size of the population is reached. It is clear that the user cost and the social marginal cost are significantly higher at the equilibrium. For very small values of N, in both the equilibrium and the optimum all commuters travel in the central departure mass, so that the equilibrium and optimum user cost are the same, as are the equilibrium and optimum marginal social cost. But, as N increases the curves corresponding to the equilibrium increase much faster than those corresponding to the optimum. The latter exhibit a rather linear form reflecting an optimal trade-off in the allocation of users between new masses (with high schedule delays) and central masses where traffic density never reaches the hypercongestion threshold ($k_j/3$ in the formulation of velocity used to develop this example). The user cost corresponding to the equilibrium has kinks located where new masses become active. This induces discontinuities in the social marginal cost curve. Shortly after a new mass becomes active, the social marginal cost decreases because the magnitude of the external cost (congestion/hypercongestion) is smaller. At the optimum, the number of new active masses increases faster than for the equilibrium. A second

mass becomes active at $N = 0.78$ for the optimum, while is occurs at $N = 1.64$ for the equilibrium. With $N = 30$, there are only 8 active masses in equilibrium, while the optimum number is 21. By creating a new mass, the user cost increases linearly, mainly through the schedule delay part since distinct masses do not overlap in traffic, while increasing the size of a given mass significantly and nonlinearly increases the user cost. Indeed, velocity drops near zero when traffic density gets close to jam capacity.

Notice that, in contrast to the smooth utility functions used in Section 3.3, the user cost corresponding to the equilibrium is monotonically increasing and is not backward bending. For each population size we obtain a unique equilibrium.[22] Indeed, with α-β-γ tastes the rush hour distance is not bounded above, as it is with logarithmic utility function, and an equilibrium exists for any population size.

We examined several other examples with different velocity functions, and different parameters for each of the functions. With smooth velocity functions, our algorithms worked well and generated the same qualitative results. With non-smooth velocity functions, such as those associated with the triangular or trapezoidal fundamental diagrams, our algorithm did not work. However, with the triangular or trapezoidal fundamental diagrams, putting aside the complication that the number of masses is integer, the social optima can be derived from first principles. In both cases, the social optimum entails all cars traveling at free-flow velocity, with the density of departure masses chosen to maximize flow consistent with this velocity. Total trip cost equals the sum of total travel time cost and total schedule delay cost. Total travel time cost is minimized when all commuters travel at free-flow velocity. Total schedule delay cost is minimized when flow is at its maximum, and, with both early and late departures, when the time early cost of the first commuter to depart equals the time late cost of the last commuter to depart. With the triangular of trapezoidal fundamental diagrams both minimums can be achieved simultaneously.

[22]To check this output we increased the population size above one thousand and the user cost curve kept the same structure.

5 Concluding Remarks

This paper undertook a preliminary investigation of the social optimum in the basic bathtub model (with endogenous trip timing). A fixed population N of *ex ante* identical commuters per unit area must travel a fixed distance on a dense network of streets from home to work over the morning rush hour. The common travel utility function is $U(t, T)$, where t is departure time and T is trip duration. Traffic congestion is described by a dynamic MFD, generated by assuming that traffic velocity at a point in time is inversely proportional to traffic density at that point in time. The social optimum problem is to choose the departure rate function to maximize social welfare, the sum of money-metric travel utilities, per unit area, or equivalently to minimize the total social costs of travel per unit area.

This problem is of interest for several reasons.

1. The standard treatment of the social optimum with endogenous trip timing uses William Vickrey's bottleneck model (1969). The bathtub model with endogenous trip timing improves on the bottleneck model by incorporating space, allowing for trips of different distances, and accommodating hypercongestion – travel on the backward-bending portion of the dynamic MFD.

2. The bathtub model provides a stepping stone towards more realistic models of the spatial dynamics of traffic congestion in metropolitan areas, for example of the equilibrium dynamics of traffic congestion along a corridor joining residential and workplace locations.

3. Enriched with realistic detail, such models would provide the conceptual basis for more sophisticated congestion pricing schemes, and, where there are constraints on congestion pricing, for the design of second-best mass transit and land use policies to mitigate traffic congestion.

Unfortunately, even the basic bathtub model with endogenous trip timing gives rise to delay-differential equations that make formal analysis difficult. The literature has responded

by considering approximations, focusing on special cases, and undertaking simulations. Approximations are more persuasive when the exact solution is known, either through formal analysis or through simulation. This paper undertook some preliminary formal analysis and developed solution algorithms that, in the limit, as the time steps shrink, give exact solutions.

The paper focused on two classes of tastes. In the first, commuter travel utility is a smooth and strictly concave function of departure time and trip duration. By exploiting some mathematical properties of the problem and using distance into the rush hour rather than time as the running variable, the algorithm developed to solve for social optima worked well. The numerical examples considered employed two smooth and strictly concave travel utility functions, (13) and (14), which derive from the specification of scheduling preferences in Vickrey (1973), in which the commuter trades off the conflicting desires to leave home later and to arrive at work earlier. In all the examples, hypercongestion does not arise and traffic density and speed evolve continuously over the rush hour but with discontinuities in the entry and exit rates at breakpoints. Unfortunately, little progress was made in analytical derivation of properties of the social optimum. Since the congestion externalities imposed by the each commuter are so spatially diffuse, it is difficult to exploit the social optimum property that the marginal social cost of all commuters is the same.

In the second class of tastes, commuter travel utility is represented by α-β-γ tastes, which are often assumed in the bottleneck model. On the assumption that the social optimum entails contiguous departure masses, good progress was made in analytical derivation of the properties of the social optimum. If $\alpha > \beta$: Hypercongestion never occurs. Since one of the departure masses arrives on time, most properties of the early morning rush hour may be derived independently of the late morning rush hour, and vice versa. In the early morning rush hour, later departure masses are more congested, and in the late morning rush hour, later departure masses are less congested, with the departure mass that arrives on time being the most congested. Utility has an inverse U-shape over the rush hour, with the commuters in the on time departure mass having the highest utility. An even simpler solution algorithm was developed for this case, which also worked well. If $\alpha < \beta$, the spatial dynamics can

become complex and hypercongestion can occur.

The differences in the numerical examples for the social optimum with smooth and strictly concave utility functions compared to those with α-β-γ tastes are stark. What account for them?

We do not have a complete explanation. We do, however have two conjectures.

1. In the previous section it was argued that, with α-β-γ tastes, the social benefit from having contiguous departure masses derives from the kink point in the schedule delay cost at t^*, the desired arrival time. Since strictly concave utility functions do not exhibit this kink point, this benefit is absent. This leads to the conjecture that a departure pattern with only contiguous departure masses does not occur with strictly concave utility functions.

2. In the previous section, it was proved that, with α-β-γ tastes, a necessary condition for hypercongestion to occur at the beginning of the rush hour is that $\beta > \alpha$. Now turn to (14). Making the transformation of variables, $a = t + T$, it can be written as

$$\hat{u}(a,T) = \frac{A_0}{a_1}\left(1 - e^{-a_1(a-T)}\right) + \frac{B_0}{b_1}\left(1 - e^{-b_1(t^{\#}-a)}\right)$$

The marginal cost of travel time is $-\partial\hat{u}(a,T)/\partial T$, while the marginal cost of time early is $\partial\hat{u}(a,T)/\partial a$. It is straightforward to show that for early arrivals the marginal cost of travel time exceeds the marginal cost of time early. This leads to the conjecture that hypercongestion at the very beginning of the rush hour does not occur with a utility function of the form (14) (the same argument applies with a utility function of the form (13)).

The following example is consistent with these two conjectures. Consider the utility function

$$U = -\alpha\, T - \beta\,(t^* - a)^2. \tag{24}$$

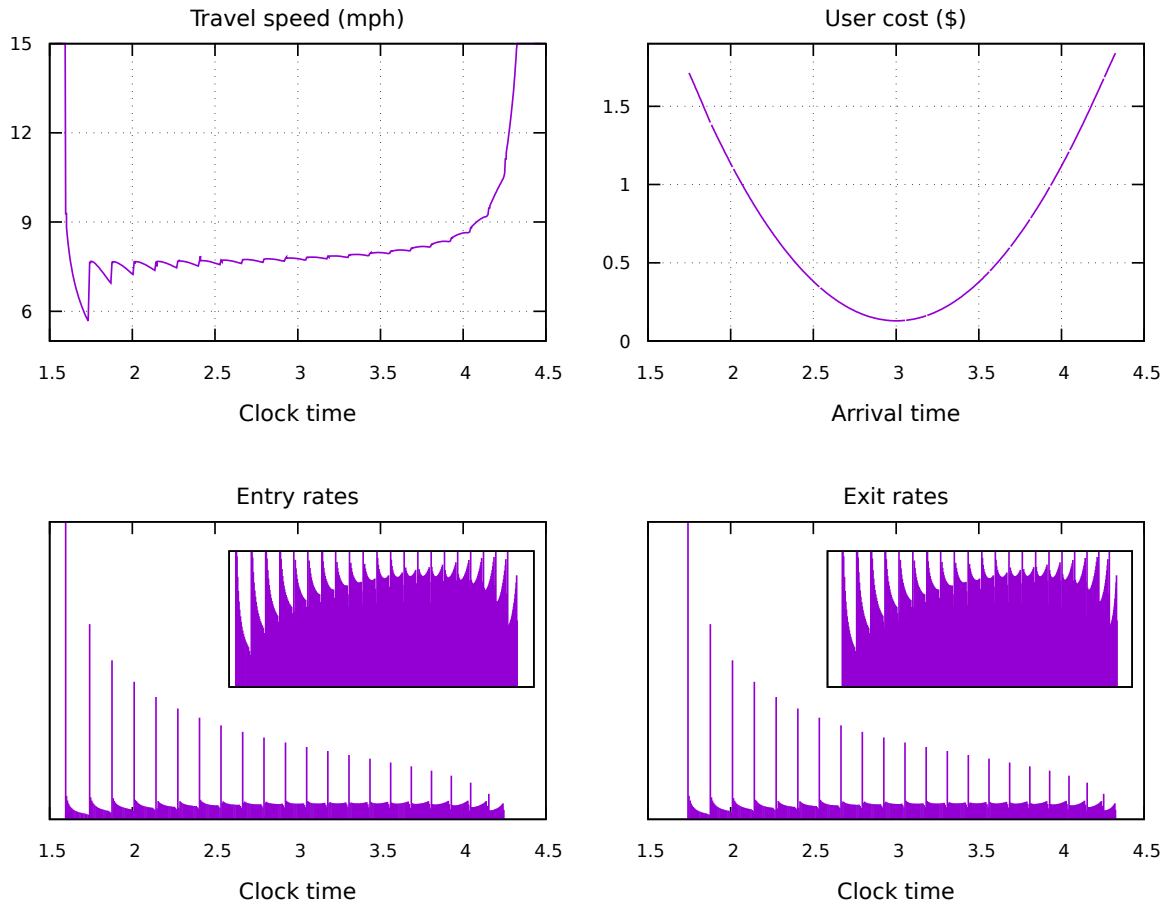


Figure 12: An illustration of hypercongestion with a continuous utility function. Greenshields' Relation is assumed for the congestion function, and the following parameter values as assumed: $\alpha = 1.0$, $\beta = 1.0$, $k_j = 1.0$, $v_f = 15.0$, $t^* = 3.0$, $L = 1.0$ and $N = 10$.

It exhibits constant marginal disutility of travel time and increasing marginal disutility of schedule delay. It is smooth and strictly concave. At peak times in the early morning rush hour, the marginal disutility of travel time exceeds the marginal disutility of time early; in the early morning shoulder of the rush hour, however, the marginal disutility of time early can exceed the marginal disutility of travel time. The results for the social optimum for a particular numerical example are displayed in Figure 12. At first glance, they appear bizarre. But they can be rationalized if the above two conjectures are correct. That departures do not all occur in contiguous departure masses is consistent with the first conjecture. Also,

since at the start of the rush hour the marginal disutility of time early exceeds the marginal disutility of travel time, that hypercongestion occurs at the very beginning of the rush hour is consistent with the second conjecture.

This example suggests that there is a wide range of qualitative departure patterns. More work is needed to understand which occurs.

The basic bathtub has considerable conceptual appeal. It checks off virtually all the boxes in terms of what one wants in a sound basic model of rush hour traffic dynamics that can be built on to develop a rich theory of the spatial dynamics of traffic congestion in a metropolitan area. It is discouraging that such an appealing model gives rise to such complicated mathematics and has such complex solution properties. It would be a mistake, however, to give up on the proper bathtub model because of these difficulties. Many papers in the current literature circumvent them by making approximations, the most common of which makes the arrival rate proportional to traffic flow.[23] It may turn out that some of these approximations simplify analysis and computation considerably without seriously compromising the properties of the bathtub model proper. But to determine this requires a reference point against which to evaluate the soundness of the approximations, and the exact solution is the appropriate reference point.

Acknowledgments

The authors would like to thank Joshua Buli, Wen-Long Jin, Robin Lindsey, Raphaël Lamotte, and session participants at a session of the International Transportation Economics Association 2018 Conference in Hong Kong (Session 31, Traffic Management: Dynamics and mechanism design) for very helpful comments on an earlier version of the paper. Our acknowledgement goes also to Antoine Benoit (a colleague of Kilani in the mathematics department) who clarified for us the definition of integrable functions and measures. Responding to the constructive critical comments and suggestions of the referees, the Associate

[23]This approximation was first proposed by Daganzo (2007) and has subsequently been adopted in the branch of the MFD literature that employs "accumulation-based" (Mariotte et al. 2017) models of rush hour traffic dynamics (without endogenous trip timing).

Editor, and the Editor resulted in a paper that is clearer, more precise, and more compact than the original version. Arnott would like to thank the Université de Lille for its hospitality during visits he made to work with Kilani, as well as the National Science Foundation for its financial support under grant no. CMMI-1629195. Kilani acknowledges financial support from Agence Nationale de la Recherche under grant no. ANR-21-HDF1-0014.

References

Arnott, R. (2013), 'A bathtub model of downtown traffic congestion', *Journal of Urban Economics* **76**, 110–121.

Arnott, R. and Buli, J. (2018), 'Solving for equilibrium in the basic bathtub model', *Transportation Research Part B: Methodological* **109**, 150–175.

Arnott, R. and Jinushi, R. (2021), Welfare economics of the basic bathtub model. Draft.

Arnott, R., Kokoza, A. and Naji, M. (2016), 'Equilibrium traffic dynamics in a bathtub model: A special case', *Economics of Transportation* **7**, 38–52.

Buli, J. (2019), Discontinuous Galerkin Methods for Hyperbolic Systems of PDEs and the Bathtub Model for Traffic Flow, PhD thesis, UC Riverside.

Daganzo, C. F. (2007), 'Urban gridlock: Macroscopic modeling and mitigation approaches', *Transportation Research Part B: Methodological* **41**(1), 49–62.

Fosgerau, M. (2015), 'Congestion in the bathtub', *Economics of Transportation* **4**(4), 241–255.

Geroliminis, N. and Daganzo, C. F. (2008), 'Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings', *Transportation Research Part B: Methodological* **42**(9), 759–770.

Geroliminis, N. and Levinson, D. M. (2009), Cordon pricing consistent with the physics of overcrowding, *in* 'Transportation and Traffic Theory 2009: Golden Jubilee', Springer, pp. 219–240.

Godfrey, J. (1969), 'The mechanism of a road network', *Traffic Engineering and Control* **8**(8), 323–327.

Lamotte, R. and Geroliminis, N. (2017), 'The morning commute in urban areas with heterogeneous trip lengths', *Transportation Research Procedia* **23**, 591–611.

Mariotte, G., Leclercq, L. and Laval, J. A. (2017), 'Macroscopic urban dynamics: Analytical and numerical comparisons of existing models', *Transportation Research Part B: Methodological* **101**, 245–267.

Small, K. A. and Chu, X. (2003), 'Hypercongestion', *Journal of Transport Economics and Policy (JTEP)* **37**(3), 319–352.

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

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

Vickrey, W. (2019), 'Types of congestion pricing models', *Economics of Transportation* **20**, 100140.

Appendix A Complementary Material to Section 4

A.1 Late Arrivals

We examine the properties of late departure masses. Late and early arrivals differ in three ways. First, and most obviously, it is typically assumed, based on empirical evidence, that the value of time late is higher than the value of time early ($\gamma > \beta$); second, while there is always an early departure mass – mass 0, which arrives at the common desired arrival time, there may not be a late departure mass; and third, a commuter in the earliest departure mass does not impose a schedule delay externality but a commuter in the latest departure mass does.

Proposition A1. *With α-β-γ tastes, in the social optimum all late departure masses are congested rather than hypercongested, and, among late departure masses, earlier masses are more congested.*

Proof. From (18c), the condition that the marginal social cost of a commuter who departs immediately after the last active departure mass arrives is weakly higher than the marginal social cost of the last active departure mass, D, is[24]

$$(\alpha + \gamma)T_f \geq (\alpha + \gamma)k_D T'(k_D) + \alpha T(k_D). \tag{25}$$

This condition establishes that the last departure mass is congested rather than hypercongested. A new departure mass is formed when population rises to that level at which (25) holds with equality.[25]

From (18c), for $i = 1, \ldots, D$, the marginal social cost of a commuter who departs in mass

[24] If mass D is hypercongested, (25) cannot hold since $\alpha T(k_D)$ is positive and $k_D T'(k_D) > T(k_D) > T_f$. Also, notice that to reduce notation we have used $-D$ to denote the first active mass and D to denote the last active mass. In general active masses are not symmetric with respect to the central mass (mass 0).

[25] For example, when $\gamma = 2\alpha$, which is often assumed in numerical examples, and when congestion is described by Greenshields' Relation with $v_f = k_j = 1$ (normalizations), a new departure mass is formed when $k_D = 0.2792$.

$i-1$ must be the same as that of a commuter who departs in mass i:

$$\alpha\Big[k_{i-1}T'(k_{i-1})+T(k_{i-1})\Big]+\gamma\sum_{i=1}^{i-1}T(k_j)+\gamma\,T'(k_{i-1})\sum_{j=i-1}^{I}k_j$$
$$=\alpha\Big[k_iT'(k_i)+T(k_i)\Big]+\gamma\sum_{j=1}^{i}T(k_j)+\gamma\,T'(k_i)\sum_{j=i}^{I}k_j, \quad (26a)$$

which, after the elimination of common terms, reduces to

$$\alpha\Big[k_{i-1}T'(k_{i-1})+T(k_{i-1})\Big]+\gamma\,T'(k_{i-1})\sum_{j=i-1}^{I}k_j$$
$$=\alpha\Big[k_iT'(k_i)+T(k_i)\Big]+\gamma\,T(k_i)+\gamma\,T'(k_i)\sum_{j=i}^{I}k_j, \quad (26b)$$

which can be rewritten as

$$\alpha\Big[k_{i-1}T'(k_{i-1})+T(k_{i-1})\Big]+\gamma\,T'(k_{i-1})\sum_{j=i}^{I}k_j$$
$$-\alpha\Big[k_iT'(k_i)+T(k_i)\Big]-\gamma\,T'(k_i)\sum_{j=i}^{I}k_j=\gamma\Big[T(k_i)-k_iT'(k_{i-1})\Big]. \quad (26c)$$

Consider $i=D$. Suppose that departure mass $D-1$ is less congested than mass D. Then the LHS of (26c) is negative. But since mass D is congested rather than hypercongested and since mass $D-1$ is less congested than mass D, then the RHS of (26c) is positive, which establishes a contradiction. Hence, mass $D-1$ is more congested than mass D. Then since the RHS is positive and since mass $D-1$ is more congested than mass D, mass $D-1$ is congested rather than hypercongested. Applying the argument recursively establishes that all late departure masses are congested rather than hypercongested, and that, among departure masses that arrive late, more central departure masses are more congested. □

We show in Appendix A.2 (Proposition A3) that the central mass is the most congested. We bring together some of the above results in

Proposition A2. *With α-β-γ tastes, if $\alpha > \beta$, hypercongestion does not occur in the social optimum, and more central departure masses are more congested.*

It follows from the above proofs that, if $\alpha > \beta$, in the social optimum both the congestion externality cost and the schedule delay externality cost are higher for more central departure masses. It follows that the time-varying toll that decentralizes the social optimum is higher for more central departure masses.

A.2 The Central Departure Mass is the Most Congested

Proposition A3. *In the social optimum of the α-β-γ model, with $\alpha > \beta$ the central departure mass is the most congested.*

Proof. If the social optimum allocation entails only early departure masses, the result follows immediately from Lemma 3 in Section 4. Consider therefore social optima with both early and late arrivals. Where N_e is the number of commuters who arrive strictly early, $N_o (= k_0)$ is the number of commuters who arrive on time, and N_l is the number of commuters who arrive strictly late, it must be the case that that

$$-\beta N_e + \gamma (N_o + N_l) > 0 \tag{27}$$

since otherwise total travel costs could be reduced by incrementally increasing the time of the first departure while holding fixed the departure pattern; this would cause each of the early departures to arrive less early, and each of the on time and late departures to arrive later. (Similarly, it must the case that $\beta (N_e + N_o) - \gamma N_l > 0$ since otherwise total travel costs could be reduced by the incrementally decreasing the time of the first departure while holding fixed the departure pattern.)

Consider adding an extra commuter to departure mass 0, holding fixed the mass' arrival time as well as the size of mass $i = -I, \ldots, -1$. This perturbation increases the time early of each commuter in masses $i = -I, \ldots, -1$, for an increase in total time early cost of $\beta N_e T'(k_0)$. It also increases the travel time of each commuter in departure mass 0 by

$T'(k_0)$ for an increase in total travel time cost of $\alpha k_0 T'(k_0)$. Finally, the direct cost of the added commuter is $T(k_0)$. Thus, the increase in total travel cost of adding this commuter is $\beta N_e T'(k_0) + \alpha(T(k_0) + k_0 T'(k_0))$.

Consider now adding an extra commuter to departure mass 1, holding fixed the mass' departure time as well as the size of masses $i = 2, \ldots, I$. This perturbation increases the time late of each commuter in masses, $i = 1, \ldots, I$, for an increase in total time late cost of $\gamma N_l T'(k_1)$. It also increases the total travel time cost of commuters in departure mass 1 by $\alpha k_1 T'(k_1)$. Finally, the direct cost of the added commuter is $(\alpha + \gamma)T(k_1)$. Thus, the increase in total travel cost is $\gamma(T(k_1) + N_l T'(k_1)) + \alpha(T(k_1) + k_1 T'(k_1))$.

At the social optimum, the cost of adding an extra commuter to departure mass 0 is the same as adding her to department mass 1. Thus,

$$\beta N_e T'(k_0) + \alpha(T(k_0) + k_0 T'(k_0)) = \gamma(T(k_1) + N_l T'(k_1)) + \alpha(T(k_1) + k_1 T'(k_1)). \quad (28)$$

Now, suppose contrary to the Proposition that $k_1 > k_0$. Then rearranging (28)

$$\alpha[(T(k_1) + k_1 T'(k_1)) - (T(k_0) + k_0 T'(k_0))] = \beta N e T'(k_0) - \gamma(T(k_1) + N_l T'(k_1)) > 0. \quad (29)$$

Since from (27) $\gamma(N_o + N_l) > \beta N_e$ and since $N_o = k_0$, $\gamma(k_0 + N_l) > \beta N_e$ so that

$$\gamma(k_0 + N_l)T'(k_0) - \gamma(T(k_1) + N_l T'(k_1)) > \beta N_e T'(k_0) - \gamma\left(T(k_1) + N_l T'(k_1)\right) > 0. \quad (30)$$

But

$$\gamma(k_0 + N_l)T'(k_0) - \gamma(T(k_1) + N_l T'(k_1)) = \gamma\left\{[N_1(T'(k_0) - T'(k_1))] + [k_0 T'(k_0) - T(k_1)]\right\} < 0; \quad (31)$$

$k_1 > k_0$ and the convexity of the function $T(k)$ imply that $T'(k_0) - T'(k_1) < 0$, while $k_1 > k_0$ and Lemma 2 of Section 4 imply the string of inequalities $T(k_1) > T(k_0) > k_0 T'(k_0)$. This establishes the contradiction. □

Appendix B Further Details on the Optimization Procedure

In what follows, we show how the computation of the optimum can be set up as a nonlinear programming problem with equality constraints defined with respect to variables m_i over the approximation grid described in Section 3.2.

Remark 1 (notation)**.** *In this appendix e_i and k_i are normalized, so that $\sum_i^n e_i = 1$. Thus, as in Eq.* (33) *below for example, k_i is multiplied by total population N to obtain traffic density.*

B.1 Evaluation of the Objective Function

Traffic density is the proportion of commuters that are in the bathtub between m_i and m_{i+1}. It is the proportion of commuters that have entered at m_i or before but that exit at m_{i+1} or after. Commuters that enter at m_{i-h} exit at m_i. This group and those that enter before are not considered. Those entering at m_{i-h+1} exit at m_{i+1}. It follows that we have to sum the commuters entering from m_{i-h+1} to m_i. Taking into account the fact that there are no entries before m_1 and after m_n, we have

$$k_i = \sum_{j=\max(1,i-h+1)}^{\min(i,n)} e_j \tag{32}$$

for $i = 1, \ldots, n + h - 1$. If Greenshields' Relation is used, then travel speed is

$$v_i = v_f \left(1 - \frac{N \cdot k_i}{k_j}\right). \tag{33}$$

If another formulation is adopted (like the generalized Greenshields' Relation or the one used to construct the illustration in Fig. 5), then Eq. (33) should be changed accordingly.

Clock-time duration of interval i (to travel from mile m_i to mile m_{i+1}) is

$$\Delta t_i = \frac{m_{i+1} - m_i}{v_i}. \tag{34}$$

The first departure occurs at mile m_1, corresponding to clock-time $t_1 = m_1/v_f$. So, clock-time t_i for $i = 1, \ldots, n + h - 1$ is given by

$$t_{i+1} = t_1 + \sum_{j=1}^{i} \Delta t_j, \tag{35a}$$

which can also be written in the recursive form

$$t_{i+1} = t_i + \Delta t_i. \tag{35b}$$

Using this information (Eqs. (32)–(35)) the social welfare, which is the objective function for the optimum, is computed as follows. The utility of a commuter who enters the bathtub at t_i is $\widetilde{u}(t_i, T_i)$, where T_i is the travel time for commuters entering at t_i. In this model, it is clear that $T_i = t_{i+h} - t_i$, so that the utility can be expressed as a function of entry time only. We then use $u(t_i) = \widetilde{u}(t_i, T_i)$. Moreover we decompose the utility into two parts, entry and exit subutilities. The former depends only on entry time, t_i, and the last on exit time, which is a function of entry time and h. Each commuter entering at t_i gets an entry utility $u_E(t_i)$. The utility obtained by all commuters entering at t_i is equal to their mass $N\, e_i$ multiplied by the level of the utility $u_E(t_i)$. The aggregate entry utility of the average commuter is then given by

$$U_E = \sum_{i=1}^{n} e_i \cdot u_E(t_i) \tag{36}$$

Commuters who enter at t_i exit at t_{i+h} and each one in this group obtains a utility $u_X(t_{i+h})$. The aggregate exit utility of the average commuter is then given by

$$U_X = \sum_{i=1}^{n} e_i \cdot u_X(t_{i+h}), \tag{37}$$

and social welfare is therefore

$$U = N\ (U_E + U_X)\,. \tag{38}$$

Then, several constraints will ensure that the solution is feasible. These constraints state that (i) traffic density is smaller than capacity, that (ii) all commuters enter the bathtub and that (iii) all departures and arrivals occur within the limits where subutilities u_E and u_X are defined.

B.2 Evaluation of the Constraints

There are three sets of constraints. In the first set, we state that traffic density never reaches jam capacity. For interval 1 this constraint is $Ne_1 < k_j$ or $e_1 < k_j/N$. For interval 2 it is $e_1 + e_2 < k_j/N$, and so on until interval h for which the constraint is $e_1 + \ldots + e_h < k_j/N$. Notice then that if this last constraint h is satisfied, then all the earlier constraints $1, \ldots, h-1$ are satisfied, and thus can be ignored. A similar argument can be stated for the constraint related to interval n, and those related to subsequent intervals ($n+1$ to $n+h-1$). Indeed the constraint related to the density in interval n is $e_{n-h+1}+e_{n-h+2}+\ldots+e_n < k_j/N$. But, since there are no entries after m_n, the next constraint (interval $n+1$) is $e_{n-h+2}+\ldots+e_n < k_j/N$, which is satisfied whenever the earlier constraint is satisfied. The constraints related to the densities in intervals $n+2, \ldots, n+h-1$ can be ignored for the same arguments. So, the only constraints taken into account are those related to intervals $i = h, \ldots, n$. Setting these

as equalities, we have

$$e_{i+n} + \sum_{j=i}^{i+h-1} e_j = \frac{k_j}{N} \qquad \text{for} \quad i = 1, \ldots, n-h+1 \tag{39}$$

where $e_{i+n} \geq \epsilon > 0$ are slack variables. The second set of constraints contains a single one stating that all the commuters enter the bathtub, or alternatively that entry rates sum up to one:

$$\sum_{i=1}^{n} e_i = 1. \tag{40}$$

The third set of constraints contains a single one stating that the latest arrival occurs before $t^{\#}$:

$$e_{2n+h+2} + t_{n+h} = t^{\#}, \tag{41}$$

where $e_{2n+h+2} \geq \epsilon' > 0$ is a slack variable. This last constraint is not required when $U_X(T)$ is defined for $t \geq t^{\#}$.

In this problem we have a total of $2n-h+2$ nonnegative decision variables, including n entry rates and $n-h+2$ slack variables, and $n-h+3$ equality constraints. For the practical implementation it may be enough to take into account only constraint (40) and let the other constraints be handled implicitly by the optimization routine. For the general case, however, it is more reliable to explicitly provide all the constraints.

B.3 The Gradient of the Objective Function

Using (32), (33) and differentiating (35) with respect to e_i we have

$$\frac{\partial \Delta t_j}{\partial e_i} = \begin{cases} \frac{v_f N}{k_j} \frac{\Delta t_j}{v_j}, & \text{if } \max(1, j-h+1) \leq i \leq \min(j, n) \\ 0, & \text{if not.} \end{cases} \tag{42}$$

Notice that the derivative does not depend on index i, so that for all i satisfying $\max(1, j-h+1) \leq i \leq \min(j, n)$, we have $\partial \Delta t_j / \partial e_i = \partial \Delta t_j / \partial e_j$, which limits the number of derivatives to compute in practice. Then, using (42) and (35), we have

$$\frac{\partial t_j}{\partial e_i} = \begin{cases} 0, & \text{if } j \leq i \\ \frac{\partial \Delta t_i}{\partial e_i} + \cdots + \frac{\partial \Delta t_{j-1}}{\partial e_{j-1}}, & \text{if } i < j \leq i+h \\ \frac{\partial t_{i+h}}{\partial e_i}, & \text{if } j > i+h. \end{cases} \tag{43}$$

Then, differentiate (36) with respect to e_i to obtain

$$\frac{\partial U_E}{\partial e_i} = u_E(t_i) + \sum_{j=i+1}^{n} e_j \cdot u'_E(t_j) \cdot \frac{\partial t_j}{\partial e_i}. \tag{44}$$

Expression (44) shows that a marginal increase in entry rate e_i increases the size of the group itself and delays the entry time for the group entering after it. To evaluate the similar impacts on exit subutility, differentiate (37) with respect to e_i to obtain

$$\frac{\partial U_X}{\partial e_i} = u_X(t_{i+h}) + \sum_{j=\max(i-h,1)}^{n} e_j \cdot u'_X(t_{j+h}) \cdot \frac{\partial t_{j+h}}{\partial e_i}, \tag{45}$$

which shows that an increase in entry rate e_i has also two impacts. It increases the size of the group itself and delays the exit of all the groups who were in the bathtub when group i have entered. Finally, summing both derivatives, we get the impact on aggregate utility

$$\frac{\partial U}{\partial e_i} = N \left(\frac{\partial U_E}{\partial e_i} + \frac{\partial U_X}{\partial e_i} \right). \tag{46}$$

B.4 The Jacobian of the Constraints

From the expression of constraints (39) and (40), all the derivatives with respect to relevant variables are equal to one. For constraint (41) the derivative with respect to e_i is equal to $\partial t_{n+h} / \partial e_i$ for all $i = 1, \ldots, n$ and equal to one for the slack variable. We now give an

expression of the Jacobian that is suitable for an implementation in a large-scale optimization package.

We consider only nonzero elements. Let $n_1 = (h+1)(n-h+1)$, and define vectors A_1, A_2 and A_3 of length $n_1 + 2\,n + 1$ as follows.

$$A_1(i) = \begin{cases} \left\lfloor \frac{i-1}{h+1} \right\rfloor + 1, & \text{if } i = 1, \ldots, n_1 \\ n - h + 2, & \text{if } i = n_1 + 1, \ldots, n_1 + n \\ n - h + 3, & \text{if } i = n_1 + n + 1, \ldots, n_1 + 2\,n + 1. \end{cases} \tag{47}$$

Let $I \equiv i \pmod{h+1}$. Then, the elements of vector A_2 are given by

$$A_2(i) = \begin{cases} n + \left\lfloor \frac{i}{h+1} \right\rfloor, & \text{if } I = 0 \text{ and } i = 1, \ldots, n_1 \\ I + \left\lfloor \frac{i}{h+1} \right\rfloor, & \text{if } I \neq 0 \text{ and } i = 1, \ldots, n_1 \\ i - n_1 & \text{if } i = n_1 + 1, \ldots, n_1 + n \\ i - n_1 - n & \text{if } i = n_1 + n + 1, \ldots, n_1 + 2\,n \\ 2\,n - h + 2 & \text{if } i = n_1 + 2\,n + 1 \end{cases} \tag{48}$$

For the last constraints, we have, for all $i = 1, \ldots, n_1 + 2\,n + 1$,

$$A_3(i + n_1 + n) = \begin{cases} \frac{\partial t_{n+h}}{\partial e_i}, & \text{for } i = 1, \ldots, n \\ 1, & \text{otherwise.} \end{cases} \tag{49}$$

The following result shows how all the elements of the Jacobian are computed using A_1, A_2 and A_3.

Lemma 4. *Let A_1, A_2 and A_3 be as defined as in* (47), (48) *and* (49), *respectively. The Jacobian of constraints* (39),(40) *and* (41) *can be computed as follows. For all $i = 1, \ldots, n_1 + 2\,n + 1$ the derivative of constraint number $A_1(i)$ with respect to variable number $A_2(i)$ is*

equal to the value of $A_3(i)$.

B.5 The Hessian of the Objective Function

From (44) we have

$$\frac{\partial^2 u_E}{\partial e_i\, \partial e_{i'}} = u'_E(t_i)\frac{\partial t_i}{\partial e_{i'}} + \sum_{j=i+1}^{n}\left(\frac{\partial e_j}{\partial e_{i'}}u'_E(t_j)\frac{\partial t_j}{\partial e_i} \;+e_j u''_E(t_j)\frac{\partial t_j}{\partial e_{i'}}\frac{\partial t_j}{\partial e_i} + e_j u'_E(t_j)\frac{\partial^2 t_j}{\partial e_i\, \partial e_{i'}}\right). \quad (50)$$

By differentiating (43) with respect to $e_{i'}$, we have

$$\frac{\partial^2 t_j}{\partial e_i\, \partial e_{i'}} = \begin{cases} \displaystyle\sum_{l=\max(i,i')}^{h+\min(i,i')} \frac{\partial^2 \Delta t_l}{\partial e_i\, \partial e_{i'}}, & \text{if } \max(i,i') \le h + \min(i,i') \\ 0 & \text{if not,} \end{cases} \quad (51)$$

and by differentiating (42) we have

$$\frac{\partial^2 \Delta t_j}{\partial e_i\, \partial e_{i'}} = \begin{cases} 2\left(\dfrac{N\, v_f}{k_j}\right)^2 \dfrac{\Delta t_j}{v_j^2}, & \text{if } \max(1, j-h+1) \le i, i' \le \min(j,n) \\ 0 & \text{if not.} \end{cases} \quad (52)$$

The computation of the second order derivatives for u_X are similar, except that t_{j+h} is used instead of t_j, and the Hessian is obtained by summing both, and scaling the matrix by N. The detailed computation of the Hessian matrix is given in Algorithm 1.

Algorithm 1: The computation of the Hessian

Data: First departure, last departure and approximation grid: M_0, $\overline{M}$, n, h ; Entry rates e_1,...,e_n ; Model parameters and subutilities: $u_E, u'_E, u''_E, u_X, u'_X$ and u''_X ; First order derivatives: $\partial t_j/\partial e_i$ for $i = 1, \ldots, n$ and $j = 2, \ldots, n+h$; Initialize second order derivatives to zero: $\partial^2 t_j/\partial e_i \partial e_{i'} = 0$ for all $i, i' = 1 \ldots, n$ and $j = 2, \ldots, n+h$.

Result: The lower elements of the Hessian matrix: $\partial^2 U/\partial e_i \partial e_{i'} = 0$ for all $i = 1, \ldots, n$ and $i' = 1, \ldots, i$.

for $i = 1$ *to* n **do**
  for $i' = \max(1, i-h+1)$ *to* $\min(i, n)$ **do**
    for $j = i$ *to* $\longrightarrow i' + h - 1$ **do**
      $\dfrac{\partial^2 \Delta t_j}{\partial e_i \partial e_{i'}} \leftarrow 2\left(\dfrac{N\, v_f}{k_j}\right)^2 \dfrac{\Delta t_j}{v_j^2}$
      if $j == 1$ **then**
        $\dfrac{\partial^2 t_{j+1}}{\partial e_i \partial e_{i'}} \leftarrow 2\left(\dfrac{N\, v_f}{k_j}\right)^2 \dfrac{\Delta t_j}{v_j^2}$
      else
        $\dfrac{\partial^2 t_{j+1}}{\partial e_i \partial e_{i'}} \leftarrow \dfrac{\partial^2 t_j}{\partial e_i \partial e_{i'}} + 2\left(\dfrac{N\, v_f}{k_j}\right)^2 \dfrac{\Delta t_j}{v_j^2}$
      end
    end
    for $j = i' + h + 1$ *to* $n + h$ **do**
      $\dfrac{\partial^2 t_j}{\partial e_i \partial e_{i'}} \leftarrow \dfrac{\partial^2 t_{i'+h-1}}{\partial e_i \partial e_{i'}}$
    end
    $$\frac{\partial^2 u_E}{\partial e_i\, \partial e_{i'}} \leftarrow u'_E(t_i)\tfrac{\partial t_i}{\partial e_{i'}} + \sum_{j=i+1}^{n}\left(\tfrac{\partial e_j}{\partial e'_i} u'_E(t_j)\tfrac{\partial t_j}{\partial e_i} + e_j u''_E(t_j)\tfrac{\partial t_j}{\partial e_{i'}}\tfrac{\partial t_j}{\partial e_i} + e_j u'_E(t_j)\frac{\partial^2 t_j}{\partial e_i\, \partial e_{i'}}\right)$$
    Let $j' \leftarrow h + \max(i-h, 1)$
    $$\frac{\partial^2 u_X}{\partial e_i\, \partial e_{i'}} \leftarrow u'_X(t_{i+h})\frac{\partial t_{i+h}}{\partial e_{i'}} + \sum_{j=j'}^{n+h}\left(\tfrac{\partial e_j}{\partial e'_i} u'_E(t_j)\tfrac{\partial t_j}{\partial e_i} + e_j u''_E(t_j)\tfrac{\partial t_{j+h}}{\partial e_{i'}}\tfrac{\partial t_j}{\partial e_i} + e_j u'_E(t_j)\frac{\partial^2 t_j}{\partial e_i\, \partial e_{i'}}\right)$$
    The Hessian elements: $\dfrac{\partial^2 u}{\partial e_i\, \partial e_{i'}} \leftarrow N\left(\dfrac{\partial^2 u_E}{\partial e_i\, \partial e_{i'}} + \dfrac{\partial^2 u_X}{\partial e_i\, \partial e_{i'}}\right)$
  end
end