Investigating Undesired Spatial and Temporal Boundary Effects of Congestion Charging

Y.E. Ge[a,b,1], Kathryn Stewart[c], B.R. Sun[b], X.G. Ban[d] and Yuandong Liu[b]

[a] *College of Transport & Communications, Shanghai Maritime University, Shanghai 201306, China*

[b] *School of Transportation & Logistics, Faculty of Infrastructure Engineering, Dalian University of Technology, Liaoning 116024, China*

[c] *Transport Research Institute, Edinburgh Napier University, Merchiston Campus, Edinburgh EH10 5DT, Scotland, U.K.*

[d] *Department of Civil & Environmental Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA*

Abstract

Two types of reported problems are related to the existing congestion charging projects that levy traffic only in a certain area within one or a few time periods during the day. One is that travellers depart earlier or later than a charging period to avoid paying full or part of the congestion charging tolls, which creates two undesired demand peaks that are often greater than available capacity. One peak comes just before the start of congestion charging and the other follows the end of it. We term this phenomenon "temporal boundary effect" of congestion charging. The other reported problem is that travellers would rather stay away from a charging zone than pay congestion charging tolls, which causes undesired congestion on those roads or paths on the edge of the charging zone. We call this phenomenon "spatial boundary effect" generated by congestion charging. This research investigates these boundary effects in the context of simultaneous route and departure time choice dynamic user equilibrium (SRD-DUE) network flows with an aim to gain new insights into congestion charging design. Numerical experiments investigating constant and time-varying congestion charging toll profiles are presented in this paper. This investigation shows that congestion charging may not be able to eliminate hypercongestion efficiently if schemes are not well-designed, and can unfortunately give rise to undesired boundary effects and that a simply-designed congestion charging scheme with small level toll or time-varying toll profiles can reduce the magnitude of boundary effects but may not be able to eliminate fully such undesired effects.

1 Introduction

In the past few decades, many congestion charging projects have been implemented and are now operational on a permanent basis around the world; including the projects in London, Singapore, California and Stockholm. These projects charge traffic only in a certain area within one or a few time periods during the day. Some only toll a fixed amount for the use of a charging zone during the charging period, such as in London. Some set a different constant toll for each time interval in the charging period, such as tolling schemes in Stockholm and Singapore (see Fig. 16.1(a) and (b) in Ge and Stewart 2010, respectively). All these are usually known as step tolling schemes. There are two types of reported problems related to these projects (see Banister 2003, TfL 2007 for example).

[1] Corresponding author.
College of Transport & Communications, Shanghai Maritime University, 1550 Haigang Avenue, Lin-Gang, Pudong, Shanghai 201306, China; Email: yege@shmtut.edu.cn; Phone: +86(0)21 3828 2301

One is that travellers depart earlier or later than a charging period to avoid paying or pay less congestion charging tolls. Hence, two undesired demand peaks have been observed and the peak demand is often greater than the associated available capacity. One comes just before the start of congestion charging and the other follows the end of it. We term this phenomenon "temporal boundary effect" arising from congestion charging. As it is known that a main purpose of traffic congestion charging is to better distribute demand for a road facility or network over time. The boundary issues of interest are the fact that the implementation of a congestion charging project may drive too many travellers to travel outside the charging period so that undesired congestion occurs just before or after the period. Ge and Stewart (2010) investigates the temporal boundary effect in a bottleneck scenario, which uses a more refined bottleneck model than that of Vickrey's. In the Vickrey model (Vickrey 1969), traffic congestion takes the form of cars queuing behind a bottleneck without occupying any physical space, which is recognised to be a key limitation of these types of bottleneck models. Ge and Stewart (2010) instead treats a bottleneck as a real road segment or link with a limited capacity, whose upstream link has a higher capacity; hence the vehicles in the bottleneck can queue up into the upstream link. Traffic is assumed to propagate through the bottleneck following the kinematic wave (KW) model of traffic flow. The cell-transmission (CT) technique is used to solve the traffic flow model. It is these settings that make the new bottleneck model readily produce temporal boundary issues. The investigation in Ge and Stewart (2010) shows that the choice of congestion charging period(s), constant toll levels and the rate of change in tolls over time all exert influence over boundary issues.

The other reported problem is that travellers would rather stay away from a charging area than pay congestion charging tolls, which brings about undesired congestion on those roads or paths just around the charging area. A key purpose of congestion charging is to make the utilisation of road facilities or networks more efficient, both spatially as well as temporally. The concern here is that there are generally more than the desired number of travellers choosing to travel on the roads or paths just around the edge of charging zones, so undesired congestion takes place in this boundary area. We call this phenomenon "spatial boundary effect" created by congestion charging.

It is of utmost importance to investigate both temporal and spatial boundary effects to gain new insights into the causes of these undesired effects and to seek ways to mitigate them and design better congestion charging schemes. In fact, in real-life traffic, these undesired boundary effects of congestion charging may bring certain secondary effects about. As for temporal boundary issues, many drivers tend to speed up (or slow down) so they can enter their targeted charging zone before (or just after) the charging begins (or ends), which gives rise to a deep concern over traffic safety. Lindsey et al. (2010) formulates a braking model to capture the braking behaviour occurring just before the termination of congestion charging when drivers try to slow down so they can enter the charging zone after the charging period and accordingly avoid paying the charging toll. As for spatial boundary issues, many drivers tend to divert around the charging zone, likely resulting in a reduction in the environmental quality just outside the charging zone. In addition, in order to avoid paying the congestion charging toll, many travellers may terminate their car journeys just outside the charging zone, which can result in heavy demand for parking around the charging zone, hence creating deterioration in the living conditions of the area. In addition to the two secondary effects of the spatial boundary effect, Banister (2003) also mentions the effects of the boundaries of charging zones on businesses, land values and rent levels inside the charging zone, which is beyond the scope of this paper. These are also discussed briefly in Seik (1997) as "Limitations of area licensing" scheme of congestion charging implemented in Singapore. This further shows the necessity of investigating undesired boundary effects of congestion charging and seeking ways

to mitigate them.

To investigate both types of undesired boundary effects with the objective of gaining new insights into congestion charging design and seeking ways to mitigate these undesirable effects, this paper takes into account both departure time and route choices on road networks. The traveller behaviour of simultaneous choice of departure times and trip routes has been widely investigated and formulated into different models, e.g. variational inequality formulation equivalent to dynamic user equilibrium (Friesz et al 1993, Ran and Boyce 1996), system dynamics model (Jin 2011), differential variational inequality (DVI) formulation for within-day dynamic user equilibrium (DUE) traffic assignment (Friesz et al 2011), etc. A review on the dynamic traffic assignment (DTA) problem is given in Peeta and Ziliaskopoulos (2001) and Iryo (2013) further discusses some desirable and non-desirable properties of DUE solutions. Different from these DUE models which require all used paths between an origin-destination (OD) pair be identical and minimized, Szeto and Lo (2006) and Ge and Zhou (2012) respectively proposed a class of dynamic user optimal (DUO) states which allows the existence of differences (fixed or variable tolerances) in travel costs among the used paths connecting an OD pair. Ge et al. (2014) further compares them with DUE states and extends the concept of DUO with variable tolerances to treatment of time-varying flows on road networks with discontinuous travel cost functions, high demand level or capacity limits. To avoid complications due to the use of DUO with fixed or variable tolerances for DTA modeling, this paper uses the DVI DUE formulation proposed in Section 4 of Friesz et al (2011) to capture both departure time and route choices on road networks.

The scenario set in this paper is at network level rather than a single bottleneck only and, similar to the work presented in Ge and Stewart (2010) and Stewart and Ge (2014), all vehicles only go through the network under study once, hence each vehicle is charged once only. It is also assumed that travellers have perfect information on traffic conditions and that every traveller simultaneously chooses a departure-time/path combination that would incur the least generalised costs to them. The generalised costs consist of travel time costs, schedule delay costs and congestion charging tolls. The total demand within the time horizon under study is known and constant for each origin-destination (OD) pair. These assumptions lead a road network to a stable state in which all travellers between the same OD pair receive identical generalised costs.

Rescheduling flexibility for both workers and non-workers that is discussed in Saleh and Farrell (2005) shall also help to resolve the boundary effects due to congestion charging. Saleh and Farrell (2005) discusses the effects of rescheduling flexibility on travellers' departure time choices in the context of congestion charging and discusses its implications on the boundary issues. The flexibility allows travellers to get to their destination earlier or later than a normal work start time. This paper takes this into account so that travellers reaching their destinations within the given arrival time window(s) will incur no schedule delay cost. One difference from Saleh and Farrell (2005) is that this paper allows all travellers to have the same rescheduling flexibility.

Before the London Congestion Charging Scheme (LCCS) came into effect on 17 February 2003, a few of auxiliary measures had been implemented, including a tougher control of parking and an improvement in public transit services. The LCCS plus tougher parking control definitely increased the cost to those who travel into the charging zone by car, which directed many of them to travel by public transit that were/are free of congestion charging. The mode shift resulting from congestion charging may be able to reduce the absolute magnitude of the boundary effects of congestion charging. These will not be further discussed in this paper but left for future work since this paper currently considers the departure time and route choice only. Furthermore, the sensitivity of travel demand to cost will not be touched upon in

this paper but it is assumed that the total demand over the given time horizon is fixed. In summary, this paper focusses on the boundary effects of congestion charging resulting from a given number of travellers simultaneously choosing their departure times and trip routes other than trip modes.

Step tolling schemes, including single- and multi-step tolls, have been used in many cities around the world and are probably the most widely-used charging scheme. This type of charging profiles is also widely discussed in the academic literature, including Arnott et al. (1990, 1993), Laih (1994, 1998), Ge and Stewart (2010), Lindsey et al. (2010), Stewart and Ge (2014), Van den Berg (2011), etc. Arnott et al. (1990, 1993) is mainly concerned with single-step or coarse tolls whereas Laih (1994, 1998) considers multi-step tolls. The LCCS is single-step, as was the initial area licensing scheme (ALS) implemented in Singapore in 1975. Multi-step toll profiles now operational in Singapore and in Stockholm and on SR 91 in California may be considered (roughly) to represent an actually-implemented version of time-varying toll profiles. This paper will show numerically that a coarse toll profile (i.e. single step or flat tolls) can readily give rise to temporal and spatial boundary effects, which negatively affects the effectiveness and efficiency of congestion charging. Although these issues can be somewhat exaggerated under the assumption of perfect traveller information, this investigation will provide some insights into congestion charging design. Numerical experiments are also conducted with time-varying toll profiles. Boundary effects from these toll profiles will be analysed in this paper. This investigation also shows that a simply-designed congestion charging scheme with small level toll or time-varying toll profiles can reduce the magnitude of boundary effects but may not be able to eliminate fully such undesired effects.

All numerical analyses of this paper were carried out on a four-link, three-node network. It is true that resulting findings would be more significant if an analysis of real cases of observed data is made on a larger road network. The investigation in this paper is to make preparations for carrying out analysis on real-sized road networks using observed data. The boundary phenomena presented in this paper are generally consistent with what have been observed in Singapore, London and Stockholm. Therefore, all findings may be considered to be applicable to real-life road networks. In addition, it is assumed that the travellers body is homogeneous. The reader who is interested in the pricing issue with heterogeneous travellers may refer to Doan et al. (2011), van den Berg and Verhoef (2011), Jiang et al. (2011), Lu and Mahmassani (2011), Wang et al. (2012) and the references therein.

This paper does not aim to optimise congestion charging tolls but to observe the effects of a given tolling profile. Therefore, determining optimal tolls is beyond the scope of this paper and we only mention a few relevant references. To seek optimal congestion charging tolls for a set of selected links, Zhang and Ge (2004) formulates a Stackelberg programming problem that consists of two levels, whose upper level captures the transportation management authority's goal (that of determining the optimal tolls) while the lower level reflects traveler choice behaviour (i.e. pursuing minimized user costs). Ban et al. (2009) uses the same framework to determine optimal tolls while taking into account travellers' risk aversion. Both references deal with the static case. This modeling framework, however, is also used commonly in optimising time-varying congestion charging tolls; see e.g., Friesz et al. (2007); Ban and Liu (2009).

This research was stimulated by the practical problem of boundary effects observed from the London Congestion Charging Scheme and elsewhere. Hence, it is not intended as an exhaustive review of the literature on traffic congestion charging. The state-of-the-art and practice of congestion charging is thoroughly reviewed in Yang and Huang (2005), Lawphongpanich et al. (2006), Lindsey (2006, 2010), and Tsekeris and Voss (2009). However, a few of significant pieces of very recent work on road pricing are worthy of mentioning here. Yang and Wang (2011) proposes to manage network mobility by means of tradable credits of travel, Wang et al. (2012) extends this concept for those scenarios with heterogeneous

users and this concept of tradable travel credits is transferred in Zhang, Yang and Huang (2011) to parking space management, i.e. tradable parking credits. Friesz et al. (2008) proposes the use of the "European-type call option" conception for pricing commuting to work along a given path for a given departure time chosen by drivers when telecommuting is available as an alternative. In the conventional approach to congestion pricing, the uncertainty in the congestion process or the risk of travellers being unable to arrive at their destinations or enter their expected roads at their expected times has usually been omitted for the sake of simplicity of the modelling treatment. The challenges from the inclusion of uncertainty in traffic congestion pricing is addressed in Yao et al. (2010), which investigates congestion derivatives for a traffic bottleneck scenario. An auction-based congestion pricing scheme has been developed in Teodorović et al. (2008) and the basic idea of the scheme is that all drivers hoping to enter a cordoned downtown area in a specific time period have to participate in an auction; the operator or traffic authority being the auctioneer who makes the decision on whether to accept or reject particular bids by the drivers. The transaction of credits/options/derivatives/auctions all involve operational costs; how the markets of these instruments can run by themselves with no money from the public funds is critical if we want such markets to emerge and to be able to stand by themselves and Nie (2011) has recently focused on this problem. The congestion pricing approach used in this paper however remains conventional, with our focus being on the boundary effects arising from some of the currently operational congestion charging projects in the world.

One contribution of this paper is to investigate properties of a state resulting from traveler choice behavior in a given road network where a congestion pricing scheme has been implemented already, which differs from the existing literature focusing on designing an optimal congestion pricing scheme (e.g. Zhang and Ge, 2004; Lindsey et al. 2012; Xiao et al. 2011, 2012; Wu and Huang, 2014). Therefore, the dynamic user equilibrium other than system optimal principle is used in this research for modeling travelers' simultaneous route and departure time choice behavior. Second, the work presented in this paper is an extension of the work on bottleneck boundary effects of congestion charging investigated in Ge and Stewart (2010). This extension shows that the inclusion of travelers' route choice is unable to eliminate or reduce significantly the temporal boundary effects of congestion charging while it creates the spatial effects. Third, this work also highlights the limitations of the existing DTA models, including the assumptions of perfect traveler information, dynamic user equilibrium principle, user homogeneity, etc.

The rest of the paper is structured as follows. Section 2 presents a dynamic user equilibrium (DUE) network flow model used for this research. Section 3 sets out the scenarios based on which numerical experiments are to be carried out. Section 4 presents and discusses numerical experiments on boundary issues. Section 5 closes the paper with some concluding remarks.

2 Methodology

This section sets out the methodology used in this paper for investigating boundary effects of congestion charging. This investigation is based on a within-day differential variational inequality (DVI) formulation for the DUE traffic assignment problem that takes into account both departure time and route choices (see Friesz et al. 2011, 2013, and Han et al. 2013 for example).

2.1 Preliminaries

Consider a traveller between an OD pair $w \in W$, where W is the set of OD pairs on a road network, who has decided to enter the network within the time horizon $[0, T]$ and wonders when to depart and

which path to take to go to his or her destination. Suppose that if (s)he leaves at time t and takes path $r \in P^w$, where P^w is the set of paths between the OD pair w, then the time-dependent journey time (s)he would experience can be denoted as $\tau_r^w(\mathbf{u}(t), t)$ —- a function of time t and a vector $\mathrm{u}(t) = (u_p(s) : \forall p \in P^w, w \in W, s \in (t_-, t])$, where $u_p(s)$ represents the rate of inflow to path $p \in P^w$ at time s and t_- denotes the earliest time at which all traffic leaving the network after t enters.

2.2 Schedule delay and generalised costs

When we model departure time choices, schedule delay should be endogenous. The following form of schedule delay is used in this research

$$c_p^s(t) = \begin{cases} \beta^w(t_l^w - t_p^a(t)) & t_p^a(t) < t_l^w \\ 0 & t_l^w \leq t_p^a(t) < t_u^w \\ \gamma^w(t_p^a(t) - t_u^w) & t_u^w \leq t_p^a(t) \end{cases} \tag{1}$$

where $t_p^a(t) = t + \tau_p^w(\mathbf{u}(t), t)$ is the actual arrival time of a traveller between the OD pair w departing along path $p \in P^w$ at time t, $[t_l^w, t_u^w]$ is the desired arrival time window for the OD pair w within which a traveller must arrive if (s)he wants to incur no schedule delay cost, and β^w and γ^w respectively represent the equivalent cost resulting from arriving earlier and later by one time unit between the OD pair w. This implies that the heterogeneity of travellers is not considered here in terms of their values of time and rescheduling flexibility since all travellers of an OD pair w share the same set of values of β^w, γ^w, α^w to be appearing later in Eq. (3) and the arrival time window $[t_l^w, t_u^w]$.

The same or similar form of schedule delay costs is used and discussed in, for example, Small (1982), Small and Verhoef (2007) and Arnott et al. (1993).

We denote a congestion charging toll at time t on path p as $c_p^c(t)$, which should satisfy the following property

$$c_p^c(t) = \begin{cases} \geq 0 & t \in [\tau_l^p, \tau_u^p] \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $[\tau_l^p, \tau_u^p]$ represents the charging period associated with path p. This paper discusses path-based congestion charging and the toll $c_p^c(t)$ is collected once a traveller enters path p at time $t \in [\tau_l^p, \tau_u^p]$, i.e. in the charging period.

If (s)he chooses path $p \in P^w$ to leave at t, we define the generalised cost a traveller would experience as $\Phi_r^w(\mathbf{u}(t), t)$, mathematically,

$$\Phi_r^w(\mathbf{u}(t), t) = \alpha^w \tau_r^w(\mathbf{u}(t), t) + c_r^s(t) + c_r^c(t) \tag{3}$$

where α^w represents the cost equivalent to one time unit of travel time for the OD pair w. This means that all travellers between an OD pair are assumed to have the same value of time.

The lower bound on achievable generalised cost levels for a user between an OD pair $w \in W$ is given by

$$\mu^w = ess\ inf\{\Phi_r^w(\mathbf{u}(t), t) : \forall r \in P^w, \forall t \in [0, T]\} \tag{4}$$

where μ^w is interpreted as the essential infinitum of $\Phi_r^w(\mathbf{u}(t), t)$ over the set of all potential paths in P^w across the time horizon, and called the *minimum generalised cost* of the OD pair w. Note that μ^w only depends on w and is independent of t and $\mathbf{u}(t)$.

2.3 Traffic conservation constraints

Suppose that the total demand between the OD pair w is known and constant and denoted as D^w. Then the traffic conservation principle requires

$$\sum_{p\in P^w}\int_0^T u_p(t)dt = D^w \qquad \forall w \in W. \tag{5}$$

Let us define

$$y^w(t) = \sum_{p\in P^w}\int_0^t u_p(t)dt \qquad \forall\, t\in[0,T], w\in W. \tag{6}$$

and then, following Friesz et al. (2008, 2011, 2013), the constraints (5) can be rewritten as a two-point boundary-value problem below:

$$\frac{dy^w(t)}{dt} = \sum_{p\in P^w} u_p(t), \quad \forall\, t\in[0,T] \tag{7a}$$

$$y^w(0) = 0, \tag{7b}$$

$$y^w(T) = D^w, \tag{7c}$$

for all $w \in W$.

In addition, the rates of inflow to a path should be nonnegative, i.e.

$$u_p(t) \geq 0 \qquad \forall\, t\in[0,T], \forall p\in P^w, w\in W. \tag{8}$$

Then, if $u^p(t)$ satisfies the above constraints (7) and (8) for all $p \in P^w, w \in W$ and $t \in [0,T]$, $\mathbf{u} = (\mathrm{u}(t) : \forall\, t\in[0,T])$ is a feasible solution.

2.4 Assumptions on DUE

Before proceeding, let us make three assumptions:

Assumption 1 (see Properties 1 and 2 in Carey et al. (2003))**.** *Traffic follows the first-in-first-out and causality principles.*

Assumption 2. *All travellers have perfect information on traffic conditions.*

Assumption 3. *All travellers choose the least costly (departure time, path) combination to travel in terms of generalised costs.*

2.5 DUE and differential variational inequalities

Then, we can have the following definition

Definition 1 (Dynamic User Equilibrium or DUE) *Under Assumptions 1-3, for a vector* $\mathbf{u}$ *satisfying the traffic conservation constraints (7) plus nonnegativity constraints (8) and nonnegative real variables* μ^w*, the pair* $(\mathbf{u}, \mu^w : w \in W)$ *represents a dynamic user equilibrium state if and only if the following conditions are satisfied for all* $t\in[0,T], r\in P^w, w\in W$

$$u_r(t) > 0 \;\Rightarrow\; \Phi_r^w(\mathrm{u}(t),t) = \mu^w \tag{9a}$$

$$u_r(t) = 0 \;\Rightarrow\; \Phi_r^w(\mathrm{u}(t),t) \geq \mu^w \tag{9b}$$

Theorem 1 in Friesz et al. (2011) formulates the within-day DUE problem as a differential variational inequality (DVI) problem as follows

Find $\mathbf{u}^*$ such that

$$\sum_{w\in W}\sum_{p\in P^w}\int_0^T \Phi_p^w(\mathrm{u}^*(t),t)\left[u_p(t)-u_p^*(t)\right]dt \geq 0 \tag{10}$$

for all $\mathbf{u}\in\Lambda$, where

$$\Lambda=\{\mathbf{u}\geq 0 : dy^w(t)/dt=\sum_{p\in P^w}u_p(t), y^w(0)=0, y^w(T)=D^w, \forall w\in W\} \tag{11}$$

which is a Hilbert space.

To investigate both temporal and spatial boundary effects of congestion, this paper uses the aforementioned Friesz et al. (2011)'s DVI formulation of the within-day DUE problem to capture the traveler behavior of simultaneous choice of departure times and trip routes. The fixed-point algorithm given in Section 6 of Friesz et al. (2011) is used in generating numerical results later in this paper. To save space, we do not repeat this algorithm here; the reader may refer to the reference for further details.

An eminent feature of this formulation (10)-(11) is that it does not include a procedure to generate $\Phi_p^w(\mathrm{u}^*(t),t)$ or essentially path travel times $\tau_r^w(\mathbf{u}(t),t)$. As in Friesz et al. (2011), this is completed by means of an exogenous dynamic network loading (DNL) procedure or model. As discussed in Carey and Ge (2011), solving for a DUE state is most commonly iterating between DNL and path inflow reassignment. Formulation (10)-(11), in essence, defines a rule for the path inflow reassignment procedure. If a solution process starts from a chosen initial path inflow set, the DNL procedure takes the inflow to each spatial path at each point in time and uploads all of these path inflows onto the network over time to obtain the travel times on all time-space paths; subsequently, feeding the newly generated travel times by DNL to the path reassignment procedure enables the latter to renew the trial path inflows. The iteration between these two procedures continues until a DUE state is found or the process terminates with no DUE solutions to be found. A DNL procedure based on the kinematic wave model of traffic or its discretised versions (e.g. Daganzo 1994, 1995) has been widely used in DTA modeling (e.g. Lo and Szeto 2002, Carey and Ge 2011, Zhong et al. 2011, Ge and Zhou 2012, Friesz et al. 2013, Stewart and Ge 2014).

3 Scenario settings

This section is to set out a scenario for numerical experiments on boundary effects due to congestion charging. The scenario is based on a simple road network as shown in Fig. 1, which has one OD pair

Figure 1: An example network

from node O to node D and is composed of four links. It has three paths: Path 1 = {links 0 and 2},

Path 2 = {links 1 and 2} and Path 3 = {link 3}. Paths 1 and 2 share link 2, which is the busiest on the network. To reduce traffic on link 2 in the peak period, it is assumed that a proposal is put forward to implement a path-based congestion charging scheme, which charges traffic for entering paths 1 and 2 in the charging period.

This charging scheme can also be regarded as a cordon-based charging scheme, in which Nodes O and D are right on the cordon and link 3 goes along the cordon, representing the edge of the congestion charging zone.

Without loss of generality, in the rest of the paper the units of the variables or parameters are often omitted.

3.1 Time horizon and charging period

The time horizon used is $[0, T]$ where $T = 35$ time units. Unless stated otherwise, the time step size used in the later numerical computing is 0.01 time units.

As customary, this research assumes that the network is initially empty. To make the scenario satisfy this assumption, a horizon should be long enough to cover a peak period plus the period during which the instantaneous demand rate rises from zero. Here, if one time unit is set to 6 minutes then the horizon will be 3.5 hours long, which implies that the horizon can start say from 6am to 9:30am. Therefore, the time horizon given here satisfies this requirement.

It is assumed that the charging period corresponds to the peak period chosen as $[10, 30]$ and that travellers' desired arrival time window is $[25, 30]$. In addition, the following parameter values are used

$$\alpha = 1.0, \beta = 0.102 \text{ and } \gamma = 0.4.$$

3.2 Link characteristics

The characteristics of each link l are listed in Table 1, including length (L), free-flow travel time (tt^{ff}), jam density (k^j) and critical density (k^c).

Table 1: Characteristics of all links on the example network

Link (l)	L_l	tt_l^{ff}	k_l^j	k_l^c
0	1.20	1.00	160.00	56.00
1	1.05	0.90	160.00	48.00
2	0.50	0.40	280.00	92.00
3	2.50	2.20	180.00	79.20

The capacity of each link is obtained by $q^{max} = v^f k^c/2$ and given in Table 2, where v^f represents average free-flow speeds and is calculated from $v^f = L/tt^{ff}$. As shown in Table 2, if traffic on both links 0

Table 2: Capacity of each link on the example network

Link (l)	0	1	2	3
q_l^{max}	33.60	28.00	57.50	45.00

and 1 reaches their respective capacities, the sum of their exit flow rates would be $33.60 + 28.00 = 61.60$, which is greater than the capacity of their only downstream link (link 2). Therefore, link 2 can be a bottleneck on the network.

These capacities can be transformed to those values equivalent to real-life ones. For example, Suppose that a time unit is equal to 6 minutes and that a passenger unit is 5 passenger cars or vehicles, then $q_0^{max} = 1680$ passenger cars/hour.

3.3 Travel demand

The total travel demand is assumed to be fixed over the time horizon, specifically

$$D = 2156 \quad \text{passenger units} \tag{12}$$

which implies that the average demand rate is $2156/35 = 61.6$ passenger units/time unit. This average rate is higher than the capacity of any one of the three paths. It is worth noting that we did not set the average rate to the sum of the capacities of Paths 1 and 2 specifically; certainly, a larger value can be chosen and will not change the later findings in this paper.

Since Paths 1 and 2 are shorter than the third one, travellers will first choose them until the resulting travel cost on either of them is greater than the free-flow cost on Path 3. Therefore, taking into consideration the schedule delay cost and the congestion effect, the assumed demand level is high enough to ensure that all three paths are to be chosen at certain times, particularly in the peak period. In addition, since the average demand rate is greater than the sum of the capacities of Paths 1 and 2 and the two paths are shorter than Path 3, the two paths are doomed to be congested in the peak period. This gives rise to the demand for congestion charging.

In addition, as mentioned before, the network is assumed to be empty at time $t = 0$.

3.4 Network Loading

For this research, a dynamic network loading (DNL) procedure is implemented based on Daganzo (1995a) and Lo and Szeto (2002) but a finite difference approximation (FDA) method proposed in Daganzo (1995b) is adopted to solve the kinematic wave (KW) model of traffic flow. The KW model consists of three parts: one is a traffic conservation or continuity equation, the second part is a flow(q)-density(k) relationship and the last one is the identity relationship, i.e. $q = vk$, where v represents traffic speed. The traffic conservation equation is

$$\frac{\partial k(x,t)}{\partial t} + \frac{\partial q(x,t)}{\partial x} = f(x,t) \tag{13}$$

where x represents the location of traffic on a link and $f(x,t)$ represents the net inflow to the link at time t and at location x. In this scenario, $f(x,t) > 0$ only if x corresponds to the entry or exit ends of a link.

The following quadratic q-k relationship is used in this paper

$$q = \begin{cases} (q^{max} - v^f k^c)(k/k^c)^2 + v^f k & 0 \le k \le k^c \\ q^{max}[1 - (k-k^c)^2/(k^c - k^j)^2] & k^c \le k \le k^j \end{cases} \tag{14}$$

If it is required that the curve (14) be smooth at $k = k^c$ then $k^c = 2q^{max}/v^f$ must hold. This relationship is also used in Ge and Stewart (2010), Carey and Ge (2011), Ge and Zhou (2011), Stewart and Ge (2011) and discussed briefly in Carey and Ge (2011).

At node M, when the receiving capacity of the downstream link is smaller than the total demand requiring entry to the link, it is assumed that 60% of the receiving capacity will be assigned to the traffic

from link 0 while the rest goes to link 1. If the outflow from either of the two upstream links is less than their assigned share, the remaining part shall be used by the traffic from the other link. The reader may refer to Section 3.2 of Daganzo (1995a) for the further details of this process and refer to Zhang, Nie and Qian (2013) for the treatment of flow at junctions.

At node O, if the demand for a path at a moment is greater than the receiving capacity of the path, the demand unable to enter shall wait at the entry of the path and the waiting time is included in their travel time. It is assumed that traffic waiting at node O to enter their respective paths does not affect each other. The use of the KW model of traffic flow in DNL enables us to observe the queueing phenomenon, in particular spillbacks. In the later numerical analysis, we omit discussion on queue lengths (or whether there are queues on the road network of interest) in the interest of clarity as consideration of this issue does not produce additional insight in this context.

In computing travel times, the method given in Ge and Carey (2004) was used. The network loading process used here is the same as one in Carey and Ge (2012) except for the link characteristics and parameter values. To make all numerical results comparable, the above settings apply to all numerical experiments in next section.

4 Numerical experiments

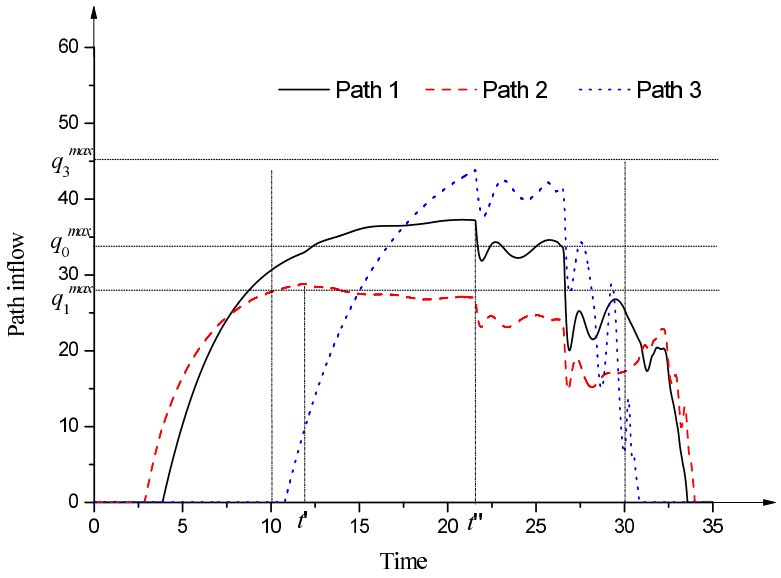
This section first shows that congestion can last a long time at node M on the network in Fig. 1 when there is no congestion charging at all. Secondly, an experiment is conducted to identify both undesired spatial and temporal boundary effects when a constant toll equal to 45% of the no-toll DUE generalised cost is collected on Paths 1 and 2 during the specified charging period. The reason why the proportion of 45% was chosen is that in the bottleneck model the first-best congestion charging toll is exactly 50% of the generalised travel costs but 50% in our experiments resulted in very large boundary peaks or fluctuations. Therefore, 45% was eventually chosen. The third and fourth experiments are to show respectively the performance of a smaller constant toll and that of time-varying toll profiles in mitigating these boundary effects.

4.1 No toll

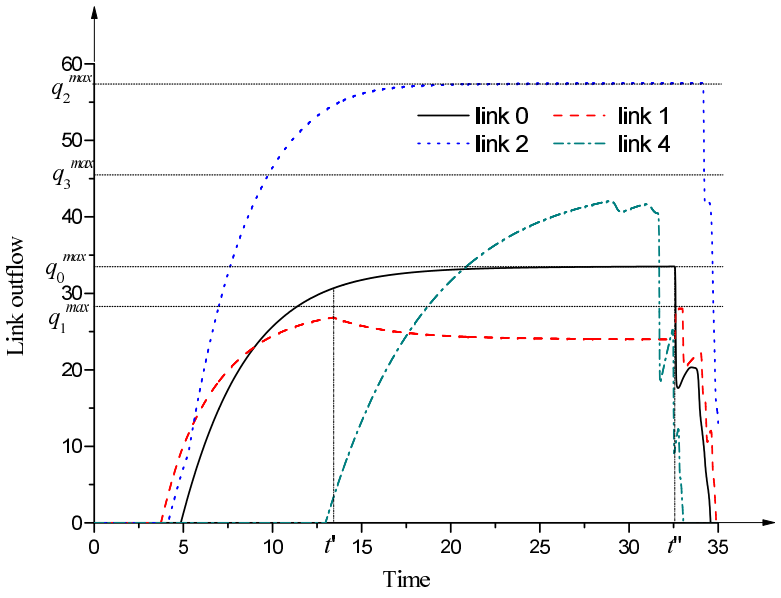
First, we examine the case with no traffic congestion charging at all. Fig. 2 gives the solution profiles at DUE, including (a) path inflow rates and (b) link outflow rates.

Fig. 2(a) shows that the rates of inflow to either path 1 or path 2 are greater than their respective first links' capacities in a certain period. This gives rise to hypercongestion. An obvious occurrence of hypercongestion happens to path 1 from $t' = 12.16$ to $t'' = 21.64$ (see Fig. 2(a)), lasting nearly 10 time units. But the hypercongestion at node M is even worse than this since it begins at $t' = 13.40$ and ends at $t'' = 32.58$ (see Fig. 2(b)), which means that the hypercongestion lasts nearly 20 time units long, i.e. more than half of the time horizon. A further investigation illustrates that the hypercongestion at node O is due to the hypercongestion at node M moving further upstream; but we will not analyse this here because it is not wholly pertinent to following discussions. For link/path 3, the demand is always lower than its capacity, so it is under-utilised.

Now there are two concerns: one is how to reduce the hypercongestion at node M or on paths 1 and 2 and the other is how to improve the utilisation of link/path 3. To do so, a proposal is put forward to charge traffic for entering paths 1 and 2 in the charging period, i.e. from $t = 10$ to $t = 30$.



(a) Path Inflow



(b) Link Outflow

Figure 2: Solution profiles of the no-toll case

At DUE, the generalised cost is equal to 3.42. It is first proposed to charge a constant toll equal to 45% of the no-toll generalised cost at DUE, which is $3.42 \times 45\% = 1.539$. The next subsection will investigate how effectively such a congestion charging scheme could mitigate the two problems raised above and what side effects it would cause.

We are aware of the sharp drops and fluctuations in the late time horizon along the solution profiles in Fig. 2. This phenomenon has appeared elsewhere in the literature on the DTA problem, which shows the complexity of the problem. Firstly, in the single link case reported in Carey and Ge (2004, 2005) the numerical experiments show that if the KW model is used for traffic loading then the link outflow rate has a sharp drop when the given profile of inflow to the link decreases smoothly. Secondly, consider the bottleneck scenario with departure time choice reported in Ge and Stewart (2010), where the dynamic flow loading component is based on the KW model; at equilibrium the profiles in all cases investigated in this reference have some sharp drops of departure rates and/or fluctuations. Thirdly, the solution profiles

from numerical experiments on road networks may be observed: using the KW model-based DNL, Carey and Ge (2012) considers the path choice only with fixed demand and uses the same road network as used in this paper; in this reference the profiles of path outflows show clear drops in outflow rates when the path inflows decrease although the inflows to all three paths change quite smoothly. In addition, such drops and fluctuations have also appeared elsewhere in the literature, e.g. Buisson et al. (1995) and Zhong et al. (2011). It is worth noting that all references mentioned in this paragraph used the KW model or its discrete version to perform dynamic network loading. The reason why such fluctuations exist is beyond the scope of this paper; hence we will not discuss this matter in further detail here.

4.2 Fixed toll on paths 1 and 2 equal to 45% of the no-toll DUE generalised cost

When a constant toll $c_p^c(t)$ is implemented in $[\tau_l^p, \tau_r^p]$ on path 1 or 2, the following technique is used to smooth out the discontinuity on the path travel cost profiles

$$c_p^c(t) = \begin{cases} \bar{T}\left[1+\sin\left(\frac{t-\tau_l^p}{\Delta}\pi\right)\right]/2 & t \in [\tau_l^p - \frac{\Delta}{2}, \tau_l^p + \frac{\Delta}{2}] \\ \bar{T} & t \in [\tau_l^p + \frac{\Delta}{2}, \tau_r^p - \frac{\Delta}{2}] \\ \bar{T}\left[1+\sin\left(\frac{\tau_r^p-t}{\Delta}\pi\right)\right]/2 & t \in [\tau_r^p - \frac{\Delta}{2}, \tau_r^p + \frac{\Delta}{2}] \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

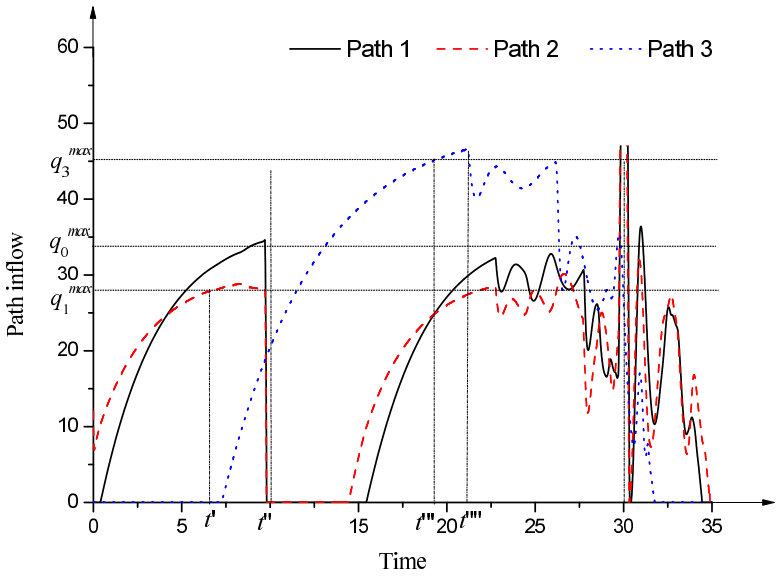
where $\bar{T}$ represents the maximum toll level and Δ is the width of the ad hoc buffer intervals on the two ends of the charging period. In this example, $\bar{T} = 1.539$ and $\Delta = 0.6$. The reason why this technique was used is to make the chosen algorithm able to carry out its task satisfactorily. Although this might have reduced the magnitude of boundary effects of congestion charging since it smooths out the change in the tolls on the two ends of the charging period, it did not affect the later discussions and findings. This technique was also used in the later numerical computing.

Fig. 3 gives the DUE solution profiles of path inflow and link outflow rates under the tolling profiles given in Eq. (15). It can be seen clearly in Fig. 3(a) that more travellers depart outside the congestion charging period and more are travelling on the no-toll path (i.e. path 3) than when no toll is collected at all on the network. Specifically, when there is no toll, only 12% travellers would like to depart before $t = 10$ whereas once the toll is introduced, nearly 23% leave before the start of charging. The implementation of the charging scheme can also make those departing after $t = 30$ increased from 6% to 9%. Overall, due to the implementation of congestion charging, travellers departing within the charging period have decreased by 14%, specifically from 82% down to 68%.

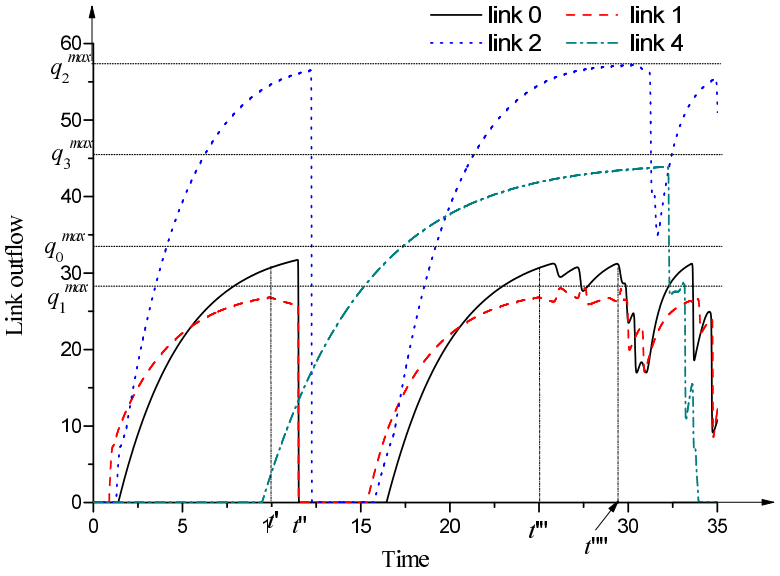
In addition, after the congestion charging is introduced, the total travellers choosing to use link/path 3 have increased from 28% to 37%, i.e. by 9%.

It should also be noted that the hypercongestion period at node M has been shortened after the implementation of the charging scheme. As shown in Fig. 3(b), hypercongestion only happens during $[t', t''] = [10.05, 11.51]$ and $[t''', t''''] = [25.00, 29.60]$ and in total lasts 6.06 time units. This means that the hypercongestion period has been reduced by $(19.18 - 6.06) \times 100/19.16\% = 68.40\%$. In addition, during the hypercongestion period, the inflow to path 1 is always lower than the capacity of the first link of the path and the inflow to path 2 is only a small amount higher than the capacity of the first link of the path for a very short while.

Clearly, the charging scheme has improved the utilisation of link/path 3 and alleviated hypercongestion on paths 1 and 2, in particular at node M. These are the benefits from the implementation of the congestion charging scheme.



(a) Path Inflow



(b) Link Outflow

Figure 3: Solution profiles given constant toll equal to 45% of the no-toll DUE generalised cost

Meantime, the scheme has also generated some undesired side effects.

4.2.1 Temporal boundary issues

As shown in Fig. 3(a), the rates of inflow to paths 1 or 2 are greater than the capacity of their respective first link at either end of the charging period. Before $t = 10$, hypercongestion at node O happens during $[t', t''] = [6.98, 9.68]$ and lasts about 2.70 time units. After $t = 30$, the inflow to both paths 1 and 2 jumps to a level much higher than their respective maximum entry rates and then fluctuates quickly. Actually, under no toll, the inflow to the two paths never exceeds their respective maximum entry rates outside the period $[10, 30]$. Hypercongestion of this kind, happening at either end of the charging period, is a temporal boundary issue caused by congestion charging.

In addition, just after the charging starts, the inflow to either of the two paths going through the charging area is zero for around 5 time units; this is because the congestion charging scheme has forced

travellers to depart earlier or later or take path 3 that does not go through the busiest link (i.e. link 2) in the charging area. This big drop in the demand for entering the charging area just after the start of charging is another kind of temporal boundary effect caused by congestion charging. In reality, the inflow to those paths going through the charging area may not drop so much (mainly due to the relaxation in practice of Assumption 2) but if a charging scheme has not been designed properly it could cause such a big drop in the travel demand for coming into the charging area just after the start of charging.

4.2.2 Spatial boundary issues

As mentioned previously, the introduction of congestion charging makes more travellers travel on link/path 3, which does not go through the charging/busy area. But the rates of inflow to link/path 3 are greater than its capacity during $[t''', t''''] = [19.10, 21.27]$. That congestion charging makes more than the desired number of travellers move onto those roads just around the edge of a charging zone to avoid paying a congestion charging toll and to create hypercongestion on these roads defines a spatial boundary issue.

Now the question is: Are these boundary effects caused simply by a too high constant congestion charging toll?

4.3 Fixed toll on paths 1 and 2 equal to 25% of the no-toll DUE generalised cost

Now the constant congestion charging toll is reduced from 45% to 25% of the no-toll DUE generalised cost and equal to $\bar{T} = 0.855$. All the rest of the settings are the same as before, including the technique smoothing out the discontinuity on the path travel cost profiles, i.e. Eq.(15). This generates the solution profiles of path inflows and link outflows as shown in Fig. 4.

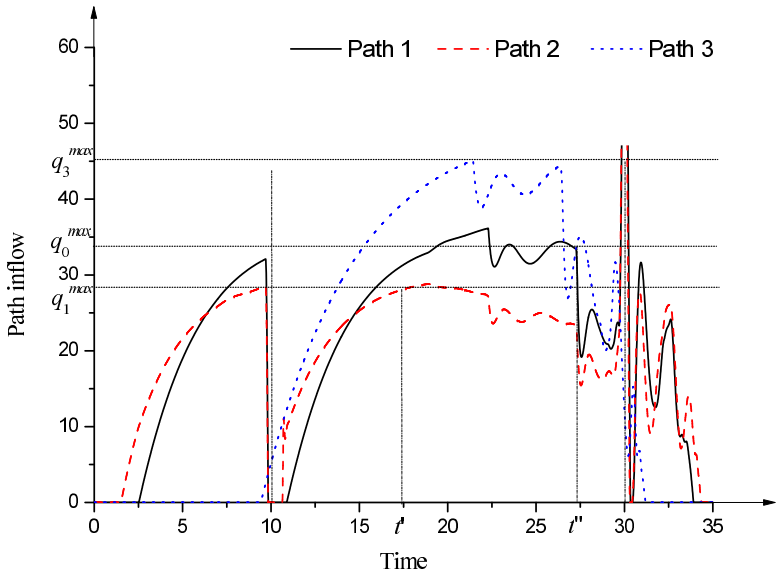
Fig. 4(a) shows that the small toll has significantly reduced the magnitude of both temporal and spatial boundary effects. But, as seen in Fig. 4(b), a long-lasting hypercongestion has returned to node M, where travellers on either of links 0 or 1 have to wait to enter the downstream link 2 and the congestion lasts slightly more than 10 time units, from $t' = 20.50$ to $t'' = 30.69$, although it is not so bad as in the case where no congestion charging is implemented at all. Even if the transportation management authority finds this problem tolerable in practice, the remaining boundary effects should still be worrying.

First, although no significant boundary effects appear before the start of charging, the rates of inflow to the two paths crossing the charging area still have sharp upward and downward fluctuations from the very end onwards of the charging period, see Fig. 4(a).

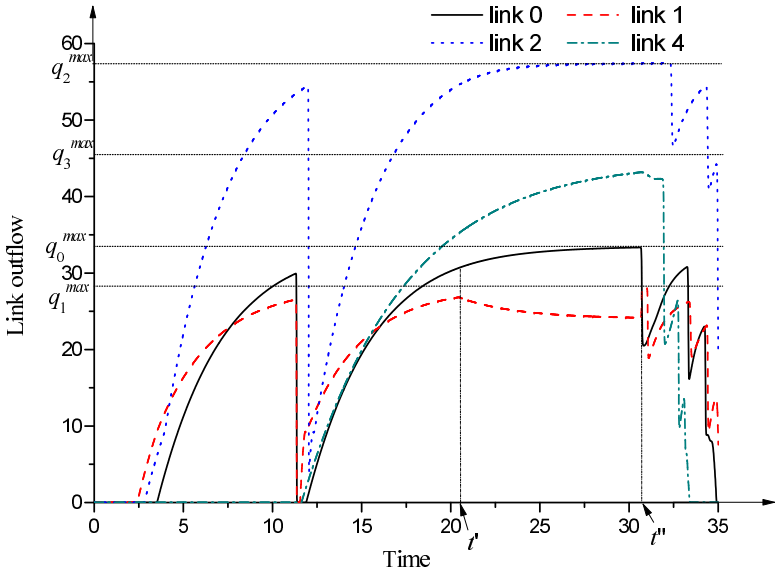
Secondly, at around $t = 10$, as have been seen before, the rates of inflow to path 1 or 2 drop to zero and stay at zero for a while; accordingly, there are sharp drops in the outflow of all four links. Although this just lasts about 1 time unit at node O, the duration in which the inflows to Paths 1 and 2 are a lot lower than their capacities is longer than 5 time units. From the perspective of road pricing design, this is unacceptable; the reason is at least twofold. Firstly, it otherwise could have saved a lot of travellers' travel time in the peak period. Secondly, this clearly results in the inefficient use of the existing road resource and would make social welfare worse off.

Therefore, a small constant toll cannot eliminate the boundary issues satisfactorily and may sacrifice some goals of implementing a congestion charging project.

Now we move on to observe the impacts of a time-varying charging profile on the boundary effects identified previously.



(a) Path Inflow



(b) Link Outflow

Figure 4: Solution profiles given constant toll equal to 25% of the no-toll DUE generalised cost

4.4 Time-varying toll

The following time-varying toll profile $c_p^c(t)$, as shown in Fig. 5, is implemented in this experiment and applies to both paths 1 and 2.

$$c_p^c(t) = \begin{cases} \bar{T} \sin\left(\frac{t-\tau_l^p}{\tau_r^p-\tau_l^p}\pi\right) & t \in [\tau_l^p, \tau_r^p] \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

The greatest toll in this case is equal to 45% of the no-toll DUE generalised cost, i.e. $\bar{T} = 1.539$. The toll starts to grow smoothly from 0 at $t = 10$ to the greatest toll at time $t = 20$ and then decreases smoothly down to 0 at $t = 30$. All the rest of the scenario settings are the same as in the previous experiments.

Fig. 6 gives the DUE solution profiles of departure flow rates along each path and link outflow rates under the time-varying congestion charging scheme. This figure shows that both temporal and

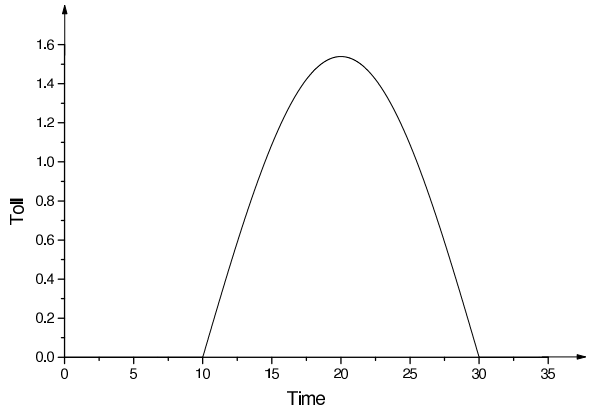


Figure 5: A time-varying toll profile

spatial boundary effects have been significantly reduced or even disappeared when the time-varying toll profile is used. In addition, the hypercongestion period has been shortened dramatically. At node O, hypercongestion within the charging period starts at $t' = 21.87$ and ends at $t'' = 27.03$ (see Fig. 6(a)); it lasts only about 5 time units, which is only half of the length of the corresponding hypercongestion period in the case with no toll at all on the network. At node M, hypercongestion happens during $[t', t''] = [25.51, 33.10]$ (see Fig. 6(b)) and lasts about 7.6 time units, just slightly more than one third of the length of the corresponding hypercongestion period in the zero-toll case. In terms of these, the achievement of the time-varying toll profile can be comparable to those of the constant toll equal to 45% of the no-toll DUE generalised cost.

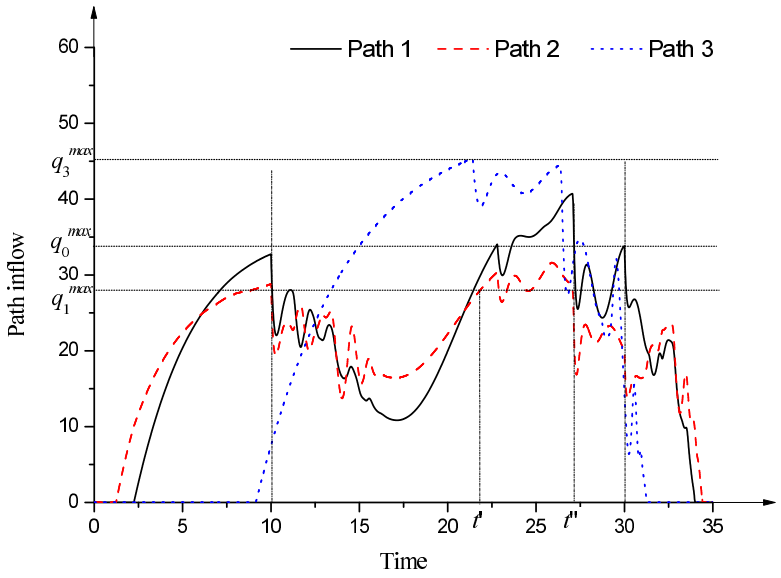
However, it is not all good news. As a matter of fact, a time-varying toll profile can not guarantee the absence of boundary issues. As we can see in Fig. 6(a), the rates of inflow to path 2 are still slightly higher than the capacity of link 1 just before the start of congestion charging although this only happens within a very short period. In addition, at both ends of the charging period, there are quite sharp demand peaks associated with paths 1 and 2. It is worth investigating impacts of the rates of change in the tolls on these fluctuations.

In addition, the rates of inflow to links 0 and 1 within the period of $[12.5, 20]$ are clearly much lower than their respective capacity, see Fig. 6(a). Therefore, the two links are under-utilised during this period under this time-varying toll profile. Fig. 6(b) shows that this phenomenon of under-utilisation looks worse with respect to the outflow rates of all four links because the period within which the rates of outflow from each link are lower than its capacity is a lot longer. How to make best of the existing capacity resource duing the busy period and reduce the travellers' schedule delay costs is another challenge facing the implementation of a time-varying toll profile.

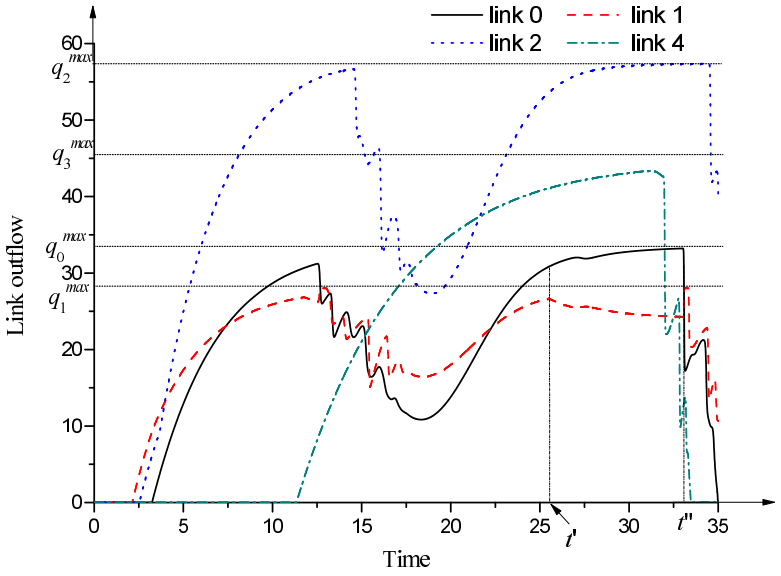
All these call for further investigation.

5 Conclusions

This paper has investigated boundary effects arising from congestion charging, which are undesired effects of congestion charging. A dynamic user equilibrium (DUE) network flow model that captures the behaviour of travellers' simultaneous choice of departure times and trip routes has been used to complete this task. The model is the within-day component of the dual-time-scale DTA model proposed in Friesz et al. (2011). At DUE, all travellers between the same OD pair experience identical generalised costs. The generalised cost consists of travel time costs, schedule delay costs and congestion charging tolls.



(a) Path Inflow



(b) Link Outflow

Figure 6: Solution profiles under time-varying congestion charging toll profile in Fig. (5)

All numerical experiments are set up on a single-OD, four-link and three-path network with a known and fixed total demand for the whole time horizon. Traffic choosing to enter paths 1 and 2 within the charging period is levied a congestion charging toll.

This investigation shows:

1. Congestion charging may not be able to eliminate hypercongestion entirely and can in turn cause undesired boundary effects, temporal and/or spatial. A temporal boundary effect arises from the fact that congestion charging may force too many travellers to depart just before the start of congestion charging or just after the charging period ends so that the demand in the short periods outside both ends of the charging period is greater than available capacity. A spatial boundary effect results from the fact that congestion charging may drive too many travellers to use those paths around the charging area and avoid paying congestion charging tolls so that the demand for these paths are greater than their respective capacities.

2. Constant tolls can readily result in both temporal and spatial boundary effects. In addition, such toll profiles can make the demand for entering the charging area drop dramatically at the very beginning of the charging period, to a level a lot lower than the available capacities, which results in wasting the road capacity resources during the busy period otherwise a large amount of travellers' schedule cost could have been reduced. Furthermore, they can also make the flow rates fluctuate a lot following the end of the charging period.

3. A scheme with a small constant toll designed simply for a congestion charging project can neither mitigate undesired boundary effects efficiently nor achieve the desired goals entirely. Therefore, a well-designed time-varying toll profile, which can efficiently reduce both temporal and spatial undesired effects of congestion charging, is more preferable than a constant one.

4. A good strategy for designing congestion charging is required to generate those schemes that not only result in no undesired boundary congestion/effects but also can effectively and efficiently achieve desired goals of congestion charging. To fulfill this requirement, a simply-picked time-varying toll profile would not be enough. Instead, it requires that a strategy to design congestion charging be developed by means of integrated network management.

As shown previously, the nature of a constant toll being introduced at a precise time means that under Assumptions 1-3 the flow effect on the network will necessarily be to produce a surge of inflow just prior to the charging period commencing and a dip in inflow just after (the effects being reversed at the termination of the charging period). In practice, Assumption 2 (perfect information) will not hold and Assumption 3 (cost minimisation) will only hold partially. DUE is an ideal state associated with Assumption 2; Ge and Zhou (2012) and Ge et al. (2014) provide a state of dynamic user optimum with variable tolerances, at which the travel costs of used paths between an OD pair may not be identical and their differences reflect travelers' tolerances. It is then of importance to investigate how the relaxation of these assumptions will affect the undesired boundary effects investigated in this paper. Another hidden assumption that all travellers share the same set of values of time (i.e. α, β and γ) may also affect boundary effects but they were not touched on in this paper since it is beyond of the intended scope of the study. Whilst it may be hypothesised that relaxing Assumptions 2 and 3 or taking account of heterogeneity in the travellers body would produce less sharp path inflow fluctuations it is unlikely that such relaxations will mitigate these boundary effects efficiently; in fact, in the real-life traffic where these assumptions are not satisfied, such undesired boundary effects of congestion charging have been recorded and discussed (e.g. TfL 2007, Salmon 2010). Therefore, an efficient removal of the undesired boundary effects will still require a well-designed time-varying toll profile, although the adverse effects reported in this paper are likely to be somewhat reduced under more realistic assumptions.

This paper focussed on the use of time-varying toll profiles to mitigate boundary effects while only the simultaneous choice of departure times and trip routes is considered. Actually, there are also other ways available for this purpose. Generally, optimising toll locations in real networks can minimise the relocation of hypercongestion. For example, it may be possible that the mitigation for spatial boundary effects would be in the scheme design such that a buffer surrounding the edge of the tolled area was introduced with decreasing tolls applied in a spatial as well as a temporal manner. However, such a scheme would by its nature be more complicated than a simple scheme and planners must therefore determine a sensible balance between technical benefit of the scheme and clarity to the users. Alternatives to charging monetary tolls within a buffer zone might be alternative network management devices such as traffic calming, partial road closures or one way systems to discourage spatial overspill.

An easy way to reduce diversions around the edge of a charging zone is to increase the size of the zone. Therefore, the size of the charging zone itself is also an un-negligible factor that can affect the boundary effects of congestion charging. It is imperative to vary and test this option in designing a congestion charging scheme. This will need to be weighted against a host of other primary factors such as where there is congestion, the availability of transit or other alternatives, and revenue generation, etc.

To mitigate temporal boundary effects, we may seek to adjust the size of a charging period or the start and/or end times of the charging period, as discussed in Ge and Stewart (2010). Some congestion charging schemes have created very large pricing periods (e.g. 6am – 10am for the morning peak period) to minimise the temporal boundary effect. This is not always efficient if a flat toll is implemented, and has likely been to generate revenues rather than meet system performance objectives per se. Therefore, if the key target of a congestion pricing project is to raise the efficiency of the transportation system other than revenues, such long charging periods should be avoided. In addition, such demand management strategies as parking fees, mode shifts, land-use planning and park-and-ride (P&R) may also be able to mitigate efficiently undesired boundary effects of congestion charging.

In summary, whilst temporal issues should be mainly resolved due to careful time-varying toll design such that there are no sudden jumps in the costs to the user, the spatial issues would require alternative scheme design other than a simple cordon. This would require the use of integrated network management such that an overspill into streets surrounding the charged area should not be adversely affected beyond a level tolerated by network planners. A good scheme of congestion charging should be able to achieve desired goals and result in as little undesired boundary effects as possible.

Acknowledgments

The first author gratefully acknowledges support from the National Natural Science Foundation of China (Grant No.: 71171026) and the fourth author's contribution to this research is partially supported by the US National Science Foundation under Grant CMMI-10555555. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the above funding bodies. This paper is based on the 8-page manuscript titled "Investigating boundary issues arising from congestion charging" presented at the 5th International Symposium on Travel Demand Management (26-28 October 2010, Aberdeen, UK). The authors would like to thank those commenting on the earlier manuscript and the presentation, including Professor William H.K. Lam of the Hong Kong Polytechnic University, Professor Gerd Sammer of the University of Natural Resources and Applied Life Sciences in Viena and Professor Hong K. Lo of the Hong Kong University of Science and Technology. They are also greatly indebted to anonymous referee for their helpful comments when this extended version was submitted to the 4th International Symposium on Dynamic Traffic Assignment (4-6 June 2012, Martha's Vineyard, USA) and then to *Transportmetrica B: Transport Dynamics*.

References

[1] Arnott R, de Palma A, Lindsey R (1990) Economics of a bottleneck. *Journal of Urban Economics* 27(1), 111–130.

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

[3] Ban XG, Liu HX (2009) A Link-Node Discrete-Time Dynamic Second Best Toll Pricing Model with a Relaxation Solution Algorithm. *Networks and Spatial Economics* 9(2): 243–267.

[4] Ban X, Lu S, Ferris MC, Liu H (2009) Risk-averse second best toll pricing. *Transportation and Traffic Theory*, Chapter 10 (W.H.K. Lam, S.C. Wong, H.K. Lo eds.), Springer, 197-218.

[5] Banister D 2003. Critical pragmatism and congestion charging in London. *International Social Science Journal* 55(176), 249–264.

[6] Buisson C, Lebacque JP, Lesort JB (1995) Macroscopic modelling of traffic flow and assignment in mixed networks. In: Pahl PJ, Werner H (eds) Proc. of the Berlin ICCCBE Conf. 1367–1374.

[7] Carey M, Ge YE (2004) Efficient discretisation for link travel time models. *Networks and Spatial Economics* 4: 269–290.

[8] Carey M, Ge YE 2005. Convergence of a discretised travel-time model. *Transportation Science* 39: 25–38.

[9] Carey M, Ge YE (2012) Comparison of methods for path inflow reassignment for dynamic user equilibrium. *Networks and Spatial Economics* 12(3), 337–376.

[10] Carey M, Ge YE, McCartney M (2003) A whole-link travel-time model with desirable properties. *Transportation Science* 37, 83–96.

[11] Daganzo CF (1995a) The cell transmission model, part II: Network traffic. *Transportation Research Part B* 29, 79–93.

[12] Daganzo CF (1995b) A finite difference approximation of the kinematic wave model of traffic flow. *Transportation Research Part B* 29(4), 261–276.

[13] Doan K, Ukkusuri SV, Lanshan H (2011) On The existence of pricing strategies in the heterogeneous single bottleneck model and its extensions. *Transportation Research Part B* 45(9), 1483-1500.

[14] Friesz TL, Bernstein D, Smith RL, Wie BW (1993) A variational inequality formulation of the dynamic network user equilibrium problem. *Operations Research* 41, 179–191.

[15] Friesz T, Han K, Neto P, Meimand A, Yao T (2013) Dynamic user equilibrium based on a hydrodynamic model. *Transportation Research Part B* 47, 102–126.

[16] Friesz TL, Kim T, Kwon C, Rigdon MA (2011) Approximate network loading and dual-time-scale dynamic user equilibrium. *Transportation Research Part B* 45(1), 176–207.

[17] Friesz TL, Kwon C, Mookherjee R (2007) A computable theory of dynamic congestion pricing. In: *Proceedings of the 17th international symposium on traffic and transportation theory*, London, 23C25 July 2007.

[18] Friesz TL, Mookherjee R, Yao T (2008) Securitizing congestion: the congestion call option. *Transportation Research Part B* 42 (5): 407–437.

[19] Ge YE, Carey M (2004) Travel time computation of link and path flows and first-in-first-out, in *Proceedings of the 4th International Conference on Traffic and Transportation Studies*, B. Mao, Z. Tian and Q. Sun (eds), 326–335, Beijing: Science Press.

[20] Ge YE, Stewart K (2010) Investigating boundary issues arising from congestion charging in a bottleneck scenario, Chapter 16 of *New Developments in Transport Planning: Advances in Dynamic Traffic Assignment*, C. Tampre, F. Viti and L. H. Immers (eds), 303–326, Publisher: Edward Elgar.

[21] Ge YE, Sun BR, Zhang HM, Szeto WY, Zhou X (2014) A comparison of dynamic user optimal states with zero, fixed and variable tolerances. *Networks and Spatial Economics*, DOI: 10.1007/s11067-014-9243-9.

[22] Ge YE, Zhou X (2012) An alternative definition of dynamic user equilibrium on signalised road networks. *Journal of Advanced Transportation*, 46(3): 236–253.

[23] Geroliminis N, Levinson DM (2009) Cordon pricing consistent with the physics of overcrowding, Chapter 11 of *Transportation and Traffic Theory*, W. H. K. Lam, S. C. Wong and H. K. Lo (eds), 219–240, Springer Science.

[24] Han K, Friesz T, Yao T (2013) Existence of simultaneous route and departure choice dynamic user equilibrium. *Transportation Research Part B* 53, 17–30.

[25] Iryo T (2013) Properties of dynamic user equilibrium solution: existence, uniqueness, stability, and robust solution methodology. *Transportmetrica B: Transport Dynamics* 1(1), 52–67.

[26] Jiang L, Mahmassani HS, Zhang K (2011) Congestion Pricing, Heterogeneous Users, and Travel Time Reliability Multicriterion Dynamic User Equilibrium Model and Efficient Implementation for Large-Scale Networks. *Transportation Research Record* 2254, 58–67.

[27] Laih CH (1994) Queuing at a bottleneck with single and multi-step tolls. *Transportation Research Part A* 28, 197–208.

[28] Laih CH (2004) Effects of the optimal step toll scheme on equilibrium commuter behavior. *Applied Economics* 36(1), 59–81.

[29] Lawphongpanich S, Hearn DW, Smith MJ (Eds.), 2006. *Mathematical and Computational Models for Congestion Charging*. Springer, Berlin.

[30] Lindsey R (2006) Do economists reach a conclusion on road pricing? The intellectual history of an idea. *Economic Journal Watch* 3(2), 292–379.

[31] Lindsey R (2010) Reforming road user charges: a research challenge for regional science. *Journal of Regional Science* 50(1), 471–492.

[32] Lindsey CR, van den Berg VAC, Verhoef E (2012) Step tolling with bottleneck queuing congestion. *Journal of Urban Economics* 72(1), 46–59.

[33] Lo H, Szeto WY (2002) A cell-based variational inequality of the dynamic user optimal assignment problem. *Transportation Research Part B* 36, 421–443.

[34] Lu CC, Mahmassani HS (2010) Modeling heterogeneous network user route and departure time responses to dynamic pricing. *Transportation Research Part C* 19(2), 320–337.

[35] Nie YM (2012) Transaction costs and tradable mobility credits. *Transportation Research Part B* 46(1): 189–203.

[36] Peeta S, Ziliaskopoulos A (2001) Foundations of dynamic traffic assignment: The past, the present and the future. *Networks and Spatial Economics* 1, 233–265.

[37] Ran B, Boyce D (1996) *Modeling Dynamic Transportation Networks: An Intelligent Transportation System Oriented Approach*, 2nd revised edition, Springer, Heidelberg.

[38] Saleh W, Farrell S (2005) Implications for congestion charging for departure time choice: work and non-work schedule flexibility. *Transportation Research Part A*, 39, 773–791.

[39] Seik FT (1997) An effective demand management instrument in urban transport: the Area Licensing Scheme in Singapore. *Cities* 14(3), 155–164.

[40] Salmon F (2010) The congestion pricing debate, cont. at http://blogs.reuters.com/felix-salmon/2010/06/04/the-congestion-pricing-debate-cont/, retrieved on 17 January 2012.

[41] Small KA (1982) The scheduling of consumer activities: Work trips. *American Economic Review* 72(3), 467–479.

[42] Small KA, Verhoef ET 2007. *The Economics of Urban Transportation.* London: Routledge.

[43] Stewart K, Ge YE (2011) Optimising time-varying network flows by low-revenue tolling with fixed and bell-shaped toll profiles, *Proceedings of the 16th International Conference of Hong Kong Society for Transportation Studies* (ISBN: 978-988-98847-9-6), 17-20 December 2011, Hong Kong, pp. 393–400.

[44] Stewart K, Ge YE (2014) Optimising network flows by low-revenue tolling under the principles of dynamic user equilibrium. *European Journal of Transport and Infrastructure Research* 14(1), 30–45.

[45] Teodorović D, Triantis K, Edara P, Zhao Y., Mladenović S (2008) Auction-based congestion pricing. *Transportation Planning and Technology* 31(4), 399–416.

[46] Transport for London (TfL) (2007) *Central London Congestion Charging: Impacts monitoring.* Fifth Annual Report.

[47] Tsekeris T, Voss S (2009) Design and evaluation of road pricing: state-of-the-art and methodological advances. *Netnomics* 10(1), 5–52.

[48] Van den Berg VAC 2012. Step-tolling with price-sensitive demand: Why more steps in the toll make the consumer better off. *Transportation Research Part A*, in press.

[49] van den Berg V, Verhoef ET (2011) Congestion tolling in the bottleneck model with heterogeneous values of time. *Transportation Research Part B* 45(1), 60–78.

[50] Vickrey WS (1969) Congestion theory and transport investment. *American Economic Review* 59, 251–261.

[51] Wu WX; Huang HJ (2014) Equilibrium and modal split in a competitive highway/transit system under different road-use pricing strategies. *JOURNAL OF TRANSPORT ECONOMICS AND POLICY), 48(1): 153-169.*

[52] Wang X, Yang H, Zhu D, Li C (2012) Tradable travel credits for congestion management with heterogeneous users. Transportation Research Part E *48(2): 426–437*

[53] Wu D, Yin Y, Lawphongpanich S (2011) Pareto-improving congestion pricing on multimodal transportation networks. European Journal of Operational Research *210(3), 660–669.*

[54] Xiao F, Qian Z and Zhang HM (2011) The morning commute problem with coarse toll and nonidentical commuters. Networks and Spatial Economics *11: 343–369.*

[55] Xiao F, Shen W and Zhang HM (2012) The morning commute under flat toll and tactical waiting. Transportation Research Part B *46: 1346–1359.*

[56] Yang H, Huang HJ (2005) Mathematical and Economic Theory of Road Pricing. *Elsevier.*

[57] Yang H, Wang X (2011) Managing network mobility with tradable credits. Transportation Research Part B *45(3): 580–594.*

[58] Yao T, Friesz TL, Wei MM, Yin Y (2010) Congestion derivatives for a traffic bottleneck. Transportation Research Part B *44(10): 1149–1165.*

[59] Zhang HM, Ge YE 2004. Modeling variable demand equilibrium under second-best road pricing. Transportation Research Part B *38, 733–749.*

[60] *Zhang HM, Nie Y, Qian Z (2013) Modelling network flow with and without link interactions: the cases of point queue, spatial queue and cell transmission model.* Transportmetrica B: Transport Dynamics *1(1), 33–51.*

[61] *Zhang X, Yang H, Huang HJ (2011) Improving travel efficiency by parking permits distribution and trading.* Transportation Research Part B *45(7): 1018–1034.*

[62] *Zhong RX, Sumalee A, Friesz TL, Lam, WHK (2010) Dynamic user equilibrium with side constraints for a traffic network: Theoretical development and numerical solution algorithm.* Transportation Research Part B *45: 1035–1061.*