Dynamic traffic assignment approximating the kinematic wave model: system optimum, marginal costs, externalities and tolls

Malachy Carey. Institute for Transport Studies, University of Leeds, Leeds LS2 9JT.
Tel: +44 28 90 203 659, Email: m.carey@its.leeds.ac.uk

Abstract

System marginal costs, externalities and optimal congestion tolls for traffic networks are generally derived from system optimizing (SO) traffic assignment models and when these are treated as varying over time they are all referred to as dynamic. In dynamic SO network models the link flows and travel times or costs are generally modelled using so-called 'whole link' models. Here we instead develop an SO model that more closely reflects traffic flow theory and derive the marginal costs and externalities from that. The most widely accepted traffic flow model appears to be the LWR (Lighthill, Whitham and Richards) model and a tractable discrete implementation or approximation to that is provided by the cell transmission model (CTM) or a finite difference approximation (FDA). These handles spillbacks, traffic controls and moving queues in a way that is consistent with the LWR model (hence with the kinematic wave model and fluid flow model). An SO formulation using the CTM is already available, assuming a single destination and a trapezoidal flow-density function. We extend the formulation to allow more general nonlinear flow density functions and derive and interpret system marginal costs and externalities. We show that if tolls computed from the DSO solution are imposed on users then the DSO solution would also satisfy the criteria for a dynamic user equilibrium (DUE). We introduce constraints on the link outflow proportions at merges and inflow proportions at diverges. We also extend the model to elastic demands and establish links with previous dynamic traffic assignment (DTA) models.

Keywords: cell transmission model; system optimum; dynamic traffic assignment; marginal costs; externalities; optimal tolls

1. Introduction

This paper is concerned with deriving system marginal costs, externalities and hence system optimizing congestion tolls for road networks when traffic flows and travel times are varying over time. In view of that it adopts a dynamic system optimizing (DSO) formulation. We present and analyse a DSO model for dynamic traffic assignment (DTA) in which the traffic flows are modelled by approximation to the widely accepted traffic flow model originated by Lighthill and Whitham (1955), Richards (1956), referred to as the LWR model. Since the latter is a differential equation model, continuous in time and space, it is not analytically or computationally tractable for general traffic network modelling and it is more convenient to approximate it by a finite difference approximation, as in Daganzo (1994, 1995a, 1995b). Daganzo (1994, 1995a) developed the cell transmission model (CTM) that approximates the LWR model when the flow-density function is assumed to be triangular or trapezoidal, as in Fig. 1. Daganzo (1995b) extended the analysis to allow general nonlinear flow-density functions as in Fig. 2 and refers to this model as a finite difference approximation (FDA) to the LWR model. For brevity we will often refer to both the CTM and FDA model as the CTM. For each of these models he showed that as the discretisation of time and space is refined to the continuous limit the model converges to a correct solution of the LWR model.

In the above papers and in various later papers the CTM in a network context is usually presented as a simulation model in which traffic at junctions and intersections merges and diverges in fixed proportions at each point in time and route choice is fixed. Later the CTM was used as the network loading component in dynamic traffic assignment models for user equilibrium (e.g. Lo (1999), Lo and Szeto (2002), Szeto and Lo (2004), Carey and Ge (2011) or see reviews of DTA such as Szeto and Lo (2006)). An important reason for using the CTM in this way is that, in traffic assignment models, route choice, and hence the proportions of traffic using the various links, are endogenously determined rather than being

prespecified. To solve the user equilibrium formulation the spatial route allocations are iteratively adjusted until an equilibrium is achieved.

Ziliaskopoulos (2000) reformulated a relaxed form of the CTM as a set of linear constraints and hence developed a linear programming model for the single-destination system optimum DTA problem for a network. The model was further analysed and applied by Waller (2000), Li *et al*. (2003), Alecsandru (2006), Ukkusuri and Waller (2008), Zeng (2009 and Lin and Liu (2010). The present paper introduces a similar system optimizing formulation, though it instead assumes that the flow-density function for each link may have a general nonlinear form rather than the triangular or trapezoidal form usually assumed in the CTM. The latter forms can be thought of as special cases of a general nonlinear form. This yields a nonlinear convex DSO model rather than a linear programme.

Though this paper is concerned with marginal costs, externalities and optimal congestion tolls or prices for road traffic it does not further pursue the various aspects of congestion pricing. For a comprehensive discussion of the mathematical and economic theory of road pricing see Yang and Huang (2005). Also, the focus in the paper deterministic rather than stochastic: some recent stochastic extensions and applications of the CTM can be found in Karoonsoontawong and Waller (2005), Alecsandru (2006), Boel and Mihaylova (2006), Szeto (2008) and Sumalee *et al*. (2010).

For various reasons we assume a general nonlinear form of flow-density function rather than the usual triangular or trapezoidal form. First, it can include the piecewise linear forms as special cases. Second, in some cases one may wish to avoid some properties of the triangular or trapezoidal form. For example, a triangular flow-density function implies that travel time as a function of flow is initially a horizontal line until it switches to backward-sloping (as in Fig. 1(b)), and a trapezoidal flow-density function has a similar implication except that the travel time function has a vertical piece before sloping backwards. Neither form allows an upward sloping travel time function, which is widely used in static traffic assignment. A further and more immediate reason for assuming a nonlinear flow-density function in this paper is that it can be assumed differentiable whereas piecewise linear forms are not. Differentiability is very convenient for the derivation and analysis of marginal costs in this paper. However, for readers who prefer a triangular or trapezoidal flow-density function, or a more general piecewise-linear flow-density function, it is worth noting that a nonlinear differentiable curve can be chosen to fit as closely as we wish to any piecewise linear curve. To obtain a smooth differentiable curve we need only assume an arbitrarily small rounding or smoothing at the break-points of the piecewise linear curve. This rounding can be assumed so small that it does not affect numerical results – is less than the working tolerance in computations.

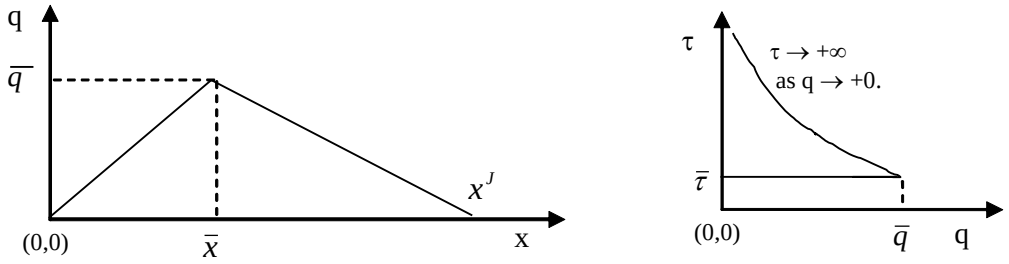

Fig. 1(a). A triangular flow-density or flow-occupancy f'n. Fig. 1(b). Corresponding link travel-time f'n.

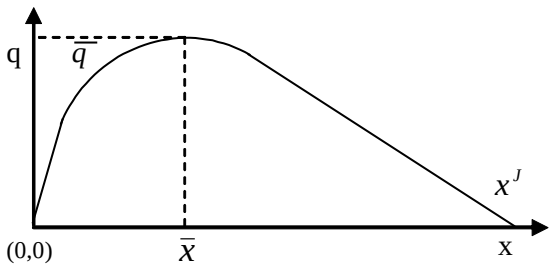

Fig. 2. A nonlinear flow-density or flow-occupancy function.

A system optimising formulation can lead to the phenomenon of "holding back" of flows on some links, which has been identified by a number of authors as a common feature in models that seek to optimize traffic flows on a network over time. By holding back some of the traffic that would otherwise enter a link it may be possible to keep the traffic density from moving onto the downward-sloping (congested) part of the flow density function. Moving onto that portion of the curve would reduce the traffic outflow and eventually cause a reduction in inflow and throughput and hence increase overall system travel times or costs. Holding back of traffic could be used to reduce or prevent that and hence can be interpreted as a desirable form of traffic flow control (as in Carey (1987) and Ziliaskopoulos (2000)). Such flow controls could potentially be implemented by variable speed controls, ramp metering or other methods associated with future intelligent traffic and transport systems.

Section 2 outlines the Daganzo (1995a) finite difference approximation to the LWR model and re-writes the max function from this as inequalities. This allows flow controls of the type mentioned in the previous paragraph. Section 3 extends this formulation from a single link to a network of such links and formulates the traffic DSO assignment problem as a convex nonlinear programme. The merge and diverge proportions at junctions are determined endogenously within the programme so as to minimize travel costs or maximise traveller net benefits. Section 4 derives, analyses and interprets optimality conditions, marginal costs, externalities and optimal tolls. Section 5 introduces constraints on merge and diverge proportions at junctions (nodes) to more realistically reflect actual constraints on these. Section 6 discusses extending to cost-elastic travel demand functions and Section 7 concludes. Appendix A considers relationships between the CTM and the Merchant-Nemhauser model and Appendix B considers other forms of cost-elastic travel demand functions for Section 6.

2. A finite difference approximation to the LWR model

The LWR traffic flow model (Lighthill and Whitam (1955) and Richards (1956) assumes that the flow at each point in space and time (z,t) depends only on the density at that point, and not at any later or earlier points, hence can be stated as a flow-density equation $q(z,t) = Q(k(z,t), z, t)$. Here, as is commonly done, we assume that the link is homogeneous over space and time: if capacity changes along a link it can be sub-divided into homogeneous links. This reduces the flow-density function to

$$q(z,t) = Q(k(z,t)) \tag{1}$$

In the LWR model this is combined with a conservation or continuity equation

$$\partial q(z,t)/\partial z = -\partial k(z,t)/\partial t \tag{2}$$

To approximate (1)-(2) by a difference equation Daganzo (1995b) discretised (1)-(2) as follows. Divide the time span into time intervals $t = 1, \ldots, T$, each of length Δt, and divide the link into $j = 1, \ldots, J$, segments or cells such that the free-flow travel time for each cell is one time interval. This satisfies the Courant-Friedrichs-Lewy (CFL) condition (Courant *et al*. (1967)) that the cell lengths travelled per time

step should not exceed one, which is a necessary condition for the convergence of a finite difference approximation to solve a partial differential equation as the step sizes go to zero. Daganzo shows that the scheme converges without explicitly referring to the CFL condition.

We can assume that the given link is homogeneous, so that all cells will be of the same length *d*. Let k_{jt} denote the cell density, which can be assumed constant along the cell length or can be taken as the mean density in the cell. The flow-density function for a cell can then be written as $q_{jt} = Q_j(k_{jt})$, but it is convenient here to work in terms flow-occupancy $x_{jt} = k_{jt} d$ rather than flow-density k_{jt}. Substituting $k_{jt} = x_{jt} / d$ in the flow-density function gives the flow-occupancy function denoted $g_j(x_{jt})$. From the latter, construct two functions $g_j^+(x_{jt})$ and $g_j^-(x_{jt})$: $g_j^+(x_{jt})$ is obtained by taking the upward sloping part of $g_j(x_{jt})$ and extending it to the right via a horizontal straight line from its peak, and $g_j^-(x_{jt})$ is obtained by taking the downward sloping part of $g_j(x_{jt})$ and extending it back to the vertical axis via a horizontal straight line from its peak. Then, as in Daganzo (1995b), for consistency with the continuous LWR model, the number of vehicles exiting from cell *j* into the next downstream cell *j*+1 in time interval *t* should satisfy

$$v_{jt} = \min\{g_j^+(x_{jt}),\ g_{j+1}^-(x_{j+1,t})\} \tag{3}$$
$$= \min\{(\text{sending capacity of cell } j \text{ in time interval } t),$$
$$(\text{receiving capacity of the next downstream cell } j{+}1 \text{ in time interval } t)\}.$$

Except for the final cell on a link, the outflow v_{jt} from cell *j* in any time interval equals the inflow $u_{j+1,t}$ to next downstream cell *j*+1 in the same interval, thus

$$v_{jt} = u_{j+1,t} \tag{4}$$

as illustrated in Fig. 3. The number of vehicles in cell *j* in time interval *t*+1 is the number present in time interval *t* plus the inflow minus the outflow in interval *t*, thus,

$$x_{j,t+1} = x_{jt} + u_{jt} - v_{jt} \tag{5}$$

Equations (3)-(5) comprise a finite difference approximation to the LWR model (1)-(2). Equation (3) can be rewritten as $v_{jt} \le g_j^+(x_{jt})$ and $v_{jt} \le g_{j+1}^-(x_{j+1,t})$ if we assume for the moment that the outflow v_{jt} is held at the maximum consistent with these inequalities, so that one or other of these inequalities is a strict equality. In that case (3)-(4) can be rewritten as

$$v_{jt} \le g_j^+(x_{jt}) \text{ and } u_{j+1,t} \le g_j^-(x_{j+1,t}) \tag{6.1}$$
$$\text{if the flow } v_{jt} = u_{j+1,t} \text{ is at the maximum consistent with (6.1).} \tag{6.2}$$

The finite difference approximation to the LWR model now consists of (4)-(6.2). The advantage of the inequalities in (6.1), for the purposes of a mathematical programming model below, is that they represent convex sets, since $g_j^+(x_{jt})$ and $g_j^-(x_{j+1,t})$ can be assumed to be concave functions. In contrast, (3) is a nonlinear equation hence represents a nonconvex set.

If the flow $v_{jt} = u_{j+1,t}$ is less than the maximum given by (6.1) then the outflow from cell *j* to *j*+1 will is less than the flow rate given by the CTM equation (3). However, this shortfall may be interpreted as a traffic control system "holding back" the "natural" flow rate, as already mentioned in the Introduction.

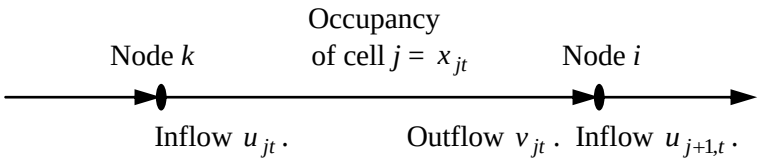

Fig. 3. Inflow, outflow and occupancy (u_{jt}, v_{jt} and x_{jt}) for cell *j*.

3. A system optimising DTA model based on a finite difference approximation to the LWR model

Consider a network consisting of a set of nodes N^O connected by a set of directed links A^O, with individual nodes and links denoted $i \in N^O$ and $j \in A^O$ respectively. Let each of the original links in the network be divided into cells, introduce an artificial node between each pair of neighbouring cells and treat each cell as a link between these neighbouring nodes. Denote this expanded set of nodes as *N* and expanded set of links as *A*, thus $N^O \subset N$ and $A^O \subset A$. Let *B*(*i*) denote the set of links immediately before node *i* (pointing into node *i*) and *A*(*i*) denote the set of links immediately after node *i* (pointing out of node *i*).

Extending the cell conservation equation (5) *to a network*. To extend (5) to the network, simply rewrite it for all cells $j \in A$ in the network, thus

$$x_{j,t+1} = x_{jt} + u_{jt} - v_{jt} \qquad \forall j \in A. \tag{7}$$

Extending the node conservation equation (4) *to a network*. To ensure conservation at all nodes $i \in N$ of the network, let the sum of the inflows to each node equal the sum of the outflows from the node, thus

$$\sum_{j \in A(i)} u_{jt} = D_{it} + \sum_{j \in B(i)} v_{jt} \qquad i \in N \tag{8}$$

where D_{it} is the exogenous travel demand from node *i* to the destination. We can assume an artificial link or cell exiting from the destination node. For the new nodes along the original links (nodes $i \in N$, $i \notin N^O$), $D_{it} = 0$ and (8) reduces to (4), i.e. $v_{jt} = u_{j't}$ where *j*' is the (single) cell immediately after cell *j*.

Extending the exit flow equations (6.1)-(6.2) *to a network*.

Applying (6.1) and (6.2) to all cells on all links we have the following. For each cell *j* the outflow v_{jt} should not exceed the sending capacity $g_j^+(x_{jt})$ of the cell, thus,

$$v_{jt} \le g_j^+(x_{jt}). \qquad \forall j \in A \tag{9.1}$$

and for each cell *j* the inflow u_{jt} should not exceed the inflow capacity or receiving capacity $g_j^-(x_{jt})$ of the cell, thus,

$$u_{jt} \le g_j^-(x_{jt}) \qquad \forall j \in A \tag{9.2}$$

and

either (9.1) or (9.2) is a strict equality $\forall j \in A.$ (9.3)

Condition (9.3) is needed to be consistent with (6.2) and hence (3). However, we wish to construct a mathematical programming model for system optimizing and an either-or constraint such as (9.3) converts a convex programme to a combinatorial problem or 0-1 integer programme, which is nonconvex. Such a programme is potentially very time consuming to solve because of the very large number of cells in the network. Fortunately, in the mathematical programme that we construct later below, condition (9.3) is frequently satisfied in a solution of the mathematical programme without having to be imposed as an explicit constraint. Furthermore, any deviation from satisfying (9.3) may be interpreted as a system optimising flow control, as noted in the Introduction.

A system optimising DTA model

We can now set up a system optimising dynamic traffic assignment model, consisting of minimising the network travel costs for all users, subject to the constraints (7)-(9.2). Let the length of each time interval be 1. Then the total time spent by users x_{jt} in cell j in time interval t is $1\,x_{jt}$, the total time spent by all users on the network is $\sum_{t=1}^{T}\sum_{j\in A} x_{jt}$ and the total cost of this user time is

$$C = \sum_{t=1}^{T}\sum_{j\in A} c_{jt}\, x_{jt} \tag{10}$$

where c_{jt} is the users' cost per unit of time spent in cell j in time interval t. To obtain system optimising flows on the network, minimise (10) subject to (7)-(9.2) and nonnegativity of all the variables. This is set out more formally as follows.

S: Minimise (10)

subject to, for all time intervals $t = 1, \ldots, T$,

$(\alpha_{jt}^{+} \geq 0)$ $\quad v_{jt} \leq g_j^{+}(x_{jt}) \quad \forall j \in A$ (11.1)

$(\alpha_{jt}^{-} \geq 0)$ $\quad u_{jt} \leq g_j^{-}(x_{jt}) \quad \forall j \in A$ (11.2)

(β_{jt}) $\quad x_{j,t+1} = x_{jt} + u_{jt} - v_{jt} \quad \forall j \in A$ (11.3)

(γ_{it}) $\quad \sum_{j\in A(i)} u_{jt} = D_{it} + \sum_{j\in B(i)} v_{jt} \quad \forall i \in N$ (11.4)

$x_{jt} \geq 0,\ u_{jt} \geq 0,\ v_{jt} \geq 0 \quad \forall j \in A.$ (11.5)

The variable in brackets before each equation (i.e., $\alpha_{jt}^{+} \geq 0$, $\alpha_{jt}^{-} \geq 0$, β_{jt} and γ_{jt}) is a dual variable or Lagrange multiplier corresponding to that equation and will be used later. S is a convex programme since the objective function (10) and the constraint sets (11.3) and (11.5) are linear and the constraints (11.1) and (11.2) represent convex sets since both are "less than or equal to" constraints with a linear l.h.s. and a concave function on the r.h.s.

In model S it is implicitly assumed that at a merge node the flows on the links pointing into the junction (node) can enter it in any proportions, subject only to flow conservation. Similarly, at a diverge node it is assumed that the flows on the links pointing out of the junction can exit from it in any proportions. However, the layout of the junction and/ or the traffic control system may impose additional restrictions on these proportions. This is ignored here but is reintroduced in Section 5. Until then it can be assumed that the layout and controls can be adjusted to accommodate whatever solution is given by Programme S.

It is interesting to consider what happens if the equations (11.2) are dropped from the above model S. Equations (11.2) are redundant if all links in the solution of S are uncongested, that is, if the cell occupancies x_{jt} are never in the downward sloping part of the exit-flow functions $g_j(x_{jt})$. Without (11.2), the above model becomes formally the same as the DTA model of Merchant and Nemhauser (1978a, 1978b) except that, in the latter, the equations (11.1) are written as strict equalities whereas here they relaxed to inequalities. The MN model with (11.1) as inequalities was introduced by Carey (1987), where it was noted that it converts the nonconvex optimisation model of MN to a convex model. The latter is formally the same as the above model, except that the MN model has usually been applied to networks with each link treated as a whole link. However, the MN model can equally well be applied after first discretising the whole links into cells and discretising time as in the above model S. Further relationships between the CTM, FDA and LWR models on the one hand and the MN model on the other are discussed in Appendix A and in Carey and McCartney (2004).

Solving programme S.

The above convex nonlinear programme S is linear except for the concave functions in (11.1) and (11.2) hence can be solved easily in various ways. It can be solved by using linear programming if we first piecewise linearise the concave functions (11.1) and (11.2). Many commercial LP (or mathematical programming) packages include a facility for automatically performing such piecewise linearization. Alternatively, the programme S can be solved using available commercial packages for solving convex programming problems with nonlinear constraint (e.g., Minos, Conopt, or other solvers available with GAMS). Or special purpose solution algorithms can be devised to take advantage of the special structure of the model. As already noted, the programme S is similar to the form of DTA model formulated in Merchant and Nemhauser (1978), except for the additional constraints (11.2). Various algorithms were devised to take advantage of the structure of the latter model and these could be extended to the present model. The model S also has another special feature which would speed up its solution: most of the links j in S were formed by discretising the original links in the network, hence are have only a single link (cells) pointing in and out of them. That simplifies the structure and greatly increases the sparsity of the matrix of constraint coefficients, which may make standard mathematical programming algorithms competitive with special purpose algorithms.

4. Properties of the model: system marginal costs, externalities and optimal tolls

To investigate the properties of solutions of Programme S we use the Kuhn-Tucker (K-T) optimality conditions which can be set out as below. These conditions are necessary and sufficient to characterise an optimal solution of S since the objective function and constraint set of S are convex.

The K-T conditions for Programme S consist of the following, for $t = 1, \ldots, T$:

Equations (11.1)-(11.5) (12.0)

Complementarity for the pairs of inequalities in (11.1) and in (11.2).

$$(u_{jt} \geq 0) \qquad -\beta_{jt} \leq -\gamma_{kt} + \alpha^{-}_{jt}, \qquad j \in A(k),\ k \in N \qquad (12.1)$$

$$(v_{jt} \geq 0) \qquad \beta_{jt} \leq \gamma_{it} + \alpha^{+}_{jt}, \qquad j \in B(i),\ i \in N \qquad (12.2)$$

$$(x_{jt} \geq 0) \qquad \alpha^{+}_{jt} g^{+}_{j}{}'(x_{jt}) + \alpha^{-}_{jt} g^{-}_{j}{}'(x_{jt}) \leq c_{jt} + (\beta_{jt} - \beta_{j,t-1}), \quad \forall j \in A \qquad (12.3)$$

Complementarity for the pairs of inequalities in (12.1)-(12.3).

Complementarity or 'complementary slackness' means that, in a solution of the K-T conditions, if either one of a pair of inequalities is a strict inequality then the other one must be a strict equality.

To interpret the above optimality conditions we first consider the system marginal costs (s.m.c's). In a constrained optimisation programme such as S, the Lagrange multiplier or dual variable associated with any constraint is the amount by which the *optimal* value of the objective function will change per unit change in the value of a constant term in the constraint, such as the D_{it} term in (11.4), while holding all other parameters fixed. This is also referred to as the system marginal cost, hence γ_{it} is the s.m.c. of increasing D_{it}. Since a unit increase in D_{it} will move through the network, governed by (10)-(11.5), until it exits at the destination, γ_{it} can be described as the s.m.c. of travelling from node *i* in time interval *t* to the destination.

In a similar way the dual variables in (12.1)-(12.2), i.e. β_{jt}, γ_{kt}, γ_{it}, α^{-}_{jt} and α^{+}_{jt}, can be interpreted as follows. For time interval *t*:

β_{jt} is the s.m.c. of adding an extra unit of traffic to cell *j* in (11.3), i.e. the s.m.c. of travelling from cell *j* to the destination.

γ_{kt} and γ_{it} are the s.m.c's per extra unit of traffic travelling to the destination from node *k* (at the *entrance* of cell *j*) and node *i* (at the *exit* of cell *j*) respectively.

α^{-}_{jt} and α^{+}_{jt} are s.m.c's incurred by the capacity restrictions (11.1) and (11.2) on entering and exiting cell *j*.

The K-T equations (12.1) and (12.2) (i.e. $\gamma_{kt} \leq \beta_{jt} + \alpha^{-}_{jt}$ and $\beta_{jt} \leq \gamma_{it} + \alpha^{+}_{jt}$) contain similar variables, hence to illustrate the difference between them, Fig. 4 places these dual variables alongside the components of the network with which they are associated (β_{jt} beside cell *j*, α^{-}_{jt} and α^{+}_{jt} beside the entrance and exit respectively of cell *j*, and γ_{kt} and γ_{it} beside the nodes at the entrance and exit respectively of cell *j*). Equation (12.1) states that the first variable in Fig. 4 (i.e. γ_{kt}, starting from the left) is $\leq$ the sum of the next two variables. Equation (12.2) states that the third variable in Fig. 4 (i.e. β_{jt}) is $\leq$ the sum of the next two variables.

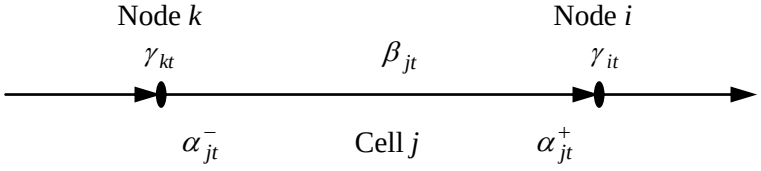

Fig. 4. Dual variables associated with cell *j*.

Proposition 1. Let P_{it} denote the set of time-space paths from node *i* to the destination setting out in time interval *t*. Then in an optimal solution of programme S
(a) the s.m.c. of traversing the utilised time-space paths P_{it} is given by the value the dual variable γ_{it} in the optimal solution of S, hence
(b) is the same for all utilised the time-space paths in P_{it} and
(c) is less than or equal to the s.m.c. of traversing any unutilised time-space path in P_{it}.

Proof. We can show this in the same way as is done for static traffic assignment but since the expressions and summations involved are much more cumbersome than for the static traffic assignment model we do not write them out here in full.
(a) As noted above, this follows from the usual interpretation of dual variables.
(b) Along any utilised time-space path P_{it} the cell and link inflows u_{jt}, outflows v_{jt} and occupancies x_{jt} must be positive which, by complementarity in the K-T conditions, means that the corresponding equations (12.1)-(12.3) will be strict equalities. In that case we can write these equations for each cell along the time-space path and, by sequential substitution, express the γ_{it} at the start of the path as a sum of cell s.m.c. terms along the path. But since the sum for each utilised path is γ_{it} and γ_{it} is independent of path (has no path subscript), the sum is the same for all paths.
(c) Along any time-space path P_{it} that is not utilised, some of the cell and link inflows u_{jt}, outflows v_{jt} or occupancies x_{jt} may be positive (as in (a)) but at least one of them must be zero, otherwise the path would be utilised. Again, by complementarity in the K-T conditions, this means that at least one of the corresponding equations (12.1)-(12.3) will be a strict *inequality*. In that case, when we perform sequential substitution from (12.1)-(12.3) along the time-space path we obtain $\gamma_{it} \leq$ the sum of the s.m.c. terms along the time-space path. ■

4.1. System marginal costs, externalities and optimal tolls

For congested road traffic, an additional or marginal traveller tends to cause an increase in travel times or costs for other users and this increase in costs is referred to as the congestion externality caused by the additional user. It is normally assumed that, when deciding whether or when to travel, each road user takes account only of the travel time or cost that they experience and does not take account of any congestion externality. The total cost caused by an additional (marginal) user of a road path, link or cell is referred to as the system marginal cost (s.m.c.) and is (the cost experienced by a marginal/ additional user) plus (the congestion externality caused by the marginal/ additional user). If a toll just equal to the externality is imposed on each user then the total cost experienced by each user (own cost + toll) becomes equal to the s.m.c. so that the user is induced to take account of the full s.m.c. when making travel decisions. In other words, the user is induced to "internalise" the externality and the resulting user equilibrium will also be a system optimum.

In static traffic assignment models the above optimal tolls are obtain directly from the solution of the model or K-T optimality conditions for the model. We can do that because there are explicit analytic expressions for the link travel times, s.m.c's and externalities and these are included in the model (in the objective function) and in the K-T optimality conditions. Unfortunately, for the CTM/ FDA based DTA model it seems that it is not possible to do that. The reason is that the K-T conditions for this model do not include or yield any variables or analytic expressions for the link travel times or costs experienced by individual users. These are not explicit in the model or its solution. Instead, the link and path travel times are obtained only by applying a computational procedure, as follows, to the solution of S after it is solved.

Computation of path travel times τ_{it}^{p} *and link travel times*. Let τ_{it}^{p} denote the travel time experienced by a user travelling from node *i* at time *t* to the destination via spatial path $p \in P_{it}$. In CTM based models, and other exit flow models, this is computed by a cumulative flow method, comparing the cumulative inflow and outflow from a path, e.g. if the *n*'th user enters a spatial path at time *t*1 and exits from it at time *t*2 then the path travel time is *t*2 – *t*1. For further details of this method of computing travel times see for Cayford, Lin and Daganzo (1997), Tong and Wong (2000) and Lo and Szeto (2002)). Link or cell travel times are computed in the same way, the only difference being in the location of the points at which the cumulative flows are computed.

Definition of DUE:

Let P_{it} denote the set of time-space paths starting from node *i* in time interval *t* and travelling to the destination. Then a flow pattern is a DUE if and only if, for each *it*

(a) the travel time/ cost experienced by users is the same for all utilised time-space paths in P_{it} and

(b) is less than or equal to the travel times/ costs for unutilised time-space paths in P_{it}.

Note that the following discussion and propositions do not provide a general DUE model based on the CTM in the absence of optimal tolls. It instead allows us only to take a DSO model based on the CTM (i.e. S) and from it derive tolls that, if imposed on users, will ensure that the DSO solution is also a DUE. A DUE methodology based on the CTM is developed in Ukkusuri (2002) and Ukkusuri and Waller (2008). In that approach the objective function is not initial fully specified but instead certain cost or penalty parameters have to be iteratively adjusted to force the solution ever closer to a DUE solution. It seems not possible to-date to develop an analytic DUE model analogous to the above CTM based SO model. As a result it has not been possible to take a DUE model based on the CTM, insert tolls in the cost function and obtain an analytic solution to examine how the tolls affect the solution.

Since the link and path travel times can be computed only after solving the Programme S, we have to take a quite different approach to obtaining optimal tolls than is followed in static assignment models.

Proposition 2.

Take a solution of Programme S and in the solution divide the set of spatial paths P_{it} from each *it* into two sets, utilized P_{it}^{u} and unutilized P_{it}^{un}. Define tolls

$$toll_{it}^{p} = \gamma_{it} - \tau_{it}^{p} \text{ for paths } p \in P_{it}^{u}$$

$$toll_{it}^{p} > \gamma_{it} - \tau_{it}^{p} \text{ (e.g. } toll_{it}^{p} = \gamma_{it} - \tau_{it}^{p} \text{ + a constant) for paths } p \in P_{it}^{un}$$

where the τ_{it}^{p} are as defined above and the γ_{it}'s are the values of the dual variables (s.m.c's) from the solution of Programme S or its dual.

If these tolls are imposed on users of the network then they will choose a user equilibrium flow pattern that is identical to the solution of Programme S, i.e. the DUE will also be a DSO.

Proof. The travel times experienced by users on paths $p \in P_{it}$ are τ_{it}^{p} hence if tolls $toll_{it}^{p} = \gamma_{it} - \tau_{it}^{p}$ are imposed on paths in P_{it}^{u} then the cost experienced by users of these paths is $\tau_{it}^{p} + (\gamma_{it} - \tau_{it}^{p}) = \gamma_{it}$ and if tolls $toll_{it}^{p} > \gamma_{it} - \tau_{it}^{p}$ are imposed on paths in P_{it}^{un} then the cost experienced by users of those paths is $\tau_{it}^{p} + toll_{it}^{p} > \gamma_{it}$. This satisfies the above definition of a DUE hence the proposition follows. ■

Proposition 3. The path s.m.c's, travel times and tolls in Proposition 2 can be decomposed into the sum of s.m.c's, travel times and tolls respectively on the time-space links along each time-space path. For example, the s.m.c's for traversing successive links along a time space path are $\gamma_{it} - \gamma_{i't'}$, $\gamma_{i't'} - \gamma_{i''t''}$, etc. , i.e. the differences between the γ_{it} 's for successive nodes. Similarly for link travel times and tolls.

Proof. The s.m.c's for traversing successive links along a time space path are $\gamma_{it} - \gamma_{i't'}$, $\gamma_{i't'} - \gamma_{i''t''}$, etc. , i.e. the differences between the γ_{it} 's for successive nodes.
From this definition of link s.m.c's it follows immediately that their sum along a time-space path is the s.m.c. γ_{it} for the initial node of the path minus the s.m.c. for the final node. The latter is zero hence the sum is simply the path s.m.c. γ_{it}. Replacing the γ 's with τ 's in the preceding sentences gives the same result for link travel times $\tau_{i't'} - \tau_{i''t''}$ and their sums on time-space paths. The tolls are the differences between the s.m.c's and the travel times hence the same result holds for those, i.e. the sum of link tolls along a time-space path equals the path toll. ■

It is worth noting a discretisation issue which arises in summing s.m.c's, externalties or tolls along a time-space path. The s.m.c. of travelling to the destination from node *i* at the *entrance* of link *j* in time interval *t* is γ_{it} and from node *i* at the *exit* of link *j* at time *e*(*jt*) is $\gamma_{i,e(jt)}$. Hence the s.m.c. of a vehicle using link *j*, entering it in time interval *t*, is ($\gamma_{it} - \gamma_{i,e(jt)}$). Note that traversing the link may take a non-integer number of time intervals, hence the exit time *e*(*jt*) may not have an integer value, hence may not correspond to exactly the beginning or end of one of the time intervals *t* in the model. Because of that, the dual variable $\gamma_{i,e(jt)}$, defined above, may not be associated with the equation (11.4) for a specific (integer) time interval *t* , hence would not be immediately available in the solution of programme S. However, in that case we can compute $\gamma_{i,e(jt)}$ by interpolating between the values of $\gamma_{it'}$ for adjacent integer times, that is, compute $\gamma_{i,e(jt)}$ by interpolating between $\gamma_{it'}$ and $\gamma_{i,t'+1}$ where $t' < e(jt) < t' + 1$, for example by using linear interpolation.

We have not shown that the path tolls in Proposition 2 or link tolls in Proposition 3 are always positive. However, in Proposition 4 below we show that if a constant (say c^R) is added to each of the tolls in Proposition 2, the resulting new tolls still satisfy Proposition 2, hence retain a DSO that is also a DUE. Thus, if any of the path tolls from Proposition 2 are negative, a c^R can be added to each path toll to ensure that all path tolls become zero or positive. For example, set c^R equal to the most negative of the path tolls. Also, this additional flat toll c^R brings in a revenue $c^R D$ where $D = \sum_{t=1}^{T} D_t$ hence c^R can be adjusted to bring in any desired level of additional revenue while retaining a DSO that is also a DUE.

Proposition 4. From Programme S construct a new programme S^R by imposing an additional cost c^R on each unit of flow exiting at the destination, i.e. add $c^R \sum_{t=1}^{T} u_{j^D t}$ to the objective function to be minimized, where j^D is an artificial link pointing out of the destination node. This:
(a) increases the total system cost by a fixed amount $c^R D$ where $D = \sum_{t=1}^{T} D_t$,
(b) makes no change in the optimal solution set of S, i.e. $\{ u_{it}^R, v_{it}^R, x_{it}^R \} = \{ u_{it}, v_{it}, x_{it} \}$,
(c) increases the s.m.c. γ_{it} at each time-space node by c^R, so that the new dual solution is $\gamma_{it}^R = \gamma_{it} + c^R$ and

(d) increases each of the optimal path tolls by c^R to $toll_{it}^{pR} = toll_{it}^{p} + c^R$.

Remarks. The above results indicate that the optimal path tolls are not unique but can be scaled up or down by adding c^R to each path toll, without affecting the DSO/ DUE solution. Since (a)-(d) indicate exactly how the optimal solution of S and its dual are affected by imposing an extra cost c^R on each unit of inflow, there is no need to need to actually solve the new programme S^R or its dual to find the new solution. They are already given by (a)-(d).

Note that the tolls $\gamma_{it} - \tau_{it}^{p}$ do not appear in the objective function of programme S; only the additional tolls c^R appear there. Also, note that Programme S is convex programme hence either has a global optimum and a single optimal solution or a convex set of solutions (all of which yield the same optimal value).

Proof. (a)-(b). All of the fixed travel demand $D = \sum_{i \in N} \sum_{t=1}^{T} D_i$ in S has to eventually exit into the artificial link j^D hence $\sum_{t=1}^{T} u_{j^D t} = D$ hence $c^R \sum_{t=1}^{T} u_{j^D t} = c^R D$. But the latter is a constant hence $c^R \sum_{t=1}^{T} u_{j^D t}$ is a constant and adding a constant to the objective function of a convex programme S has no effect on an optimal solution of S, except that the optimal value of the objective function is increased by a fixed amount $c^R D$.

(c). Consider the dual of Programme S. Since S is a convex programme there is no duality gap, that is, the optimal value of S and its dual are equal. The objective function of the dual is $\sum_{i \in N} \sum_{t=1}^{T} D_{it} \gamma_{it}$. We have seen that adding $c^R \sum_{t=1}^{T} u_{j^D t}$ to the objective function of S increases its optimal value by a constant $c^R D = c^R \sum_{i \in N} \sum_{t=1}^{T} D_{it}$ hence increases the optimal value of the dual programme by the same amount, thus $\sum_{i \in N} \sum_{t=1}^{T} D_{it} \gamma_{it}^{R} = c^R \sum_{i \in N} \sum_{t=1}^{T} D_{it} + \sum_{i \in N} \sum_{t=1}^{T} D_{it} \gamma_{it}$. But $\gamma_{it}^{R} = \gamma_{it} + c^R$ solves this equation hence is an optimal solution of the dual.
(d). From Proposition 2 the optimal path toll is $\gamma_{it} - \tau_{it}^{p}$ (i.e. the path s.m.c. γ_{it} minus the path travel time). τ_{it}^{P} is computed from the solution of S and since the latter is unchanged by introducing c^R (see (b)) the path travel times are also unchanged, i.e. $\tau_{it}^{pR} = \tau_{it}^{p}$. Also, from (c), the new s.m.c's are $\gamma_{it}^{R} = \gamma_{it} + c^R$. Hence the new optimal tolls are $toll_{it}^{pR} = \gamma_{it}^{R} - \tau_{it}^{pR} = (\gamma_{it} + c^R) - \tau_{it}^{p} = toll_{it}^{p} + c^R$. ■

The DUE in the above discussion and propositions is an "ideal" rather than "instantaneous" UE. In the DTA literature, two different forms of UE have been generated, namely ideal and instantaneous. In both cases the UE is defined as above, that is, the travel times on utilised paths are equal and are less than or equal to the those for unutilised paths. However, the travel times are defined differently in the two cases. In instantaneous UE the travel times for links on a path are those that obtain at the time of entry to the path while in ideal UE they are the times that will be experienced by travellers when they arrive at those links. The ideal UE is thus more realistic. Instantaneous UE arises because in some network models traffic that enters a link or path at time *t* can exit instantaneously at the end of the link or path at the same time *t*. The DUE in the above discussion and propositions is an ideal DUE since the travel times on each link are those experienced when the traffic arrives at that link. Ideal versus instantaneous for the CTM and FDA model is discussed further in Appendix A.

4.2. Relationships between components of system marginal costs for paths

S.m.c's are used above to obtain optimal congestion tolls, but there are also other possible uses for the s.m.c's obtained from the solution of programme S. For example, in time interval *t* the s.m.c. of travelling to the destination from nodes *k* or *i* are γ_{kt} and γ_{it} respectively hence the s.m.c. of letting a vehicle (a marginal unit of traffic) enter at node *k* instead of node *i* in time interval *t* is ($\gamma_{kt} - \gamma_{it}$). Similarly, the s.m.c. of letting a vehicle enter at node *i* in time interval $t - \tau$ instead of in time interval *t* is ($\gamma_{i,t-\tau} - \gamma_{it}$).

In 4.1 we considered the path s.m.c. that is incurred by an additional unit of traffic entering at node *i*, which is given by γ_{it} the dual variable associated with the conservation equation (11.4) for node *i*. In the discussion below it is instead convenient to consider a path as starting from a link or cell *j* pointing into node *i*. The s.m.c. for this path is given by the dual variable β_{jt} associated with the conservation equation (11.3). We can interpret β_{jt} and γ_{it} as follows, for time interval *t*:

β_{jt} = is the s.m.c. of an additional unit of traffic entering cell *j* and traveling from there to the destination.

γ_{it} = is the s.m.c. of an additional unit of traffic entering at node *i* (at the exit of cell *j*) and traveling from there to the destination.

The discussion in Section 4.1 was concerned mainly with the paths and the original links rather than the cells into which these are divided. However, the discussion below refers mainly to the cells.

For traffic entering cell *j* in time interval *t*–1, the path s.m.c. is $\beta_{j,t-1}$ and by rearranging the K-T condition (12.3) we can express this as

$$\beta_{j,t-1} = c_{jt} + \beta_{jt} - \alpha_{jt}^{+} g_j^{+}{}'(x_{jt}) - \alpha_{jt}^{-} g_j^{-}{}'(x_{jt}) \qquad (12.3')$$

It is unusual for both α_{jt}^{+} and α_{jt}^{-} to be nonzero for a cell *j*. A nonzero α_{jt}^{+} means that (11.1) is binding, i.e., the outflows v_{jt} from cell j are on the nondecreasing (or upward sloping) part $g_j^{+}{}'(x_{jt})$ of the flow-density function. A nonzero α_{jt}^{-} means that (11.2) is binding, i.e., the inflows u_{jt} to cell *j* are on the nonincreasing (or downward sloping) part $g_j^{-}{}'(x_{jt})$ of the flow-density function. It is more usual that only one of (11.1) and (11.2) is binding for a cell *j*. Assuming only one is binding then complementarity implies that either α_{jt}^{+} or α_{jt}^{-} will be zero, which reduces (12.3') to

$$\beta_{j,t-1} = c_{jt} + \beta_{jt} - \alpha_{jt}^{+} g_j^{+}{}'(x_{jt}) \qquad (13.1)$$

if $u_{jt} < g_{j'}^{-}(x_{j't})$ or

$$\beta_{j,t-1} = c_{jt} + \beta_{jt} - \alpha_{jt}^{-} g_j^{-}{}'(x_{jt}) \qquad (13.2)$$

if $v_{jt} < g_j^{+}(x_{jt})$. From (12.2), $v_{jt} > 0$ implies $\alpha_{jt}^{+} = \beta_{jt} - \gamma_{it}$ and, from (12.1), $u_{jt} > 0$ implies $\alpha_{jt}^{-} = -\beta_{jt} + \gamma_{kt}$. Substituting these for α_{jt}^{+} and α_{jt}^{-} above gives

$$\beta_{j,t-1} = c_{jt} + \beta_{jt}[1 - g_j^+{}'(x_{jt})] + \gamma_{it} g_j^+{}'(x_{jt}) \tag{14.1}$$

or

$$\beta_{j,t-1} = c_{jt} + \beta_{jt}[1 + g_j^-{}'(x_{jt})] - \gamma_{kt} g_j^-{}'(x_{jt}) \tag{14.2}$$

Equation (13.1) states that:
(The s.m.c. of travelling from cell *j* to the destination, setting out a time *t*– 1) = (the s.m.c. if setting out one time interval later, at time *t*) + (the cost c_{jt} of waiting that extra time interval) – (the cost or penalty that would have had to be paid to enter the cell in time step *t*, but does not have to be paid since the traffic is already in the cell).
Equation (13.2) can be interpreted similarly, but note that $g_j^-{}'(x_{jt})$ in (13.2) is non-positive, since $g_j^-(x_{jt})$ is the nonincreasing or downward sloping part of the flow-density function, hence this penalty term is added rather than subtracted in (13.2). This is because in (11.2) an additional unit in cell *j* increases x_{jt} so that if $g_j^-(x_{jt})$ is downward sloping this reduces or restricts the inflow u_{jt} on the l.h.s. of (11.2) which imposes a system cost given by the dual variable for (11.2) namely α_{jt}^-.

To help interpret the optimality conditions (14.1)-(14.2) we use the following lemmas.

Lemma 1. Let $g_j(x)$ have the following usual properties:
(a) $g_j(x)$ is a nonnegative concave function which starts from the origin (x, $g_j(x)$) = (0,0), and
(b) $g_j(x_{jt}) \le x_{jt}$, that is, the amount exiting from a cell/ cell in time interval *t* can not exceed its current occupancy x_{jt}. Then
(c) the gradient $g_j'(x)$ at the origin is ≤ 1 and (d) $0 \le g_j'(x) \le 1$.
Proof. (c) follows immediately from the assumptions $g_j(x) = x$ at the origin and $g_j(x) \le x$ for $x \ge 0$.
(d) follows immediately from (c) and the concavity of $g_j(x)$ for all $x \ge 0$. ■

Lemma 2. (a) The r.h.s. of (14.1) is *c* plus a weighted average (convex combination) of β_{jt} and γ_{jt}.
(b) The r.h.s. of (14.2) is *c* plus a weighted average (convex combination) of β_{jt} and γ_{kt}.
Proof. (a). From Lemma 1(c), $0 \le g_j^+{}'(x_{jt}) \le 1$, hence $0 \le (1 - g_j^+{}'(x_{jt})) \le 1$ and (a) follows.
(b). $g_j^-{}'(x_{jt})$ is negative hence (14.2) can be rewritten as $\beta_{j,t-1} = c_{jt} + \beta_{jt}[1 - |g_j^-{}'(x_{jt})|] + \gamma_{kt}|g_j^-{}'(x_{jt})|$ where $|.|$ denotes the absolute value. Almost all relevant empirical evidence indicates that the gradient of the upward sloping (congested) part $g_j^-{}'(x_{jt})$ of the flow-density function is less steep than the upward sloping part $g_j^+{}'(x_{jt})$. This implies $|g_j^-{}'(x_{jt})| \le g_j^+{}'(x_{jt})$, hence $0 \le |g_j^-{}'(x_{jt})| \le 1$ since $0 \le g_j^+{}'(x_{jt}) \le 1$, hence also $0 \le (1 - |g_j^-{}'(x_{jt})|) \le 1$, and (b) follows. ■

An interpretation of the marginal cost equations (14.1)-(14.2)

The dual variable $\beta_{j,t-1}$ on the l.h.s. of (14.1) and (14.2) is the s.m.c. of getting from cell *j* to the destination, for traffic that enters cell *j* in time interval *t*-1. Equations (14.1) and (14.2) express this s.m.c. as a sum of three components, and we will interpret these three components in turn below.

Because of equation (11.3), traffic that enters a cell *j* in time interval *t*-1 is included in the traffic x_{jt} on the cell only in the next time interval *t*, and becomes eligible to exit from the cell only in that time interval (constrained by (11.1)-(11.2)). That ensures that traffic entering a cell in any time interval can not exit from it until the next time interval. The cost of remaining in the cell for this single time interval is c_{jt}, the first term on the r.h.s. of (14.1) and (14.2).

The second and third terms on the r.h.s. of (14.1) and (14.2) have very natural interpretations, as follows. Consider (14.1), which assumes that $v_{jt} \le g_j^+(x_{jt})$ is a strict equality and $v_{jt} \le g_{j'}^-(x_{jt})$ is slack. For traffic x_{jt} present in cell *j* in time interval *t*, two things occur. Some traffic exits from the cell in time interval *t* (an amount $v_{jt} = g_j^+(x_{jt})$) and some remain behind on the cell (an amount $x_{jt} - g_j^+(x_{jt})$). Hence, of any "marginal" increment of the traffic x_{jt} in the cell, the amount that exits in time interval *t* is the first derivative of $g_j^+(x_{jt})$, i.e. $g_j^+{}'(x_{jt})$, and the amount that remains behind is the first derivative of $x_{jt} - g_j^+(x_{jt})$, i.e. $1 - g_j^+{}'(x_{jt})$. Then:

(a) For traffic that remains behind in cell *j* in time interval *t* the s.m.c. of travelling to the destination is β_{jt} hence the s.m.c. for a marginal increment $1 - g_j^+{}'(x_{jt})$ in the traffic remaining in the cell is β_{jt} $(1 - g_j^+{}'(x_{jt}))$.

(b) For traffic that exits from the link *j* in time interval *t*, the s.m.c. of travelling to the destination is the s.m.c. of travelling from the exit node *k* of link *j* to the destination, which is given by γ_{jt}, the dual variable associated with equation (11.4). It follows that the s.m.c. for a marginal increment $g_j^+{}'(x_{jt})$ in the amount exiting the link in time interval *t* is $\gamma_{jt}(g_j^+{}'(x_{jt}))$.

Thus for a unit of traffic that enters the link in time interval *t*-1, some exits from the link in time interval *t*, incurring an s.m.c. of $\gamma_{jt}(g_j^+{}'(x_{jt}))$, and some remains in the cell in time interval *t*, incurring an s.m.c. of β_{jt} $(1 - g_j^+{}'(x_{jt}))$. Summing these two s.m.c. components, plus the cost c_{jt} already explained above, gives equation (14.1).

5. Introducing constraints on flows at merges and diverges in Programme S

In Programme S it is assumed that when there are two or more links pointing into a node, traffic is free to exit from these links in any proportions that will reduce travel costs. However, that is not always possible. For example, if traffic light timings are fixed, then each out-link is allocated only a fraction of the time. We can assume that the traffic light timings can be adjusted to match or facilitate the outflow patterns obtained from the solution of Programme S, so that the solution of S can be interpreted as also designing network flow controls to match the solution. But even in that case there are limits to how much the timings can be adjusted. It is not feasible to reduce signal green times below certain limits. Also, if the merge junction is not signalised, the proportions exiting from the various arms of a merge are determined by the junction layout, numbers of lanes, lane widths, etc., which are not variable in Programme S. These

issues are further discussed in Lin and Liu (2010) which indicates the importance of modeling priorities at merges in the CTM and sets out ways to do this. Similarly, at diverges, traffic may have preferences or requirements for choosing among out-links from the diverge node and additional constraints would be needed to enforce these in Programme S.

We consider how to include such features in the Programme S and what effect that may have on the optimality conditions and their interpretation. For simplicity we follow Daganzo (1995a) in considering a merge node with a single out-link and a diverge node with a single in-link. We assume that the merge node has two or more in-links and the diverge node has two or more out-links. For illustration we assume certain systems at merges and diverges but other control systems are possible. Also, for simplicity below, we use the same notation *j* to denote a link and also to denote the first or last cell of that link.

Merges

Consider a merge junction *i* consisting of two or more in-links $j \in A(i)$ and a single out-link j'. The receiving capacity of the first cell of the out-link j' sets the current throughput capacity of the junction, i.e. $u_{j't} \le g^-_{j'}(x_{j't})$. Let the junction inflows be controlled so that a fraction ρ_{ij} of the time or space is allocated to each in-link $j \in B(i)$ so that $\sum_{j \in B(i)} \rho_{ij} = 1$. This restricts the outflow from the last cell of each in-link to $v_{jt} \le \rho_{ij} u_{j't}$. To introduce this behaviour into Programme S we need only add the following equations to the constraint set of S,

$$v_{jt} = \rho_{ij} u_{j't} \qquad \text{for } j' \in A(i), \text{ all } j \in B(i) \text{ and all } i \in N^M \in N \tag{15}$$

where N^M is the set of merge nodes.

Let λ_{jt} be the dual variables associated with constraints (15) when these constraints are inserted in Programme S. Since the constraints (15) are equalities the λ_{jt}'s can be positive or negative. Inserting (15) in Programme S adds λ_{jt} to the r.h.s. of the K-T equations (12.1) and (12.2). In (12.1) α^-_{jt} is replaced by $\alpha^-_{jt} - \lambda_{jt}$ and in (12.2) α^+_{jt} is replaced by $\alpha^+_{jt} + \rho_{ij}\lambda_{j't}$. In view of that, the new K-T conditions can be interpreted in the same way as before, as in Section 4, except that the α^-_{jt} and α^+_{jt} variables are now extended to include λ_{jt} or $\rho_{ij}\lambda_{j't}$ This is not surprising since, like the α^-_{jt} and α^+_{jt} variables, the λ_{jt} variables are associated with entry and exit from each link.

Since λ_{jt} appears in the K-T conditions it affects the values of the γ_{it}'s which are s.m.c's used in deriving externalities and tolls. For example, equation (12.1), i.e. $\gamma_{kt} \le \beta_{jt} + \alpha^-_{jt}$, becomes $\gamma_{kt} \le \beta_{jt} + \alpha^-_{jt} - \gamma_{jt}$. This in turn affects the values of the externalities and tolls computed in Section 4.1. Though the values are changed the rest of the discussion and results in 4.1 concerning externalities and tolls is unchanged.

Note that (15) adds to the externalities and tolls for the same reason that any congested facility imposes an externality (an additional cost) on other potential users. If (15) is a binding constraint then the flows v_{jt}

or $u_{j't}$ in (15) must have squeezed out other potential users who will thus have to choose a less desirable (more costly) time-space path. λ_{jt} is the s.m.c's that this imposes.

Diverges

In programme S it is assumed that, at a diverge node, traffic is willing and able to exit from the in-link into any or all of the out-links in any proportions. For a single traffic type and a single destination that may be true. However, even in that case drivers may have definite preferences or requirements among out-links. Daganzo (1995a) considers traffic at a diverge node with a single in-link, and assumes that the traffic has fixed preferences among the out-links hence exits to these in fixed proportions. He assumes that traffic also respects first-in-first-out so that if the flow into an out-link is restricted (congested) or blocked that also holds back the traffic for the other out-links. To introduce that behaviour here, let the above exit proportions at a diverge node *i* be ρ_{ij} for all $j \in A(i)$, $\sum_{j \in A(i)} \rho_{ij} = 1$. Let the inflow to the diverge node be $v_{j't}$ so that the inflow to each out-cell is $u_{jt} = \rho_{ij} v_{j't}$, and summing these satisfies the conservation equation $\sum_{j \in A(i)} u_{jt} = v_{j't}$ since the ρ_{ij}'s sum to 1. To introduce this behaviour into Programme S we need only add the following equations to the constraint set of S,

$$u_{jt} = \rho_{ij} v_{j't} \qquad \text{for } j' \in B(i), \text{ all } j \in A(i) \text{ and all } i \in N^D \in N \tag{16}$$

where $N^D \in N$ is the set of diverge nodes.

Let μ_{jt} be the dual variables associated with constraints (16) when these constraints are inserted in Programme S. Similar comments can be made about these as for the λ_{jt}'s for merges above. Since the constraints (16) are equalities the μ_{jt}'s can be positive or negative. Inserting (16) in Programme S adds μ_{jt} to the r.h.s. of the K-T equation in (12.1) and (12.2). In (12.1) α_{jt}^- is replaced by $\alpha_{jt}^- - \mu_{jt}$ and in (12.2) α_{jt}^+ is replaced by $\alpha_{jt}^+ + \rho_{ij}\mu_{j't}$. In view of that, the new K-T conditions can be interpreted in the same way as before, except that the α_{jt}^- and α_{jt}^+ variables are now extended to include μ_{jt} or $\rho_{ij}\mu_{j't}$. Again this is not surprising since again, like the α_{jt}^- and α_{jt}^+ variables, the μ_{jt} variables are associated with entry and exit from each link.

As for merges, the new terms μ_{jt} or $\rho_{ij}\mu_{j't}$ in the K-T equations (12.1) and (12.2) affect the values of the s.m.c. variables γ_{it} which in turn affects the values of the externalities and tolls in Section 4.1, but again, though the values are affected the rest of the discussion and results in 4.1 is unchanged.

6. Extending to cost-elastic travel demands

In the discussion so far we have assumed that the travel demands D_{it} are fixed, that is, they do not depend on any of the other variables in the problem. However, in practice the origin-destination (O-D) travel demands may depend on the travel times or costs experienced or perceived by users, and the latter travel times or costs depend on the traffic flows. Such travel demands are usually referred to as elastic or cost-elastic and can be introduced into Programme S by a slight extension of the well-known method used in models for static traffic assignment. Recall that, in the latter, elastic travel demands are introduced by

taking the cost minimising objective function and replacing it with the maximizing the sum of the integrals of the inverse travel demand functions minus the travel costs.

Thus, to introduce elastic demands, let the O-D travel demand from node *i* in time interval *t* be $D_{it} = d_{it}(c_{it})$ where c_{it} is the cost experienced or perceived by a user travelling from node *i* to the destination, and let the inverse of this be $c_{it} = c_{it}(D_{it})$ where $c_{it}(.)$ denotes $d_{it}^{-1}(.)$. (In Appendix B we consider a different form of elastic demand.) The integral of this inverse function summed over all demand nodes and time intervals is $I = \sum_{t=1}^{T} \sum_{i \in N^O} \int_{D_{it}=0}^{+\infty} c_{it}(D_{it})\, dD_{it}$, which can be interpreted as a measure of benefit to travellers. To maximise net benefit (i.e. the above travel benefit function minus the travel costs (10)) proceed as follows: add the negative of the benefit function (i.e. $-I$) to the objective function of Programme S, treat D_{it} as a variable rather than a constant in constraints (11.4) and leave Programme S otherwise unchanged.

Now consider how the above introduction of elastic demands affects the K-T optimality conditions for Programme S, which are set out at the beginning of Section 4. Since the demand functions $D_{it} = d_{it}(c_{it})$ can be assumed decreasing or non-increasing in c_{it}, the inverse functions $c_{it} = c_{it}(D_{it})$ are decreasing or non-increasing in D_{it}, hence the integral *I* is a concave function and the negative of the integral is a convex function. It follows, as before, that the K-T conditions are necessary and sufficient to characterise an optimal solution of Programme S. The K-T conditions are the same as before except that there is now an additional set of conditions,

$$(D_{it}) \qquad \gamma_{it} = c_{it}(D_{it}), \qquad \forall\ i \in N^{O} \tag{17}$$

Inverting the latter gives $D_{it} = d_{it}(\gamma_{it})$ as intended. We saw in Proposition 1 that the s.m.c. of traversing any utilised time-space path from node *i* to the destination, setting out in time interval *t*, is given by the dual variable γ_{it}. Thus, (17) states that the travel demand D_{it} at each origin node increases up to the point where the s.m.c. γ_{it} of an additional trip is just equal to the travel cost $c_{it}(D_{it})$ that travellers are willing to incur to sustain that level of demand $D_{it} = d_{it}(\gamma_{it})$.

7. Concluding remarks

Above we set out a system optimising model for a traffic network in which the flows within links are handled by a finite difference approximation to the LWR model. From the model we show how to obtain system marginal costs (s.m.c's) and congestion externalities for each link and path of the network and discuss relationships among these. To obtain optimal tolls for paths or links it turns out that we can not follow the usual approach that is well-known from static assignment models, since neither the CTM or FDA based model nor their solution include any explicit expressions or variables for the link travel times experienced by users. These travel times are obtained by a computational procedure only after solving the CTM or FDA based model. We show that if tolls computed from the solution of the system optimising model are imposed on users then the system optimising solution also satisfies the criteria for a user equilibrium, that is, given these tolls, the DUE is also a DSO. For simplicity, the model as initially formulated does not handle restrictions on link outflow proportions at merges or link inflow proportions at diverges but these are introduced later in Section 5. We also extend the model to allow cost responsive travel demands, that is, let the travel demands realised at each origin node, at each point in time, depend on the current costs of travelling from there to the destination. The results obtained for the fixed demand case continue to hold. In an appendix we also set out interesting links between the cell-transmission

model, the present CTM or FDA based system optimising model and one of the oldest models developed for dynamic traffic assignment, namely the Merchant-Nemhauser model.

Acknowledgements

The author would like to thank two anonymous reviewers for their thoughtful, helpful comments and thank the UK Engineering and Physical Sciences Research Council for supporting this research through grant number EP/G051879/2. An earlier version of this paper, Carey (2006), was presented at the *First International Symposium on DTA* (DTA2006).

References

Alecsandru, C.D. 2006. A stochastic mesoscopic cell-transmission model for operational analysis of large-scale transportation networks. PhD Thesis, Dept. of Civil and Environmental Engineering, Louisiana State University. http://etd.lsu.edu/docs/available/etd-07132006-094645/ (20 June 2011).

Beard, C. and Ziliaskopoulos, A. 2006. A system optimal signal optimization formulation. Paper presented at the 85th TRB Annual Meeting, Washington, DC

Boel, R. and Mihaylova, L. 2006. A compositional stochastic model for real time freeway traffic simulation. *Transportation Research Part B* 40, 319–334.

Carey, M. 1987. Optimal time-varying flows on congested networks. *Operations Research* 35(1), 56-69.

Carey, M. 2006. Traffic assignment on networks with time-varying flows, while approximating continuum flows on links. Paper presented at the *First International Symposium on DTA* (DTA2006), University of Leeds, 21-24 June.

Carey, M. and Ge, Y.E. 2011. Comparison of methods for path flow reassignment for dynamic user equilibrium. *Networks and Spatial Economics*. DOI: 10.1007/s11067-011-9159-6, April 2011.

Carey, M. and McCartney, M. 2004. An exit-flow model used in dynamic traffic assignment. *Computers & Operations Research* 31(10), 1583-1602.

Cayford, R., Lin, W.-H. and Daganzo, C.F. 1997. The NETCELL Simulation Package: Technical Description. University of California, Berkeley, California PATH Research Report UCB-ITS-PRR-97-23. ISSN 1055-1425. http://www.ce.berkeley.edu/~daganzo/netcellm.pdf (20 June 2011).

Courant, R., K. Friedrichs and Lewy, H. 1967. On the partial difference equations of mathematical physics", *IBM Journal*, March 1967, pp. 215-234. English translation of the 1928 German original.

Daganzo, C.F. 1994. The cell-transmission model: a simple dynamic representation of highway traffic. *Transportation Research Part B* 28(4), 269-287.

Daganzo, C.F. 1995a. The cell-transmission model, Part II: Network traffic. *Transportation Research Part B* 29(2), 79-93.

Daganzo, C.F. 1995b. A finite difference approximation of the kinematic wave model of traffic flow. *Transportation Research Part B* 29(4), 261-276.

Karoonsoontawong, A. and Waller, S.T. 2005. Comparison of system- and user-optimal stochastic dynamic network design models using monte carlo bounding techniques. *Transportation Research Record*, No. 1923, 91-102.

Li, I.Y., Ziliaskopoulos, A.K. and S.T. Waller, S.T. 2003. A decomposition scheme for system optimal dynamic traffic assignment models. *Networks and Spatial Economics* 3(4), 441-455.

Lin, W.-H. and Liu, H. 2010. Enhancing realism in modeling merge junctions in analytical models for system-optimal dynamic traffic assignment. *IEEE Transactions on Intelligent Transportation Systems* 11(4), 838-845.

Lighthill, M. J. and Whitham, G.B. 1955. On Kinematic waves. I: Flow movement in long rivers II: A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society A* 229, 281-345.

Lo, H.K. 1999. A dynamic traffic assignment formulation that encapsulates the cell transmission model. In A. Ceder (ed.) *Transportation and Traffic Theory*, Pergamon, Oxford, pp. 327-350.

Lo, H.K and Szeto, W.Y. 2002. A cell-based variational inequality formulation of the dynamic user optimal assignment problem. *Transportation Research Part B* 36(5), 421-443.

Merchant, D. K. and Nemhauser, G. L. 1978a. A model and an algorithm for the dynamic traffic assignment problem. *Transportation Science* 12(3), 183-199.

Merchant, D. K. and Nemhauser, G. L. 1978b. Optimality conditions for a dynamic traffic assignment model. *Transportation Science* 12(3), 200-207.

Richards, P.I. 1956. Shock waves on the highway. *Operations Research* 4, 42-51.

Sumalee, A., Zhong, R., Pan, T., Iryo, T. and Lam, W.H.K. 2010. Stochastic cell transmission model for traffic network with demand and supply uncertainties. Presented at The Third International Symposium on Dynamic Traffic Assignment (DTA2010), 29-31 July 2010, Takayama, Japan.

Szeto, W.-Y. 2008. Stochastic cell transmission model under demand and supply uncertainties and its network application. Presented at The Second International Symposium on Dynamic Traffic Assignment (DTA2008), 18-20 June 2008, Katholieke Universiteit Leuven, Belgium.

Szeto, W.Y. and Lo, H.K. 2004. A cell-based simultaneous route and departure time choice model with elastic demand. *Transportation Research Part B* 38(7), 593-612.

Szeto, W.Y. and Lo, H.K. 2006. Dynamic traffic assignment: properties and extensions. *Transportmetrica* 2(1), 31-52.

Tong, C.O and S.C. Wong 2000. A predictive dynamic traffic assignment model in congested capacity-constrained road networks. *Transportation Research Part B* 34(8), 625-644.

Ukkusuri, S. 2002. Linear Programs for the User Optimal Dynamic Traffic Assignment Problem. *Master Thesis*, University of Illinois at Urbana-Champaign.

Ukkusuri, S. and Waller, S. T. 2008. Linear programming models for the user and system optimal dynamic network design problem: formulations, comparisons and extensions. *Networks and Spatial Economics* 8(4), 383-406.

Waller, S. T. 2000. Optimization and control of stochastic dynamic transportation systems: formulations, solution methodologies, and computational experience. Ph.D. Dissertation, Northwestern University.

Yang, H. and Huang, H.-J. 2005. *Mathematical and Economic Theory of Road Pricing*. Elsevier Science.

Zeng, H. 2009. Efficient algorithms for the cell based single destination system optimal dynamic traffic assignment problem. PhD Thesis, University of Arizona, USA.

Ziliaskopoulos, A.K. 2000. A linear programming model for the single destination system optimum dynamic traffic assignment problem. *Transportation Science* 34(1), 37-49.

Appendix A: *Comparing Programme S with the M-N model.*

The Merchant and Nemhauser (1978a, 1978b) model (MN model) is seldom mentioned in connection with the CTM or FDA model, but comparing them sheds interesting light on both, in particular on the "instantaneous" versus "ideal" system optimum, FIFO and causality for both models. Here we compare a mathematical programming version of the FDA model ((10)-(11.5)) and CTM with versions of the MN model. Comparisons of other aspects of these models can be found in Carey and McCartney (2004).

The objective function and constraints of an optimisation version of the FDA model are set out in (10)-(11.5). An optimisation version of the CTM is basically the same, except that in that, for the CTM, $g_j^+(x_{jt})$ and $g_{j'}^-(x_{jt})$ are straight lines, with positive and negative slope respectively. The objective function and constraints of the MN model are the same as (10)-(11.5) but with (11.2) deleted.

The MN model has usually been applied to "whole links", though it has often been remarked that it it more accurate or appropriate to first divide the links into small segments and then apply the MN model. Hence suppose that:

(a) Before applying the MN model, time and space (link lengths) are divided into small segments as in the CTM or FDA models, i.e., divide time into small intervals (say 1 second each) and divide each link into segments or "cells" such that the free flow travel time in each cell is exactly one time interval.]

(b) Let the exit function $g_j(x_{jt})$ consist of only the increasing, or nondecreasing, part $g_j^+(x_{jt})$, as was always assumed in the MN model.

(c) Consider a single destination network, as was assumed in the MN model.

Applying assumption (b) to the FDA model and to the CTM means that the "min" functions in each of these becomes a strict equality, i.e., it reduces to the strict equality version of (11.1), without (11.2). But that reduces the Programme S ((10)-(11.5)) to the original MN model. Or in other words, assumptions (a)-(c) make the MN model the same as an optimisation version of the FDA model or CTM.

"Instantaneous" versus "ideal" system optimum.

It has often been remarked that the MN model or versions of it yield an "instantaneous" rather than an "ideal" system optimum. Ideal and instantaneous traffic assignment are defined in Section 4.1 above. If the MN model is a special case of the CTM (or FDA), then how can it yield an "instantaneous" DSO while the CTM/ FDA based Programme S yields an "ideal" DSO? We will see that the MN model yields an instantaneous DSO only if we treat time in the MN model as continuous rather than as discrete time intervals, while treating links as whole links. To see this, proceed as follows. Consider an increase in the exogenous demand D_{it} at node *i* in time interval *t*. This increases the inflows u_{jt} to the some out-links from the node (via (11.4)), which increases the occupancy $x_{j,t+1}$ of these links in the next time interval *t*+1 (via (11.3)), which increases outflows $v_{j,t+1}$ from these links in time interval *t*+1 (via (11.1)). Thus some of the traffic that enters link *j* at time *t* and will exit at time *t*+1, hence traverses the link in a single time interval. If the time intervals are made sufficiently small or continuous then traffic traverses the link instantaneously even though the link may be long. Proceeding in this way from link to link yields an instantaneous outflow at the destination, which is why the solution of the continuous-time MN model is referred to as yielding an "instantaneous" DSO. However, if we coordinate the discretisation of time and space so that each link (cell) takes at least one time interval to traverse then the time delay between entry and exit from a link will reflect the true travel time. That is, the solution of the MN model can yield an "ideal" DSO. Whether the MN model yields an instantaneous DSO or an ideal DSO depends on how we discretise time and space prior to applying the model.

FIFO and "causality" in the MN model and the CTM and FDA model

It is sometimes stated that the MN model does not satisfy FIFO. If the MN model is a special case of the FDA model (or CTM)), then how can it violate FIFO while the others do not? The answer is, if appropriately interpreted, the MN model does not violate FIFO. The reason why it is sometimes said to violate FIFO is that inflow to a link increases the amount of traffic x_{jt} in the link and it seems that some

of this traffic instantly becomes eligible to exit from the link due to equation (11.1), i.e. outflow $v_{tj} \le g_j(x_{jt})$. This is sometimes explained by saying that, in the MN model, the traffic entering a link "instantly spreads out uniformly along the whole link", so that some of it becomes eligible to exit immediately together with traffic that entered earlier, thus violating FIFO. However, an alternative interpretation of $v_{tj} = g_j(x_{jt})$ is that, when additional traffic enters the link, the sequence order of traffic on the link is maintained and the traffic x_{jt} on the link instantly spreads (readjusts) itself uniformly along the link, retaining its FIFO sequence order. That ensures that traffic exits in the same order that it entered, hence is consistent with FIFO. However, this interpretation implies that the MN model exhibits at least a minor violation of "causality". An traffic assignment model is said to violate causality if traffic at any point in time in the model affects the behaviour of traffic that entered earlier, instead of affecting only traffic that entered later. The above interpretation of the MN model means that traffic entering link causes the existing traffic on the link to adjust elastically forward or backward on the link so that it exits sooner or later than it otherwise would have, which violates causality. If the links and time intervals in the MN are long, then this causality violation can be significant. Conversely, as the discretisation (of links and time) in the model is refined, the causality violations are reduced and in the continuous limit they do not occur. Not that this same causality violation is also present in the CTM and FDA model. These models are defined as having a very fine discretisation of space and time, but with rougher discretisation they would exhibit a slight violation of causality. This simply illustrates that the fact that any discrete approximation in almost any model tends to introduces some loss of accuracy.

Appendix B *Introducing an aggregate demand function for each node*

In Section 6 we assumed a separate demand function for each node *i* in each time interval *t*. It is sometimes assumed instead that only the travel demand function relates only to the aggregate travel demand $D_i = \sum_{t=1}^{T} D_{it}$ at each node *i*, thus $D_i = d_i(c_i)$. In that case we can extend the discussion in Section 6 as follows. The inverse demand function is then $c_i = c_i(D_i)$ where $c_i(.)$ denotes $d_i^{-1}(.)$ and the integral of this, summed over all demand nodes and time intervals, becomes $I = \sum_{t=1}^{T} \sum_{i \in N^O} \int_{D_i=0}^{+\infty} c_i(D_i)\, dD_i$. As before, add (–*I*) to the objective function of Programme S, but also include an extra set of constraints,

$$(\eta_i) \qquad D_i = \sum_{t=1}^{T} D_{it} \qquad \forall\ i \in N^{\mathrm{O}} \tag{18}$$

and treat D_{it} as a variable in (18) and (11.4). The K-T conditions (17) corresponding to the demand variable D_{it} now become

$$(D_{it}) \qquad \gamma_{it} = \eta_i, \qquad \forall\ i \in N^{\mathrm{O}} \tag{19}$$

which means γ_{it} is the same for all *t* hence reduces to γ_i. There is also a new set of K-T condition corresponding to the aggregate demand variable D_i, thus

$$(D_i) \qquad \eta_i = c_i(D_i), \qquad \forall\ i \in N^{\mathrm{O}} \tag{20}$$

Combining (19) and (20) gives $\gamma_i = c_i(D_i)$ which is similar to (17), i.e. $\gamma_{it} = c_{it}(D_{it})$, hence the discussion and interpretation following (17) above continues to hold in the present case. A significant difference between (17) and $\gamma_i = c_i(D_i)$ is that the latter is independent of *t* and inverting gives $D_i = d_i(\gamma_i)$ so that the demand is the same in each utilised time interval *t*. This arises here and in other

DTA models when users have only an aggregate demand over time (i.e. $c_i = c_i(D_i)$ subject to (18)) and no preference for setting out at any particular time *t*. If the travel time (here the travel s.m.c.) was higher for entering at certain times, then users would simply keep switching to other entry times until the entry costs are the same at all entry times.

The travel demands $D_i = d_i(\gamma_i)$ indicate that users have no preference as to when they start their journeys and the model S contains no preferences as to when they end their journeys. That is typically considered unrealistic for the journey to work, since travellers usually have a desired or planned work start time or time interval. That can be introduced here in the usual way by inserting costs or penalties per unit time that the users are early or late at their destination. These costs can be attached to the inflows u_{j^*t} to a destination link j^*, for example by adding an a cost $(t_a - t)c_a$ for arriving per minute earlier than time t_a and a cost $(t - t_b)c_b$ per minute later than time t_b. This will not significantly affect the discussion or analysis in the rest of this paper. It simply means that in the expressions for path marginal costs γ_{it} and β_{jt} the cost associated with last link j^* will be $(t_a - t)c_a$ or $(t - t_b)c_b$ instead of being c_{j^*t} or zero.