Improved road usage through congestion-free route reservations

Charalambos Menelaou
Ph.D. student*
Phone: +357-22-893459
cmenel02@ucy.ac.cy

Stelios Timotheou
Research Associate*
Phone: +357-22-893459
timotheou.stelios@ucy.ac.cy

Panayiotis Kolios
Research Associate*
Phone: +357-22-893459
pkolios@ucy.ac.cy

Christos G. Panayiotou
Associate Professor*
Phone: +357-22-892298
christosp@ucy.ac.cy

*KIOS Research Center for Intelligent Systems and Networks
University of Cyprus
P.O.Box 20537
CY-1678 Nicosia
Cyprus
Fax: +357-22-893455

A Paper Submitted for Presentation at the 2017 Annual Meeting of the Transportation Research Board and Publication in the Transportation Research Record

Total words = 4980 + 250*8 + 500 (6 Figures + 2 Tables + References) = 7480
November 15, 2016
ABSTRACT

Road traffic congestion occurs as demand exceeds the capacity of particular road segments. Its consequences (including, travelling delay, fuel consumption, and emission of pollutants) have a major impact on cities, and thus interest for new, and innovative solutions to the problem remain high despite great efforts made to alleviate the problem. Interestingly, recent literature has shown that better management and control mechanisms could significantly curb its effects.

One such mechanism is elaborated in this work, where route reservations are made across congestion-free road segments. The proposed route reservation scheme assumes that vehicles communicate their origin-destination pairs to a controller that identifies and reserves road segments to be traversed in both the space and time dimension. The scheme is mathematically formulated and two algorithms with complementary objective functions are discussed. In the first case, vehicles are routed through road segments that do not exceed their critical density while minimizing the destination arrival time. In the second case, road segments below critical density are used while reservations are made in such a way to balance traffic across the available alternative routes. Analytical and simulations results demonstrate the great benefits that can be realized by applying the proposed solutions.
INTRODUCTION

Road traffic congestion is the lead cause of various negative consequences such as driver frustration, environmental pollution and non-predictable vehicle travel times. Despite the great effort to alleviate its effects, the problem still pertains due to the high demand experienced along particular road segments while a large portion of the rest of the network remains under-utilized (1). Interestingly, real-time state information on road conditions would result to oscillations in the network state since all rational drivers would opt to traverse through non-congested road segments. These key insights emphasize that congestion can be significantly alleviated through effective control.

One such mechanism to achieve congestion alleviation is through efficient vehicle routing. Vehicle routing involves minimizing the transit time of vehicles from source to destination. To achieve this, various mobility characteristics need to be taken into account such as critical density and maximum capacity of different links, or even mathematical functions that describe the relationship speed-density-flow. However, accurate knowledge of these characteristics for each network link is very difficult if not impossible to achieve. One way to overcome this issue is through the use of the Macroscopic Fundamental Diagram (MFD) (2) which relates space-mean flow, density and speed of large regions of a network. An MFD is in the free-flow regime when the density of its region is below the critical density, otherwise it is in the congested regime (3). It should be emphasized that in the free-flow regime, vehicle dynamics can be estimated using free-flow speed conditions. On the other hand, slight changes on the region state within the congested regime can result in large and often unpredictable variations of speed, so that the behavior of the road network cannot be effectively estimated or predicted (4).

Therefore, simple control mechanisms can be employed to maintain traffic flow below the critical density of each road segment and improve the network utilization. As such, this work investigates a novel route reservation scheme which ensures that each vehicle is scheduled to transit only along congestion-free road segments. The proposed scheme discretizes the time horizon into slots and a control unit keeps track of the number of vehicles that have reserved slots along the different road segments. Whenever a route request is received by the controller, a new path from the origin to the destination is computed based on the slot availability. Road segments become temporarily unavailable whenever reservations reach the critical density and are reconsidered only when the allocated time slots have elapsed.

Evidently today’s information and communication technologies can effortlessly support such a scheme, while the emergence of connected and autonomous vehicles can have significant gains from the proposed solution. For simplicity, in this work we consider a centralized control unit that assumes responsibility of all route reservations. Such schemes are well suited in the era of autonomous and connected vehicles which allow vehicle-infrastructure interaction and autonomous decision making by the vehicles. In the considered centralized scheme, a vehicle sends out a reservation request to a central control unit indicating its origin-destination pair. In return, the control unit responds with the path that should be followed and reserves the road segments during the time-slots that
are expected to be traverse.

Two route reservation algorithms are derived which constitute the contribution of this paper. The first algorithm seeks to navigates vehicles through non-congested road segments while minimizing the destination arrival time for each vehicle. This problem had been shown to be NP-complete (5) and the proposed algorithm can achieve an optimum solution in pseudo-polynomial time. The second algorithm seeks to navigate vehicles through congestion-free road segments while minimizing the load variance of the overall traffic. Traffic load balancing can lead to better network performance as it improves the homogeneity of the network, alleviating in this way various unwanted phenomena, such as spillbacks. Another important aspect of the proposed schemes is that vehicles are allowed to wait at their origin before commencing their trip, if this action leads to smaller overall cost. Waiting at the origin is considered for both aforementioned algorithms.

Allowing waiting only at the origin means that these vehicles do not occupy space in the transportation network; in this way, our approach removes waiting in congested situations. In addition, having free-flow conditions has significant advantages:

- Increased road utilization as there is no capacity drop due to congestion.
- Elimination of the adverse effects of congestion such as unnecessary fuel consumption, lost time, and health problems.

The rest of the paper is organized as follows. Section 3 discusses the relevant literature, while Section 4 introduces the route reservation architecture. Section 5 mathematically describes the route reservation problem and Section 5 formulates the two distinct problems called the Earliest Destination Arrival Time problem (EDAT) and the Traffic Load Balancing problem (TLB). Sections 6 and 7 introduce the algorithms for the solution of EDAT and TLB, respectively. Simulation results included in Section 8 demonstrate the potential benefits that can be earned from applying the proposed solutions. Finally, Section 9 concludes this work and discusses future work.

RELATED WORK

Currently, gating and perimeter control approaches constitute the state-of-the-art congestion mitigation mechanisms. *Gating control* aims to regulate the amount of traffic that resides inside a homogeneous region by allowing external traffic to enter if the critical density of the region’s MFD has not been reached (6, 7); this is achieved using street closures and traffic signal control. To avoid the extensive amount of data needed for the characterization of regions with distinct MFDs, a reduced MFD was employed for gating which was constructed using real-time measurements (8). *Perimeter and boundary control* is applied in multi-region networks to regulate traffic exchange between regions with homogeneous MFD characteristics and the outside world (9, 10). Similar to gating, vehicles are allowed to enter within homogeneous regions only if the critical density has not been reached. Since control is made at the level of specific regions, the approach is simple to implement with existing road infrastructure (11). Unfortunately, no control
mechanisms are in place within particular regions, and thus congestion can occur anywhere within the region. This is especially true for traffic that originates from within the region.

Recent literature also considers routing techniques using online or offline traffic state estimation. Traffic state estimation either through static or stochastic models (as elaborated in (12)) has been used to guide routing decisions based on online predictions of the travel times along each road segment and predictions of the speed conditions for each vehicle at future time-slots, (13) (14). In their majority, scheduling decisions are made through shortest-travel-time paths according to these state estimates, without considering the negative effects that may occur when the selected road segments exceed their critical density (i.e., congestion is not considered in the decisions) (15). Hence, these algorithms do not consider the changes in traffic state when scheduling decisions are made, and thus there is no guarantee that traffic state will not experience congestion. There is also the possibility of a capacity drop along particular road segments that experience congestion in which case the capacity can be reduced up to 15%, contributing further to the unpredictability of vehicle speeds and network dynamics (16). An additional disadvantage of using traffic state estimation in routing, is that estimation is required for all links of the network, which is usually not the case as state estimation is only available for the main road links. Finally, the authors of (17) utilize the space-time network for simultaneous route guidance and traffic signal optimization.

Recent research efforts aim towards the integration of routing guidance in MFDs. In (18), it was shown that the shape of MFD and the size of hysteresis loops can be influenced by vehicle accumulation and redistribution of network traffic via appropriate traveller information. In addition, the authors (19) developed routing algorithms that distributed traffic load across a larger network area to significantly improve the overall travelling time. An approximate dynamic traffic assignment model to establish dynamic stochastic user equilibrium conditions was also proposed in (20), illustrating that such approaches outperform other traffic assignment solutions. Finally, a route choice strategy was developed in (21) to alleviate congestion in urban areas, by considering the effect of aggregated regional and partially known sub-regional information.

Unlike the aforementioned literature, the proposed scheme makes sure that route reservations are made along road segments that have not reached their critical density and to do so, time slots are used to reserve individual routes. Evidently, time-slot reservations are not new in the literature. Similar schemes have been used by Air Traffic Management and Control Systems (ATM/ATC) in much the same way to solve the popular ground holding problem. Specifically, to increase their runway capacity, airport ground control allocates specific time slots for each airplane that requests take-off or landing. Time slots are shared among arrival and departure flights and airplanes are instructed to follow their schedules without any delays or deviations. In doing so, the airport efficiency is significantly improved as shown in (22) and (23).
The proposed architecture is used to support efficient route reservations while preventing congestion by ensuring that the traffic of each segment is below critical density. The proposed architecture is depicted in Figure 1, showing a Road Site Unit (RSU) that acts as a central controller responsible of monitoring the utilization of each road segment and for reserving routes for arbitrary origin-destination pairs. Reservation decisions are made by the route-reservation algorithms running at the RSU and routes are communicated to requesting vehicles which are in turn responsible to traverse them. As indicated above, free-flow conditions are maintained by the reservation algorithms. The RSU may impose a waiting period at the origin if the traveling cost is minimized by doing so.

A vehicle entering the particular region of the network, sends its origin-destination pair to the RSU. The RSU replies with the time that the vehicle should start its journey and the road segments that should be traversed. Thereafter, the vehicle is responsible of travelling along the allocated route within the time constrains imposed without any deviations. As Figure 1 illustrates, a vehicle requests a route from nodes O (Origin) to (Destination) D and the RSU replies with particular road segments that the vehicle should follow as indicated by the red line in the figure. As new route requests are issued by soon-to-be-departing vehicles, the RSU uses its existing reservation registry and the route-reservation algorithm to make new reservations.

While multiple MFDs can exist across large regions (especially in urban settings as shown in (10) (24)), the assumption in this work is that a homogeneous MFD region is investigated. Homogeneous regions are those where all road segments have similar characteristics (e.g. critical density and traffic demand distribution); note that the existence of...
homogeneous regions has been observed using empirical data in different cities, e.g. San Francisco (11).

3 PROBLEM FORMULATION

The road network is considered as a graph $G = (V, E)$ with vertices $V$, $N_V = |V|$, representing the road junctions and edges $E$, $N_E = |E|$, representing the road segments. We assume that the road network is within a homogeneous static traffic region with parameters of the MFD diagram $\rho^C$, $\rho^I$ and $u_f$, representing the critical density corresponding to the maximum flow, jam density and free-flow speed, respectively. The region’s MFD is a priori computed based on historical measurements.

Each road segment $(i, j) \in E$, $\{i, j\} \in V$ is described by parameters $\lambda_{ij}$, $N_{ij}$, and $\rho_{ij}$ which denote the length, number of lanes, and jam density respectively. Let also $\rho_{ij}(t)$ denote the instantaneous density of the road segment at time $t$. In addition, we define the critical density of link $(i, j)$ as $\rho_{ij}^C = (\rho^C / \rho^I)\rho_{ij}^f$ which represents the density at which maximum flow is achieved for the specific road segment, so that for $\rho_{ij}(t) \leq \rho_{ij}^C$ the vehicles travel with free-flow speed $u_f$. Time is discretized into time-slots of duration $T$ such that the number of time-slots required to traverse road segment $(i, j)$ is $\tilde{c}_{ij} = \lfloor \lambda_{ij} / u_f / T \rfloor$, where $[z]$, is the nearest integer to $z$. In the free-flow region the speed ranges from the free-flow speed to the speed-at-capacity; here, we have assumed for simplicity free-flow speed for the entire free-flow region, as the speed-at-capacity is usually approximately equal to the free-flow speed.

The proposed reservation architecture requires the monitoring of the cumulative number of expected (based on reservations) vehicle arrivals and departures at road segment $(i, j) \in E$ up to time $t$, $\alpha_{i,j}(t)$ and $\beta_{i,j}(t)$, respectively. In addition, it requires to monitor the total number of active vehicle reservations ($n_{ij}(t) = \alpha_{i,j}(t) - \beta_{i,j}(t)$) of road segments $(i, j)$ for time-slot $t$ in order to keep track of the admissibility of road segments. We denote the admissibility of road segment $(i, j)$ at time-slot $t$ with $x_{ij}(t)$, and let $x_{ij}(t) = 1$ if a vehicle starting from road junction $i$ at time $t$ can traverse road segment $(i, j)$ without making the total reserved density larger than the critical density at any point within the traversal time, and $x_{ij}(t) = 0$ otherwise. Mathematically $x_{ij}(t)$ can be defined as:

$$x_{ij}(t) = \begin{cases} 1, & \text{if } n_{ij}(\tau) / (\lambda_{ij}N_{ij}) \leq \rho_{ij}^C, \forall \tau = t, \ldots, t + \tilde{c}_{ij} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where the quantity $n_{ij}(\tau) / (\lambda_{ij}N_{ij})$ denotes the density of road segment $(i, j)$ at time $\tau$ due to the reservations. The admissibility of road segments is an important aspect of the reservation scheme as it allows vehicles to always traverse through non-congested links at free-flow speed, leading to congestion-free routing. Nonetheless, it introduces the additional challenge of dealing with non-admissible road segments.

There are two alternative options to consider in case a vehicle encounters a non-admissible road segment in its shortest path. The first option is to wait at its origins until
all links of the shortest-path are admissible. The second option is to find an alternative admissible path. Interestingly, a combination of the above options leads to better solutions (i.e., wait for a short period at the origin and then take an alternative path).

Considering the above notation, the cost of traversing a road segment \( c_{i j}(t) \) can be expressed as follows:

\[
c_{i j}(t) = \begin{cases} 
\bar{c}_{i j}, & \text{if } x_{i j}(t) = 1 \\
\infty, & \text{if } x_{i j}(t) = 0 \text{ and } i \neq O \\
\bar{c}_{i j} + w, & \text{if } x_{i j}(t) = 0 \text{ and } i = O
\end{cases}
\]  

(2)

where, \( w \) denotes the number of time slots that a vehicle waits at the origin junction \( O \).

This work investigates two problems to address route-reservation under admissible/non-admissible road segments and where waiting is allowed at the origin junction. The first problem, called the Earliest Destination Arrival Time (EDAT), aims to find the path arriving at the destination at the earliest time. The second problem, called Traffic Load Balancing (TLB), aims to find a path that provides a good tradeoff between early destination arrival and traffic load balancing. Although arriving at the destination at the earliest possible time is highly desirable, balancing the traffic creates robustness against travel time estimation inaccuracies due to acceleration/deceleration and queuing at intersections. These two problems are mathematically described below.

**Earliest Destination Arrival Time (EDAT) problem**

Given an origin-destination \((O - D)\) pair, the time-stamp \( t_0 \) at which the routing request is made, and the reservation states \( x_{i j}(t) \), \((i, j) \in E, \forall t \geq t_0 \) then the EDAT problem requests the earliest-destination-arrival-time (from \( O \) to \( D \)). Let \( p_k \) denote the \( k \)-th path from source \( O \) to destination \( D \) denoted as \( p_k = (v_0^k, v_1^k), (v_2^k, v_3^k), ... (v_{L_k}^k, v_{L_k}^k) \), where \( v_j^k \in V \) is the \( j \)-th visited node in the \( k \)-th path, with \( v_0^k = O \) and \( v_{L_k}^k = D \), and \( L_k \) is the length of the path \( p_k \). Additionally, let \( w \) and \( d_j^k(t_0) \) denote the waiting time at the origin junction and the earliest arrival time at junction \( j \), respectively. Then, the earliest arrival time to each node of the path can be expressed as:

\[
d_{v_0}^k = t_0 + w, \quad w \geq 0
\]

(3)

\[
d_{v_1}^k = d_{v_0}^k + c_{v_0, v_1}(d_{v_0}^k)
\]

(4)

\[
\vdots
\]

(5)

Hence, the EDAT problem can be expressed in compact form as:

\[
d_D = \min_{w, p_k} \d_D^k \]

(6)
Traffic Load Balancing (TLB) problem

Given an origin-destination \((O - D)\) pair, the time-stamp \(t_0\) of the routing request, the reservation state \(x_{ij}(t)\) of each road segment and the total number of active reservations of each road segment \(n_{ij}(t)\), the TLB problem requests the path (from \(O\) to \(D\)) that minimizes the variance of densities across the network. Let \(p_k\) denote the \(k\)-th path from source \(O\) to destination \(D\) defined in the same way with the EDAT problem. Let also \(T_H = a \cdot d_D^*\) denote the time-horizon for the TLB problem, where \(d_D^*\) is the solution to problem of the EDAT problem (equation 6) and \(a \geq 1\) is a constant that defines the tradeoff between achieving the earliest destination arrival for the particular vehicle and load balancing the traffic. Based on the initial conditions of the network, we define the mean \(\mu_0\), second moment \(M_0\), and variance \(\sigma_0^2\) of the densities of the network for the time-horizon considered as:

\[
\mu_0 = \frac{1}{N_S} \sum_{(i,j) \in E} \sum_{\tau=0}^{T_H} \frac{n_{ij}(\tau)}{b_{ij}}
\]  

(7)

\[
M_0 = \frac{1}{N_S} \sum_{(i,j) \in E} \sum_{\tau=0}^{T_H} \left( \frac{n_{ij}(\tau)}{b_{ij}} \right)^2
\]  

(8)

\[
\sigma_0^2 = \frac{1}{N_S} \sum_{(i,j) \in E} \sum_{\tau=0}^{T_H} \left( \frac{n_{ij}(\tau)}{b_{ij}} - \mu_0 \right)^2 = M_0 - \mu_0^2
\]  

(9)

where \(b_{ij} = l_{ij}N_{ij}\) and \(N_S = (T_H - t_0 + 1)N_E\).

When a vehicle with path \(p_k\) and waiting time at the origin \(w\) enters the network, the number of reservations \(n_{ij}(t)\) at the corresponding road segments is increased by one for the occupancy period of each segment. To obtain values for the mean \(\mu_k(t)\), second moment \(M_k(t)\), and variance \(\sigma_k^2(t)\) of path \(p_k\) when the destination is reached at time \(t\), i.e. \(t = d^*_D\), we consider the amount of change occurring on (7)-(9); this is achieved by increasing the value of \(n_{ij}(\tau)\) by one for each road segments included in path \(p_k\), yielding the following expressions:

\[
\mu_k(t) = \mu_0 + \frac{1}{N_S} \sum_{(i,j) \in p_k} \sum_{\tau=d^*_k}^{d^*_k+1} \left( \frac{n_{ij}(\tau)+1}{b_{ij}} - \frac{n_{ij}(\tau)}{b_{ij}} \right) = \mu_0 + \frac{1}{N_S} \sum_{(i,j) \in p_k} c_{i,j}(d^*_k) \frac{c_{i,j}(d^*_k)}{b_{ij}}
\]  

(10)

\[
M_k(t) = M_0 + \frac{1}{N_S} \sum_{(i,j) \in p_k} \sum_{\tau=d^*_k}^{d^*_k+1} \left( \left( \frac{n_{ij}(\tau)+1}{b_{ij}} \right)^2 - \left( \frac{n_{ij}(\tau)}{b_{ij}} \right)^2 \right)
\]  

(11)

\[
\sigma_k^2(t) = M_k(t) - (\mu_k(t))^2
\]  

(12)

One important observation is that we can define the mean, second moment and variance of the path \(p_k\) up to node \(v_{l-1}\) reached at time \(t\), i.e. \(t = d^*_v\), denoted as \(\mu_{k,l}(t), M_{k,l}(t)\) and \(\sigma_{k,l}(t)\), respectively, based on the associated quantities \(\mu_{k,l-1}, M_{k,l-1}\) and
\( \sigma_{k,l-1}^2 \) and the incurred increment due to the increase of \( n_{ij}(\tau) \). In particular, simple mathematical calculations yield the expressions:

\[
\mu_k(v_i^k, t) = \mu_k(v_{i-1}^k, d_{i-1}^k) + \Delta \mu_k(v_i^k, t) = \mu_k(v_{i-1}^k, d_{i-1}^k) + \frac{1}{N_S} \frac{c_{ij} d_{ij}^k}{b_{ij}^k}
\]

\[
M_k(v_i^k, t) = M_k(v_{i-1}^k, d_{i-1}^k) + \Delta M_k(v_i^k, t) = M_k(v_{i-1}^k, d_{i-1}^k) + \frac{1}{N_S} \sum_{\tau=d_{i-1}^k+1}^{d_i^k} \left( \left( \frac{b_{ij}^k}{n_{ij}^k(\tau)} + 1 \right)^2 - \left( \frac{b_{ij}^k}{n_{ij}^k(\tau)} \right)^2 \right)
\]

\[
\sigma_k^2(v_i^k, t) = \sigma_k^2(v_{i-1}^k, d_{i-1}^k) + \Delta \sigma_k^2(v_i^k, t) = \sigma_k^2(v_{i-1}^k, d_{i-1}^k) + \frac{2}{N_S} \left( n_{ij}^k(\tau) \right)^2 \Delta \sigma_k^2(v_i^k, t) = (\Delta \mu_k(v_i^k, t))^2
\]

Based on the above discussion, the TLB problem can be defined as

\[
\min_{w, p_k} \sigma_k^2(D, d_D^k), \text{ such that } T_H / a \leq d_D^k \leq T_H.
\]

Formulation (16) aims to minimize the spatio-temporal variance of traffic densities in the network provided that the time required to reach the destination is not higher than a percentage \((a - 1)100\%\), with respect to the earliest destination arrival time. In this context, other measures can also be considered such as the weighted sum between mean and variance (25), or some reliability shortest path measure combined with variance, e.g. (26, 27).

**EDAT ALGORITHMIC SOLUTION**

The objective of the **EDAT** problem is to minimize the destination arrival time. To solve this problem, a directed acyclic graph is build using a space-time network. The space dimension contains indices of the road junctions and the time dimension contains consecutive time slots. Each node replica in the space-time network is identified by the index of the road junction and a specific time slot. Edges on this network represent road segments and the length of each edge reflects the time necessary to travel between adjacent junctions.

To construct a directed graph on this network, edges are assessed based on the reachability of nodes from the origin, and their admissible capacity of edge \((i, j)\), using variables \(d_i(t)\) and \(x_{ij}(t)\), respectively. Specifically, variable \(d_i(t)\) determines if node \(i\) of edge \((i, j)\) is reachable; indicated when \(d_i(t) < \infty\). Thereafter, edge \((i, j)\) is considered admissible when \(x_{ij}(t) = 1\). In the process, a topological ordering is imposed for all nodes in the graph. Also, in this directed graph no direct cycles exist, and thus it is easy to indicate reachability from the origin using merely \(d_i(t) < \infty\). In the case when both conditions are satisfied, edge \((i, j)\) is added on the graph. Specifically, if junction \(i\) is
Data: \( G(V,E) \), \( n_{ij}(t) \), \( O \), \( D \), \( t_0 \), \( x_{ij}(t) \)

Result: Returns Earliest Destination Arrival Time Route

/* Algorithm Initialization */

\[
t = t_0 - 1;
\]

\[
d_i(t) = \infty \forall t, i \in V;
\]

\[
d_D^* = \infty;
\]

\[
d_O(t) = 0, \forall t;
\]

/* Algorithm Execution */

while \( t < d_D^* \) do

\[
(i,j) \in E \ do
\]

\[
t = t + 1;
\]

if \( (((i == D) OR (j == D)) AND (d_D(t) < d_D^*)) \) then

\[
d_D^* = d_D(t);
\]

else

\[
if ((x_{ij}(t) == 1) AND (d_i(t) < \infty)) then
\]

\[
d_j(t) = d_i(t) + (t + c_{ij});
\]

previous\([j][t+c_{ij}] = i;\]

end

end

end

Trace back the optimal path \( p^* \) starting from previous\([D][d_D^*];\)

Return\((p^* \ and \ d_D^*);}\)

Algorithm 1: EDAT algorithmic solution

reachable from the originating junction \( O \) and \( x_{ij}(t) = 1 \), then a directed edge from \( i \) at time \( t \) to junction \( j \) at time \( t + c_{ij} \) is added to the graph. The whole process repeats until that time when \( D \) becomes reachable, i.e. \( d_D(t) < \infty \) for any \( i \) and \( t \) (edge \((i,D)) \). It should be noted here that since the earliest destination arrival time route is required, the algorithm stops when \( D \) becomes reachable and traces back the nodes in the space-time network that resulted to this route. Algorithm 1 depicts the steps of the aforementioned algorithmic procedure to compute the EDAT.

Example 1

To better understand the EDAT algorithmic procedure consider the example illustrated in Figure 2 where edge lengths reflect the traversal times for specific road segments. In this example, \( t_0 = 0 \) and the admissibility along different road segments is given as follows:

\[ x_{OC}(2) = x_{BE}(1) = x_{BE}(2) = x_{CD}(1) = x_{CD}(2) = 0. \]

Figure 3 shows the graph constructed by the EDAT algorithm. The space dimension of each node indicates the junction index while the time dimension indicates the replica of the junction created over time. As before, the two variables assess the reachability of nodes and the admissibility of edges in the graph. As illustrated in the figure,
FIGURE 2: Example Network $G(V,E)$.

FIGURE 3: EDAT Algorithmic Solution.

The first column contains only edges emerging from the origin $O$ since all other junctions until time $t = 0$ are not reachable. Similarly, the second column contains edges emerging from junctions $O$, $B$, and $C$ since at $t = 1$ these nodes have been reached from origin $O$.

The dashed-line edges represent road segments that can alternatively be selected without affecting the cost of the solution. The algorithm selects the road segments that were identified first and discard any subsequent arrivals to the specific node. Following the same procedure at the fifth column, the destination is reached and the algorithm terminates. Grid-shaded nodes shown in the graph depict those nodes selected in the route $(O \to B \to C \to D$ and $w = 1)$.

The devised algorithm results to an optimal solution and executes in pseudo-polynomial time in the discrete time case, since the state space is not known until the algorithm converges with complexity $O(d_D^* N_E)$. To see why an optimal solution is found,
notice that the minimum cost of state \((i, t)\) is equal to \(t\) if the state is reachable and \(\infty\) otherwise. Hence, a path achieving reachability of state \((i, t)\) provides the minimum cost to that state and ensures reachability of all predecessor states forming the path. Therefore, if state \((D, \bar{t})\) is reachable though path \(p\), then all states forming \(p\) are also reachable (with minimum cost) and the \textit{optimal substructure property} applies (28). The reachability of all states is examined for increasing \(t\) and thus the optimal solution is found at time \(d^*_D\) which is the earliest time at which \(D\) is reachable.

8 TLB ALGORITHMIC SOLUTION

Solving the TLB problem to optimality is a challenging task because variance does not adhere to the optimal substructure property (6), i.e. an optimal solution \textit{cannot} be constructed efficiently from optimal solutions of its subproblems. To see why this is true, consider a road network consisting of two paths arriving at junction \(i\) and a road segment directly connecting junction \(i\) to the destination \(D\), \((i, D)\). Let the mean and variance of the two paths \(p_1\) and \(p_2\) up to junction \(i\) be given by \(\mu_1(i, t), \sigma_1^2(i, t)\) and \(\mu_2(i, t), \sigma_2^2(i, t)\), respectively. Additionally, let \(\sigma_1^2(i, t) < \sigma_2^2(i, t)\) such that the optimal path to \(i\) is \(p_1\). If the optimal substructure property holds then the minimum variance at the destination must utilize path \(p_1\). However, using Eq. (15), it can be easily shown that the optimal path to the destination is through \(p_2\) when \(\mu_2(i, t) > \mu_1(i, t) + (\sigma_2^2(i, t) - \sigma_1^2(i, t))/(2\Delta \mu(i, t))\), where \(\Delta \mu(i, t) = c_{i,D}(t)/(N_S b_{i,D})\) which confirms that the optimal substructure property does not hold for TLB problem.

Based on the above discussion, a dynamic programming algorithm similar to Algorithm 1 cannot be developed to optimally solve TLB. One approach to address this issue is to consider dynamic programming with an additional dimension in the state of the time-expanded graph associated with the origin waiting time so that \(\mu_k(i, t)\) is constant at one particular state; however, this significantly increases the complexity of the problem which is not desirable. An alternative approach is to approximate the variance with a new metric that has good load balancing performance and satisfies the optimal substructure property.

Towards this direction we consider the second moment of the network densities as the cost metric. This metric provides a good approximation of the variance when the length of the paths reaching the destination are of approximately equal length.

The solution of the TLB problem is outlined in Algorithm 2. Similar to Algorithm 1, a directed acyclic time-expanded graph is constructed with states \((i, t)\) indicating that junction \(i\) is reached at time \(t\) and minimum cost of reaching the state equal to \(M(i, t) = \min_{p_k} M_k(i, t)\). The initialization of the algorithm is similar to Algorithm 1, while the main body consists of two blocks. The first examines if the destination has been reached with a better cost than \(M^*_{D, \bar{t}}\), in which case the destination cost is updated; parameter \(d^*_D\) maintains the time-slot with the best cost to the destination. The second block computes the cost of reaching junction \(j\) through \(i\) based on Eq. (12) and updates it when it is better than the current cost, if \(i\) is reachable from the origin \(O\) and \((i, j)\) is admissible. To backtrack the best path to the destination, a predecessor list is maintained through matrix \textit{previous}, where expression \textit{previous}[\(j\)][\(t + \bar{c}_{ij}\)]\(= i\) indicates that state \((j, t + \bar{c}_{ij})\) is reached through state \((i, t)\). The complexity of TLB algorithm is equal to \(O(T_H \sum_{(i,j) \in E} \bar{c}_{ij})\) due
\textbf{Data:} $G(V,E), n_{ij}(t), O, D, t_0, x_{ij}(t), T_H, b_{ij}, N_S, M_0$

\textbf{Result:} Returns a load balancing route

\begin{verbatim}
/* Algorithm Initialization */
t = t_0 - 1;
M(i,t) = \infty, \forall t, i \in V;
/* Origin junction is reachable */
M(O,t) = M_0, \forall t;
/* Minimum second moment of D */
M^*_D = \infty;
/* Algorithm Execution */
for t = t_0, t_0 + 1, \ldots, T_H do
    for (i,j) \in E do
        t = t + 1;
        if ((i == D) OR (j == D) AND (M(D,t) < M^*_D)) then
            /* Arrival time at D */
d_D = t;
            M^*_D = M(D,t);
        end
        if ((x_{ij}(t) == 1) AND (M(i,t) < \infty)) then
            M_{temp}(j,t + \tilde{c}_{ij}) = M(i,t) + \frac{1}{N_S} \sum_{t}^{t + \tilde{c}_{ij}} \left(\frac{n_{ij}(t)+1}{b_{ij}}\right)^2 - \left(\frac{n_{ij}(t)}{b_{ij}}\right)^2
        end
    end
end
Trace back the optimal path $p^*$ starting from $previous[D][d_D]$;
Return($p^*$ and $M^*_D$);
\end{verbatim}

\textbf{Algorithm 2:} TLB Algorithmic Execution

1. to the iteration over time and the summation that appears in the computation of $M_{temp}$.
2. Note that other research works have also used space-time graphs to solve vehicle routing
3. problems e.g. (29).
4. To illustrate the execution of Algorithm 2, let us revisit Example 1 aiming to
5. solve the TLB problem with $a = 1.25$. In this case, the constructed DAG corresponding
6. to the Algorithm 2 is shown in Figure 4. The dashed green lines indicate edges from
7. candidate paths that have not produced minimal cost, rather than alternative solutions.
8. Comparing Algorithms 1 and 2, the TLB-based algorithm examines solutions up to $T_H =$
5 (= 4 × 1.25) rather than $d_D^s = 4$. In addition, the optimal solution provided by Algorithm 2 (illustrated by the red nodes) involves a different path and waiting time compared to the EDAT solution ($O \rightarrow C \rightarrow D$ and $w = 3$ versus $O \rightarrow B \rightarrow C \rightarrow D$ and $w = 1$).

SIMULATION SETUP AND RESULTS

The road network under consideration is a 1.8km$^2$ homogeneous region of downtown San Francisco. The spatial compactness and homogeneity of this area was first established in (30) while a similar region is used by (10). The selected region consists of 99 road junctions and 208 single-lane road segments with lengths varying from 100m to 400m. The network is imported into the SUMO micro-simulator (31), where vehicle mobility is simulated according to the Krauss car following model (32). The car-following model parameters are set as follows: vehicle length 5m, maximum speed 15m/s, acceleration 2.5m/s$^2$, deceleration 4.5m/s$^2$, and minimum gap distance 2.5m. Simulation time was set to 1 hour and only vehicles that completed their journeys were considered in the results while no overtaking was allowed (ensuring FIFO property i.e., $t + \bar{c}_{ij}(t) \leq (t + 1) + \bar{c}_{ij}(t + 1)$, $\forall (i, j) \in E$). Monte Carlos simulations were constructed (10 realizations) whereas both algorithms (EDAT and TLB) were compared against the traditional behavior (TB) experienced within the network when no control mechanism is applied and each vehicle travels from the origin to the destination along the shortest (in terms of distance) path. Finally, vehicles follow strictly their reservation routes, but not their reservation times. The reason is that in the conducted simulations, the route reservation scheme makes all reservations
and routing decisions based on computed travel times without consideration of the actual network state at the time a vehicle request arrives; hence, travel times may vary from those computed by the proposed algorithms due to various sources of uncertainty such as driver imperfection and junction priorities.

By injecting flow into the region, the MFD and its characteristics, including $u_c$ and $\rho_{ij}$, can be identified for use in the two algorithms. For this reason, a 6-hour scenario was simulated within which for the first four hours the input flow was set to 2000veh/h and incrementally increased by 2000veh/h for the next three hours. Thereafter, the input flow was set to 4000veh/h and 2000veh/h for the next two hours, respectively. Fig. 5 (a) depicts the MFD for the uncontrolled scenario (TB). The calibrated model indicated by the solid yellow line is derived through the automated calibration method proposed in (33) for the single-regime Van Aerde model (34). According to the generate model the selected parameters for both algorithms are: $u_f = 47\text{ km/h}$, $u_c = 40.5\text{ km/h}$, $\rho^d = 1050\text{ veh}$ and $\rho_{ij}^C = 40\text{ veh/km/\text{lane}}$ (i.e., around 40% of the region’s total density).

Figs. 5 (b) and (c) depict the resulting MFD when the TLB and EDAT algorithms are employed respectively, demonstrating that congestion is alleviated from both algorithms. Additionally, the total volume of flow, density and speed are illustrated in Figs. 5 (d), (e) and (f) as a function of the simulation time, respectively. Comparing these three figures it is demonstrated that as both algorithm operate the density decreases (near 330veh for TLB and near 410veh for EDAT compared to more than 700veh for TB as shown in Fig. 5 (d)). Thus, the link densities are always maintain below the critical capacity (near 350veh) for both algorithms even when traffic demand is high (e.g., for simulation time 200-240min). Observing Fig. 5 (e) the travelling speed manages to remain high for both algorithms and thus the flow is similar to that of TB Fig. 5 (f). Indicatively, Fig. 5 (e) depicts that in high demand periods TLB outperforms EDAT since the travelling speed is always maintained near $u_f$.

For the results presented hereafter, Monte Carlo simulations were executed with random $O-D$ pairs and for flow rates varying between 1000 – 8000veh/h over a duration of two hours. Figures 6 (a) and 6(b) show the average number of vehicles that reach their destination and the average vehicle travel time as a function of the different flow rates, respectively. Specifically, the dashed lines in Figure 6 (a) represent the average number of vehicles that have finished their journey within the simulation time and the scattered plots are the realizations obtained by each simulation run. Similarly, the dashed lines in Figure 6 (b) represent the value of the average travel time for the different realizations and the scattered plots represent the average travel time across all Monte Carlo simulations.

As illustrated in Figures 6(a) and 6(b), at low flow rates EDAT and TB behave similarly while TLB slightly lags in terms of the average number of vehicles that reach their destination within the simulation time. On the other hand, TLB outperforms EDAT and TB in high flow rates, where a larger number of vehicles reach their destination and results in a more robust and shorter travel times. In either case, both algorithms perform much better than TB. Comparing EDAT against TLB it is clear that better network utilization can be achieved in the case where EDAT is applied at low flow rates while TLB
Figure 6(c) illustrates the travel time distribution for all vehicles that managed to reach their destination during the simulation time when a flow rate of 8000 veh/hour is used. As clearly shown, EDAT and TLB greatly outperform TB. In numbers, the mean employed at higher flow rates.
FIGURE 6: (a) Number of vehicles towards to the route end; (b) Average travel time ($t \rightarrow s$); (c) Travel time distribution (8000veh/h)

travel time for EDAT is 139.7s, for TLB is 118.9s and for TB is 2160.8s. The standard deviation for EDAT is 86.65, for TLB is 58.2 and for TB is 2762 demonstrating that as congestion of the road segments increases, TLB is more stable and accurate than the other two solutions. Additionally as shown in the figure, TLB is highly resilient to the increase in flow rate since travel times do not significantly deviate. Finally Figure 6(c) clearly indicate that TLB achieves better travel times while eliminating spillbacks.

CONCLUSIONS

This work proposes a novel route reservation scheme which aims to prevent congestion by restricting the traffic density in different road segments. For this scheme, the earliest destination arrival time problem and the traffic load balancing problem are examined. Both problems are solved by developing dynamic programming algorithms on directed acyclic time-expanded graphs. Simulation results demonstrate the superiority of both algorithms compared to the traditional traffic behavior with no reservations achieving substantial improvements in terms of road utilization and travel times. In addition, results demonstrate the complementary nature of the developed algorithms, as the earliest destination arrival time algorithm provides better results for low congestion conditions and the traffic load balancing algorithm for high congestion conditions.
Future work will investigate congestion-free vehicle routing in static and dynamic multi-region traffic networks. Also it is important to examine how reservation-based congestion-free routing compares to other approaches such as dynamic traffic assignment and user equilibrium such as (35). Another direction of work regards the examination of scenarios with both normal vehicles and vehicles that follow the route reservation scheme. In this setting one could examine the performance as a percentage of normal vehicles and the effect of traffic state uncertainty.

8 ACKNOWLEDGMENTS

This work is partially funded by the European Research Council Advanced Grant FAULT-ADAPTIVE (ERC-2011-AdG-291508).

REFERENCES


Menelaou, Timotheou, Kolios and Panayiotou 19


[10] Aboudolas, K. and N. Geroliminis, Perimeter and boundary flow control in multi-
reservoir heterogeneous networks. Transportation Research Part B: Methodologi-


[12] Kaysi, I., M. Ben-Akiva, and H. Koutsopoulos, Integrated approach to vehicle rout-
ing and congestion prediction for real-time driver guidance. 1408, 1993.

ligent vehicle-highway systems application*. Journal of Intelligent Transportation

Information, 2014.

Based on Routing Game Theory. In Transportation Research Board 94th Annual
Meeting, 2015, 15-3613.


[17] Li, P., P. Mirchandani, and X. Zhou, Solving simultaneous route guidance and traffic
signal optimization problem using space-phase-time hypernetwork. Transportation

[18] Mahmassani, H. S., M. Saberi, and A. Zockaie, Urban network gridlock: Theory,
characteristics, and dynamics. Transportation Research Part C: Emerging Technolo-

[19] Knoop, V., S. Hoogendoorn, and J. Van Lint, Routing strategies based on macro-
scopic fundamental diagram. Transportation Research Record: Journal of the

[20] Yildirimoglu, M. and N. Geroliminis, Approximating dynamic equilibrium condi-

[21] Yildirimoglu, M., M. Ramezani, and N. Geroliminis, Equilibrium analysis and route
guidance in large-scale networks with MFD dynamics. Transportation Research


