Network Reserve Capacity under Influence of Traveler Information

Y. E. Ge\textsuperscript{1}; H. M. Zhang\textsuperscript{2}; and William H. K. Lam\textsuperscript{3}

Abstract: This paper provides a hierarchical framework for studying the impact of traveler information on network reserve capacity. The framework comprises a two-level mathematical program—the upper-level program maximizes the reserve capacity multiplier subject to a link capacity constraint, and the lower-level program generates user equilibrium flow patterns under the influence of traveler information. The two-level program is solved by a genetic algorithm-based solution method. Numerical results indicate that the reserve capacity of a road network does not increase monotonically with the increase of information level. The implications of this finding, including its dependence on the characteristics of the road network, are discussed.

DOI: 10.1061/(ASCE)0733-947X(2003)129:3(262)

CE Database subject headings: Transportation networks; Information systems; Travel demand; Traffic capacity; Algorithms.

Introduction

The concept of capacity is widely used in the planning, design, and operation of transportation facilities [for example, Webster and Cobbe (1966); Allsop (1972); Yagar (1974), (1985); Wong (1996); Wong and Yang (1997)]. The capacity of a freeway section, for example, is used in deciding the level of service of an existing freeway facility and, if the level of service is below a certain level, the number of lanes to be added to that freeway to obtain the desired level of service. For a freeway section, the definition of capacity is straightforward: it is the maximum sustainable flow that can be accommodated by the freeway section in a suitably chosen time interval (usually 15 min). The concept of system capacity, however, is much more complex. This is because many competing factors come into play in the definition of system capacity. Take a two-way signalized intersection as an example. Suppose that the two traffic streams receive $g_1$, $g_2$ green times, respectively, and the saturation flow rates for the two approaches are $s_1$ and $s_2$, respectively. Moreover, the signal has cycle length $C_y$ and lost time $R$ (for simplicity $R$ is set to zero; then $C_y = g_1 + g_2$). Given these settings, the maximum flow rate through the intersection is given by

$$Q_n = \frac{g_1}{c_y} s_1 + \frac{g_2}{c_y} s_2 = \frac{g_1}{g_1 + g_2} (s_2 - s_1) \quad (1)$$

It is clear from Eq. (1) that an intersection does not have a fixed capacity. The capacity of an intersection, as defined by Eq. (1), depends on the selection of green splits ($g_1, g_2$). For example, the capacity of the intersection would be $s_1$ if $g_2 = 0$ and $s_2$ if $g_1 = 0$. One therefore cannot discuss intersection capacity without first specifying green splits.

The capacity of a road network with many origin-destination (O-D) pairs, like that of an intersection, is also variable and cannot be discussed without specifying an O-D demand pattern. This is because there are many ways to utilize the link capacities of the road network. For example, one can obtain different network capacities by maximizing the flows of different O-D pairs through a network, just as one can obtain different intersection capacities through assigning different green times to different approaches in an intersection.

To extend the capacity concept to a multiple O-D network, we introduce the concept of network reserve capacity. The reserve capacity of a network is defined as the additional demand that can be accommodated by a road network without changing the physical characteristics of the road network and without exceeding the capacities of all its links. Moreover, the additional demand is apportioned to each O-D pair proportionally to its original demand.

A network can have positive, zero, or negative reserve capacity. A positive reserve capacity indicates that the network still has slack capacities in all its links, while a zero reserve capacity implies that some critical links have reached saturation and become bottlenecks of the road network. A negative reserve capacity, on the other hand, indicates that the network is already overloaded and congestion exists in the network. As such, network reserve capacity can be viewed as a networkwide level of service indicator and can be used to establish efficient policies for traffic restraint, demand growth control, and transportation network planning and management.

Just as O-D patterns affect network reserve capacity, drivers’ route choice behavior is also expected to affect network reserve capacity because it changes the distribution of traffic in the net-
work. In this paper, we study the effects of traveler information on network reserve capacity. To do this, a bilevel programming formulation is developed to investigate the effect of various levels of driver information on network reserve capacity. This program takes into account user behavior and the quality of traveler information through a logit-based assignment model and is solved by a stochastic search method that uses genetic algorithms (GAs), which are suited for finding global optimal solutions of complex combinatorial optimization problems. Two example networks are used to demonstrate both the solution procedure and the impact of traveler information on network reserve capacity. The results, perhaps counterintuitive, are carefully explained, and their implications for more general networks are discussed in detail.

The remainder of this paper is organized as follows. The section on formulation that follows presents a (bilevel) hierarchical optimization formulation of the network reserve capacity problem. The section after that develops a GA-based solution method to solve the problem formulated in the next section and is followed by sections that report the numerical results from two examples and discuss the impact of network characteristics on the reserve capacity, and finally a concluding section.

**Formulation**

Consider a transportation network represented by a directed graph $G=(N,L)$, where $N$ is a set of nodes, and $L$ is a set of directed links. A node represents either an origin, a destination, or an intersection. For each intersection $j$, $L_j$ denotes the set of links connected to it, and $L_j$ the number of links connected to $j$. All the signalized intersections in the network compose a set $J$. In the following, $r$ denotes an origin node and $s$ a destination node; $(r,s)$ denotes an O-D pair. To facilitate presentation, we have provided a notation at the end of the paper.

To measure how much additional demand can be accommodated by a road network, a common multiplier can be applied to an existing O-D trip table, and the sum of the O-D trips (existing plus additional) is assigned to the network, subject to flow conservation and capacity constraints. In this process, it is essential to take into account the redistribution of traffic in the network. In fact, when the demand of the network is increased to an amount close to its full capacity, traffic congestion has substantial impacts on the distribution of flow over the network. The congestion effect would result in a change in traffic flow pattern, and hence a different value of the reserve capacity. To take all these effects into account, a general formulation of network reserve capacity is formulated as follows:

**Leader Problem (Upper Level)**

maximize \( \mu \) \hspace{1cm} (2)

subject to

\[ x_a(\mu) \leq p_at_a^a; \quad \forall a \in L \] \hspace{1cm} (3)

where the link flow pattern \( x = (\ldots x_a, \ldots)^T \) is determined by solving the following network equilibrium problem:

**Follower Problem (Lower Level)**

\[ f_{rs}^a \geq 0, \quad \forall r,s,k \in K^{rs} \] \hspace{1cm} (6)

and the definitional constraints

\[ x_a = \sum_{rs} \sum_{k \in K^{rs}} f_{rs}^a u_{ak}, \quad \forall a \in L \] \hspace{1cm} (7)

where Eq. (3) requires that the ratio of the flow on link \( a \) \((x_a)\) over its capacity \( c_a \) not exceed the maximum tolerable degree of saturation \( p_a \) to ensure a specified level of service \((p_a \) serves a function similar to the volume-to-capacity ratio in highway capacity analysis\). In this framework, the upper-level problem maximizes the network reserve capacity (multiplier) subject to the link capacity constraints, while the lower-level problem makes the link flow pattern satisfy the user equilibrium constraint. As can be seen from Eq. (5), there is an implicit assumption that \( \mu \gg -1 \). In fact, the above optimization formulation can ensure that this condition holds. Let the resultant value of \( \mu \) be \( \mu^* \); then \( \mu^* > 0 \) indicates that the network has reserve capacity, and \( \mu^* < 0 \) indicates that the network is overloaded (at least on one link or O-D pair).

Capacity constraints can generally be classified into two categories: physical constraints and environmental constraints (Minister of Transport 1963). The former reflect the unique characteristics of each road network, and, if they are violated, there would be a high risk of failure of the system. Conversely, the latter are imposed by the system manager on some links or specific parts of the urban transportation network [for example, central business districts (CBDs), urban freeways] to avoid significant negative effects caused by traffic to the environment. In the above formulation, \( p_a \) in Eq. (3) can be considered as an environmental constraint imposed by the manager on link \( a \in L \), while \( c_a \)—the capacity of a link—can be considered as a physical constraint unique to link \( a \). In general, \( p_a \) is assigned a small value for those links in the CBDs and a large value for others in the suburbs.

The physical capacity of links controlled by signals is a function of signal timing, as is the reserve capacity of the signalized road network. Consider an intersection controlled by adaptive or actuated signal systems. The capacity of each approach of the intersection is determined by its saturation flow rate and green split:

\[ c_a = \lambda_a s_a, \quad \forall a \in L_j, j \in J \] \hspace{1cm} (8)

where \( \lambda_a \) denotes the proportion of green time received by link \( a \) (or referred to as approach \( a \) of intersection \( j \)), and \( s_a \) denotes the saturation flow rate of approach \( a \). These proportions, often referred to as green splits, are required to satisfy the following feasibility constraints:

\[ \lambda_{a \min} \leq \lambda_a \leq \lambda_{a \max}, \quad \forall a \in L_j, j \in J \] \hspace{1cm} (9)

\[ \sum_{a \in L_j} \lambda_a = B_j, \quad \forall j \in J \] \hspace{1cm} (10)

where \( \lambda_{a \min} \) and \( \lambda_{a \max} \) are two constants representing, respectively, the minimum and maximum allowable green splits for approach \( a \) (obviously \( \lambda_{a \max} = \lambda_{a \min} = 0 \)), and \( B_j \) is the difference between the cycle length at intersection \( j \) and the sum of all lost time and pedestrian phases at this intersection to the cycle length; thus \( B_j < 1 \). Clearly

\[ \lambda_{a \max} = B_j - \sum_{b \in L_j, b \neq a} \lambda_{b \min}, \quad \forall a \in L_j, j \in J \] \hspace{1cm} (11)
Furthermore, average travel time on signalized links is a function of both link flow $x_a$ and green split $\lambda_a$ (here we only consider separable link cost functions):  
\[
\tau_a = \tau_a(x_a, \lambda_a), \quad \forall a \in L, j \in J
\]

Previous studies on reserve capacity all assume that travelers’ route choice behavior observes the deterministic user equilibrium (UE) principle, that is, the Wardropian first principle. Characteristic features of the Wardropian UE principle are that travelers behave identically and all have perfect information about network travel cost. Obviously, these two assumptions are somewhat restrictive and idealistic. In this paper we relax the second assumption of perfect knowledge of travel cost and study the case where travelers have incomplete information on travel costs. Travelers’ current perception of the state of the network usually comes from two sources: experience gained through previous trips, and ATIS. Although ATIS provides travelers with quality traffic information, it cannot completely eliminate travel time uncertainty. We assume that this uncertainty is captured by a random term, $\varepsilon_k$, and the estimated travel time on route $k \in K^r$ by ATIS is  
\[
\Phi^r_k = \phi^r_k + \varepsilon^r_k, \quad \forall k \in K^r
\]

where $\phi^r_k =$ actual travel time on route $k \in K^r$ for a given flow pattern. The travelers’ route choice between competitive alternatives can then be modeled using discrete choice theory where the principle of utility maximization is used to explain travelers’ choice behavior. It can be seen from Eq. (13) that the probability of choosing a specific route, given actual travel time, is dependent on the probability distribution of the random term. Currently, there exist two kinds of basic assumptions regarding the probability distribution of $\varepsilon^r_k$: logit based and probit based. The former assumes that $\{\varepsilon^r_k, k \in K^r\}$ are independently and identically distributed Gumbel variates, and the latter assumes they are multivariate normally distributed variates. The use of the probit model requires numerical integration of a multidimensional normal distribution, while the logit model can be evaluated directly. Therefore, for simplicity we adopt the first assumption, which leads to the following route choice probability:  
\[
p^{rs}_k = \exp\{\theta \Phi^r_k\} \sum_{l \in K^r} \exp\{-\theta \Phi^r_l\}, \quad \forall k \in K^r
\]

where $\theta =$ positive scaling factor, also known as the dispersion parameter in the logit model, and is related to the variance of ATIS-provided travel time. This route choice behavior can be realized in the SUE formulation of Fisk (1980):  
\[
\text{minimize } \sum_{a \in L} \int_0^{r_s} \tau_a(w) \, dw + \frac{1}{\theta} \sum_{rs} \sum_{k \in K^r} f^r_k \ln f^r_k
\]

subject to
\[
\sum_{k \in K^r} f^r_k = q^r_s, \quad \forall r, s
\]

\[
f^r_k \geq 0, \quad \forall r, s, k \in K^r
\]

\[
x_a = \sum_{rs} \sum_{k \in K^r} f^r_k \delta_{ak}, \quad \forall a \in L
\]

where $\ln f^r_k$ is defined as zero at $f^r_k = 0$. For $0 < \infty$, the above problem is strictly convex: hence, the equilibrium route flow pattern that solves the problem is unique. In contrast, when $\theta = \infty$ (that is, in the UE state), equilibrium route flows are generally not unique. We want to point out that the use of the dispersion parameter $\theta$ to capture the information level of ATIS is not new. Examples of such usage in the literature include Koutsopoulos and Lotan (1990), van Vuren and Watling (1991), Maher and Hughes (1995), and Yang (1998). It can be shown that this dispersion parameter $\theta$ in the logit model is inversely proportional to the standard deviation of ATIS-supplied route travel costs; in other words, it is related to the variation of travel costs perceived by travelers. When the parameter has a high value, the variance is small, indicating that ATIS provides an accurate estimate of route travel times. At the limiting case of $\theta \to \infty$ ATIS provides exact information of route travel times, and the road users can choose their minimal travel time route with certainty. In other words, the SUE solution in this case approaches the UE one.

On the other hand, when the parameter value is low, the variance is high, indicating that many routes would be chosen. In the limit of $\theta \to 0$, the shares of flow on all routes between any one O-D pair will be equal, regardless of route travel times. Therefore, it is reasonable to consider the parameter $\theta$ as an indicator of the quality of information provided by ATIS. Thus, we provide the following framework to study the impact of ATIS on network reserve capacity, with $\theta$ capturing the quality of ATIS information.

**Leader Problem (Upper Level)**

\[
\text{maximize } \mu
\]

subject to
\[
\begin{align*}
x_a \leq & \rho_a, \quad \forall a \in L, j \in J \\
\rho_a \leq & \rho_a c_a, \quad \forall a \in L \cup j \in J L_j \\
\lambda^\min_a \leq & \lambda_a \leq \lambda^\max_a, \quad \forall a \in L, j \in J \\
\sum_{a \in L} \lambda_a = & B_j, \quad \forall j \in J
\end{align*}
\]

where the flow pattern $x$ is determined by the following SUE assignment model:

**Follower Problem (Lower Level)**

\[
\text{minimize } \sum_{a \in L} \int_0^{r_s} \tau_a(w, \lambda_a) \, dw + \frac{1}{\theta} \sum_{rs} \sum_{k \in K^r} f^r_k \ln f^r_k
\]

subject to
\[
\sum_{k \in K^r} f^r_k = (\mu + 1) q^r_s, \quad \forall r, s
\]

\[
f^r_k \geq 0, \quad \forall r, s, k \in K^r
\]

\[
x_a = \sum_{rs} \sum_{k \in K^r} f^r_k \delta_{ak}, \quad \forall a \in L
\]

where the upper level models the reserve capacity of a signalized road network, while the lower level captures travelers’ route choice behavior under ATIS.

In transportation studies, this formulation is usually known as a bilevel program (Migdalas 1995), while in game theory it is called the hierarchical or Stackelberg equilibrium model [for example, Basar and Olsder (1995)]. Note that if signalized links are not considered, constraints Eqs. (20)–(22) of the upper level problem become constraints Eq. (3) in Eqs. (2)–(7). In a signal-
ized network, traffic flows on those links approaching any signalized intersection \( j \) depend on both the multiplier \( \mu \) and the green splits \( \lambda^g_j \), that is, \( x_{a_j} = x_{a_j}(\mu, \lambda^g_j) \), \( a \in L_j \). Without signals, link flows depend solely on multiplier \( \mu \), that is, \( x_a = x_a(\mu) \). Also, if nonseparable link cost functions are considered, the lower level corresponds to an asymmetric network equilibrium problem [for example, Sheffi (1985); Meneguzzo (1995)].

**Genetic Algorithm-based Solution Method**

The general hierarchical or bilevel mathematical program is a nonconvex, nondifferential optimization problem that has certain combinatorial features in its constraints. As such, it is computationally very difficult to solve, especially if one wishes to compute a globally optimal solution. Currently, sensitivity analysis-based methods are widely used to solve such problems in transportation studies [for example, Yang (1997)]. Such methods, however, are computationally expensive because they require the computation of derivatives and matrix inversions. Moreover, methods of this kind cannot guarantee the convergence to the global optimum.

In recent decades, stochastic search techniques such as simulated annealing (SA) and GAs [for example, Michakewicz (1996)] began to see applications in the solution of complex transportation problems. Ge (1999), for example, developed a GA-based method for solving a bilevel hierarchical optimization formulation oriented to the combined signal control and traffic assignment problem. Yin (2000) reported on his GA-based solution procedure developed specially for bilevel programming. In this section we develop a GA-based solution method for the problem described by Eqs. (19)–(26). Before proceeding, it is essential to give an introduction to GAs.

GAs are search methods that mimic the evolution process in nature and usually consist of the following algorithmic steps:

1. **Choose a criterion (that is, an objective function) to judge the fitness of a solution.** A solution is also known as an individual, or chromosome or string in GA literature. Each chromosome corresponds to a candidate solution of the problem. In our problem, Eqs. (19)–(26), a vector composed of decision variables \( \mu \) and \( \lambda^g_a(a \in L_j, j \in J) \) makes up a chromosome. The elements of a chromosome are referred to as genes of the chromosome. Genes are usually coded as bit or decimal strings.
2. **A population of chromosomes is initialized subject to certain constraints.**
3. **Each individual is checked against the fitness criterion, and the fittest individuals (candidate solutions) are selected for reproducing the next generation of candidate solutions.** This corresponds to a search through a space of potential solutions. For the problem in Eqs. (19)–(26), the process searches for the optimal solution on the space defined by Eqs. (20)–(26). It should be noted that in the search process a GA maintains a population of candidate solutions rather than a single solution.
4. **Generate new individuals (or chromosomes, or candidate solutions) through genetic operations.** Two genetic operations, known as crossover and mutation, are often used to generate offspring (for example, new solutions). Crossover refers to the process where two chromosomes (parents) exchange their genetic material at randomly selected locations to create two new chromosomes (offspring). Mutation offers a chromosome an opportunity to shuffle the positions of its bits to generate a new chromosome.
5. **The process stops if a suitable solution is found, or the computation time expires. Otherwise, go back to Step 3.**

Let us consider the problem at hand. As stated above, the string of potential solutions consists of \( \mu \) and \( \lambda^g_a(a \in L_j, j \in J) \), each of whose elements is a gene. Bit or decimal strings both are suitable for a chromosomal representation of a solution to the problem. For example, the gene \( \lambda^g_a \) may correspond to 00110011 (an 8-bit binary substring) or 0.2 (a decimal real number) in a chromosome. For ease of implementation, those genes corresponding to the same intersection are arranged in successive locations in the chromosome.

The implementation of constraints on potential solutions is an important question to be considered, both in initializing a population and in designing an objective (evaluation) function. Constraints can be implemented by imposing moderate penalties on individuals that violate them or by creating individuals directly satisfying them by means of a decoding procedure (or decoder). In the leader problem of Eqs. (19)–(26), there are two types of constraints: the link capacity constraints, and the feasible constraints of green splits.

Two different techniques are introduced to deal with the two types of constraints, respectively. An exterior penalty method is used to penalize those individuals violating the capacity constraints. For the feasible constraints of green splits, we build a “decoder” that intelligently avoids developing an illegal individual from the chromosome (Ge 1999). A decoder is a problemspecific operator, and its use usually can preserve feasibility of the solutions. Consider a single signalized intersection \( j \) with \( L_j \) approaches: there is an \( L_j \)-gene substring \((s_{1j}^g, \ldots, s_{5j}^g)\) in the chromosome corresponding to the \( L_j \) green splits \( \lambda^g_a(a \in L_j) \). Note that green splits associated with a junction should satisfy Eqs. (21)–(22). When the decoding procedure is carried out, the gene \( s_{dj}^g \) in the substring \((s_{1j}^g, \ldots, s_{5j}^g)\) corresponding to \( \lambda^g_a \) is first mapped to \( \eta_a \) in the following range: \( 0 \leq \eta_a \leq \lambda^g_{a_{\max}} - \lambda^g_{a_{\min}} \). In fact, if \( s_{dj}^g \) is, for example, 00110011, let

\[
\eta_a = \frac{2^5 + 2^4 + 2^1 + 2^0}{255} = 0.2(\lambda^g_{a_{\max}} - \lambda^g_{a_{\min}})
\]

and then the \( \eta_a \) generated above falls in \([0, \lambda^g_{a_{\max}} - \lambda^g_{a_{\min}}]\). Further let

\[
\lambda_a = \lambda^g_{a_{\min}} + \eta_a \left( B_j - \sum_{d \in L_j} \lambda^g_{a_{d_{\min}}} \right) \bigg/ \sum_{d \in L_j} \eta_d \forall a \in L_j
\]

It can be shown that the green splits generated from the above expression satisfy the constraints Eqs. (21)–(22). Such a decoder can “translate” each individual in the population into a possible solution satisfying the constraints Eqs. (21)–(22). The decoder not only plays an important role in the initialization of a population, but is also a crucial component of genetic operators (that is, crossover, mutation). In this solution procedure, the traditional genetic operators are adopted. However, any change of one gene corresponding to some green splits is certain to cause a chain reaction arising from Eq. (22), which makes the other green splits related to the same intersection change simultaneously. Therefore, it is essential to call for the “decoder” procedure to avoid the violation of the feasibility of green splits.

To penalize those individuals violating the capacity constraints Eq. (20), we define an evaluation function as
eval(μ, λ, a ∈ Uj ∈ J Lj) = \frac{1}{μ + 1} + γ \\
\times \sum_{a ∈ Uj ∈ J Lj} \left[ \max \{x_a - ρ_aλ_a, 0\}\right]^{m_γ} \\
+ \xi \sum_{a ∈ Uj ∈ J Lj} \left[ \max \{x_a - ρ_aλ_a, 0\}\right]^{m_ξ}
(27)

where γ and ξ are penalty parameters that are required to be positive numbers, and m_γ and m_ξ are powers of the penalty terms. Usually m_γ ≥ 2 and m_ξ ≥ 2. In the above penalty function, the first term transforms the maximization problem into the minimization one, and adding 1 to μ precludes zero from appearing in the denominator; the last two terms are the penalties.

The generic algorithm solution procedure for solving the problem in Eqs. (19)–(26) is summarized below.

Step 1: Initialization
- Randomly generate an initial population (0) each of whose individuals satisfies the feasible constraints Eqs. (21)–(22) of green splits with the aid of the decoder procedure developed above. Let population generation n = 0.

Step 2: Evaluation
- Decoding: Transform the genotype of each individual in population (n) into its phenotype, that is, potential solution (including multiplier and green splits), by means of the decoder.
- Traffic assignment: Solve the lower-level problem using, for example, the method of successive averages (MSA) in terms of each of the potential solutions corresponding to population (n), and then produce the corresponding link flow patterns.
- Evaluation: Substitute each potential solution corresponding to population (n) and the corresponding link flow pattern into the evaluation Eq. (27). This produces fitness of each individual in population (n).

Step 3: Reproduction/Selection
- Reproduce a new population (n + 1) from population (n) according to the distribution of fitness.

Step 4: Genetic Operators
- Crossover: Cross two individuals chosen from population (n + 1) with a specified crossover probability, p_c. The newly born offspring replace the locations of their parents in population (n + 1).
- Mutation: Select one individual from population (n + 1) with a specified probability, p_m, to mutate and then place it in the location of the original individual.

Step 5: Convergence Test
- If the termination criterion is met, accept the best individual in population (n + 1) as the (approximate) solution to problem in Eqs. (19)–(26) and stop; otherwise, set n := n + 1 and return to Step 2.

The termination criterion is either (1) no further improvement in fitness values between two successive generations; (2) a preset number of generations is reached; or (3) a preset amount of running time is consumed.

Numerical Results

This section provides two numerical examples to demonstrate the effects of ATIS information quality on network reserve capacity. The first example uses a network consisting of one origin and one destination with two parallel links/routes, and the second uses the same network but adds an intersection to one of its links. The intersection is controlled by an actuated signal. The following Bureau of Public Roads (BPR)-type link travel time function is used throughout the analysis:

\[ \tau_a = \tau_0^a \left[ 1 + \alpha \left( \frac{x_a}{c_a} \right)^\beta \right], \quad \forall a \]

(28)

where \( \tau_0^a, x_a, \) and \( c_a \), respectively, denote the free flow travel time, traffic flow, and practical capacity of link \( a \), and \( \alpha \) and \( \beta \) are parameters to be calibrated. Commonly used values for \( \alpha \) and \( \beta \) are \( \alpha = 0.15 \) and \( \beta = 4 \). Capacities on those links leading into signalized intersections are determined by Eq. (8). In addition, \( \rho_a \) is fixed to be 0.95 for all links.

Example 1

Consider a small network that consists of two nodes and two parallel links (a two-link network). The travel demand between the origin and the destination is assumed to be \( q = 6 \) (vehicle units per unit of time). Here it is assumed that vehicles are counted in units (for example, pairs, dozens, and so on). The free-flow travel times on the two links are \( \tau_0^1 = 3 \) (unit times) and \( \tau_0^2 = 2.75 \) (unit times), and the link capacities are \( c_1 = 5 \) (vehicle units per unit of time) and \( c_2 = 4.5 \) (vehicle units per unit of time); that is, link 2 is faster but has smaller capacity than link 1 (one can think of link 2 as a freeway link and link 1 a collection of arterial links put into one).

The numerical results are shown in Fig. 1. Fig. 1(a) shows that network reserve capacity, contrary to expectations, monotonically decreases rather than increases as the quality of ATIS information (indicated by parameter \( \theta \)) becomes higher. The maximum network reserve capacity is reached at \( \theta = 0 \), when the two links have identical flow, that is, the maximum allowable flow rate of link 2. It is a known fact that the SUE solution in logit assignment converges to the UE solution as the dispersion parameter \( \theta \) approaches infinity. Fig. 1(a) indicates that the minimum network reserve capacity occurs at \( \theta = \infty \), that is, the UE solution.

The reason why such a situation exists can be found from the investigation of the changes in traffic flows and in travel times on both links. Figs. 1(b and c) reveal much of the reason. Note that as the quality of ATIS information increases, more road users recognize that link 2 is the faster route and choose it over link 1. Because a major portion of the demand is assigned to link 2, its capacity is filled up quickly, allowing no further increase of total travel demand. When the quality of ATIS information is low, however, traffic demand is more evenly distributed among the two routes, allowing more demand to be served until flow on one of the links reaches the capacity of that link.

In the above example, link 2 is faster than link 1 while its capacity is also lower. We also investigated the configuration where the fast route (link 2) also has larger capacity: \( c_1 = 4 < 4.5 = c_2 \). The results of this case are shown in Fig. 2. A prominent feature of this result is that the network reserve capacity no longer decreases monotonically with the increase of information quality, but in fact reaches the maximum at \( \theta = 0.4198 \). The reserve capacity increases before \( \theta \) reaches 0.4198 and decreases monotonically afterward. Moreover, flows on both links can simultaneously reach their maximum allowable flow rates. Here \( \theta = 0.4198 \) represents the most desired information level for fully utilizing the capacity of the sample network.
Example 2

In this example we study the effects of ATIS information on network reserve capacity in signalized road networks. While link capacities of unsignalized road networks are treated as constants, those of signalized networks are variables that depend on traffic signal timing. The sample network used in this example is shown in Fig. 3. It has a signalized intersection controlled by an actuated signal. There are two O-D pairs in this network, \( (A \to B) \) and \( (C \to D) \). The O-D demands are \( q^{AB} = 6.0 \) (vehicle units per unit of time) and \( q^{CD} = 3.6 \) (vehicle units per unit of time). Free-flow travel times \( \tau_{af} \), link capacities \( c_{a} \), and saturation flow rates \( s_{a} \) are given in Table 1.

The numerical results are shown in Fig. 4. Fig. 4(a) illustrates the pattern of change in network reserve capacity in the signalized network. Similar to the first case of the previous example, this...
The figure indicates that the reserve capacity of the signalized road network decreases monotonically as the quality of ATIS information improves. This result can be explained in the same way as in the previous example: better information allows a large portion of demand to use the fast route and thus saturate the weakest link of that route (that is, the link with smallest capacity), making it impossible to accommodate more traffic.

The fact that the supply of traffic information may decrease network reserve capacity may seem paradoxical, yet such paradoxes abound in network flow problems. An example is Braess’s paradox, where the addition of a link to a road network would make every traveler worse off (Sheffi 1985). In fact, network reserve capacity, like system travel time, is another performance measure of a road network. Therefore, the explanation of Braess’s paradox also applies to the reserve capacity paradox; that is, the reason such paradoxes arise is because users minimize their own travel cost rather than the system cost.

**Discussion**

As seen from the numerical results of the first example, the distribution of link capacities has an impact on the pattern of change in network reserve capacity. To find out more on this, we investigate the conditions under which both links simultaneously reach their maximum allowable flow rates, that is, $x_a = \rho c_a$ ($a = 1, 2$). Without loss of generality we assume that $r_a$ equals $r$ for all $a$.

Consider the flow on link 1, which is given by

$$f_1 = q \frac{\exp(-\theta \tau_1)}{\exp(-\theta \tau_1) + \exp(-\theta \tau_2)}$$

(29)

when the logit route choice model is used. The travel times on the two links are determined by Eq. (28). If both links simultaneously reach their maximum allowable flow rates, then

$$x_a = \rho c_a, \quad a = 1, 2$$

(30)

**Table 1. Link Data for Example 2**

<table>
<thead>
<tr>
<th>Link $a$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau^0_a$ (units of time)</td>
<td>3.45</td>
<td>2.0</td>
<td>1.15</td>
<td>3.0</td>
</tr>
<tr>
<td>$c_a$ (vehicle units per unit of time)</td>
<td>7.5</td>
<td>—</td>
<td>6.0</td>
<td>—</td>
</tr>
<tr>
<td>$s_a$ (vehicle units per unit of time)</td>
<td>—</td>
<td>10.5</td>
<td>—</td>
<td>10.0</td>
</tr>
</tbody>
</table>

Note: Numbers in first row represent link numbers: 1 = AFB; 2 = AE; 3 = EB; 4 = CED.

**Fig. 3. Network of example 2**

**Fig. 4. Changes in network states of signalized network shown in Fig. 3:** (a) pattern of change in reserve capacity multiplier; (b) change in traffic flows; (c) change in travel time; (d) change in green splits of link AE
Substituting Eq. (28) and the above expression, Eq. (30), into Eq. (29), we have
\[ \theta(t_2^0 - t_1^0) = -\frac{1}{1 + \alpha \rho^\beta} \ln \frac{c_2}{c_1}. \]  
(31)

There are five cases to discuss about Eq. (31):

Case 1: \( t_1^0 > t_2^0 \) and \( c_1 > c_2 \)

This corresponds to the first case of the first example where the slower route has greater capacity. In fact, this situation is often observed in real-world road networks, where a freeway is often regarded as a fast route with smaller capacity when compared with a collection of parallel streets whose collective capacity is larger but average travel speed is lower. Under such a situation both links (routes) can never reach their respective capacity simultaneously because such a condition requires \( \theta < 0 \). For networks of such structure the inferior routes are usually underutilized. In fact, the greater the capacity difference between the two routes and the better quality the ATIS information, the greater the underutilization of the capacity of the slower route.

Case 2: \( t_1^0 > t_2^0 \) and \( c_1 < c_2 \)

This corresponds to the second case of the first example where the longer link (route) also has smaller capacity. In this case it is possible for both links to reach their respective maximum allowable flow rates simultaneously, provided that

\[ \theta = -\frac{1}{(t_2^0 - t_1^0)(1 + \alpha \rho^\beta)} \ln \frac{c_2}{c_1} > 0 \]

When this occurs, the network reserve capacity also reaches its maximum.

Case 3: \( t_1^0 = t_2^0 \) and \( c_1 = c_2 \)

In this case the two links are identical and the reserve capacity is independent of \( \theta \), that is, the quality of ATIS information.

Case 4: \( t_1^0 = t_2^0 \) and \( c_1 \neq c_2 \)

Under such conditions, Eq. (31) does not hold, which implies that the two links do not reach their respective capacity flow no matter how good/bad the ATIS information quality is. That is, there will always be unused capacity on the route with larger capacity.

Case 5: \( t_1^0 \neq t_2^0 \) and \( c_1 = c_2 \)

In this case, Eq. (31) holds only when \( \theta = 0 \). Thus in general the link with a longer travel time is certain to have unused capacity.

Similar cases also arise from the second example—the one with signalized links. For brevity they are not discussed here.

These results strongly indicate that the reserve capacity of a road network is closely linked to the capacity of some critical links, usually those that attract much of the flow because they are on the shortest routes of many O-D pairs. Better ATIS information, on the other hand, tells travelers more accurately which are the shortest routes for them, and in so doing it expedites the saturation of the critical links, leading to lower reserve capacity.

Information uniformly applied to all travelers, therefore, is not a desirable strategy to accommodate more travel demand in a given road network without lowering its level of service. Since one cannot expect road users to choose slower routes over faster ones unless certain incentives/disincentives are provided (such as through congestion pricing), selective information, or information applied to only a subset of travelers that achieves a better utilization of unused capacities of inferior routes, may have a better chance both to increase reserve capacity and to reduce total system travel cost, a claim supported by earlier studies investigating the impact of market penetration of ATIS on congestion management [for example, Mahmassani and Herman (1990); Arnott et al. (1991); and Koutsopoulos and Lotan (1990)].

Conclusions

Through a bilevel formulation, we studied in this paper the effect of ATIS information on network reserve capacity. The findings are that (1) ATIS information, when applied uniformly to all travelers, tends to reduce network reserve capacity as the quality of information improves; and (2) this trend depends on network configuration and characteristics. For some networks a certain level of information supply actually leads to increased reserve capacity. Upon a thorough analysis of this dependency in a small network example, it is concluded that the provision of information invariably leads to the early saturation of links on fast routes, preventing the further increase of travel demand (due to capacity constraints) despite slower routes that still have unused capacity. Thus selective information, or information for a subset of travelers in the network that encourages travelers to use some of the slow routes, may produce a better distribution of traffic in the network, hence allowing the network to accommodate more traffic demand without sacrificing level of service.

Acknowledgments

This research was supported partially by the Research Grants Council of the Hong Kong Special Administrative Region (Project No. PolyU 5029/98E) when the first writer worked as a research assistant in early 1999 in The Hong Kong Polytechnic University, and partially by the Institute of Transportation Studies at the University of California at Davis. The support from both institutions is gratefully acknowledged. The careful, thorough, and helpful comments of the anonymous referees are also gratefully acknowledged.

Notation

The following symbols are used in this paper:

- \( c_a \) = practical capacity of link \( a \in L \);
- \( f_{ik}^s \) = flow rate along route \( k \) from origin \( r \) to destination \( s \);
- \( j \) = signalized intersection in network under consideration;
- \( K_{rs} \) = set of routes connecting O-D pair \( (r,s) \);
- \( q_{rs}^0 \) = existing travel demand between O-D pair \( (r,s) \);
- \( x \) = column vector of all link flows (that is, link flow pattern);
- \( x_a \) = flow on link \( a \in L \);
- \( n_a \) = 1 if route \( k \in K_{rs} \) uses link \( a \), and 0 otherwise;
- \( \mu \) = multiplier to existing O-D trip table indicating network reserve capacity, referred to as reserve capacity multiplier;
- \( \rho_a \) = maximum acceptable degree of saturation for link \( a \in L \), then \( \rho_a c_a \) is maximum allowable flow rate on link \( a \in L \); and
- \( \tau_a \) = travel time on link \( a \in L \).
References