Overview
Traffic network modeling is a broad domain covering various problems that aid transportation design, planning, operations and policy. A key problem in traffic network modeling is to analyze the highway system performance.

The analysis of highway system performance is built upon traffic assignment, which is a process of allocating the given origin-destination (OD) demand to an analytical representation of the traffic network under some path choice principles. The traffic assignment process incorporates the demand (represented by the OD demand matrix) and supply (represented by the traffic network) characteristics that interact through a certain path choice mechanism. The outputs of this interaction are the link (path) flows and corresponding link (path) travel times.

Through this process, decision-makers are able to predict the paths taken by travelers and the resulting traffic volumes on the highway system. The outputs of the network modeling can be used to assess the deficiencies of the existing highway system and evaluate alternative proposals to improve it.

The rest of this chapter will discuss traffic network modeling, focusing on each component. The next section discusses the process to translate a highway system into a topological structure. Section 3 provides approaches for modeling the travel delay in a highway system. Section 4 describes how an OD demand matrix is obtained for traffic network modeling. Section 5 introduces traffic assignment in terms of the fundamental path decision principles and their extensions. Section 6 summarizes the network design problem to improve the existing highway system.

1. Network Representation
A highway traffic network is modeled as a directed graph consisting of a set of links (or arcs) to represent the roads carrying traffic flows, and a set of nodes (or vertexes) to represent the intersections. Nodes and links in the graph are indexed to identify them, and have associated attributes to represent their specific characteristics, such as capacities and travel times. This section focuses on the topological representation of the traffic network and associated attributes of nodes and links.
Intersection Representation

An intersection of roads in a highway traffic network is typically represented as a node in a graph. Figure 3.1a illustrates a typical representation of a four-legged intersection. Nodes that generate trips are called origins, while nodes that absorb trips are called destinations. A node can be an origin, a destination, both or neither.

Detailed movements at an intersection can be represented by a set of nodes connected by a set of links that represent the vehicular movement directions at the intersection. Figure 3.1b illustrates the representation of a four-legged intersection when detailed vehicular movements are considered.

In most traffic network problems, nodes have no associated attributes. In some specific problems, e.g., dynamic traffic assignment, nodes may have capacities that are associated with the control strategies imposed at intersections (Tampère et al., 2011).

Road Representation

A road in a highway traffic network is typically represented as a directed link whose direction is that of the traffic movement. A link is associated with several attributes, such as capacity, free flow speed (speed when the road is empty), flow, and travel time. Some link attributes, such as number of lanes and link length, can also be implicitly represented by other attributes such as link capacity and free flow travel time.

An important link attribute is the link performance function. It determines the link travel time given the link flow. Typically, the link travel time is illustrated as increasing with link flow due to congestion. A widely used link performance function is the Bureau of Public Roads (BPR) function:

$$ t_a = t_0 \left[ 1 + \alpha \left( \frac{x_a}{c_a} \right)^\beta \right] $$

(1)
where $t_a^0$, $x_a$ and $c_a$ represent free flow travel time, link flow and capacity of link $a$, respectively. Parameters $\alpha$ and $\beta$ determine the link travel time increase ratio. In general, $\alpha = 0.15$ and $\beta = 4$, indicating that link travel time increases in proportion to the fourth power of the volume-to-capacity ratio.

Figure 3.2 shows a plot of the BPR function, where link flow can exceed link capacity. However, consistent with the real world, the relationship between link travel time and link flow is more appropriately represented by a backward-bending curve as illustrated in Figure 3.3, where link flow decreases as congestion increases and link flow cannot exceed link capacity.

In Figure 3.3, as two different travel time values correspond to a link flow value, it is difficult to analytically represent this relationship in a simple form. Hence, BPR type link performance functions are used for modeling purposes when queuing in links is not a major concern.
When modeling congestion is important, various approximation approaches (such as point queues, spatial queues, etc.) or simulation tools are used. We will discuss how to model travel delay under congestion in Section 2.

**Topological Structure**

The topological structure of a traffic network is established through the analytical representation of the relationships between links and nodes. There are two types of representation: node-link incidence matrix and node-node adjacency matrix.

The topological structure of a network can be fully described by the node-link incidence matrix, defined as:

\[
A = \begin{cases} 
1 & \text{if link } a \text{ starts at node } i \\
-1 & \text{if link } a \text{ ends at node } i \\
0 & \text{otherwise.}
\end{cases}
\]

where \(A_{ia}\) denotes the entry in \(i\)th row and \(a\)th column of the node-link incidence matrix.

For example, suppose we have a directed network with four nodes connected by four links as shown in Figure 3.4. Then, the link-node incidence matrix \(A\) describes the topological structure of the network.

\[
A = \begin{bmatrix}
1 & 1 & 1 & 1 \\
-1 & 1 & 2 & \\
-1 & -1 & 3 & \\
-1 & & & 4
\end{bmatrix}
\]

Another network topology representation is the node-node adjacency matrix. Each element \(B_{ij}\) in the matrix indicates whether node \(i\) is adjacent to node \(j\). Thereby, the node-node adjacency matrix \(B\) also describes the topological structure of the network shown in Figure 3.4.

![Figure 3.4 A directed network](image-url)
Path Representation

A path in a highway traffic network is a sequence of links or nodes that connect an origin to a destination. Trips generated from an origin arrive at destinations through paths. Therefore, path flows represent the numbers of vehicles traveling on specific paths. The path flow representation allows the translation of OD demands into a link flow pattern.

One way to represent paths in a network is the link-path incidence matrix. Let $H$ denote a link-path incidence matrix. Each entry of the matrix denotes whether a link lies on a path. That is, $H_{apw} = 1$ if link $a$ lies on path $p$ connecting OD pair $w$, and $H_{apw} = 0$ otherwise. Each row of the matrix corresponds to one of the links and each column of the matrix corresponds to one of the paths. Thereby, the 1s in the row for link $a$ indicate which paths pass through that link, and the 1s in the column for path $p$ indicate which links are on path $p$.

The link-path incidence matrix can translate path flows into link flows. Here, a path (or link) flow is the volume (in number of vehicles per hour) of traffic traversing along a particular path (or link). Denote $x_a$ as the flow on link $a$ and $f_p$ as the flow on path $p$. As $H_{apw} = 1$ when path $p$ passes through link $a$ and $H_{apw} = 0$ when it does not, the total flow through link $i$ can be computed as:

$$x_i = \sum_{w \in W} \sum_{p \in P_w} H_{apw} f_p$$

where $P_w$ represents the set of paths connecting OD pair $w$. The above equation shows that a link flow is the summation of flows on all paths that pass through that link. Let $\mathbf{x}$ be a vector of link flows and $\mathbf{f}$ be a vector of path flows. The above equation can be written as:

$$\mathbf{x} = \Delta \mathbf{f}$$

(2)

In addition, an OD-path incidence matrix, denoted by $\Phi$, can be constructed to establish the relationship between path flows and OD demands. Each row of the matrix corresponds to one of the OD pairs and each column of the matrix corresponds to one of the paths. Then, $\Phi_{wp}$ if path $p$ connects OD pair $w$, and $\Phi_{wp}$ otherwise. Therefore,

$$d_w = \sum_{p \in P} \Phi_{wp} f_p$$

Let $d$ be a vector of OD demands. Then, we have a succinct representation of the above equation in matrix form as:

$$\mathbf{d} = \Phi \mathbf{f}.$$  

(3)
2. Modeling Travel Delay in Traffic Network

The BPR function (Equation 1) is unrealistic because it does not consider link interactions, capacity constraints and junction signals. To model the travel delay due to congestion and signals as well as link interactions and capacity constraints, various symmetric and asymmetric link performance functions were proposed in the literature (Nagurney, 1993).

A link performance function, including traffic signal delay, can be formulated by incorporating a signal delay function \(D(x,\lambda)\), as:

\[
t(x,\lambda) = t_s + D(x,\lambda)
\]

where \(x\) is the link flow, and \(\lambda\) is the green split. In particular, Webster’s delay formula can be used to compute the signal delay:

\[
D(x,\lambda) = \frac{x}{\lambda s(\lambda s - x)}
\]

where \(s\) is the saturation flow rate. However, Webster’s delay formula (5) only describes the undersaturated traffic condition, which may not be suitable when congestion occurs. For the saturated traffic condition, the Akcelik function, which includes a uniform delay term and an overflow delay term, can be used to compute the signal delay:

\[
D(y,\lambda) = \frac{0.5L(1-\lambda)}{(1-\lambda)y} + 900T \left[ y - 1 + \sqrt{(y - 1)^2 + \frac{8(y - 0.5)}{MT}} \right]
\]

where \(L\) is signal cycle length, \(y\) is the volume-to-capacity ratio, \(T\) is the study duration of the flow period, and \(M\) is the capacity.

The above functions only model the average cost of traversing a link and cannot capture on-and-off signal change in reality. In addition, both delay functions (Equation 5) and (Equation 6) do not capture the delay due to the vehicular queue. To capture the travel delay due to either on-and-off signal change or vehicular queue, traffic flow models are required to formulate a link performance function. These traffic flow models can be categorized as whole link models, queuing models and continuum traffic flow models.

When a whole link model is used, a general time-dependent link performance function considering traffic flow dynamics can be expressed as a function of the number of vehicles on a link \(x(\tau)\), inflow rate to a link \(u(\tau)\) and outflow rate \(v(\tau)\) (Carey, 2001):

\[
t(\tau) = f \left( x(\tau), u(\tau), v(\tau) \right)
\]

where \(t(\tau)\) is the delay or the link traversal time of the vehicle entering a link at time \(\tau\). The simplest form of a delay function only depends on the traffic volume. Accordingly, the traffic flow dynamics can be written as:

\[
\begin{align*}
\dot{x}(\tau) &= u(\tau) - v(\tau) \\
t(\tau) &= f \left( x(\tau) \right) \\
v(\tau + t(\tau)) &= u(\tau) / \left[ 1 + i(\tau) \right].
\end{align*}
\]
The link travel time function in (8) can be specified as a linear delay function:

\[ f(x(\tau)) = \alpha + \beta x(\tau), \quad \alpha > 0, \beta > 0 \]  

or a piece-wise linear function:

\[ f(x(\tau)) = \begin{cases} \alpha, & \text{if } x(\tau) < \alpha / \beta \\ x(\tau) / \beta, & \text{otherwise} \end{cases} \]

or any smooth and convex functions which satisfy the following conditions:

\[ f''(x) \text{ is continuous}, \]

\[ f(0) = \alpha \]

\[ f'(x) > 0, f''(x) > 0 \text{ for all } x > 0, \text{ and} \]

\[ f'(x) \to \beta \text{ as } x \to \infty. \]

When the point-queue model is adopted, the link performance function can be represented as the free-flow travel time plus the queuing delay:

\[ x(\tau) = t_0 + \frac{q(\tau)}{M} \]  

where \( q(\tau) \) is the queue length at time \( \tau \) and \( M \) is the capacity of the bottleneck. The point-queue model can be replaced by the spatial queue model or the double-queue model to compute the queue length.

If a continuum traffic flow model (e.g., LWR model) is used to describe traffic flow dynamics, then link travel time can be determined by a specified speed-density relationship, for instance, the modified Greenshields equation:

\[ v(\tau) = v_j + \left( v_0 - v_j \right) \left( 1 - \frac{k(\tau)}{k_j} \right)^\alpha \]

where \( v(\tau) \) is the space-mean speed at time \( \tau \), \( v_0 \) is the minimum speed, \( v_j \) is the free-flow speed, \( k(\tau) \) is the link density at time \( \tau \), \( k_j \) is the jam density, and \( \alpha \) is a user-specific parameter. Then the link travel time can be computed by \( t(\tau) = l/v(\tau) \) where \( l \) is link length.

3. OD Demand Matrix

An origin-destination (OD) demand matrix represents the demand aspects of the traffic network. Each element in the matrix denotes the travel desires from an origin to a destination. When an OD matrix is given, demand can be assigned onto the traffic network, and a link flow pattern will be produced.
Unfortunately, obtaining the OD demand matrix is not easy. In the classical four-step urban transportation planning model (Martin and McGuckin, 1998), the first three steps (trip generation, trip distribution, and mode split) are focused on determining the OD demand matrix. Traditional methods of estimating the OD matrix rely on large-scale surveys like home interview survey, roadside interview, and license plate method. These surveys are costly in terms of monetary expense, time, and manpower. Through surveys, data related to social and economic attributes of households are collected. After the socioeconomic data are collected, the trip generation process is performed to estimate the number of trips generated from each traffic analysis zone. Statistical regression is used to model the trip generation as a function of the socioeconomic attributes of households.

Based on the numbers of trips generated from traffic analysis zones, several methods can be used to assign the trips to different destinations, including the gravity, growth factor, and entropy-maximization models (Ortuzar and Willumsen, 2011). The gravity model is used extensively in the trip distribution process because it accounts for the attributes of the highway system and land-use characteristics. The gravity model for trip distribution is based on the concept that the attractiveness of a location declines with increasing distance (or travel time). The mathematical representation of the gravity model is as follows:

\[ T_{ij} = K_i K_j T_i T_j f(C_{ij}) . \]

where \( T_0 \) denotes the total number of trips between origin \( i \) and destination \( j \), \( T_i \) is the total number of trips departing from origin \( i \), \( T_j \) is the total number of trips attracted to destination \( j \), \( C_{ij} \) denotes the travel cost between origin \( i \) and destination \( j \), \( K_i \) and \( K_j \) are balancing factors, and \( f(\cdot) \) is a distance decay function. The output of the trip distribution process is a trip table. The summation of row \( i \) of the trip table is the total number of trips generated from origin \( i \), and the summation of row \( j \) of the trip table is the total number of trips attracted to destination \( j \). Note that the trips can be completed using one or more of several modes (auto, transit, biking, walking, etc.).

To estimate the demand by mode, a mode split process is needed. Mode choice of highway network users can be modeled by performing discrete choice analysis (Ortuzar and Willumsen, 2011). Utility maximization and maximum likelihood estimation can be used to split trips to different travel modes. After the mode split process, the vehicular OD demand matrix is obtained which can be used in traffic assignment.

When traffic count information is available on the links, the OD demand matrix can be estimated through mathematical programming. A general optimization formulation for OD estimation using traffic counts can be represented as:

\[ \min \gamma_1 F_1(T, \hat{T}) + \gamma_2 F_2(x, \hat{x}) , \]

subject to the link flows, \( x \), as the result of a traffic assignment process. In the formulation, \( \hat{T} \) is the target OD matrix, and \( \hat{x} \) the observed traffic counts with \( F_1 \) and \( F_2 \) being distance measure functions. The objective function (equation 14) can be specified using maximum likelihood, generalized least squares, or Bayesian Inference (Cascetta and Russo, 1997).

4. Traffic Assignment

This section focuses on a core issue of traffic network modeling: how is the travel demand between different origins and destinations distributed over the network? The travel demand is
loaded onto the network based on a principle by which travelers choose a path to complete their trips. The simplest path choice principle is all-or-nothing, where all travelers between an OD pair choose the same path (for example, based on the lowest travel cost). However, the all-or-nothing principle is limiting in that capacity constraints are ignored, implying that congestion is ignored.

The most widely used principles for traffic assignment are the user optimal and system optimal principles. They have been further extended by considering demand elasticity, perception errors, traffic dynamics, etc. This section will start with the user optimal and system optimal path decision principles, followed by their extensions known as elastic-demand traffic assignment, stochastic traffic assignment, and dynamic traffic assignment.

**Fundamental Path Decision Principles**

The path decision of traveler is impacted by many factors, such as travel time, fuel consumption, toll, comfort, convenience, etc. All these factors that influence the decision can be ascribed monetary values and summed to a generalized travel cost. As travel time is the most significant factor in traffic network modeling, it is typically used to represent the generalized travel cost.

Two fundamental path decision principles of highway network users were introduced by Wardrop (1952):

First principle: The journey times of all routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route.

Second principle: At equilibrium the average journey time is minimum.

The equilibrium characterized by Wardrop’s first principle is called “user equilibrium” as it corresponds to the state in which travelers seek to minimize their own travel cost, whereas the equilibrium characterized by Wardrop’s second principle is called “system optimum” as it corresponds to the state in which the total cost in the traffic network system is minimized.

The next few sections discuss how these two fundamental principles are rigorously formulated as mathematical models. A special example, known as the Braess paradox, is used to illustrate the difference between them.

**Traffic Assignment Based on the User Optimal Principle**

User optimal traffic assignment follows Wardrop’s first principle. It is based on the assumptions that all travelers are homogeneous and fully rational, and they have perfect knowledge of travel costs. Travelers make path choice decisions based on average travel cost. The solution of user optimal traffic assignment corresponds to the Nash Equilibrium, namely, travelers selfishly compete to travel on the shortest paths and continue to modify their path choice until they minimize their travel cost. User equilibrium is reached when all travelers stop changing their paths and none can reduce travel cost by unilaterally changing path.

Beckmann et al. (1956) rigorously formulated user equilibrium as an equivalent constrained optimization problem. In the formulation, the objective function is equal to the sum of the integrals of the link cost functions, integrated between zero and the current link flows. Mathematically, it has the form:

\[
\min_x z(x) = \sum_{a \in A} \int_0^{x_a} t_a(\omega) \, d\omega
\]  

(15)
subject to link flow definitional constraints (Equation 2), demand conservation constraint (Equation 3), and nonnegativity constraints $f^w_p$, for all $p \in P_w$. This nonlinear programming formulation can be applied to highway traffic networks with separable link cost functions, such as the BPR function (Equation 1).

The first-order optimality conditions for the optimal solution to the constrained optimization problem (Equation 15) are:

$$f^w_p\left(c^w_p - \pi^w\right) = 0, \quad \forall p \in P_w, \; w \in W$$  \hspace{1cm} (16)

$$c^w_p - \pi^w \geq 0, \quad \forall p \in P_w, \; w \in W$$  \hspace{1cm} (17)

$$\sum_{p \in \mathcal{P}_w} \Phi_{wp} f^w_p = d_w \quad \forall w \in W$$  \hspace{1cm} (18)

$$f^w_p \geq 0 \quad \forall p \in P_w, \; w \in W$$  \hspace{1cm} (19)

$$\pi^w \geq 0, \; d^w \geq 0 \quad \forall w \in W$$  \hspace{1cm} (20)

where $c^w_p = \sum_{s \in A} \Delta_{sp} t_s$ represents the path cost and $\pi^w$ represents the minimal path cost for OD pair $w$. The mathematical formulation (Equations 16–20) is called the nonlinear complementarity problem (NCP).

Nagurney (2004) rephrased user equilibrium as: “user travel costs on used paths for each OD pair are equalized and minimal.” Based on this statement, the user equilibrium flows must satisfy:

$$c_1^w = c_2^w = \ldots = c_p^w = \pi^w = \ldots = c_{p+1}^w \quad \forall p \in P_w, \; w \in W$$  \hspace{1cm} (21)

where flows on paths 1, 2, . . . , $p$ have positive values while flows on paths $p + 1, \ldots, P$ are zero. It has been shown that path flows satisfying conditions (Equation 21) must satisfy the NCP (Equations 16–20), and vice versa. Thus, the NCP (Equations 16–20) is also used as an alternative formulation for user optimal traffic assignment.

In addition, based on conditions (Equation 21), it can be shown that user equilibrium flows $f^*$ must satisfy an inequality:

$$(f - f^*)^T c(f^*) \geq 0 \quad \forall f \in \Omega$$  \hspace{1cm} (22)

where $\Omega$ represents the feasible path flow set, i.e.,

$$\Omega = \left\{ f \mid \Phi f = d, f \geq 0 \right\}.$$  \hspace{1cm}

Inequality (Equation 22) is called the variational inequality problem (VIP). Both the NCP and VIP models are more general than the optimization formulation for user equilibrium, as they can be applied to traffic networks with non-separable link cost functions.

**Traffic Assignment Based on the System Optimal Principle**

The system optimal traffic assignment follows Wardrop’s second principle. It addresses travelers’ path choice decisions from the system perspective, namely, travelers choose their paths so as to
minimize the total travel time in the highway system. Therefore, this principle may not be consistent with the more reasonable behavioral notion that travelers seek to minimize their individual path travel times.

In system optimal traffic assignment, travelers are assumed to make path choice decisions based on the marginal travel cost, instead of the average cost. Therefore, the mathematical models for user equilibrium can be revised by replacing the average cost with the marginal cost to formulate the system optimal state. Accordingly, the constrained optimization formulation for system optimal traffic assignment is:

\[
\min z(x) = \sum_{a \in A} t_a(x_a) \cdot x_a = \sum_{a \in A} \int_0^{\tilde{t}_a} \tilde{t}_a(\omega) \, d\omega
\]

subject to link flow definitional constraints (Equation 2), demand conservation constraint (Equation 3), and nonnegativity constraints \( f_p^w \geq 0 \) for all \( p \in P_w \). In the formulation, \( \tilde{t}_a \) represents the marginal travel time, i.e., \( \tilde{t}_a(x_a) = t_a(x_a) + t_a'(x_a) \cdot x_a \).

Similarly, an NCP can be developed for system optimal traffic assignment as:

\[
f_p^w \cdot (\tilde{c}_p^w - \tilde{\pi}_w) = 0, \quad \forall p \in P_w, w \in W
\]

\[
\tilde{c}_p^w - \tilde{\pi}_w \geq 0, \quad \forall p \in P_w, w \in W
\]

\[
\sum_{p \in P_w} \Phi_{wp} f_p^w = d_w \quad \forall w \in W
\]

\[
f_p^w \geq 0, \quad \forall p \in P_w, w \in W
\]

\[
\tilde{\pi}_w \geq 0, \quad d \geq 0 \quad \forall w \in W
\]

where \( \tilde{c}_p^w = \sum_{a \in A} \Delta_{at} \tilde{t}_a \) represents the marginal path cost and \( \tilde{\pi}_w \) represents the minimal marginal path cost for OD pair \( w \). And a VIP model for system optimal traffic assignment can be formulated as:

\[
(f - f^*)^T \tilde{c}(f^*) \geq 0 \quad \forall f \in \Omega
\]

**The Braess Paradox**

This section presents the well-known Braess paradox (Braess, 1969; Braess et al., 2005) to illustrate the difference between the concepts of user equilibrium and system optimum. By intuition, we expect that building more roads can enhance the highway system performance. However, the Braess paradox shows a contradiction in which the addition of a link causes the increase of average travel cost.

The classical Braess network is shown in Figure 3.5, where link 1 (from node 1 to node 2), link 2 (from node 1 to node 3), link 3 (from node 2 to node 4) and link 4 (from node 3 to node 4) form the basic network (Figure 5(a)); while link 5 (from node 2 to node 3) is a new link to be added (Figure 5(b)). The link performance functions are listed beside the links. The total demand traveling from origin node 1 to destination node 4 is 6.
Before the new link is added, there are two alternative paths: path 1 consists of node sequence 1: 2–4 and path 2 consists of node sequence 1: 3–4. In user equilibrium, 3 units of flow would choose path 1 and the rest choose path 2, resulting in the path travel cost \((10 \times 3) + (3 + 50) = 83\) time units. After link 5 is added, a new alternative path can be chosen by travelers. The new path consists of node sequence 1 \ldots 2: 3–4. In the new user equilibrium, all three paths are used, each of which carries two units of flow. The resulting path travel cost is \((10 \times 4) + (2 \times 50) = (10 \times 4) + (2 + 10) + (10 \times 4) = 92\) time units. Thus, each traveler has a higher travel cost after the new road is built. It illustrates that building more roads may not enhance the highway system performance.

The reason for the occurrence of the Braess paradox is due to the selfishness in travelers’ path choice behavior. Pas and Principio (1997) analyzed the relationship between the occurrence of the Braess paradox and the parameters in link performance functions and demand level. They showed that the Braess paradox occurs only when the travel demand is within a certain range.

### User Equilibrium with Elastic Demand

The standard User Equilibrium (EU) network assignment problem can be viewed as a model of traveler path choice regardless of the congestion level. However, travelers may adjust not only their path but also trip destination, mode, and departure time to respond to congestion. The elastic demand equilibrium assignment problem can be viewed as a model to describe the choice of not to travel, in addition to path choices.

Elastic demand equilibrium models consider the case when the OD trip rate is influenced by the level of service on the network. As the congestion level increases, travelers may change their travel mode, shift the departure time, alter the trip destination, or even cancel the trip. Therefore, the trip rate can be assumed to be a monotonically decreasing function of the shortest travel cost of an OD pair.

Let \(D_w(\omega)\) denote the demand function for OD pair \(w\). It determines the value of demand as \(d_w = D_w(\pi_w)\). As \(D_w(\omega)\) is a monotonically decreasing function, its inverse function \(D_w^{-1}(\omega)\) exists. Thus, the shortest travel cost between OD pair \(w\) is \(\pi_w = D_w^{-1}(d_w)\). Applying the inverse demand function \(D_w^{-1}(\omega)\) to the nonlinear programming model (Equation 15), the mathematical programming formulation for user equilibrium with elastic demand is as follows:

\[
\min_{x,d} z(x) = \sum_{\alpha \in A} \int_0^{x_{\alpha}} t_{\alpha}(\omega) \, d\omega - \sum_{w \in W} \int_0^{D_w^{-1}(\omega)} D_w^{-1}(\omega) \, d\omega
\]  

\[(30)\]
subject to link flow definitional constraints (Equation 2), demand conservation constraint (Equation 3), and nonnegativity constraints

\[ f_p^w \geq 0 \quad \forall p \in P_w \quad \text{and} \quad d_w \geq 0 \quad \forall w \in W. \]

**Stochastic Traffic Assignment**

The standard user equilibrium network assignment problem relies on the assumption of perfect knowledge of travel cost. However, in the real world, users may perceive travel costs differently leading to variability in response, which entails the need to model travelers’ perception errors.

Under stochastic traffic assignment, travelers choose their paths according to their perceived generalized costs. Utility maximization theory is used to model their path choice decision, where a rational traveler will choose the path with the minimum perceived generalized cost. The perceived generalized cost includes a deterministic component of travel cost and a random error term that captures variations in travel cost perception, tastes of individual trip making, as well as measurement errors. Therefore, the perceived travel cost (or disutility) can be formulated as:

\[ C_p^w = \bar{c}_p^w + \varepsilon_p^w, \quad \forall p \in P_w, w \in W \tag{31} \]

where \( C_p^w \) represents the perceived path travel cost, \( \bar{c}_p^w \) represents the deterministic (or measured) path travel cost, and \( \varepsilon_p^w \) represents the random error. Then, the probability that an individual driver will choose path \( p \) from the path set \( P \) is determined by:

\[
\begin{align*}
\text{Prob} \left( p \mid P_w \right) &= \text{Prob} \left( C_p^w < C_q^w, \quad \forall p \in P_w, q \neq p \right) \\
&= \text{Prob} \left( \varepsilon_p^w - \varepsilon_q^w < \bar{c}_q^w - \bar{c}_p^w, \quad \forall p \in P_w, q \neq p \right). \\
\end{align*}
\]

(32)

The joint cumulative distribution function of \( \varepsilon_p^w - \varepsilon_q^w \) is evaluated at \( \bar{c}_q^w - \bar{c}_p^w \).

If the distribution of \( \varepsilon_p^w \) is known, Equation 32 can be used to compute the probability of an individual’s path choice. Two probability distributions are widely used for the error term \( \varepsilon_p^w \). If \( \varepsilon_p^w \) follows the multinomial normal distribution, a probit model is obtained. However, the multinomial probit model does not have an analytically closed form. Solving the probit model relies on approaches such as the Monte Carlo simulation.

Another distribution used in stochastic traffic assignment is the Gumbel distribution, which has a cumulative density function as:

\[ F(x) = \exp \left\{ -\exp \left( x \right) \right\}. \tag{33} \]

Let \( \xi_p^w = -(1/\theta) \varepsilon_p^w \) be an independently and identically distributed Gumbel random variable, where \( \theta \) represents a dispersion parameter related to the variance of the Gumbel random error term. Then, the path choice probability has a closed form:

\[ \text{Prob} \left( p \mid P_w \right) = \frac{\exp \left( -\theta \xi_p^w \right)}{\sum_{q \in P_w} \exp \left( -\theta \xi_q^w \right)}. \tag{34} \]
For a given OD pair \( w \), the corresponding flow on path \( p \in P_w \) is determined by

\[
f_p^w = d^w \cdot \text{Prob} \left( p \mid P_w \right).
\]

As \( \sum_p \text{Prob} \left( p \mid P_w \right) = 1 \), it ensures that \( \sum_p f_p^w = d^w \).

To identify a solution in stochastic traffic assignment, Daganzo and Sheffi (1977) introduced the Stochastic User Equilibrium (SUE) principle.

**Stochastic User Equilibrium principle:** At stochastic user equilibrium, no traveler can improve his or her perceived travel time by unilaterally changing routes.

Based on the SUE principle, Fisk (1980) proposed a nonlinear programming model for SUE:

\[
\min_z (x) = \sum_{a \in A} \int_{0}^{x_a} t_a(\omega) \; d\omega + \frac{1}{\theta} \sum_{w \in W} \sum_{p \in P_w} f_p^w \ln f_p^w
\]

subject to link flow definitional constraints (Equation 2), demand conservation constraint (Equation 3), and nonnegativity constraints, \( f_p^w \geq 0 \) for all \( p \in P_w \). Sheffi and Powell (1982) proposed an unconstrained minimization program for the SUE problem as follows:

\[
\min_z (x) = - \sum_{w \in W} d^w S_w \left[ C^w(x) \right] + \sum_{a \in A} x_a(t_a(x)) - \sum_{a \in A} \int_{0}^{x_a} t_a(\omega) \; d\omega
\]

where

\[
S_w \left[ C^w(x) \right] = E \left[ \min_{\omega \in \Omega} \left\{ C^w_{\omega} \mid C^w(x) \right\} \right]
\]

Equation 37 defines the expected perceived travel cost for OD pair \( w \), which depends on the perceived costs of the entire path set. As the expectation cannot be evaluated readily, the Monte Carlo simulation is used for solving this unconstrained optimization problem.

**Dynamic Traffic Assignment**

All models discussed heretofore are based on the assumption that path flow is present simultaneously on all links of the path. These models, called static traffic assignment models, focus on representing average or steady-state conditions over an analysis time period that is long enough to allow all traffic flows to arrive at their destinations. However, these models cannot account for the variations of travel times and flows, because travel times and flows on links and paths are constant over the analysis period. Therefore, they are unable to analyze phenomena related to time-dependent traffic conditions, such as oversaturated traffic flow, queue spillback, dynamic routing, and peak spreading.

As active traffic management strategies require the accurate modeling of the within-day traffic dynamics and traveler adaptation to traffic dynamics, advanced traffic analysis tools are needed to predict network conditions and analyze network performance in both the planning and operational stages. Dynamic traffic assignment (DTA) (Peeta and Ziliaskopoulos, 2001) models can be used to aid decision-making for operational, construction, or demand management actions.
that are likely to induce a temporal or spatial traffic pattern shift. DTA models can interact
with other traffic management or control systems, to provide travel mode, departure time, path,
and other traffic information.

DTA follows the same process as static traffic assignment. However, DTA requires time-
dependent OD matrices, and the time-dependent flow-density relationship (also known as
fundamental diagram of traffic flow) needs to be specified for each link in the network. In DTA,
travel demand enters the network over time, characterized by the time-dependent OD matrix.
In addition, as the fundamental diagram is used in DTA models, travel time and flow on a link
are more accurately associated with traffic density, such that the backward bending curve for
travel time illustrated in Figure 3.3 can be captured.

The path decision principles used in DTA can be viewed as extensions of Wardrop’s first
and second principles and the SUE principle. Demand elasticity and perception errors can
also be considered. For example, the path decision principle for user equilibrium DTA can be
defined as:

User equilibrium DTA principle: The routes chosen by those departing at the same time
between the OD pair should have equal experienced travel time.

Modeling user equilibrium DTA is significantly more complex than modeling static traffic
user equilibrium assignment as the temporal dimension needs to be included in the model.
Analytical models have been developed by formulating user equilibrium DTA as a constrained
optimization program, variational inequality system, and nonlinear complementarity problem.
Constraints in these analytical approaches include flow conservation constraints, flow propagation
constraints, boundary constraints, and nonnegativity constraints, all of which are time-dependent.
The flow propagation constraints in analytical DUE approaches rest on various link performance
models, such as point-queue model, whole link model, and cell transmission model. Another
class of DTA models is the simulation-based DTA models, where the traffic flow modeling is
represented using a traffic simulator. They circumvent some of the analytical modeling challenges
associated with traffic flow modeling, and are hence more commonly used for DTA deployment
in practice.

5. Network Design Problem

The highway network design problem seeks to optimize the improvement of a highway network
with respect to a system objective while considering the path choice behavior of network users
and deployment constraints. A typical network design problem includes four main components:
objectives, decision variables, constraints, and underlying path choice behavior. A comprehensive
review on highway network design problems is provided by Farahani et al. (2013)

A highway network design problem generally includes two objectives: (1) at the system level
the objective is to optimize the overall system performance; (2) at the user level, the individual
network users seek to minimize their travel costs. The decision-making at the system level will
affect the responses of network users’ path choices. If the decision-maker knows how network
users respond to a decision he/she makes, then the network design problem is a Stackelberg
game from a game theory perspective.

A Stackelberg game is usually formulated as a bi-level programming problem, whose applica-
tions in transportation have been reviewed by Clegg et al. (2001). The upper-level problem
is formulated from the perspective of the leader (decision-maker). The upper-level problem has
the following generic structure:
subject to \( G(x, y(x)) \leq 0 \), where \( y(x) \) is a solution of the lower level (follower) optimization problem for any fixed \( x \):

\[
\min_{y \in Y} f(x, y)
\]

subject to \( g(x, y) \leq 0 \).

The structure of the bi-level programming problem typically does not maintain convexity even under the condition that both the upper and lower problems are convex. Therefore, it may have multiple local optima and is difficult to solve.

As the lower-level problem represents the path choice behavior of highway network users, it can be formulated as a fixed demand user equilibrium (Equation 15), user equilibrium with elastic demand (Equation 30), stochastic user equilibrium (Equation 35), or dynamic user equilibrium. If the lower-level program employs nonlinear complementarity (Equations 16–20) or variational inequality (Equation 22) to represent users’ path choice behavior, then the network design problem is formulated as a mathematical program with equilibrium constraints (MPEC) and mathematical program with complementarity constraints (MPCC), respectively.

The upper-level problem represents the decision-maker’s objective. It can be categorized as discrete or continuous network design problems, depending on whether or not the decision variables are integer. Discrete network design relates to the topology of the highway network; for example, building new roads or road closure. Continuous network design addresses the parameterization of the network, for example, capacity expansion, road pricing, and signal timings. Some network design problems contain both integer and real decision variables. Such mixed network design problems address both the topology and parameterization of the network, for example, adding new infrastructure and expanding capacity.

The main constraints in the upper-level program include budget constraints, resource availability constraints, capacity constraints, upper and lower bounds, etc. Objectives of the upper-level programming problem include minimizing total system travel time or construction cost, or maximizing social welfare, total revenue, the ratio of total revenue to total cost, or the network reserve capacity, etc.

References


