Discussion Paper No. 130

TRAFFIC EQUILIBRIA ANALYSED VIA GEOMETRIC PROGRAMMING

by

Michael A. Hall and Elmor L. Peterson

February, 1975

Traffic Equilibria Analysed via Geometric Programming [1,2,3]

by

Michael A. Hall[4] and Elmor L. Peterson[5]

Abstract. The "traffic-assignment problem" consists of predicting "Wardrop-equilibrium flows" on a roadway network when origin-to-destination "input flows" are specified. The "demand-equilibrium problem" consists of predicting those input flows that place the network in a state of "economic equilibrium" when the input flows are related via given travel-demand (feedback) curves to the resulting Wardrop-equilibrium origin-to-destination "travel costs".

The traffic-assignment problem is treated as a special case of the demand-equilibrium problem (the case in which the travel-demand curves are graphs of constant functions); and the demand-equilibrium problem is formulated and studied in the context of (generalized) "geometric programming". In particular, existence, uniqueness and characterization theorems are obtained via the "duality theory" of geometric programming by introducing appropriate "extremality con-

---

1. Prepared for the "International Symposium on Traffic Equilibrium Methods", held at the Centre de Recherche sur les Transports of the Université de Montréal, November 21 through November 23, 1974.

2. The authors are indebted to Professor E.K. Morlok of the University of Pennsylvania for directing their attention to the traffic equilibrium problems analysed herein.

3. Research partially sponsored by the National Science Foundation under Grant #GU 3851 to the Urban Systems Engineering Center at Northwestern University, and also by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant #AFOSR-73-2516. The United States Government is authorized to reproduce and distribute reprints of this paper for Governmental purposes notwithstanding any copyright notation hereon.

4. Bell Laboratories, P.O. Box 2020, New Brunswick, New Jersey 08903.

5. Department of Industrial Engineering/Management Sciences and Department of Mathematics, Northwestern University, Evanston, Illinois 60201.

ditions" and their corresponding "dual variational principles" (sometimes called "complementary variational principles"). These dual variational principles and extremality conditions also lead to new computational algorithms that show promise in the analysis of relatively large networks (such as those in relatively large urban or metropolitan areas).

All of this is done for a relatively flexible model in which (1) each "commodity" is permitted to preselect the only feasible "paths" (from its given origin to its given destination) on which some of its input flow can be assigned, (2) each link travel cost need only be nondecreasing and unbounded from above as a function of the total traffic flow on the link, and (3) each commodity input flow, as specified by the given commodity's travel-demand (feedback) curve, need only be nonincreasing and not approach $+\infty$ without its corresponding Wardrop-equilibrium origin-to-destination travel cost approaching $-\infty$. Moreover, analogies are drawn with "monotone network problems" (which arise in various physical contexts, such as the analysis of electric and hyraulic networks, but involve only a single commodity).

No prior knowledge of traffic equilibria, geometric programming, or monotone network theory is required.

TABLE OF CONTENTS

1. Introduction. In studying the traffic-assignment problem we consider "multicommodity networks" (i.e. roadway networks) on which each commodity (i.e. each traffic type as distinguished by origin, destination, and perhaps other criteria, such as vehicle class) distributes its own input flow (specified by its own travel demand) over certain preselected feasible network paths (subject of course to any capacity constraints) in such a way that a Wardrop-equilibrium [41] is achieved. A Wardrop-equilibrium occurs when the total origin-to-destination travel cost per unit of flow (e.g. the total origin-to-destination travel time [per unit of flow]) is the same on each path used by a given commodity and is not greater than that on each path feasible for but unused by the given commodity. In particular, we suppose that each network link contributes to each such cost per unit of flow a term that (due to traffic congestion) is actually a nondecreasing function of the total flow on that link and approaches plus infinity at the link's total capacity flow (which may itself be plus infinity).

This multicommodity problem of predicting Wardrop-equilibrium flows (i.e. the traffic-assignment problem) is somewhat similar to, but more complicated than, the single-commodity problem of predicting "equilibrium flows on monotone networks" (e.g. "Kirchoff-Ohm equilibrium flows" on non-linear electric or hydraulic networks). In fact, Charnes and Cooper [6] have attempted to make appropriate extensions of the classical variational principles [9,10,3,21,2,34] that reduce such single-commodity equilibrium problems to equivalent optimization problems. However, their extensions are at best imprecise (e.g. vague assertions about how Wardrop-equilibria can be treated as a special case of "Nash-equilibria" are made without a definition of the "game" under consideration). Moreover, the models treated are extremely unrealistic (e.g. one-way streets are not included, and traffic on a given two-way street must always experience the same flow "resistance" in both directions). Although Dafermos and Sparrow [7,8] have made such extensions precise and more realistic (without relying on game-theoretic concepts), a more thorough study of the Wardrop-equilibrium problem is possible. However, to carry out such a study with maximal simplicity and ease, the theory of monotone networks should be generalized by replacing certain vector spaces with (more general) polyhedral cones. Actually, the required generalization is already available in the form of (generalized) geometric programming [25,26, 28,29,32].

In addition to relaxing some of the assumptions made by previous workers, our approach to the Wardrop-equilibrium problem provides extra insight and additional algorithms via the modern duality theory [14,25,26,28,29,32,35,36] of mathematical programming. In doing so, it also subsumes both the work of Murchland [22] and the very recent work of Evans [12] by providing the first reasonably definitive treatment of the most fundamental questions having to do with the existence, uniqueness and properties of Wardrop-equilibria. Finally, it should be noted that our approach is fundamentally different from that taken by Rosenthal [37], which emphasizes the game-theoretic aspects of discrete multicommodity networks on which each player controls an indivisible unit of flow.

In discussing the demand-equilibrium problem we assume in the context of the preceding Wardrop-equilibrium problem that the input flow of a given commodity (given by its own travel demand) is actually related via a given nonincreasing travel-demand (feedback) curve to the resulting origin-to-destination travel cost per unit of flow experienced by the given commodity when the network is in a state of Wardrop-equilibrium. In particular then, a demand-equilibrium occurs when the network is in a state of Wardrop-equilibrium and the resulting origin-to-destination travel cost per unit of flow experienced by a given commodity produces via its own travel-demand curve the same input flow already specified for the given commodity. It is, of course, clear that this demand-equilibrium problem is far more complicated than the corresponding Wardrop-equilibrium problem; in fact, it is obvious that this demand-equilibrium problem reduces to just the corresponding Wardrop-equilibrium problem when the travel-demand curves are the graphs of constant functions that are identical to the specified input flows. Nevertheless, the mathematics required to treat the demand-equilibrium problem is no more sophisticated than that required to treat the Wardrop-equilibrium problem. Geometric programming does both jobs with equal ease.

It seems that the demand-equilibrium problem was first identified as essentially a mathematical programming problem by Beckmann [1]. Although Beckmann's work has spawned numerous papers on the algorithmic treatment of more realistic models (as reviewed by Nguyen [23]), it seems that there has been relatively little progress of a more fundamental nature. In fact, it seems that this paper provides the first reasonably definitive treatment of the more fundamental questions having to do with the existence, uniqueness and properties of demand-equilibria -- a task that is carried out for even more realistic and flexible models. In doing so, this paper also subsumes both the work of Murchland [22] and the very recent work of Evans [12] by providing a more direct and simple mechanism through which modern duality theory can be applied to the study of demand-equilibria. In addition to providing considerably more insight, this mechanism also serves as the basis for a variety of new algorithms, some of which may be especially suited to the analysis of relatively large networks (such as those in relatively large urban or metropolitan areas) when used in conjunction with appropriate geometric programming "decomposition principles" [27,28,31]. Finally, it should be noted that there are alternative (nonprogramming) approaches to the study of demand-equilibria -- the earliest ones having been summarized by Potts and Oliver [33], with the more recent ones having been surveyed by Ruiter [38].

2. The Model. A roadway network can be conveniently represented by a "directed graph" (consisting of "nodes" and "directed links") on which there is multi-commodity flow.

Each commodity (i.e. each traffic type) is associated with a (fictitious) "return link" over which only that particular commodity flows. Such a return flow takes place only in the prescribed return link direction, which is of course from the commodity's given destination (node) back to its given origin (node). For the example network shown in Figure 1, there are a total of three commodities: a single commodity with origin-to-destination pair (1,5) has return link 1, and two commodities with the same origin-to-destination pair (2,5) have return links 2 and 3.

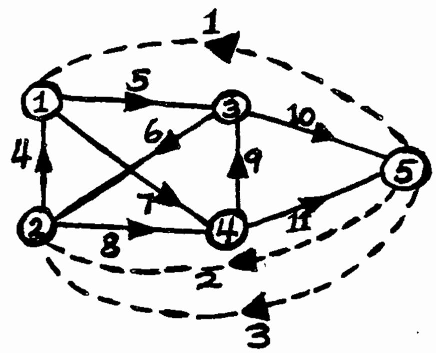


Figure 1.

We show the return links of the network with broken lines and the "real links" with unbroken lines. Each real link represents a collection of unidirectional lanes over which any of the given commodities might flow from a certain intersection (node) to an adjacent intersection (node). Such a (multicommodity) flow can take place only in the prescribed real link direction, which is of course dictated by local traffic regulations. For the example network shown in Figure 1, flow can take place on real link 5 only from node 1 to node 3, etc. Thus, two-way streets are represented by at least two real links.

In the most general model to be studied here, we consider a directed graph with a total of n links. In particular,

$$\text{the return links are numbered } 1,2,\ldots,r,$$

and

$$\text{the real links are numbered } r+1,\ r+2,\ldots,n.$$

Hence, there are a total of r commodities and a total of $(n-r)$ real links over which the r commodities can flow. For convenience we number the r commodities so that

commodity i has return link i for $i = 1,2,\ldots,r$.

With each commodity i is associated the family of all possible "paths" over which that particular commodity might conceivably flow (over real links from its given origin to its given destination). For the example network shown in Figure 1, the family of all possible paths over which commodity 2 might conceivably flow consists of: the path over links 4,5, and 10; the path over links 4, 7, and 11; the path over links 4,7,9, and 10; the path over links 8,9, and 10; and the path over links 8 and 11. Note, however, that links 8,9,6,8, and 11 do not constitute such a path; the reason is that a path (by definition) can not "cross" itself. This rejection of such possible "routes" is realistic because a commodity would obviously be foolhardy to follow one of them. Moreover, it is clear that a given commodity's possible route family is finite if and only if it includes only paths, in which event it is of course identical to the given commodity's possible path family. Since a possible path family is obviously always finite, and since a possible route family is clearly almost always infinite, this rejection of possible routes that are not paths is also a mathematical nicety. Needless to say, we assume that the network is sufficiently "connected" to guarantee that the possible path family for each commodity i is not empty.

It is obvious that two or more different commodities exhibit the same possible path family if and only if they have the same origin and destination. For the example network shown in Figure 1, commodities 2 and 3 (and only commodities 2 and 3) exhibit the same possible path family (the one enumerated in the preceding paragraph). To eliminate this ambiguity, each possible path is extended into a possible "circuit" by appending to each such path the given commodity's return link. For the example network shown in Figure 1, the possible path over links 8 and 11 is extended into the possible circuit around links 8, 11, and 2 when it is to be associated with commodity 2, but is extended into the possible circuit around links 8, 11, and 3 when it is to be associated with commodity 3. Naturally, the family of all circuits obtained by appending a given commodity's return link to the end of each of its possible paths is termed the commodity's possible circuit family.

There are numerous reasons why a given commodity usually eliminates from consideration some of the paths (circuits) in its possible path (circuit) family. For example, a complete scanning of them all might be unrealistic for even moderate sized networks (as found in moderate sized urban or metropolitan areas); in which event only those that seem to show the most promise of being reasonably "short" can actually be considered. Moreover, local traffic regulations may not permit a given commodity (i.e. a given traffic type, such as trucks) on certain real links; in which event only those of its possible paths that do not include such real links can actually be considered. Finally, other factors (such as personal safety, driving pleasure, etc.) may cause a given commodity to eliminate some of its possible paths from further consideration. In any event, we assume that each commodity i selects at least one of its possible paths for further consideration; and we term the resulting (not necessarily proper) subfamily of its possible path family its feasible path family. Needless to say, we also term the corresponding subfamily of its possible circuit family its feasible circuit family.

For our purposes, the most convenient way to represent a path (circuit) is to first associate each network link k with the k'th component of the vectors in n-dimensional Euclidean space E_n. Then, a given path (circuit) can be represented by the vector whose k'th component is either 1 or 0, depending respectively on whether link k is or is not part of the given path (circuit). For the example network shown in Figure 1, the possible path over links 8 and

11 is represented by the vector (0,0,0,0,0,0,0,1,0,0,1) in E_{11}. Moreover, its extension to the corresponding possible circuit around links 8, 11, and 2 for commodity 2 is represented by the vector (0,1,0,0,0,0,0,1,0,0,1) in E_{11}, while its extension to the corresponding possible circuit around links 8, 11, and 3 for commodity 3 is represented by the vector (0,0,1,0,0,0,0,1,0,0,1) in E_{11}.

In the most general model to be studied here, we suppose that commodity i has a nonempty feasible circuit (path) family that is enumerated by the integer index set

$$[i] \stackrel{\Delta}{=} \{m_i, m_i+1, \ldots, n_i\} \qquad \text{for} \qquad i = 1,2,\ldots,r,$$

where

$$1 = m_1 \le n_1, \qquad n_1 + 1 = m_2 \le n_2, \qquad \ldots, \qquad n_{r-1} + 1 = m_r \le n_r \stackrel{\Delta}{=} m.$$

Thus, there is a total of m feasible circuits over which traffic can flow; and given a feasible circuit j (namely, an integer in the circuit index set $\{1,2,\ldots,m\}$) there is a unique commodity i (in the commodity index set $\{1,2,\ldots,r\}$) such that $j \in [i]$, which means that commodity i (and only commodity i) flows over the feasible circuit j. Moreover, the vector δ^j representing a given feasible circuit j has components

$$\delta_k^j = \begin{cases} 1 \text{ when } k = \text{that } i \text{ for which } j \in [i] \\ 0 \text{ when } k \neq \text{that } i \text{ for which } j \in [i] \text{ but } 1 \le k \le r \\ 1 \text{ when } r+1 \le k \le n \text{ and real link } k \text{ is part of circuit } j \\ 0 \text{ when } r+1 \le k \le n \text{ and real link } k \text{ is not part of circuit } j \end{cases}$$

Now, a possible circuit flow is just a vector

$$z \in E_m$$

whose j'th component z_j is simply the input flow on circuit j of that commodity i for which $j \in [i]$. Since each commodity can flow only in the given link directions, it is obvious that each possible circuit flow z must also satisfy the (vector) inequality

$$z \geq 0. \tag{1}$$

Of course, each such z generates a <u>possible total flow</u>

$$x \triangleq \sum_{j=1}^{m} z_j \delta^j \in E_n, \tag{2}$$

whose k'th component x_k is clearly the resulting total flow of all commodities on link k for $k = 1,2,\ldots,n$.

Given that

d_i is the total (non-negative) input flow of commodity i

and that

b_k is the total capacity of real link k (which may be $+\infty$),

a <u>feasible circuit flow</u> is just a possible circuit flow z that generates a <u>feasible total flow</u> x, namely, a possible total flow x such that

$$x_k = d_k \qquad \text{for} \qquad k = 1,2,\ldots,r \tag{3}$$

while

$$0 \leq x_k < b_k \qquad \text{for} \qquad k = r+1, r+2, \ldots, n. \tag{4}$$

(Even though conditions (1) and (2) clearly imply that the inequality $0 \leq x_k$ in condition (4) is redundant, it is explicitly included to indicate the interval over which certain cost functions c_k of x_k are to be defined.) Although a given feasible circuit flow z generates a unique feasible total flow x, it is obvious that a given feasible total flow x can generally be generated by more than a unique feasible circuit flow z; in fact, the set of all such feasible circuit flows z is clearly identical to the set of all solutions z to (the linear) conditions (1-2).

It is worth noting that the preceding definition of feasible flows (as well as all definitions and theorems to follow) does not invoke "nodal conservation laws" -- though it is easy to see that such laws are in fact implicitly satisfied. Actually, all roadway network models that are explicitly based on nodal conservation laws (in fact, many previous roadway network models) are erroneous in a very fundamental way. For example, although a flow (0,0,0,1,1,1,0,0,0,0,0) of commodity 1 on the network shown in Figure 1 obviously satisfies the nodal conservation laws, it is clearly in no real sense a feasible commodity flow. It is fortunate though that such extraneous flows are automatically eliminated by certain Wardrop-equilibrium solution techniques when the travel cost per unit of flow on each real link is strictly positive -- as demonstrated in [18].

On each real link k we suppose that the travel cost per unit of flow (e.g. the link's node-to-node travel time [per unit of flow]) perceived by each commodity i is a positive nondecreasing function c_k (of only the total flow x_k) that is continuous at 0 and approaches $+\infty$ at b_k; that is,

$$(H_1) \qquad c_k(x_k) \le c_k(x_k') \text{ when } 0 \le x_k \le x_k' < b_k \qquad \text{for} \qquad k = r+1, r+2, \ldots, n,$$

$$(H_2) \qquad 0 < c_k(0) = \lim_{x_k \to 0^+} c_k(x_k) \qquad \text{for} \qquad k = r+1, r+2, \ldots, n,$$

$$(H_3) \qquad \lim_{x_k \to b_k^-} c_k(x_k) = +\infty \qquad \text{for} \qquad k = r+1, r+2, \ldots, n.$$

Needless to say, none of these hypotheses (H) are very restrictive in the context of roadway networks.

2.1. Wardrop-equilibria. We assume that the (real link) capacities b_k are sufficiently large to handle the commodity input flows d_k; that is, we assume that there exists at least one feasible circuit flow z. Actually, for nontrivial roadway networks of practical interest, there are usually infinitely many feasible circuit flows z.

By definition, the only feasible circuit flows z that place a given roadway network in a state of Wardrop-equilibrium are those z that generate a feasible total flow x for which there are both (fictitious) return link travel revenues per unit of flow

$$y_k \in E_1 \qquad \text{for} \qquad k = 1,2,\ldots,r \tag{5a}$$

and real link travel costs per unit of flow

$$y_k = c_k(x_k) \qquad \text{for} \qquad k = r+1, r+2, \ldots, n \tag{5b}$$

such that the resulting revenue-cost per unit of flow vector

$$y = (y_1, y_2, \ldots, y_n)$$

satisfies the following "inner product" conditions

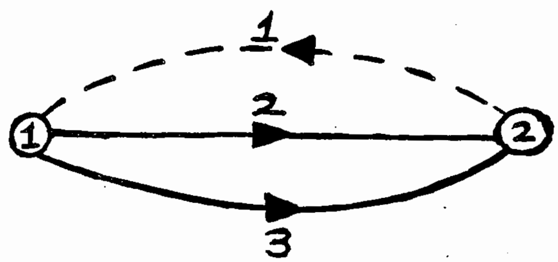
$$\langle \delta^j, y \rangle \begin{cases} = 0 \text{ if } z_j > 0 \\ \geq 0 \text{ if } z_j = 0 \end{cases} \qquad \text{for } j = 1,2,\ldots,m. \tag{6}$$

To properly interpret the preceding definition, note from the defining formula for the feasible circuit vectors δ^j that condition (6) simply asserts the following traffic situation: for each of the feasible circuits actually used by a given commodity i (namely, each circuit $j \in [i]$ for which $z_j > 0$) the total origin-to-destination travel cost per unit of flow is the same, in fact, just $-y_i$ (the negative of the corresponding "return link travel revenue per unit of flow" y_i), which in turn does not exceed the total origin-to-destination travel cost per unit of flow for each of the feasible circuits not used by the given commodity i (namely, each circuit $j \in [i]$ for which $z_j = 0$).

Each vector z that satisfies conditions (1-6) is termed a Wardrop-equilibrium circuit flow, and each such flow z generates a Wardrop-equilibrium total flow x via equation (2). Although a given Wardrop-equilibrium circuit flow z generates a unique Wardrop-equilibrium total flow x it is obvious that a given Wardrop-equilibrium total flow x can generally be generated by more than a unique Wardrop-equilibrium circuit flow z; in fact, the set of all such Wardrop-equilibrium circuit flows z is clearly identical to the set of all solutions z to both (the linear) conditions (1-2) and the "complementary slackness" conditions

$$\text{either } \langle \delta^j, y \rangle = 0 \text{ or } z_j = 0 \quad \text{for} \quad j = 1,2,\ldots,m. \tag{7}$$

<u>There are roadway networks that satisfy our basic hypotheses</u> (H) <u>and for which there are feasible circuit flows</u> z <u>but for which there are no Wardrop-equilibria.</u> A simple example is described by Figure 2.



$$d_1 = 4 \quad c_2(x_2) = \begin{cases} 6 & \text{for } 0 \le x_2 < 1 \\ 7 & \text{for } 1 \le x_2 < 5 \\ 2 + x_2 & \text{for } 5 \le x_2 \end{cases} \qquad c_3(x_3) = \begin{cases} 6.5 & \text{for } 0 \le x_3 < 5 \\ 1.5 + x_3 & \text{for } 5 \le x_3 \end{cases}$$

<u>Figure</u> 2.

Note that the lack of continuity of c_2 at $x_2 = 1$ seems to cause the lack of Wardrop-equilibria for the preceding example. Actually, continuity of all cost per unit of flow functions c_k is, in general, sufficient to guarantee the existence of Wardrop-equilibria -- a fact that turns out to be a corollary to a more general existence theorem.

The more general existence theorem is concerned with (Wardrop) "quasi-equilibria", which seem relevant to the prediction of traffic on roadway networks with discontinuous cost per unit of flow functions c_k, and which coincide with Wardrop-equilibria in the case of roadway networks with continuous c_k. Although all models known (by us) to be in use by traffic scientists and engineers involve continuous c_k, we suspect that in many cases more accuracy could be obtained by modeling with discontinuous c_k. For instance, link 2 of the network described by Figure 2 may model a roadway artery with two traffic control lights that are synchronized and timed relative to the speed limit in such a way that: when $x_2 < 1$ each vehicle moves at the speed limit and is stopped by only one of the lights; when $x_2 > 1$ each vehicle moves at less than the speed limit (due to congestion) and is stopped by both lights; when $x_2 = 1$ some vehicles are stopped by only one of the lights while other vehicles are stopped by both lights.

2.2. Quasi-equilibria. To define quasi-equilibria, we first imitate the theory of monotone networks [21,2,34] by embedding the graph of each cost per unit of flow function c_k in a (continuous) "complete nondecreasing curve"

$$\Gamma_k \triangleq \{(x_k,y_k) \mid \text{either } x_k = 0 \text{ and } y_k \le \lim_{\chi_k \to 0^+} c_k(\chi_k), \text{ or}$$

$$0 < x_k < b_k \text{ and } \lim_{\chi_k \to x_k^-} c_k(\chi_k) \le y_k \le \lim_{\chi_k \to x_k^+} c_k(\chi_k)\} \quad \text{for} \quad k=r+1,r+2,\ldots,n.$$

Each curve Γ_k is "nondecreasing" in the sense that, for each $(x_k,y_k) \in \Gamma_k$, and each $(x_k',y_k') \in \Gamma_k$, either $x_k \le x_k'$ and $y_k \le y_k'$, or $x_k' \le x_k$ and $y_k' \le y_k$. Each nondecreasing curve Γ_k is "complete" in the sense that it cannot be embedded in a (properly) larger nondecreasing curve. (In general, a complete nondecreasing curve Γ can be described geometrically as an infinite continuous curve that crosses each of the lines with slope -1 exactly once.)

It is natural to treat each Γ_k as a multivalued or point-to-set function γ_k on $[0,b_k)$ with functional values

$$\gamma_k(x_k) \triangleq \{y_k \mid (x_k,y_k) \in \Gamma_k\} \text{ for each } x_k \in [0,b_k) \quad \text{for} \quad k = r+1,r+2,\ldots,n,$$

each of which is a nonempty closed interval. For a given $x_k \in (0,b_k)$ the set $\gamma_k(x_k)$ is a singleton $\{y_k\}$ if and only if the cost per unit of flow function c_k is continuous at x_k, in which case $y_k = c_k(x_k)$. Thus, if c_k is continuous on $[0,b_k)$

then c_k differs from γ_k only in that $\gamma_k(0) = \{y_k \le c_k(0)\}$, a mathematical difference that turns out to be inconsequential in the determination of Wardrop-equilibria from quasi-equilibria.

By definition, the only feasible circuit flows z that place a given roadway network in a state of <u>quasi-equilibrium</u> are those z that generate a feasible total flow x for which there are both (fictitious) <u>return link quasi-revenues per unit of flow</u>

$$y_k \in E_1 \quad \text{for} \quad k = 1,2,\ldots,r \tag{5a'}$$

and <u>real link quasi-costs per unit of flow</u>

$$y_k \in \gamma_k(x_k) \quad \text{for} \quad k = r+1, r+2, \ldots, n \tag{5b'}$$

such that the resulting <u>quasi-revenue-cost per unit of flow vector</u>

$$y = (y_1, y_2, \ldots, y_n)$$

satisfies the following inner product conditions

$$\langle \delta^j, y \rangle \begin{cases} = 0 & \text{if } z_j > 0 \\ \ge 0 & \text{if } z_j = 0 \end{cases} \quad \text{for} \quad j = 1,2,\ldots,m \tag{6'}$$

To properly interpret the preceding definition, note from the defining formula for the feasible circuit vectors δ^j that condition (6') simply asserts the following traffic situation: for each of the feasible circuits actually used by a given commodity i (namely, each circuit $j \in [i]$ for which $z_j > 0$) the total origin-to-destination quasi-cost per unit of flow is the same, in fact just $-y_i$ (the negative of the corresponding "return link quasi-revenue per unit of flow" y_i), which in turn does not exceed the total origin-to-destination quasi-cost per unit of flow for each of the feasible circuits not used by the given commodity i (namely, each circuit $j \in [i]$ for which $z_j = 0$).

Each vector z that satisfies conditions (1-4,5',6') is termed a quasi-equilibrium circuit flow, and each such flow z generates a quasi-equilibrium total flow x via equation (2). Although a given quasi-equilibrium circuit flow z generates a unique quasi-equilibrium total flow x, it is obvious that a given quasi-equilibrium total flow x along with the corresponding quasi-revenue-cost per unit of flow vector y can generally be generated by more than a unique quasi-equilibrium circuit flow z; in fact, the set of all such quasi-equilibrium circuit flows z is clearly identical to the set of all solutions z to both (the linear) conditions (1-2) and the complementary slackness conditions

$$\text{either } \langle \delta^j, y \rangle = 0 \text{ or } z_j = 0 \qquad \text{for} \qquad j = 1,2,\ldots,m. \tag{7'}$$

It is easy to show that each Wardrop-equilibrium flow is a quasi-equilibrium flow and that, when all cost per unit of flow functions c_k are continuous, each quasi-equilibrium flow is a Wardrop-equilibrium flow. However, if at least one c_k is discontinuous, there can be quasi-equilibrium flows that are not Wardrop-equilibrium flows. For instance, the roadway network described by Figure 2 clearly has a (unique) quasi-equilibrium flow $x^* = z^* = (4,1,3)$ even though we have already noted that it has no Wardrop-equilibrium flow. Actually, in view of our previous interpretation of the discontinuity of c_2 in terms of traffic-control lights, it certainly seems reasonable to expect that (4,1,3) constitutes a (stable) traffic equilibrium. In any event, we shall henceforth be concerned (without any loss of generality) with the existence and properties of quasi-equilibria.

2.3. Demand-equilibria. We now suppose that the input flow d_i of a given commodity i is not necessarily fixed, but is instead related via a given complete nonnegative nonincreasing (feedback) curve to the origin-to-destination quasi-cost per unit of flow $-y_i$ experienced by the given commodity i when the network is in a state of quasi-equilibrium. In essence then, we suppose that each d_i is related via a complete nonnegative nondecreasing curve to its corresponding return link quasi-revenue per unit of flow y_i. Since it is not difficult to see that each complete nondecreasing curve has an "inverse" that is also a complete nondecreasing curve, we can equivalently suppose that

(H_4) each y_i is related via a complete nondecreasing curve Γ_i to its corresponding d_i,

where

(H_5) the "domain" of each such Γ_i is a (not necessarily proper) subinterval I_i of the interval of all nonnegative real numbers.

Moreover, we also suppose that

(H_6) the "range" of each such Γ_i consists of all real numbers (equivalently, a given d_i does not approach $+\infty$ without its corresponding y_i approaching $+\infty$)

Needless to say, none of these hypotheses (H) are very restrictive in the context of roadway networks.

As with each real link curve Γ_i, it is natural to treat each return link curve Γ_i as a multivalued or point-to-set function γ_i on I_i with function-values

$$\gamma_i(d_i) \triangleq \{y_i \mid (d_i, y_i) \in \Gamma_i\} \text{ for each } d_i \in I_i \quad \text{for} \quad i = 1, 2, \ldots, r,$$

each of which is a nonempty closed interval. For a given $d_i \in I_i$ the set $\gamma_i(d_i)$ is a singleton $\{y_i\}$ if and only if the "travel demand" from commodity i results in the given input flow d_i for only a single origin-to-destination quasi-cost per unit of flow $-y_i$.

By definition, the only feasible circuit flows z that place a given roadway network in a state of (quasi-)demand-equilibrium are those z that generate a feasible total flow x with corresponding feasible input flows

$$x_k = d_k \in I_k \qquad \text{for} \qquad k = 1, 2, \ldots, r \qquad (3'')$$

for which there are both (fictitious) return link quasi-revenues per unit of flow

$$y_k \in \gamma_k(d_k) \quad \text{for} \quad k=1,2,\ldots,r \tag{5a''}$$

and real link quasi-costs per unit of flow

$$y_k \in \gamma_k(x_k) \quad \text{for} \quad k=r+1,r+2,\ldots,n \tag{5b''}$$

such that the resulting quasi-revenue-cost per unit of flow vector

$$y=(y_1,y_2,\ldots,y_n)$$

satisfies the following inner product conditions

$$\langle \delta^j,y\rangle \begin{cases} =0 \text{ if } z_j>0 \\ \geq 0 \text{ if } z_j=0 \end{cases} \quad \text{for} \quad j=1,2,\ldots,m. \tag{6''}$$

To properly interpret the preceding definition, note that it differs from the definition of quasi-equilibria only in that the input flows d_i can now vary over I_i, and the symbol E_1 in condition (5a') has been replaced by the symbol $\gamma_k(d_i)$ in condition (5a"). Consequently, a demand-equilibrium occurs when the network is in a state of quasi-equilibrium and the input flow d_i of a given commodity i is related via its own travel-demand curve Γ_i to the resulting origin-to-destination quasi-cost per unit of flow $-y_i$ experienced by the given commodity i.

Each vector z that satisfies conditions (1,2,3",4,5",6") is termed a demand-equilibrium circuit flow, and each such flow z generates a demand-equilibrium total flow x via equation (2). Although a given demand-equilibrium circuit flow z generates a unique demand-equilibrium total flow x, it is obvious that a given demand-equilibrium total flow x along with the corresponding quasi-revenue-cost per unit of flow vector y can generally be generated by more than a unique demand-equilibrium circuit flow z; in fact, the set of all such demand-equilibrium circuit flows z is clearly identical to the set of all solutions z to both (the linear) conditions (1-2) and the complementary slackness conditions

$$\text{either } \langle \delta^j, y\rangle = 0 \text{ or } z_j = 0 \quad \text{for} \quad j = 1,2,\ldots,m. \tag{7''}$$

It is, of course, clear that the problem of predicting those quasi-equilibrium flows that result from fixed input flows d_k is a special case of the problem of predicting demand-equilibrium flows -- the case in which each return link function γ_k simply has a singleton domain $\{d_k\}$ and a single function value $\gamma_k(d_k) = E_1$ for $k = 1,2,\ldots,r$. Consequently, we shall henceforth concentrate (without any loss of generality) on the existence and properties of demand-equilibria.

3. The Key Characterization. First, we consider the irreducible commodity set

$$I \triangleq \{i \mid 1 \le i \le r \text{ and } I_i \text{ contains at least one number } d_i^+ > 0\}$$

and the (resulting) irreducible circuit (path) family

$$J \triangleq \bigcup_I [i].$$

If I (and hence J) is empty, it is clear that the only feasible circuit flow $z = 0$ places the given network in a state of demand-equilibrium (with a feasible total flow $x = 0$ and a [rather meaningless nonunique] corresponding quasi-revenue-cost per unit of flow vector y with components $y_k = 1$ for $k = 1,2,\ldots,r$, and $y_k = c_k(0)$ for $k = r+1, r+2,\ldots,n$). Consequently, we shall henceforth assume (without any real loss of generality) that I (and hence J) is not empty.

Now, we incorporate the feasibility conditions (1,2) into a set

$$X \triangleq \{x = \sum_J z_j \delta^j \mid z_j \ge 0 \text{ for } j \in J\}, \tag{8}$$

which is obviously a "convex polyhedral cone generated by" the irreducible circuit (path) vectors δ^j, $j \in J$. We also introduce its "dual cone"

$$Y \triangleq \{y \in E_n \mid 0 \le \langle x, y\rangle \text{ for each } x \in X\}, \tag{9a}$$

which clearly has the representation formula

$$Y = \{y \in E_n \mid 0 \le \langle \delta^j, y\rangle \text{ for each } j \in J\}. \tag{9b}$$

This representation formula (9b) shows that Y is also a convex polyhedral cone (whose generators can be computed in any given situation with the aid of a linear-algebraic algorithm due to Uzawa [40]).

By definition, the network-equilibrium conditions (for the general demand-equilibrium problem) are:

(I)	$x \in X$	and	$y \in Y$
(II)		$0 = \langle x,y \rangle$	
(III)	$y_k \in \gamma_k(x_k)$	for	$k = 1,2,\ldots,n.$

Each vector (x,y) that satisfies these conditions is termed a network-equilibrium vector.

The following theorem is the key to almost everything that follows.

Theorem 1. Each demand-equilibrium total flow x along with the corresponding quasi-revenue-cost per unit of flow vector y produces a network-equilibrium vector (x,y). Conversely, each network-equilibrium vector (x,y) produces a demand-equilibrium total flow x along with a corresponding quasi-revenue-cost per unit of flow vector y.

Proof. First, note from the defining equation (8) for X that conditions (1,2,3",4,5") are equivalent to the first part of condition (I) and all of condition (III). Consequently, we can complete our proof by showing that condition (6") is equivalent to the second part of condition (I) and condition (II).

Toward that end, simply note from the defining equation (8) for X that

$$\langle x,y \rangle \equiv \sum_J z_j \langle \delta^j, y \rangle \text{ for } x \in X. \tag{10}$$

It is then an immediate consequence of this identity (10) that condition (6") implies the second part of condition (I) and condition (II). On the other hand, it is an immediate consequence of the representation formula (9b) for Y and this identity (10) that condition (6") is also implied by the second part of condition (I) and condition (II). q.e.d.

It is, of course, a consequence of (the key) Theorem 1 that we can henceforth concentrate (without any loss of generality) on the existence and properties of network-equilibrium vectors (x,y). There are at least two reasons why such a concentration pays extremely high dividends: 1. the network-equilibrium conditions (I-III) are mathematically simpler than the demand-equilibrium conditions (1,2,3",4,5",6"), in that the latter also explicitly involve demand-equilibrium circuit flows z; and 2. the network-equilibrium conditions (I-III) turn out to be essentially the "extremality conditions" for (generalized) geometric programming and hence provide a mechanism through which the powerful theory of (generalized) geometric programming can be applied to the study of traffic equilibria.

The next section is devoted to actually constructing those geometric programming "dual problems" whose corresponding extremality conditions coincide with the network-equilibrium conditions (I-III).

4. Dual Variational Principles. Imitating the theory of monotone networks, we first "integrate" each complete nondecreasing point-to-set function γ_k. We do so by computing the (Riemann) integral of any real-valued (point-to-point) function c_k whose domain coincides with the domain of γ_k and whose function values $c_k(x_k) \in \gamma_k(x_k)$. Of course, for $k = r+1, r+2, \ldots, n$, the given real link travel cost per unit of flow function c_k is a legitimate choice for c_k. In any event, our monotonicity hypotheses (H_1) and (H_4) along with monotone function theory guarantee that such integrals exist and do not depend on the particular function-value choices $c_k(x_k) \in \gamma_k(x_k)$. Consequently, we denote such integrals by the symbol $\int^{x_k} \gamma_k(s)ds$ even though γ_k is generally not a real-valued function.

Our monotonicity hypotheses (H_1) and (H_4) along with monotone function theory also guarantee that each such integral $\int^{x_k} \gamma_k(s)\, ds$ gives a "convex function" g_k whose domain

$$C_k \overset{\Delta}{=} \text{the (interval) domain of } \gamma_k \text{ along with any endpoint } e_k \text{ not in the domain of } \gamma_k \text{ but for which the improper integral} \int^{e_k} \gamma_k(s)\,ds \text{ converges,} \tag{11a}$$

and whose function-values

$$g_k(x_k) \overset{\Delta}{=} \int_{\alpha_k}^{x_k} \gamma_k(s)\,ds \tag{11b}$$

where

α_k is any convenient point in the domain of γ_k for which $0 \in \gamma_k(\alpha_k)$ (the existence of such an α_k being guaranteed by hypotheses (H_2) and (H_6), with hypothesis (H_2) actually implying that $\alpha_k = 0$ for $k = r+1, r+2, \ldots, n$). (11c)

Each such convex function g_k is known to be "closed" in that its "epigraph"

$$(\text{epi } g_k) \overset{\Delta}{=} \{(x_k, y_k) \in E_2 \mid x_k \in C_k \text{ and } g_k(x_k) \le y_k\}$$

is a (topologically) closed subset of E_2.

On the other hand, each closed convex function g_k with domain $C_k \subseteq E_1$ has a (point-to-set) "subderivative" function ∂g_k with function-values

$$\partial g_k(x_k) \overset{\Delta}{=} \{y_k \mid g_k(x_k) + y_k(x_k' - x_k) \le g_k(x_k') \text{ for each } x_k' \in C_k\},$$

at least one of which is nonempty. Needless to say, subderivatives are actually generalized derivatives -- in that a convex function g_k with domain $C_k \subseteq E_1$ is differentiable at a point $x_k \in C_k$ if and only if $\partial g_k(x_k)$ is a singleton $\{y_k\}$, in which case the derivative $g_k'(x_k) = y_k$. However, for our purposes, the key fact is that the subderivative function ∂g_k for an arbitrary closed convex function g_k with domain $C_k \subseteq E_1$ is a complete nondecreasing point-to-set function (corresponding to a complete nondecreasing curve) -- whose integral is just g_k (plus, of course, an arbitrary "constant of integration"). In particular then, the defining equations (11) for our functions g_k imply that

$$\text{the domain of } \gamma_k \text{ is identical to the set of all points } x_k \text{ for which } \partial g_k(x_k) \text{ is not empty,} \tag{12a}$$

and

$$\gamma_k(x_k) = \partial g_k(x_k) \text{ for each } x_k \text{ in the domain of } \gamma_k. \tag{12b}$$

The reader is probably already familiar with the more special relation between (complete) nondecreasing continuous real-valued functions γ_k and their (closed) convex differentiable integrals g_k. Although a proof for the more general relation just described herein is not difficult, it is omitted because of its lengthiness and the fact that it can be found in [34]. However, as a simple (but important) example of the more general relation, note that when γ_k has a singleton domain $\{d_k\}$ and a single function-value $\gamma_k(d_k) = E_1$, the integral g_k also has the singleton domain $\{d_k\}$ but the single function-value $g_k(d_k) = 0$; from which it is obvious that $\partial g_k(d_k) = E_1$.

The optimization problem that serves as the basis for one of our dual variational principles involves the functions g_k in the guise of a function g whose domain

$$C \triangleq \mathop{\times}_{1}^{n} C_k \tag{13a}$$

and whose function-values

$$g(x) \triangleq \sum_{1}^{n} g_k(x_k). \tag{13b}$$

For obvious reasons this function g is said to be "separable" into the functions g_k. Moreover, it is not difficult to show that g <u>inherits the convexity and closedness of the</u> g_k.

From the cone X and the function g we now construct the following optimization problem.

Problem A. Using the "feasible solution" set

$$S \stackrel{\Delta}{=} X \cap C,$$

calculate both the "problem infimum"

$$\Phi = \inf_{x \in S} g(x)$$

and the "optimal solution" set

$$S^* \stackrel{\Delta}{=} \{x \in S \mid \Phi = g(x)\}.$$

The fact that X is a cone means that problem A is a (generalized) "geometric programming problem". (Justification for this terminology is given in [28,29,32]). Moreover, the fact that X is actually a convex polyhedral cone and the fact that g is a closed convex separable function imply that problem A is a closed convex separable programming problem with (essentially) linear constraints.

It is, of course, obvious that each feasible total flow x is also a feasible solution to problem A. On the other hand, it is clear that problem A can have a feasible solution x that is not a feasible total flow only if C_k is actually larger than the domain of γ_k for at least one k.

Imitating the theory of monotone networks, we also integrate each complete nondecreasing point-to-set "inverse function" γ_k^{-1}, whose function-values

$$\gamma_k^{-1}(y_k) \stackrel{\Delta}{=} \{x_k \mid (x_k, y_k) \in \Gamma_k\} \text{ for each } y_k \in E_1 \quad \text{for} \quad k = 1,2,\ldots,n,$$

each of which is a nonempty closed interval. In particular, each integral $\int^{y_k} \gamma_k^{-1}(t)\, dt$ gives a closed convex function h_k whose domain

$$D_k \stackrel{\Delta}{=} E_1, \tag{14a}$$

and whose function-values

$$h_k(y_k) \stackrel{\Delta}{=} \int_0^{y_k} \gamma_k^{-1}(t)dt. \tag{14b}$$

Naturally, each such h_k has the property that

$$\gamma_k^{-1}(y_k) = \partial h_k(y_k) \text{ for each } y_k \text{ in } E_1. \tag{15}$$

As a simple (but important) example, note that when γ_k has a singleton domain $\{d_k\}$ and a single function-value $\gamma_k(d_k) = E_1$, its inverse γ_k^{-1} has domain E_1 and a constant function-value $\gamma_k^{-1}(y_k) \equiv \{d_k\}$, so $D_k = E_1$ and $h_k(y_k) \equiv d_k y_k$; from which it is obvious that $\partial h_k(y_k) = \{h_k'(y_k)\} \equiv \{d_k\}$.

The optimization problem that serves as the basis for our other dual variational principle involves the functions h_k in the guise of a function h whose domain

$$D \stackrel{\Delta}{=} \overset{n}{\underset{1}{\times}} D_k \tag{16a}$$

and whose function-values

$$h(y) \stackrel{\Delta}{=} \sum_1^n h_k(y_k). \tag{16b}$$

Of course, this function h is separable, and h <u>inherits the convexity and closedness of the</u> h_k.

From the cone Y and the function h we now construct the following optimization problem.

<u>Problem B</u>. Using the feasible solution set

$$T \stackrel{\Delta}{=} Y \cap D,$$

calculate both the problem infimum

$$\Psi \stackrel{\Delta}{=} \inf_{y \in T} h(y)$$

and the optimal solution set

$$T^* \stackrel{\Delta}{=} \{y \in T \mid \Psi = h(y)\}.$$

The fact that Y is a cone means that problem B is a (generalized) geometric programming problem. Moreover, the fact that Y is actually a convex polyhedral cone and the fact that h is a closed convex separable function imply that problem B is a closed convex separable programming problem with (essentially) linear constraints.

Problem B has important features not possessed by problem A. In particular, note that relations (14a) and (16a) imply that

$$D = E_n, \tag{17}$$

which in turn clearly implies that

$$T = Y.$$

Since each cone Y contains at least the zero vector, we infer that problem B has at least one feasible solution (even when problem A has no feasible solutions).

We shall eventually see that "geometric programming duality theory" reduces the study of traffic equilibria to a study of either problem A or problem B. Actually, we shall then see that problems A and B should be studied simultaneously in order to provide the most insight into traffic equilibria. In essence, problem A describes traffic equilibria entirely in terms of total flows x, while problem B describes traffic equilibria entirely in terms of quasi-revenue-cost per unit of flow vectors y.

An important ingredient in our simultaneous study of problems A and B is the fact that the cones X and Y defined by equations (8) and (9), respectively, constitute a pair of (convex polyhedral) "dual cones"; that is

$$Y = \{y \in E_n \mid 0 \le \langle x,y \rangle \text{ for each } x \in X\}, \tag{18}$$

while

$$X = \{x \in E_n \mid 0 \le \langle x,y \rangle \text{ for each } y \in Y\}. \tag{19}$$

Of course, equation (18) is just a repetition of the defining equation (9a) for Y; and it is obvious from equation (18) that X is a subset of the right-hand side of equation (19). To show in turn that the right-hand side of equation (19) is a subset of X, and hence establish equation (19), simply use the representation formula (9b) for Y in conjunction with both the "Farkas lemma" [11] and the defining equation (8) for X.

Another important ingredient in our simultaneous study of problems A and B is the fact that <u>the functions</u> g_k <u>and</u> h_k <u>defined by equations</u> (11) <u>and</u> (14), <u>respectively</u>, <u>constitute a pair of (closed convex) "conjugate functions"</u>; that is,

$$D_k = \{y_k \mid \sup_{x_k \in C_k} [y_k x_k - g_k(x_k)] \text{ is finite}\} \tag{20a}$$

and

$$h_k(y_k) = \sup_{x_k \in C_k} [y_k x_k - g_k(x_k)], \tag{20b}$$

while

$$C_k = \{x_k \mid \sup_{y_k \in D_k} [x_k y_k - h_k(y_k)] \text{ is finite}\} \tag{21a}$$

and

$$g_k(x_k) = \sup_{y_k \in D_k} [x_k y_k - h_k(y_k)]. \tag{21b}$$

It is obvious from the defining equation (14a) for D_k that equation (20a) can be established simply by showing that D_k(i.e. E_1) is a subset of the right-hand side of equation (20a). To do so, first observe from our hypotheses (H_3) and (H_6) along with the defining formulas for γ_k that for a given y_k in D_k (i.e. E_1) there exists at least one x_k in the domain of γ_k such that y_k is in $\gamma_k(x_k)$. Now, observe from equation (12b) that such a y_k must also be in $\partial g_k(x_k)$, which in turn implies via an elementary manipulation of the defining inequality for $\partial g_k(x_k)$ that

$$\sup_{x_k' \in C_k} [y_k x_k' - g_k(x_k')] = [y_k x_k - g_k(x_k)].$$

In addition to showing that y_k is in the right-hand side of equation (20a) and hence establishing equation (20a), the preceding displayed equation also establishes equation (20b). The reason is that the defining equations (11) and (14) for $g_k(x_k)$ and $h_k(y_k)$, respectively, imply that

$$x_k y_k = g_k(x_k) + h_k(y_k) \text{ when } y_k \in \gamma_k(x_k)$$

(i.e. the [signed] area $x_k y_k$ of the rectangle with sides $[0,x_k]$ and $[0,y_k]$ is the sum of the [signed] areas $g_k(x_k)$ and $h_k(y_k)$ when $y_k \in \gamma_k(x_k)$). Similar, though somewhat more complicated, steps can be used to establish equations (21a) and (21b). However, for complete proofs of both equations (20) and equations (21) in increasing degrees of generality see [4,9,34].

It is also important to know that <u>the functions</u> g <u>and</u> h <u>defined by equations</u> (13) <u>and</u> (16), <u>respectively, inherit the conjugacy of the</u> g_k <u>and</u> h_k; that is,

$$D = \{y \in E_n \mid \sup_{x \in C} [\langle y,x\rangle - g(x)] \text{ is finite}\} \tag{22a}$$

and

$$h(y) = \sup_{x \in C} [\langle y,x\rangle - g(x)], \tag{22b}$$

while

$$C = \{x \in E_n \mid \sup_{y \in D} [\langle x,y\rangle - h(y)] \text{ is finite}\} \tag{23a}$$

and

$$g(x) = \sup_{y \in D} [\langle x,y\rangle - h(y)]. \tag{23b}$$

In fact, it is easy to see that equations (13), (16), (20) and (21) imply equations (22) and (23).

The most fundamental properties of problems A and B are not induced by the special nature of roadway networks -- only by the (conical) duality of X and Y and the (functional) conjugacy of g and h. We henceforth concentrate on such properties [28,29], some of which also depend on the separability of g and h.

Notice how problem B can be obtained directly from problem A: simply replace the convex polyhedral cone X with its (convex polyhedral) "dual" Y, and replace the closed convex separable function g with its (closed convex separable) "conjugate transform" h. The symmetry of (conical) duality demonstrated by equations (18) and (19) together with the symmetry of (functional) conjugacy demonstrated by equations (22) and (23) imply that the problem obtained by applying the same transformation to B is again A. This symmetry justifies the terminology "(geometric) dual problems" for problems A and B. Needless to say, it also induces a symmetry on the theory that relates A to B -- in that <u>each statement about</u> A <u>and</u> B <u>(whose proof uses only the duality of</u> X <u>and</u> Y <u>together with the conjugacy of</u> g <u>and</u> h) <u>automatically produces an equally valid "dual statement" about</u> B <u>and</u> A (<u>obtained by interchanging the symbols</u> X <u>and</u> Y, <u>the symbols</u> C <u>and</u> D, <u>and the symbols</u> g <u>and</u> h). However, to be concise, each dual statement will be left to the reader's imagination.

Unlike the usual min-max formulations of duality in mathematical programming (e.g. in linear programming), both problem A and its dual problem B are minimization problems. The relative simplicity of this min-min formulation will soon become clear, but the reader who is accustomed to the usual min-max formulation must bear in mind that a given duality theorem generally has slightly different statements depending on the formulation in use. In particular, a theorem that asserts the equality of the min and max in the usual formulation asserts that the sum of the mins is zero (i.e. $\Phi + \Psi = 0$) in the present formulation. (The reader interested in the precise connections between the various formulations of duality in mathematical programming should consult [26] and the references cited therein).

By definition, the "extremality conditions" (for separable geometric programming) are:

(I)	$x \in X$	and	$y \in Y$
(II)		$0 = \langle x,y \rangle$	
(III)	$y_k \in \partial g_k(x_k)$	for	$k = 1,2,\ldots,n.$

Each vector (x,y) that satisfies these conditions is termed an <u>extremal vector</u>.

From equations (12) we immediately see that <u>each network-equilibrium vector</u> (x,y) <u>is an extremal vector, and each extremal vector</u> (x,y) <u>is a network-equilibrium vector</u>. Consequently, <u>we can henceforth concentrate (without any loss of generality) on the existence and properties of extremal vectors</u> (x,y). The main reason for doing so is that the following powerful theory of (generalized) geometric programming is particularly suited to the study of extremal vectors.

5. <u>Duality Theory</u>. Proofs for the most fundamental duality theorems actually exploit the conjugacy of g and h indirectly via the following (Young-Fenchel) "conjugate inequality"

$$\langle x,y \rangle \le g(x) + h(y) \text{ for each } x \in C \text{ and each } y \in D.$$

This inequality is an immediate consequence of either equations (22) or equations (23); and its equality state can be conveniently characterized in terms of the "subgradient" set

$$\partial g(x) \triangleq \{y \in E_n \mid g(x) + \langle y, x' - x \rangle \le g(x') \text{ for each } x' \in C\}.$$

In particular, elementary (algebraic) manipulations of the defining inequality for $\partial g(x)$ show that $y \in D$ and $\langle x,y\rangle = g(x) + h(y)$ when $y \in \partial g(x)$. On the other hand, equally elementary (algebraic) manipulations show that $y \in \partial g(x)$ when $\langle x,y\rangle = g(x) + h(y)$. Hence,

$$\langle x,y\rangle = g(x) + h(y) \text{ if and only if } y \in \partial g(x);$$

from which it follows via the symmetry of (functional) conjugacy that

$$y \in \partial g(x) \text{ if and only if } x \in \partial h(y).$$

Finally, in the event that g and h are separable (e.g. in the present roadway network case), it is clear that

$$y \in \partial g(x) \text{ if and only if } y_k \in \partial g_k(x_k) \text{ for each } k = 1,2,\ldots,n;$$

and it is, of course, equally clear that

$$y_k \in \partial g_k(x_k) \text{ if and only if } x_k \in \partial h_k(y_k) \quad \text{for} \quad k = 1,2,\ldots,n.$$

The following duality theorem relates the dual problems A and B directly to the extremality conditions (I-III) -- and hence to traffic equilibria. This theorem is also basic to all succeeding duality theorems.

Theorem 2. If x and y are feasible solutions to the dual problems A and B respectively (in which case the extremality conditions (I) are satisfied), then

$$0 \le g(x) + h(y),$$

with equality holding if and only if the extremality conditions (II) and (III) are satisfied; in which case x and y are optimal solutions to problems A and B respectively.

Proof. The defining inequality for Y and the conjugate inequality for h show that $0 \leq \langle x,y\rangle \leq g(x) + h(y)$, with equality holding in both of these inequalities if and only if the extremality conditions (II) and (III) are satisfied; in which case a subtraction of the resulting equality $0 = g(x) + h(y)$ from each of the inequalities $0 \leq g(x') + h(y)$ and $0 \leq g(x) + h(y')$ (which are valid for arbitrary feasible solutions x' and y') shows that x and y are actually optimal solutions. q.e.d.

The fundamental inequality given by Theorem 2 implies important properties of the dual infima Φ and Ψ.

Corollary 2A. If the dual problems A and B both have feasible solutions, then

(i) the infimum Φ for problem A is finite, and

$$0 \leq \Phi + h(y) \text{ for each } y \in T,$$

(ii) the infimum Ψ for problem B is finite, and

$$0 \leq \Phi + \Psi$$

The proof of this corollary is, of course, a trivial application of Theorem 2.

The strictness of the inequality $0 \leq \Phi + \Psi$ in conclusion (ii) plays a crucial role in almost all that follows. In fact, dual problems A and B that have feasible solutions and for which $0 < \Phi + \Psi$ are said to have a "duality gap" of $\Phi + \Psi$. Although duality gaps do occur in (generalized) geometric programming, we shall eventually see that they do not occur in the present context of traffic equilibria. This lack of duality gaps is extremely fortunate because of their highly undesirable properties. In particular, we shall soon see that they weaken the bond between the dual problems A and B and imply that the extremality conditions have no solutions (i.e. there are no "extremal vectors" even though problems A and B may have optimal solutions). They also tend to destroy the possibility of using the inequality $0 \le g(x) + h(y)$ to provide an "algorithmic stopping criterion".

Such a criterion results from specifying a positive tolerance ϵ so that the numerical algorithms being used to minimize both $g(x)$ and $h(y)$ are terminated when they produce a pair of feasible solutions $x^{\dagger}$ and $y^{\dagger}$ for which

$$g(x^{\dagger}) + h(y^{\dagger}) \le \epsilon.$$

Because conclusion (i) to Corollary 2A along with the definition of Φ shows that

$$-h(y^{\dagger}) \le \Phi \le g(x^{\dagger}),$$

we conclude from the preceding tolerance inequality that

$$\Phi \le g(x^{\dagger}) \le \Phi + \epsilon.$$

Hence, Φ can be approximated by $g(x^{\dagger})$ with an error no greater than ϵ (and, dually, Ψ can be approximated by $h(y^{\dagger})$ with an error no greater than ϵ). This stopping criterion can be useful even though it does not estimate the proximity of $x^{\dagger}$ to an optimal solution x^* (and, dually, even though it does not estimate the proximity of $y^{\dagger}$ to an optimal solution y^*). In any event, it is clear that the algorithms being used need not terminate unless $\Phi + \Psi < \epsilon$, a situation that is guaranteed for each tolerance ϵ only if $0 = \Phi + \Psi$ (i.e. there is no duality gap).

The equality characterization of the fundamental inequality given by Theorem 2 implies the key relations between extremal vectors (x,y) and optimal solutions $x \in S^*$ and $y \in T^*$.

<u>Corollary</u> 2B. <u>If</u> (x^+,y^+) <u>is an extremal vector</u> (i.e. <u>if</u> (x^+,y^+) <u>is a solution to the extremality conditions</u> (I-III)), <u>then</u>

(i) $x^+ \in S^*$ and $y^+ \in T^*$,

(ii) $S^* = \{x \in X \mid 0 = \langle x, y^+ \rangle, \text{ and } x_k \in \partial h_k(y_k^+) \text{ for } k = 1,2,\ldots,n\}$

(iii) $T^* = \{y \in Y \mid 0 = \langle x^+, y \rangle, \text{ and } y_k \in \partial g_k(x_k^+) \text{ for } k = 1,2,\ldots,n\}$

(iv) $0 = \Phi + \Psi$.

On the other hand, if the dual problems A and B both have feasible solutions and if $0 = \Phi + \Psi$, then (x,y) is an extremal vector if and only if $x \in S^*$ and $y \in T^*$.

The proof of this corollary comes from a trivial application of Theorem 2 along with the (previously mentioned and easily established) conjugate transform relations $\partial g_k(x_k) \subseteq D_k$ and $\partial h_k(y_k) \subseteq C_k$.

Corollary 2B readily yields certain basic (nontrivial) information about traffic equilibria. In particular, the first part of Corollary 2B asserts that if there is at least one network-equilibrium vector (x^+, y^+), then there is no duality gap. On the other hand, the second part of Corollary 2B clearly implies that if there is no duality gap, then the set E of all network equilibrium vectors (x,y) is just the cartesian product $S^* \times T^*$ of the optimal solution sets S^* and T^*. Consequently, when network equilibria exist, each demand-equilibrium total flow x^* (namely, each $x^* \in S^*$) can be paired with each equilibrium quasi-revenue-cost per unit of flow vector y^* (namely, each $y^* \in T^*$) to produce a network equilibrium vector $(x^*, y^*) \in E$. Moreover, network equilibria exist if and only if hypotheses (H) are sufficiently strong to guarantee the absence of a duality gap along with the existence of optimal solutions $x^* \in S^*$ and $y^* \in T^*$. Actually, the existence of network equilibria is established in section 9 by using the network classification and reduction given in section 7 to show that $0 = \Phi + \Psi$ and neither S^* nor T^* is empty.

Computationally, it is important to note that conclusions (i-iii) of Corollary 2B provide methods for calculating all network-equilibrium vectors $(x,y) \in E$ from the knowledge of a single network-equilibrium vector

$(x^+,y^+) \in E$ (or even from just the knowledge of a single optimal solution $x^* \in S^*$, or a single optimal solution $y^* \in T^*$). Of course, the (nonempty) set $Z(x,y)$ of all demand-equilibrium circuit flows z that generate a given network-equilibrium vector $(x,y) \in E$ can be calculated simply by calculating the set of all solutions z to the feasibility conditions (1,2) and the complementary slackness conditions (7") -- as explained in the next-to-last paragraph of subsection 2.3.

It is also important to note that conclusions (ii-iii) of Corollary 2B imply that both the set S^* (of all demand-equilibrium total flows x^*) and the set T^* (of all equilibrium quasi-revenue-cost per unit of flow vectors y^*) are convex polyhedral sets, and hence so is the set

$$E = S^* \times T^*$$

(of all network-equilibrium vectors (x^*,y^*)). The reason is that $\partial h_k(y_k^+)$ and $\partial g_k(x_k^+)$ are closed intervals (as previously observed); and convex polyhedralness is preserved under cartesian products. Of course, the set Z of all demand-equilibrium circuit flows z has the representation formula

$$Z = \bigcup_{(x,y) \in E} Z(x,y)$$

where each set

$$Z(x,y) \triangleq \{z \in E_m \mid \text{conditions (1,2) and (7") are satisfied}\}$$

is clearly a bounded convex polyhedral set.

6. Uniqueness Theorems. It is worth noting that various uniqueness theorems result from the representation formula $E = S^* \times T^*$ by imposing strict monotonicity and/or continuity hypotheses on the travel cost per unit of flow functions c_k. In particular, if there is a demand-equilibrium total flow x^* such that a given c_k is strictly increasing on a (relative) neighborhood of x_k^*, then every other demand-equilibrium total flow x^{**} exhibits the same unique total flow $x_k^{**} = x_k^*$ on link k. The reason is that strict monotonicity of c_k on a (relative) neighbothood of x_k^* implies

that $\gamma_k(x_k^*)$ and $\gamma_k(x_k^{**})$ do not intersect when $x_k^* \neq x_k^{**}$; so a given y_k^* can not be in both $\gamma_k(x_k^*)$ and $\gamma_k(x_k^{**})$ unless $x_k^* = x_k^{**}$. On the other hand, if there is a demand-equilibrium total flow x^* for which $x_k^* > 0$ and such that the given c_k is continuous at x_k^*, then each equilibrium quasi-revenue-cost per unit of flow vector y^* exhibits the same unique quasi-cost per unit of flow $y_k^* = c_k(x_k^*)$ for link k. The reason is that continuity of c_k at x_k^* implies that $\gamma_k(x_k^*)$ contains a unique element $c_k(x_k^*)$ when $x_k^* > 0$. Finally, if there is a unique quasi-cost per unit of flow y_k^* for each (real) link k for which there is at least one demand-equilibrium total flow x^* such that $x_k^* > 0$ (e.g. if each c_k is continuous on its domain $[0, b_k))$, then each equilibrium quasi-revenue-cost per unit of flow vector y^* exhibits the same unique quasi-revenue per unit of flow y_i^* for a given (return) link i if $x_i^* > 0$ for each demand-equilibrium total flow x^* (e.g. if 0 does not belong to the domain I_i of γ_i, as is the case when $I_i = \{d_i\}$ and $d_i > 0$). The reason is that the preceding observation along with conditions (6") and the finiteness of the feasible circuit family [i] clearly imply that y_i^* can have at most a finite number of values; while the convexity of T* clearly implies that y_i^* can have infinitely many values if it has more than one value.

Other uniqueness theorems result from the representation formula $E = S^* \times T^*$ by imposing comparable hypotheses on the travel demand curves Γ_i. In particular, if there is a demand-equilibrium total flow x^* and a (return) link i such that $\gamma_i(x_i^*)$ and $\gamma_i(x_i^{**})$ do not intersect when $x_i^* \neq x_i^{**}$, then every other demand-equilibrium total flow x^{**} clearly exhibits the same unique total (input) flow $x_i^{**} = x_i^*$ on (return) link i. On the other hand, if there is a demand-equilibrium total flow x^* and a (return) link i such that $\gamma_i(x_i^*)$ contains a unique element y_i^*, then each equilibrium quasi-revenue-cost per unit of flow vector y^* clearly exhibits the same unique quasi-revenue per unit of flow y_i^* for (return) link i.

Even when there is a unique network-equilibrium vector (x^*, y^*), there can obviously be infinitely many demand-equilibrium circuit flows z*. In fact, uniqueness of z* in that case clearly requires a positive independence of those circuit vectors δ^j for which $\langle \delta^j, y^* \rangle = 0$ and $z_j^* > 0$. It is clear that such circuit vectors δ^j are less likely to be positively independent, the more the corresponding circuits "overlap" (i.e. the more possibility there is for congestion). In any event, it is easy to construct simple examples where neither x*, y* nor z* have any unique components. Such constructions are, however, left to the reader's imagination.

7. <u>Network Classification and Reduction</u>. In this section the family of all (roadway) networks is partitioned into a family of "canonical networks" and a family of "degenerate networks". Each degenerate network has a "reduced form" that is canonical, and each canonical network is its own reduced form.

This mapping of all networks onto canonical networks induces another partitioning of the family of all networks, namely, a partitioning into an infinite collection of "equivalence classes". This collection of equivalence classes is in one-to-one correspondence with the family of all canonical networks. Corresponding to a given canonical network is the equivalence class that consists of the given canonical network and all degenerate networks that have the given canonical network as a reduced form.

Networks can be classified as canonical or degenerate simply by inspection. If a network is degenerate, the inspection yields its reduced form, which is just a "subnetwork" of the given degenerate network. Moreover, the geometric programming problems A and $\mathcal{A}$ associated with a network N and its reduced form $\mathcal{N}$, respectively, have the same infima Φ and φ. Likewise, the geometric programming problems B and $\mathcal{B}$ associated with a network N and its reduced form $\mathcal{N}$, respectively, have the same infima Ψ and ψ. Furthermore, the optimal solution set S*(T*) for problem A (B) can be generated from the optimal solution set $\mathcal{S}^*(\mathcal{T}^*)$ for problem $\mathcal{A}(\mathcal{B})$; and hence the extremal vector set $E = S^* \times T^*$ for network N can be generated from the extremal vector set $\mathcal{S} = \mathcal{A}^* \times \mathcal{T}^*$ for network $\mathcal{N}$. Thus, the family of all reduced forms (canonical networks) is the most important family from a computational point of view. It is also the most important family from a theoretical point of view, because the absence of duality gaps and the existence of dual optimal solutions can readily be proved for all networks N once they are established for all canonical networks $\mathcal{N}$.

Roughly speaking, a network N is canonical if each commodity i has at least one feasible input flow $d_i^+ > 0$ and if each link k in the network N is part of at least one circuit j (for some commodity i for which $j \in [i]$). In essence then, a network N is canonical if it is impossible to infer from the feasible input flows and the feasible circuit families alone that some link k (possibly a return link) is unusable (i.e. $x_k = 0$ for each feasible total flow x). Naturally, a degenerate network N is then a network that contains unusable links k (some of which may be return links), each of which is to be pruned away (along with the corresponding commodity k in the case of an unusable return link k) when constructing its reduced form $\mathcal{N}$.

To give a precise classification of the most general network N (corresponding to the demand-equilibrium problem described in section 2), use the "irreducible circuit family" J introduced at the beginning of section 3 to define the <u>irreducible link set</u>

$$K \triangleq \{k \mid 1 \leq k \leq n \text{ and there exists a } j \in J \text{ for which } \delta_k^j = 1\}.$$

Then, network N is said to be <u>canonical</u> if $K = \{1,2,\ldots,n\}$, and <u>degenerate</u> if $K \subset \{1,2,\ldots,n\}$.

Since we have already assumed (at the beginning of section 3) that the "irreducible commodity set" I (and hence J) is not empty, we immediately see that K is not empty. Consequently, <u>deleting those links</u> $k \notin K$ <u>from network</u> N <u>produces a (nonempty) subnetwork</u> $\mathcal{N}$ of N, which is termed the <u>reduced form</u> of N. Naturally, the only commodities that flow on $\mathcal{N}$ (over the links $k \in K$) are those commodities $i \in I$; but the feasible circuit family for such a commodity i is clearly still [i]. Needless to say, each link $k \in K$ retains the same "characteristic functions" γ_k, and hence $\mathcal{N}$ inherits from N all of the hypotheses (H) along with all of the properties previously derived therefrom. In addition, $\mathcal{N}$ has properties not generally possessed by N -- the most important ones stemming from the obvious fact that $\mathcal{N}$ is canonical and its own reduced form.

Two simple observations are crucial to ultimately showing that essentially nothing is lost in replacing network N with its reduced form $\mathcal{N}$. The first observation is that the defining equation (8) for X along with the defining equation for K clearly implies that

$$\text{each vector } x \in X \text{ has components } x_k = 0 \text{ for } k \notin K. \tag{24}$$

The second observation is that identity (10) for X along with the defining equation for K clearly implies that

the vector y^- in E_n with components

$$y_k^- \triangleq \begin{cases} 0 & \text{for } k \in K \\ -1 & \text{for } k \notin K \end{cases} \tag{25}$$

is orthogonal to each vector $x \in X$,

and hence $y^- \in Y$.

In replacing network N with its reduced form $\mathcal{N}$ it is helpful to employ the linear transformation

$$\mathcal{L}: E_n \to E_{\mathcal{n}},$$

defined so that the image $\mathcal{v} = \mathcal{L}(v)$ is obtained from v simply by deleting the k'th component of v for each $k \notin K$. Thus, $\mathcal{n}$ is equal to the number of links in $\mathcal{N}$; and, unless otherwise stated, a vector represented by a script symbol $\mathcal{v}$ resides in the image space $E_{\mathcal{n}}$ of $\mathcal{L}$. However, it is convenient not to relabel the indices on the components of $\mathcal{v}$; consequently, the j'th component of $\mathcal{v}$ may be $\mathcal{v}_k = v_k$ where $j < k$.

It is clear that the geometric programming problem $\mathcal{A}$ associated with the reduced network $\mathcal{N}$ is constructed from the cone

$$\mathcal{X} = \mathcal{L}(X) \tag{26}$$

and the function g whose domain

$$\mathcal{C} = \underset{K}{\times} C_k \tag{27a}$$

and whose function-values

$$g(x) = \sum_K g_k(x_k). \tag{27b}$$

<u>Problem</u> $\mathcal{A}$. Using the feasible solution set

$$\mathcal{S} \triangleq \mathcal{X} \cap \mathcal{C},$$

calculate both the problem infimum

$$\varphi \triangleq \inf_{x \in \mathcal{S}} g(x)$$

and the optimal solution set

$$\mathcal{S}^* \triangleq \{x \in \mathcal{S} \mid \varphi = g(x)\}.$$

It is equally clear that the geometric programming problem $\mathcal{B}$ associated with the reduced network $\mathcal{N}$ is constructed from the cone

$$\mathcal{Y} = \mathcal{L}(Y) \tag{28}$$

and the function h whose domain

$$\mathcal{D} = \underset{K}{\times} D_k \tag{29a}$$

and whose function-values

$$h(y) = \sum_K h_k(y_k). \tag{29b}$$

<u>Problem</u> $\mathcal{B}$. Using the feasible solution set

$$\mathcal{T} \triangleq \mathcal{Y} \cap \mathcal{D},$$

calculate both the problem infimum

$$\psi \triangleq \inf_{y \in \mathcal{T}} h(y)$$

and the optimal solution set

$$\mathcal{T}^* \triangleq \{y \in \mathcal{T} \mid \psi = h(y)\}.$$

We have already observed that the reduced form $\mathcal{N}$ inherits from network N all of the hypotheses (H) along with all of the properties previously derived therefrom. Consequently, the "reduced form" $\mathcal{A}$ inherits from problem A all of the previously derived properties of A; and the "reduced form" $\mathcal{B}$ inherits from problem B all of the previously derived properties of B. In addition, problems $\mathcal{A}$ and $\mathcal{B}$ have important properties not generally possessed by problems A and B, respectively. However, such properties can be fully exploited only after problems A and B are more completely related to their reduced forms $\mathcal{A}$ and $\mathcal{B}$, respectively.

The following theorem brings to light the most important relations between problem A and its reduced form $\mathcal{A}$.

Theorem 3. Problem A has feasible solutions if and only if its reduced form $\mathcal{A}$ has feasible solutions, in which case

(i) $$\varphi = \Phi$$

(ii) $$\mathcal{S}^* = \mathcal{L}(S^*)$$

(iii) $$S^* = \mathcal{L}^{-1}(\mathcal{S}^*) \cap S$$

Proof. If $x \in \mathbf{X}$, then $\mathcal{x} = \mathcal{L}(x) \in \mathcal{X}$ by virtue of equation (26). If, in addition, $x \in C$, then $\mathcal{x} \in \mathcal{C}$ and $g(\mathcal{x}) = g(x)$ by virtue of equations (27) and relation (24), along with both our choice of α_k in equations (11) and the definition of K in terms of J and I. These two facts establish the following lemma.

Lemma a. If $x \in S$, then $\mathcal{x} = \mathcal{L}(x) \in \mathcal{S}$, and $g(\mathcal{x}) = g(x)$.

Now, if $\mathcal{x} \in \mathcal{X}$, then $\mathcal{x} = \mathcal{L}(x)$ for some $x \in \mathbf{X}$ by virtue of equation (26). Moreover, relation (24) asserts that $x_k = 0$ for each $k \notin K$; so the definition of $\mathcal{L}$ shows that x is uniquely determined by $\mathcal{x}$. If, in addition, $\mathcal{x} \in \mathcal{C}$, then $x \in C$ and $g(x) = g(\mathcal{x})$ by virtue of equations (27) and relation (24), along with both our choice of α_k in equations (11) and the definition of K in terms of J and I. These two facts establish the following lemma.

<u>Lemma</u> b. <u>If</u> $\mathcal{x} \in \mathcal{S}$, <u>then there is a unique</u> $x \in S$ <u>such that</u> $\mathcal{x} = \mathcal{L}(x)$, <u>and</u> $g(x) = \mathcal{g}(\mathcal{x})$.

To prove the first assertion of Theorem 3, first observe that Lemma a implies that the reduced form $\mathcal{A}$ has feasible solutions when problem A has feasible solutions, and then observe that Lemma b implies the converse.

To prove conclusion (i) of Theorem 3, use Lemma a to infer that $\varphi \leq \Phi$, and then use Lemma b to infer that $\Phi \leq \varphi$.

To establish conclusion (ii) of Theorem 3, first let $\mathcal{x} \in \mathcal{L}(S^*)$. Then $\mathcal{x} = \mathcal{L}(x)$ for some $x \in S$ such that $g(x) = \Phi$. Thus, from Lemma a and conclusion (i) we deduce that $\mathcal{x} \in \mathcal{S}^*$, which shows that $\mathcal{S}^* \supseteq \mathcal{L}(S^*)$. Now, let $\mathcal{x} \in \mathcal{S}^*$, which means that $\mathcal{x} \in \mathcal{S}$ and $\mathcal{g}(\mathcal{x}) = \varphi$. Then, from Lemma b and conclusion (i), we deduce the existence of an $x \in S$ for which $\mathcal{x} = \mathcal{L}(x)$ and $g(x) = \Phi$. Thus, $x \in S^*$ and hence $\mathcal{x} \in \mathcal{L}(S^*)$, which establishes the relation $\mathcal{S}^* \subseteq \mathcal{L}(S^*)$ and consequently completes our proof of conclusion (ii).

To prove conclusion (iii), first observe that the relation $S^* \subseteq \mathcal{L}^{-1}(\mathcal{S}^*) \cap S$ follows immediately from conclusion (ii) because $S^* \subseteq S$. Now, suppose that $x \in \mathcal{L}^{-1}(\mathcal{S}^*) \cap S$, which means there is an $\mathcal{x} \in \mathcal{S}^*$ such that $\mathcal{L}(x) = \mathcal{x}$. Then $\mathcal{g}(\mathcal{x}) = \varphi$, and $g(x) = \mathcal{g}(\mathcal{x})$ because of Lemma b. These two equations and the assumption that $x \in S$ imply that $x \in S^*$ by virtue of conclusion (i). Hence, $S^* = \mathcal{L}^{-1}(\mathcal{S}^*) \cap S$, and thus our proof of Theorem 3 is complete.

The following theorem brings to light the most important relations between problem B and its reduced form $\mathcal{B}$.

<u>Theorem</u> 4. <u>Both problem</u> B <u>and its reduced form</u> $\mathcal{B}$ <u>have feasible solutions, and</u>

(i) $$\psi = \Psi$$

(ii) $$\mathcal{T}^* = \mathcal{L}(T^*)$$

(iii) $$T^* = \mathcal{L}^{-1}(\mathcal{T}^*) \cap T.$$

Proof. As previously noted, the existence of feasible solutions to problem B is a direct consequence of equation (17) and the fact that each cone $T = Y$ contains at least the zero vector. Since the reduced form $\mathcal{B}$ inherits all of the previously derived properties of B, we infer that problem $\mathcal{B}$ also has a nonempty feasible solution set $\mathcal{T} = \mathcal{Y}$. This proves the first assertion of Theorem 4.

If $y \in T$ (i.e. $y \in Y$), then $\mathcal{y} = \mathcal{L}(y) \in \mathcal{Y}$ (i.e. $\mathcal{y} \in \mathcal{T}$) by virtue of equation (28). Moreover, $\mathcal{h}(\mathcal{y}) \leq h(y)$ by virtue of equations (29), along with the first part of hypothesis (H_2), the construction of γ_k^{-1} from c_k, and both equation (14b) and the definition of K in terms of J and I. These two facts establish the following lemma.

<u>Lemma</u> c. <u>If</u> $y \in T$, <u>then</u> $\mathcal{y} = \mathcal{L}(y) \in \mathcal{T}$, <u>and</u> $\mathcal{h}(\mathcal{y}) \leq h(y)$.

Now, if $\mathcal{y} \in \mathcal{T}$ (i.e. $\mathcal{y} \in \mathcal{Y}$), then $\mathcal{y} = \mathcal{L}(y)$ for some $y \in Y$ (i.e. $y \in T$) by virtue of equation (28). Moreover, relation (25) clearly implies that such a y is not uniquely determined by $\mathcal{y}$ and can in fact be chosen so that $y_k \leq 0$ for $k \notin K$. If such a y is chosen, then $h(y) = \mathcal{h}(\mathcal{y})$ by virtue of equations (29), along with the first part of hypothesis (H_2), the construction of γ_k^{-1} from c_k, and both equation (14b) and the definition of K in terms of J and I. These facts establish the following lemma.

<u>Lemma</u> d. <u>If</u> $\mathcal{y} \in \mathcal{T}$, <u>then there is at least one</u> $y \in T$ <u>such that</u> $\mathcal{y} = \mathcal{L}(y)$, <u>and</u> $h(y) = \mathcal{h}(\mathcal{y})$.

We can now prove conclusions (i), (ii), and (iii) of Theorem 4 by using Lemmas c and d in the same manner that Lemmas a and b were used to establish conclusions (i), (ii), and (iii) of Theorem 3. We leave the details to the reader and hence consider the proof of Theorem 4 to be complete.

Needless to say, Theorem 3 shows that <u>problem $\mathcal{A}$ is equivalent to problem</u>

A, and Theorem 4 shows that *problem* $\mathcal{B}$ *is equivalent to problem* B. Consequently, *network* $\mathcal{N}$ *is equivalent to network* N.

8. *Boundedness Theorems*. It is worth noting that boundedness of the extremal vector set E (consisting of all network-equilibrium vectors (x,y)) is intimately related to boundedness of the dual optimal solution sets S* and T*. In particular, the representation formula $E = S^* \times T^*$ clearly implies that E *is bounded if and only if* S* *and* T* *are both bounded*. Moreover, we shall soon see that hypotheses (H) guarantee that S* *is always bounded*, *while* T* *is bounded only if network* N *is canonical*. Needless to say, an immediate corollary to these facts is that E *is bounded only if* N *is canonical*. Moreover, it is worth noting that the boundedness of S* implies via conditions (1,2) and the non-negativity of the feasible circuit vectors δ^j that *the set* Z (*consisting of all demand-equilibrium circuit flows* z) *is always bounded*.

To show that S* is bounded, it is of course sufficient to show that g(x) *approaches* $+\infty$ *when some component* x_q *of* $x \in C$ *becomes unbounded* (because $g(x^*) \equiv \Phi$ for each $x^* \in S^*$). Now, using the defining equation (13a) for C along with the defining equation (11a) for C_k, we infer from the defining equations for (the domains of) γ_i and γ_k that each such component x_q is bounded from below by 0. Moreover, our hypotheses (H) and the defining equations (11b) and (11c) for $g_k(x_k)$ clearly imply that $g_k(x_k) \geq 0$ for each $x_k \in C_k$ while $g_q(x_q)$ approaches $+\infty$ when x_q approaches $+\infty$. The desired conclusion is now a direct consequence of the defining equation (13b) for g(x).

To show that T* is unbounded if network N is degenerate (and T* is not empty), simply note that the non-negativity of the inverse functions γ_k^{-1} (as implied by hypotheses $(H_1 - H_3)$ and hypothesis (H_5)) along with both the defining equations (14) and (16) for h and equation (17) for D imply that *any positive* multiple of the vector y^- described by relation (25) produces a vector in T* when added to a vector in T*. Consequently, canonicality of N is a necessary condition for boundedness of T* (and hence E) -- though we conjecture that boundedness of T* (and hence E) is not characterized by canonicality of N. However, characterizations of the boundedness of T* (and hence E) can probably be obtained via refinements of the general results given in [30].

9. The Main Existence Theorem. The following fundamental theorem shows that our hypotheses (H) are sufficiently strong to guarantee the existence of traffic equilibria.

Theorem 5. If there is at least one feasible circuit flow z (i.e. if there is at least one vector z that satisfies conditions (1,2,3'',4)), then there is at least one extremal vector (x^+,y^+).

Proof. As previously observed, the existence of a feasible circuit flow z implies the existence of a feasible solution x to problem A. Since we have also previously observed that problem B has at least the feasible solution 0, we deduce from Corollary 2A that

the infima Φ and Ψ are both finite.

Now, the fact that Y is nonempty and convex (as inferred from equation (9)) and the fact that $D = E_n$ (as asserted by equation (17)) clearly imply via Theorem 6.2 on page 45 of [36] that problem B has a feasible solution $y^o \in (\text{ri}\, Y) \cap (\text{ri}\, D)$, where (ri W) denotes the "relative interior" of the given set W. From the existence of such a feasible solution y^o and from the preceding displayed statement, we deduce via (the geometric programming version of Fenchel's) Theorem 5 in subsection 3.1.4 of [28] that $0 = \Phi + \Psi$ and S* is not empty.

In view of the final assertion of Corollary 2B, we now need only show that T* is not empty to complete our proof. To do so, we first note from conclusion (ii) to Theorem 4 that it is sufficient to show that $\mathcal{T}^*$ is not empty. Toward that end, observe that Theorems 3 and 4 along with the preceding displayed statement imply that

the infima φ and ψ are both finite.

Now, the defining equations for I and J along with the feasibility conditions (2,3") obviously imply that the feasible circuit flow z has components $z_j = 0$ for each $j \notin J$. Consequently, the definition of network $\mathcal{N}$ clearly implies that the vector $\mathcal{z}$ obtained from z by deleting component z_j for each $j \notin J$ is a feasible circuit flow for $\mathcal{N}$. Moreover, it is clear that $\mathcal{x} = \mathcal{L}(x)$ where $\mathcal{x}$ and x are the total flows on $\mathcal{N}$ and N respectively resulting from $\mathcal{z}$ and z respectively.

From the defining equation for I and the feasibility conditions (2,3",4) it is obvious that a slight (possibly zero) perturbation of $\mathcal{z}$ can be used to produce a feasible $\mathcal{z}'$ with strictly positive components. Moreover, it is equally obvious that a slight (possibly zero) perturbation of $\mathcal{z}'$ can then be used to produce a feasible $\mathcal{z}^o$ with strictly positive components such that the resulting total flow $\mathcal{x}^o$ has components $\mathcal{x}_k^o \in (\mathrm{ri}\, \mathcal{C}_k)$ for each $k \in K$. Now, it is a consequence of the strict positivity of $\mathcal{z}^o$ along with equations (8) and (26) for X and $\mathcal{X}$ respectively that $\mathcal{x}^o \in (\mathrm{ri}\, \mathcal{X})$, by virtue of the linearity of $\mathcal{L}$ and Theorem 6.6 on page 48 in [36]. Moreover, it is a consequence of the relations $\mathcal{x}_k^o \in (\mathrm{ri}\, \mathcal{C}_k)$ along with equation (27a) for $\mathcal{C}$ that $\mathcal{x}^o \in (\mathrm{ri}\, \mathcal{C})$, by virtue of the first formula at the top of page 49 in [36].

From the existence of this feasible solution $\mathcal{x}^o \in (\mathrm{ri}\, \mathcal{X}) \cap (\mathrm{ri}\, \mathcal{C})$ and from the preceding displayed satement, we deduce via the (unstated) dual of (the geometric programming version of Fenchel's) Theorem 5 in subsection 3.1.4 of [28] that $0 = \varphi + \psi$ and $\mathcal{T}^*$ is not empty. q.e.d.

10. Computational Considerations. There are at least four different approaches to computing traffic equilibria (x,y), but only (very limited versions of) two of them have formed the basis for all computer algorithms [5,13,15,20] known to us.

The first approach [5,13,20] seems to be the only one that is now widely used in practice. In essence, it "simulates" the behavior of drivers by using "shortest path algorithms" (as described, for example, in [16,19]) to iteratively solve in heuristic ways the defining equations (1-6) for Wardrop equilibria. Even though such algorithms have (to the best of our knowledge) been used only when all cost per unit of flow functions c_k are continuous and strictly increasing, it is widely known that they frequently fail to converge. Of course, such numerical difficulties are likely to be compounded when discontinuous c_k are present and demand equilibria are being sought.

The second approach is (to the best of our knowledge) now being set forth for the first time (even in the special context of computing Wardrop equilibria when all c_k are continuous and strictly increasing). It uses any algorithm (possibly those in [24]) that can solve the network-equilibrium conditions (I-III). This approach is analogous to a well-known (rather successful) approach used in analysing electric and hydraulic networks. In the context of such networks, condition (I) constitutes the "Kirchoff current and potential (conservation) laws", condition (II) is implied by (I) and hence is redundant, and condition (III) is just "Ohm's law".

The third approach is essentially a simplification, amplification, and generalization of the variational approach considered (in the context of Wardrop equilibria) by Dafermos and Sparrow [7,8] and computerized by many workers (e.g., see [15]). It uses any algorithm that can solve the closed convex separable optimization problem A. Actually, any such algorithm that exploits the absence of nonlinear constraints (in particular, any of the most powerful convex programming algorithms [42] presently available for solving linearly-constrained problems) can be used. To do so, simply replace the only (conceivably) nonlinear constraints $x_k \in C_k$ with the (generally weaker) linear constraints

$$x_k = d_k \in \text{closure of } I_k \qquad \text{for} \qquad k = 1,2,\ldots,r \qquad (3''')$$

and

$$0 \le x_k \le b_k \qquad \text{for} \qquad k = r+1, r+2, \ldots, n. \qquad (4''')$$

This reformulation does not alter the infimum Φ and the optimal solution set S^* for problem A, because the defining equations (11a) and (11b) for C_k and $g_k(x_k)$ respectively clearly imply via the defining equation (13b) for $g(x)$ that $g(x)$ approaches $+\infty$ when at least one component x_q approaches some point not in C_q. Of course, once a single optimal solution $x^* \in S^*$ has been computed, the set E of all traffic equilibria (x^*,y^*) can be computed via the technique outlined after Corollary 2B. This approach is analogous to a classical variational approach [39,10,3,21,2,34] used in analysing electric and hydraulic networks. In the context of such networks, problem A consists of minimizing the "content" of the network (which is frequently just the power dissipated in the network), subject to only the Kirchoff current (conservation) law.

The fourth approach is essentially a simplification, amplification, and generalization of the variational approach considered by Murchland [22] and more recently by Evans [12]. It uses any algorithm that can solve the closed convex separable linearly-constrained optimization problem B (in particular, any of the most powerful convex programming algorithms [42] presently available for solving linearly-constrained problems). Needless to say, once a single optimal solution $y^* \in T^*$ has been computed, the set E of all traffic equilibria (x^*,y^*) can be computed via the technique outlined after Corollary 2B. This approach is also analogous to a classical variational approach [39,10,3,21,2,34] used in analysing electric and hydraulic networks. In the context of such networks, problem B consists of minimizing the "co-content" of the network (which is frequently just the power dissipated in the network), subject to only the Kirchoff potential (conservation) law.

Unlike the first and second approaches, the only nonlinear functions to be reckoned with in the third and fourth approaches (the functions g_k and h_k respectively) are both continuous and convex. In fact, a multivalued γ_k(e.g. a discontinuous c_k) induces only a lack of differentiability in an otherwise differentiable g_k; and a multivalued γ_k^{-1} (e.g. a c_k that is constant over at least one nontrivial interval) induces only a lack of differentiability in an otherwise differentiable h_k. Naturally, the third and fourth approaches can be carried out in unison to provide an algorithmic stopping criterion, as explained after Corollary 2A. However, that criterion may not be of much use because it provides direct information only about the degree of convergence of g(x) and h(y) -- not the desired degree of convergence of x and y. Nevertheless, the degree of convergence of x and y can frequently be deduced from the degree of convergence of g(x) and h(y) through an analysis of the rates of change of the γ_k and the γ_k^{-1} (e.g. the second derivatives of the g_k and the h_k when they exist). In any event, the fact that $D = E_n$ (as asserted by equation (17)) obviously endows the fourth approach with a numerical advantage over the third approach (because $C \neq E_n$). Moreover, the immediate availability of the data required in the representation formula (9b) for Y (i.e., the generators δ^j for X) endows the fourth approach with still another numerical advantage over the third approach (because comparable data required in the corresponding (unstated) representation formula for X can be obtained only from the generators for Y, which, to the best of our knowledge, can be computed only with the aid of a rather intricate linear algebraic algorithm due to Uzawa [40]).

Actually, network planners need only know T* to uncover the possible "bottlenecks" in a given network. In fact, it is clear from linear programming theory that they need only know the "extreme points" and "recession directions" of the convex polyhedral set T* to uncover such bottlenecks. Of course, the same information can also be extracted from the extreme points of the convex polyhedral set S* (which has no recession directions because of its boundedness). Moreover, it is worth noting that either set of extreme points (and recession directions) can be computed without computing the set Z (of all demand-equilibrium circuit flows) -- simply by using either the second, third, or fourth approaches. Consequently, the first approach is, no doubt, the least attractive approach for network planning, even though it is the only approach that is now widely used.

Finally, each of the preceding approaches should clearly be used on the reduced network $\mathcal{N}$ when network N is not canonical. Moreover, the decomposition principles given in [27,28,31,32] should probably be used in conjunction with the preceding approaches when $\mathcal{N}$ is large but "sparse" (e.g. when $\mathcal{N}$ results from modeling a metropolitan roadway network in which various identifiable subnetworks are only weakly linked to one another).

References

1. Beckmann, M.J., McGuire, C.B., Winsten, C.B., Studies in the Economics of Transportation, Yale Univ. Press, New Haven, Conn., (1965).

2. Berge, C., Ghouila-Houri, A., Programming, Games, and Transportation Networks, Wiley, New York, (1963).

3. Birkhoff, G., Diaz, J.B., "Nonlinear Network Problems", Quart. Appl. Math., 13 (1956), 431-443.

4. Boas, R.P., Marcus, M., "Inverse Functions and Integration by Parts", Amer. Math. Month., 81(1974), 760-761

5. Bureau of Public Roads, Traffic Assignment Manual, U.S. Gov't. Printing Office, Washington, (1964).

6. Charnes, A., Cooper, W.W., "Multicopy Traffic Network Models", Proc. Sym. on Theory of Traffic Flow, held at General Motors Research Lab., Warren, Michigan, (1961).

7. Dafermos, S.C., "An Extended Traffic Assignment Model with Application to Two-Way Traffic", Trans. Science, 5, (1971), 367-389.

8. Dafermos, S.C., Sparrow, F.T., "The Traffic Assignment Problem for a General Network", Jour. Res. Nat. Bur. Standards, 73B, (1969), No. 2, 91-117.

9. Diaz, J.B., Metcalf, F.T., "An Analytic Proof of Young's Inequality", Amer. Math. Month., 77 (1970), 603-609.

10. Duffin, R.J., "Nonlinear Networks IIa", Bull. Amer. Math. Soc., 53, (1947), 963-971,

11. Duffin, R.J., Peterson, E.L., Zener, C., Geometric Programming, John Wiley and Sons, New York, New York, (1967).

12. Evans, S.P., "Some Models for Combining the Trip Distribution and Traffic Assignment Stages in the Transport Planning Process", paper presented at the Intern. Sym. on Traffic Equilibrium Methods, Centre de Recherche sur les Transports, Univ. de Montréal, (1974).

13. Federal Highway Administration, Urban Transportation Planning, Washington, (1972), III-1 to III-20.

14. Fenchel, W., "Convex Cones, Sets and Functions", Mathematics Dept. Mimeographed Lecture Notes, Princeton University, (1951).

15. Florian, M., Nguyen, S., "Recent Experience with Equilibrium Methods for Studying Congested Urban Road Networks", paper presented at the Inter. Sym. on Traffic Equilibrium Methods, Centre de Recherche sur les Transports, Univ. de Montréal, (1974).

16. Ford, L.K., Fulkerson, D.R., Flows in Networks, Princeton University Press, Princeton, New Jersey, (1962).

17. Hall, M.A., "Civil Engineering Applications of (Generalized) Geometric Programming", Jour. of Eng. Optimization, to appear in Vol. 1.

18. ___________, "On the Definition of Feasible Flows on Roadway Networks", in preparation.

19. Hu, T.C., Integer Programming and Network Flows, Addison-Wesley, Reading, Mass., (1969).

20. Huber, M., et al., "Comparative Analysis of Traffic Assignment Techniques with Actual Highway Use", N.C.H.R.P. report No. 58.

21. Minty, G.J., "Monotone Networks", Proc. of Royal Soc., London, Ser. A 257, (1960), 194-212.

22. Murchland, J.D., "Road Network Traffic Distribution in Equilibrium", paper presented at the Conf. on Math. Meths. in the Econ. Sciences, Mathematisches Forschungsinstitute, Oberwolfach, (1969).

23. Nguyen, S., "A Unified Approach to Equilibrium Methods for Traffic Assignment", paper presented at the Inter. Sym. on Traffic Equilibrium Methods, Centre de Recherche sur les Transports, Univ. de Montréal, November, 1974.

24. Ostrowski, A.M., Solution of Equations and Systems of Equations, 2nd Ed., Academic Press, New York, New York, (1966).

25. Peterson, E.L., "Symmetric Duality for Generalized Unconstrained Geometric Programming", SIAM Jour. Appl. Math., 19, (1970), 487-526.

26. ___________, "Mathematical Foundations of Geometric Programming", Appendix to "Geometric Programming and Some of Its Extensions", Optimization and Design, (Avriel, Rijckaert, Wilde, eds.), Prentice-Hall, Englewood Cliffs, N.J., (1973), 244-289.

27. ___________, "The Decomposition of Large (Generalized) Geometric Programming Problems by Tearing", Proc. NATO Inst. Large Scale Systems, (Himmelblau, ed.), North Holland, Amsterdam, (1973), 525-545.

28. ___________, "Geometric Programming - A Survey", SIAM Review, to appear.

29. ___________, "Geometric Programming - A Handbook", Chapter V of Handbook of Operations Research, (Elmaghraby, Moder, eds.), Van Nostrand-Reinhold, Princeton, N.J., to appear.

30. ___________, "On the Existence of Recession Directions in Geometric Programming", in preparation.

31. ___________, "The Decomposition of Large (Generalized) Geometric Programming Problems by Geometric Lagrangians", in preparation.

32. ___________, The Mathematical Foundations of Convex and Nonconvex Programming, in preparation.

33. Potts, R.B., Oliver, R.M., Flows in Transportation Networks, Academic Press, New York, N.Y., (1972).