Returns to Scale in Networks

by

Marvin Kraus*

June 2006

Keywords: Networks, congestion, returns to scale, congestion pricing

*Department of Economics, Boston College, Chestnut Hill, MA 02467, USA. E-mail: kraus@bc.edu. I am grateful to Richard Arnott for helpful discussions.

Returns to Scale in Networks

1. Introduction

When demand increases in a congestible network, there are typically a variety of ways that a network authority can provide additional capacity. In a highway network, existing roads can be widened or new ones added. In a bus network, not only is an increase in service frequency or bus size possible, but also in the density of the route network.

This paper is concerned with the determination of the degree of local returns to scale in the cost function for the network's outputs when there are multiple margins – including the network's density – along which the network authority can make adjustments to capacity. The special importance of the degree of local returns to scale is that it determines whether budgetary balance can be achieved under optimal pricing and capacity provision. In particular, when the degree of local returns to scale is equal to one, the receipts from congestion charges for an optimally priced and designed network are exactly equal to its capacity costs, resulting in exact budgetary balance.[1] First for a highway network, then for a bus network, we prove:

Theorem. Under the provision of a cost-minimizing network, the degree of local returns to scale is the same along all margins for adjusting capacity, both singly and in combination. This includes the network's density.

The theorem applies to any type of network. This will hopefully be clear from the intuition we provide. We will see that the theorem is essentially an Envelope Theorem result and that the intuition lies in the Envelope Theorem.

A particular implication of the theorem is that, subject to the optimality of the initial network, its density can be held fixed in evaluating local returns to scale. This was argued by Kraus (1981) in the context of highway networks, but, with the exception of one simple network model, without formal proof. The objective of the present paper is proof of the more general theorem.

There are two important applications of the theorem. One is that if a network can be assumed

to be cost-minimizing, then local returns to scale can be estimated along any margin. An example of this type of application is Kraus (1981), which estimated the degree of local scale economies in urban highway network capital costs holding network density fixed. The second is that if cost minimization cannot be assumed, and if indeed it turns out that the estimated degree of local returns to scale along different margins is different, then one has evidence of cost inefficiency.

The next section sets out our basic model. It is well suited to highway networks and can be adapted to other types of networks. We illustrate this in Section 3 for an adaptation to bus networks. Section 4 concludes.

2. Model and Analysis

The network consists of two types of elements, nodes and links. The number of each is finite. The set of nodes is denoted by N, the set of links by L. Each element of N is either an origin, a destination, or a junction of links. Each link in L is associated with (exactly) two distinct nodes in N and provides a directed connection between these nodes.[2] There can be multiple links that connect a directed pair of nodes.

Let r and s be any two distinct elements of N. Then by a path from r to s, we mean a sequence of links $\{a_1,\ldots,a_K\}$ such that there exist $K+1$ distinct nodes $n_0, n_1,\ldots,n_K$, where $n_0 = r$, $n_K = s$, and a_i provides a connection from n_{i-1} to n_i. The set of all paths in the network is denoted by P.

Let r and s be two distinct nodes such that there is a path from r to s. Then the ordered pair of nodes $w = (r,s)$ will be said to be connected, and the set of all such w will be denoted by W. W is the set of all origin-destination (O-D) pairs.[3]

For each origin-destination pair $w \in W$, there is some nonnegative demand y_w and a subset P_w of P, consisting of those paths of the network associated with w. The output y_w has to be produced using the paths in P_w. For $p \in P_w$, the part of y_w produced using p is denoted by q_p. We have

$$\sum_{p \in P_w} q_p = y_w \qquad \forall w \in W. \tag{1}$$

Let δ_{ap} be the binary variable defined by $\delta_{ap} = 1$ if link $a \in L$ is contained in path $p \in P$, and otherwise defined by $\delta_{ap} = 0$. Denoting the flow (i.e., load) on link a by x_a, we have

$$x_a = \sum_{p \in P} \delta_{ap} q_p \qquad \forall a \in L. \tag{2}$$

Each link a has some nonnegative capacity k_a, which is set by a network authority (NA). The NA can exclude link a from the network by setting $k_a = 0$. For any $k_a > 0$, link a is part of the network. Thus, L is the set of potential links, with the set of actual links being some endogenously determined subset of L.

Let k be a vector whose components are all of the k_a's. Similarly, let x be a vector whose components are the x_a's. Denoting user cost (the cost incurred by a user for a single use) on link a by f_a, we assume that f_a depends not only on x_a and k_a, but on the flows and capacities of other links. We denote this by $f_a(x, k)$, which we take to be a C^1 function satisfying $\partial f_a(\cdot)/\partial x_a > 0$, $\partial f_a(\cdot)/\partial k_a < 0$. Aggregate user costs on link a are

$$c_a(x, k) \equiv x_a f_a(x, k). \tag{3}$$

$f_a(x, k)\big|_{k_a = 0}$ is assumed to be finite, but sufficiently large that $k_a = 0$ leads to path flows in which $x_a = 0$. From (3), it follows that $c_a(x, k)\big|_{k_a = 0} = 0$.

Finally, the cost that the network authority incurs for capacity provision is $g(k)$. We take this to be a C^1 function satisfying $\partial g(\cdot)/\partial k_a > 0$.

Full Long-Run Optimum

By a full long-run optimum, we mean vectors of link and path flows, x and q (q is a vector whose components are the q_p's), and a vector of link capacities, k, such that the total system cost expression

$$\sum_{a \in L} c_a(x, k) + g(k) \tag{4}$$

is at a minimum subject to (1), (2) and nonnegativity restrictions on q and k.[4]

Let y be a vector whose components are all of the y_w's. When the solution to the preceding problem is unique in link flows and capacities ($x(y)$ and $k(y)$, respectively), then we can write the problem's value function

$$\sum_{a \in L} c_a(x(y), k(y)) + g(k(y)) \equiv C(y). \tag{5}$$

$C(y)$ is the network's long run cost function. Our concern in what follows is with the associated elasticity of scale function

$$E(y) \equiv \frac{\lambda}{C(\lambda y)} \cdot \frac{dC(\lambda y)}{d\lambda}\bigg|_{\lambda = 1}. \tag{6}$$

Its reciprocal, $S(y) = 1/E(y)$, is the conventional measure of the degree of local returns to scale in terms of the cost function (see, e.g., Bailey and Friedlaender (1982)). As we will see below, $S(y)$ gives the ratio of the total cost of output to the total value of output under marginal cost pricing.[5]

For expositional purposes, we now proceed to break up the problem into stages, with x and q optimized conditional on k at the first stage, and k optimized at the second stage.[6] With k taken as given, the problem at the first stage is:

$$\min_{x,q} \quad \sum_{a \in L} c_a(x,k)$$

$$\text{s.t.} \quad \sum_{p \in P_w} q_p = y_w \qquad \forall w \in W \tag{1}$$

$$x_a = \sum_{p \in P} \delta_{ap} q_p \qquad \forall a \in L \tag{2}$$

$$q_p \geq 0 \qquad \forall p \in P.$$

In what follows, we will assume that a solution to this problem exists and that optimal link flows are given by the C^1 function $x(y,k)$.[7]

The problem at stage 2 is simply

$$\min_{k \geq 0} \quad \phi(y,k) \tag{7}$$

where, for each value of k,

$$\phi(y,k) \equiv \sum_{a \in L} c_a(x(y,k), k) + g(k) \tag{8}$$

is one of the network's short run cost functions. Since $c_a(x,k)$, $g(k)$ and $x(y,k)$ are all C^1 functions, so is $\phi(y,k)$, and the first-order conditions for (7) are the Kuhn-Tucker conditions that, $\forall a \in L$,

$$\frac{\partial \phi}{\partial k_a} \geq 0, \qquad k_a \geq 0, \tag{9a}$$

$$k_a \frac{\partial \phi}{\partial k_a} = 0. \tag{9b}$$

We assume that a solution to (7) exists, and that optimal capacities are given by the C^1 function $k(y)$.

Under this approach, the network's long run cost function arises as the value function for (7):

$$C(y) = \phi(y, k(y)), \tag{10}$$

which is nothing more than the envelope of the network's family of short run cost functions. Since $\phi(y,k)$ and $k(y)$ are both C^1 functions, so is $C(y)$, ensuring that the elasticity of scale at a point is well-defined (see (6)).

Remark. One can now appreciate the role of the various smoothness assumptions we have made. They are indispensable to having a well-defined elasticity of scale at a point. Had we decided not to stage the problem, slightly weaker assumptions would have been possible, but would have come at the expense of the expositional ease that will now ensue.

From (6),[8]

$$E(y) = \frac{1}{C(y)} \sum_{w \in W} y_w \frac{\partial C}{\partial y_w}. \tag{11}$$

Meanwhile, from (10),

$$\frac{\partial C}{\partial y_w} = \frac{\partial \phi}{\partial y_w} + \sum_{a \in L} \frac{\partial \phi}{\partial k_a} \frac{\partial k_a}{\partial y_w}. \tag{12}$$

Thus,

$$E(y) = \sum_{w \in W} \frac{y_w}{C(y)} \frac{\partial \phi}{\partial y_w} + \sum_{a \in L} \sum_{w \in W} \frac{y_w}{C(y)} \frac{\partial \phi}{\partial k_a} \frac{\partial k_a}{\partial y_w}. \tag{13}$$

The first term on the right-hand-side of (13) is the elasticity of scale of short run costs. In the second term,

$$\sum_{w \in W} \frac{y_w}{C(y)} \frac{\partial \phi}{\partial k_a} \frac{\partial k_a}{\partial y_w} \tag{14}$$

is the contribution to the elasticity of scale from the network authority using k_a as a margin of adjustment. The key thing to note is that (14) is equal to zero $\forall a \in L$. This is clearly the case whenever $\partial \phi / \partial k_a = 0$, which, from (9b), holds whenever $k_a > 0$. But what about when $\partial \phi / \partial k_a > 0$? Not only is k_a equal to zero in this case, but the solution for k_a is deep in the corner, resulting in $\partial k_a / \partial y_w = 0 \ \ \forall w \in W$. Thus, (14) equals zero in this case also.

To see the meaning of this result, let θ_a be a zero-one variable and, with θ as the vector of the θ_a's, define

$$e(y, \theta) = \sum_{w \in W} \frac{y_w}{C(y)} \frac{\partial \phi}{\partial y_w} + \sum_{a \in L} \theta_a \sum_{w \in W} \frac{y_w}{C(y)} \frac{\partial \phi}{\partial k_a} \frac{\partial k_a}{\partial y_w}. \tag{15}$$

$1/e(y, \theta)$ is a generalized local returns to scale measure in which θ_a is an indicator variable which takes on the value of 1 or 0 according to whether or not k_a is used as a margin of adjustment in evaluating local returns to scale. Note that $e(y, \theta)$ has $E(y)$ as a special case (when $\theta_a = 1$ $\forall a \in L$) and that our result about (14) implies that $e(y, \theta)$ is independent of θ. We state this as:

Theorem 1. For a network at a full long-run optimum, $e(y, \theta)$ is independent of θ and equal to $E(y)$. That is, it is immaterial which of a network authority's possible margins of adjustment are used in evaluating local returns to scale. In particular, it is unnecessary in evaluating local returns to scale to consider adding new links to the network.

Conditional Cost Minimum

What if the congestion externalities imposed by users of the network are left unpriced, individuals choose paths according to their private costs, and, conditional on these choices, the network authority sets cost-minimizing capacities. We will refer to this as a conditional cost minimum and show that the previous result holds.

In the absence of congestion pricing, the price of using link a is its user cost $f_a(x, k)$. The price of using path p is the sum of the prices of its component links or

$$\sum_{a \in L} \delta_{ap} f_a(x, k). \tag{16}$$

For an O-D pair w, the lowest of its path prices is

$$\min_{p \in P_w} \sum_{a \in L} \delta_{ap} f_a(x, k). \tag{17}$$

Individuals act as price-takers, taking link and path prices as given. The only paths in P_w that they use are those that have prices equal to the minimum in (17). Given k, the conditions for link and path flows to be in equilibrium consist of (1), (2) and the following pair of conditions $\forall w \in W$:

$$q_p \left(\sum_{a \in L} \delta_{ap} f_a(x, k) - \min_{p \in P_w} \sum_{a \in L} \delta_{ap} f_a(x, k) \right) = 0 \qquad \forall p \in P_w \tag{18a}$$

$$q_p \geq 0 \qquad \forall p \in P_w. \tag{18b}$$

The key condition here is (18a). It implies that if $q_p > 0$ for some $p \in P_w$, then p must be one of w's lowest-price paths. It also implies that if $p \in P_w$, but is not one of w's lowest-price paths, then $q_p = 0$.

We assume that a solution to the system of equilibrium conditions exists, and that equilibrium link flows are given by the C^1 function $\tilde{x}(y, k)$. $\tilde{x}(y, k)$ corresponds to $x(y, k)$ of the full long-run optimum problem.

The problem faced by the network authority is similar to the previous stage 2 problem. It is simply

$$\min_{k \geq 0} \quad \tilde{\phi}(y,k), \tag{19}$$

where

$$\tilde{\phi}(y,k) \equiv \sum_{a \in L} c_a(\tilde{x}(y,k), k) + g(k). \tag{20}$$

From here, we proceed just as in the previous section, next invoking the Kuhn-Tucker conditions for (19). Under existence, uniqueness and smoothness assumptions analogous to those for (7), the theorem of the previous section holds for a network at a conditional cost minimum.

3. A Model of a Bus Network

Consider a road network through which individuals make trips on buses. We define nodes, links and paths as before. Unlike in the previous section, we take all link capacities as given. Thus, the road network is given, and the NA optimizes bus service.

Each path of the network is a potential bus line. A bus line is therefore associated with a sequence of nodes, and these are its stops. Note also that a bus line is directional (e.g., no. 40, inbound), with the set of all possible bus lines represented by P.

Given an O-D pair $w = (r,s)$, we now take P_w to be the set of bus lines that an individual can use to travel from r to s. These not only include bus lines that start at r and terminate at s, but those for which r and/or s are in between the line's extremities, with s coming sequentially after r.

Given any $p \in P_w$, where $w = (r,s) \in W$, we define q_{pw} to be the number of individuals who use bus line p to travel from r to s. We assume that the only way to travel is by bus and hold off for the time being on introducing the complication of a trip involving a possible transfer from one bus line to another. Under this assumption, (1) becomes

$$\sum_{p \in P_w} q_{pw} = y_w \qquad \forall w \in W. \tag{21}$$

Given any bus line p, let L_p be the set of all links a that make up p, and for any link $a \in L_p$, define x_{ap} to be the passenger load on bus line p over link a. We can express x_{ap} as

$$x_{ap} = \sum_{w \in W_p} \delta_{apw} q_{pw} \qquad \forall a \in L_p \text{ and } p \in P, \tag{22}$$

where for any bus line p, W_p is the set of all O-D pairs served by p, and for any $w = (r,s) \in W_p$ and link $a \in L_p$, δ_{apw} is an indicator variable equal to one if link a is part of the subpath of p from r to s and zero otherwise.

For simplicity, we will assume that the only cost that a passenger incurs is a crowding cost. For link a of bus line p, we write this as $f_{ap}(x_{ap}, k_p)$, where k_p is the frequency of bus service that the NA provides to bus line p. If $k_p = 0$, line p is excluded from the network of bus routes. $f_{ap}(\cdot)$ is assumed to satisfy $\partial f_{ap}(\cdot)/\partial x_{ap} > 0$, $\partial f_{ap}(\cdot)/\partial k_p < 0$, and we define

$$c_{ap}(x_{ap}, k_p) \equiv x_{ap} f_{ap}(x_{ap}, k_p). \tag{23}$$

Writing the cost that the NA incurs for providing bus service as $g(k)$ (k is now a vector whose components are the k_p's), a full long-run optimum now consists of a set of link and path flows (the x_{ap}'s and q_{pw}'s, respectively) and a vector of service frequencies, k, such that the total system cost expression

$$\sum_{p \in P} \sum_{a \in L_p} c_{ap}(x_{ap}, k_p) + g(k) \tag{24}$$

is at a minimum subject to (21), (22) and the nonnegativity restrictions

$$q_{pw} \geq 0 \qquad \forall w \in W_p \text{ and } p \in P \tag{25}$$

$$k_p \geq 0 \qquad \forall p \in P. \tag{26}$$

Staging the problem as in the previous section, we can again write the stage 2 problem as (7), where $\phi(y,k)$ is again the value function for stage 1. Thus, under the same assumptions as in the previous section and with obvious adjustments to (13) and (15) – k_p's instead of k_a's and indicator variables θ_p – Theorem 1 holds for the present model of a bus network. It implies that it is unnecessary in evaluating local returns to scale to consider adding new service routes to the network.

Transfers

A passenger is now permitted to make up to a single transfer.[9] In other respects the model is unchanged.

Given an O-D pair $w = (r, s)$, the set of bus lines that an individual can use to travel from r to s without making a transfer is again denoted by P_w.

By a pair of transfer-compatible paths for an O-D pair $w = (r, s)$ we mean an ordered pair of paths $\tilde{p} = (p', p'')$ which have exactly one common node – the transfer point[10] – such that r is a node of p' which comes sequentially before the transfer point, while s is a node of p'' which comes sequentially after the transfer point. The set of all pairs of transfer-compatible paths for w is denoted by A_w. Also, for any $w = (r, s)$ and any $\tilde{p} = (p', p'')$ in A_w, the number of individuals who travel from r to s by transferring from p' to p'' is denoted by $Q_{\tilde{p}w}$. Then (21) generalizes to

$$\sum_{p \in P_w} q_{pw} + \sum_{\tilde{p} \in A_w} Q_{\tilde{p}w} = y_w \qquad \forall w \in W. \tag{27}$$

In order to generalize (22), we let $\Delta_{ap\tilde{p}w}$ be an indicator variable which is one if either p is the first path that makes up $\tilde{p}$ and link a is between r and the transfer point or p is the second path that makes up $\tilde{p}$ and a is between the transfer point and s, and zero otherwise. Then (22) becomes

$$x_{ap} = \sum_{w \in W_p} \delta_{apw} q_{pw} + \sum_{w \in B_p} \sum_{\tilde{p} \in A_w} \Delta_{ap\tilde{p}w} Q_{\tilde{p}w} \qquad \forall a \in L_p \text{ and } p \in P, \tag{28}$$

where B_p is the set of all O-D pairs served by p through a transfer.

As before, the problem is to minimize (24). The only difference is that (27) and (28) replace (21) and (22). This has no effect on the applicability of Theorem 1. Nor would allowing for multiple transfers.

4. Conclusion

This paper has presented a theorem:

Theorem. Under the provision of a cost-minimizing network, the degree of local returns to scale is the same along all margins for adjusting capacity, both singly and in combination. This includes the network's density.

Seeing that the theorem is general is no more difficult than understanding the Envelope Theorem. The key is to model the capacity of a potential link or serviceable path as a control variable subject to a nonnegativity restriction. Regardless of whether the optimal value of the control variable is initially positive or zero, its value can be held fixed in evaluating the cost function's elasticity of scale. Either costs are stationary with respect to the control variable or the solution for the control variable is deep in the corner.

References

1. E.E. Bailey and A.F. Friedlaender, Market structure and multiproduct industries, *Journal of Economic Literature* 20 (1982), 1024-1048.
2. M. Kraus, Scale economies analysis for urban highway networks, *Journal of Urban Economics* 9 (1981), 1-22. Reprinted in H. Mohring, ed., *The Economics of Transport*, Edward Elgar, 1994.
3. H. Mohring and M. Harwitz, *Highway Benefits: An Analytical Framework*, Northwestern University Press, 1962.
4. A. Nagurney, *Network Economics*, Kluwer, 1993.
5. K.A. Small, Economies of scale and self-financing rules with non-competitive factor markets, *Journal of Public Economics* 74 (1999), 431-450.

Footnotes

1. The basic result for an isolated highway was derived by Herbert Mohring and appears in Mohring and Harwitz (1962).
2. Thus, a two-way road is represented by separate links for each direction.
3. This does not rule out the possibility that the demand for a certain O-D pair is zero.
4. Equation (2) ensures that x is nonnegative whenever q is.
5. It should be noted that, since user costs are included in both the total cost of output and the total value of output, this ratio is different than the ratio of total capacity costs to receipts from congestion charges.
6. The first-stage problem has been studied extensively (see, e.g., Nagurney (1993)). In most network models, link capacities are exogenous and there is no second-stage problem.
7. Because it has both y and k as arguments, there should be no confusion between this function and the function $x(y)$ introduced earlier.
8. To see that $E(y)$ gives the ratio of the total value of output to the total cost of output under marginal cost pricing, simply substitute price for marginal cost where the latter appears in (11). The result holds regardless of whether factor markets are competitive (Small (1999)).
9. Allowing for multiple transfers complicates the model, but introduces nothing new as far as Theorem 1 is concerned.
10. For simplicity, transfers between lines with multiple intersection points are ignored.