92-25

UNIVERSITY OF CALIFORNIA, SAN DIEGO

DEPARTMENT OF ECONOMICS

EFFICIENT TRANSPORTATION ROUTING AND NATURAL MONOPOLY IN THE AIRLINE INDUSTRY: AN ECONOMIC ANALYSIS OF HUB-SPOKE AND RELATED SYSTEMS

BY

ROSS M. STARR

AND

MAXWELL B. STINCHCOMBE

DISCUSSION PAPER 92-25
JUNE 1992

EFFICIENT TRANSPORTATION ROUTING AND NATURAL MONOPOLY IN THE AIRLINE INDUSTRY: AN ECONOMIC ANALYSIS OF HUB–SPOKE AND RELATED SYSTEMS†

by

Ross M. Starr and Maxwell B. Stinchcombe

May, 1992

Department of Economics 0508
University of California at San Diego
La Jolla, CA 92093–0508.

ABSTRACT

We investigate transportation network structures that result from optimization problems in the face of given cost and demand functions and given constraints on the networks. Hub–spoke or related systems are preferable to point–to–point systems when the number of connecting routes constitutes a significant component of network cost. Further, these systems are optimal under a variety of cost and demand configurations, and typically demonstrate large, pervasive economies of scale. The scale economies arise because each additional city included in the network increases traffic density to all other cities. Pervasive scale economies indicate that the industry is a natural monopoly.

† Many thanks to Vince Crawford for helpful comments on an early version of this paper, to Robert Horwitz and Joel Sobel for references and discussions, and to Hal White for several useful conversations.

1. Introduction

In 1978 the deregulation of the U.S. airline industry began. By the spring of 1980, airlines were virtually free to determine their route structures and prices. On the basis of an initial analysis of the benefits of "hubbing," American Airline's Robert Crandall was the first to aggressively pursue this strategy.[1] Since that time all major domestic carriers have adopted interlocking hub–spoke route structures. There have been several investigations of the implied cost savings ([4], [5], [6], [16], [18]). Here we investigate the theory of optimal route structures and find that the hub–spoke and related structures are optimal under a variety of assumptions. Typically, these systems also demonstrate large and pervasive scale economies. These scale economies arise from the interaction across spokes of demand in the network structure with the decreasing average passenger costs at the level of individual routes. The implications of this observation for the industrial organization of the airline system are strong. Consistently decreasing average cost characterizes a 'natural monopoly.' We would expect the systemic scale economy to give rise to a highly concentrated industry focusing on a relatively small number of hubs.

As is well understood, hub–spoke systems allow higher traffic densities, that is, more passengers per plane, than would be possible in a system of direct flights. This higher density is achieved by funneling passengers through the hub airport. This concentrates the passengers to each destination on a single aircraft, rather than spreading them over the several aircraft that would be required in a system of direct flights. As the airline's marginal cost of a passenger on a plane with space is very small, any arrangement that

[1] See [21] for a historical survey of airline regulation and practices through to the late 1980's and an analysis of how the industry technology and cost structure adapted to its different regulatory environments. [7] analyzes the effects of indivisibilities and scale economies at the level of the individual route. [15] and [20] contain a more current perspective, while [1] contains an account of the deregulation of the airlines, and [12] contains a superb account of the deregulation of another industry with network characteristics, telecommunications.

increases traffic density reduces average cost per passenger–trip.

We will formalize this argument in a combinatorial model of transportation route structure. Declining average cost on each route traveled is roughly a necessary and sufficient condition for hub–spoke route structures to dominate direct point–to–point route structures. Further, under a wide range of conditions and constraints, hub–spoke and related multi–hub systems are optimal. This is both intuitive and consistent with the literature.[2]

The comparison between point–to–point route structures and hub–spoke systems is instructive: It takes $n(n-1)$, or approximately n^2, connecting lines to directly link n distinct points in both directions; it takes $2(n-1)$, or approximately $2n$, connecting lines to link the same points in both directions by connecting through a common hub. The requisite number of lines grow quadratically in a point–to–point system, and only linearly in a hub–spoke system. It's that simple — hub–spoke airline route structures economize on the number of route segments needed to link a large number of origin–destination pairs. When the number of routes is a significant determinant of costs, hub–spoke systems dominate point–to–point systems.

To examine the overall optimality of hub–spoke and related systems, we must consider different cost structures. We begin by characterizing the optimal networks in a polar cases where hub–spoke systems are not optimal.

- SMALL FIXED COSTS – COSTS NEARLY LINEAR IN PASSENGER MILES: If total transportation costs varied linearly with the number of passenger–miles then for any array of origins and destinations, the cost minimizing schedule would be direct point–to–point. Any system other than point–to–point would be inefficient because some passengers would be obliged to travel along indirect routes, increasing the number of passenger miles

[2] When the work of this paper was essentially complete we became aware of [11]. The same basic questions are answered there for more general specifications of the demand and cost systems. Our simpler structures afford us both more complete characterizations of the solutions, and the possibility of solving these problems under a wider variety of constraints.

travelled. This observation is formalized as Theorem A.

Theorem A shows that when it is the marginal costs per passenger mile that dominate, costs as well as benefits grow quadratically in the number of cities connected. Theorems B–D cover the case where fixed costs dominate the cost function. In most of these cases, hub–spoke or related multi–hub systems arise as solutions, and whatever solution arises, costs are essentially linear in the number of cities.

- SMALL VARIABLE COSTS — COSTS NEARLY DETERMINED BY ROUTES: If all transportation costs were fixed costs per vehicle–trip, independent of the number of passengers, then the optimal network uses at most $2(n-1)$ routes. If the fixed costs are also independent of the pairs of destinations, the minimum cost network is achieved with the minimal number of vehicle trips to move among any n points. This is n, the cyclical route structure $i_1 \to i_2 \to \ldots \to i_n \to i_1$ for some permutation of the cities in the network. These observations are formalized as Theorem B.

Cyclical routes structures (milk runs) have an obvious drawback from the point of view of the passengers — their trips may involve as many as $n-1$ distinct flights (e.g. a passenger wanting to get from i_2 to i_1 must pass through $i_3, i_4, \ldots, i_n$ on the way.) Our proxy for this sort of cost is the constraint that passengers need take no more than m different flights in order to complete their travel plans. If m is 1, the only feasible network route structure is a system of direct flights. When $m = 2$, we are led to hub–spoke networks.

- SMALL VARIABLE COSTS — MAXIMUM OF ONE INTERMEDIATE STOP REQUIRED: When the variable costs are low, fixed costs depend only on the number of vehicle trips, not on the pairs of destinations, and there are 4 or more cities to connect, then optimal networks are hub–spoke networks with one central hub. This is formalized in Theorem C. Here the number of routes is $2(n-1)$.

If the variable costs are small and $m > 2$, that is, more than 2 changes of plane are allowed, then optimal network structures include a variety of interlocking small cycles combined with hubs. Again, from the point of view of the passenger, cyclical patterns have obvious drawbacks, drawbacks that competitors might be motivated to take advantage of. Our proxy for this sort of competition is the constraint that the network route structure be symmetric, if there are direct flights from i to j, then there are direct flights from j to i.

- SMALL VARIABLE COSTS — SYMMETRIC ROUTE STRUCTURES: If we require that the route structure economize on fixed costs and be symmetric, then the graph of the solution is a tree. Trees represent routes structures with many hubs and no cycles. Depending on the cost structure there will be more or less concentration at a small number of hubs. This is formalized as Theorem D. Here again, the number of routes is $2(n-1)$.

To us, the important comparison between Theorem A, covering the case of low fixed costs, and Theorems B through D, covering the case of high fixed costs, is the emergence of a system–wide scale economy in the high fixed cost case. In the high fixed cost case, costs are linear in the number of cities connected, $2(n-1)$ or better, and benefits are quadratic, $n(n-1)$. Now, a small scale economy is implicit in the statement of the problem, the declining average cost on each airline route. One might reasonably expect scale economies to be fully realized once each single route achieves efficient passenger loads on each flight. On the contrary, scale economies are consistently present — a many–fold growth in the number of cities in a hub–spoke system can result in a more than proportionate increase in measures of welfare (eg. monopoly profit or consumer plus producer surplus).

The system–wide scale economy comes from the interaction of travel demand, route structure, and the route–level declining average cost. There are strong spillover effects between routes to a given hub. Each route generates traffic for the others since from each origin there are passengers for each destination through the hub.[3] Hence the

[3] [21, Fig. 2 and 3] shows that American Airlines had discovered that the cost of adding cities to a hub–spoke route system are linear in the number of cities while benefits are quadratic.

organizational structure of scheduling and pricing will be complicated by the interaction between spokes. In a hub–spoke system each city generates travel demand from that city to each of the others in the system. Adding cities increases traffic density (reduces average cost) to and from each of the other cities. Overall, the system is characterized by diminishing average cost over a wide range of activity levels. The scale economy is systemic and large.

The implications of this observation for the industrial organization of the airline system are strong. Consistently decreasing average cost characterizes a natural monopoly. We would expect the systemic scale economy to give rise to a highly concentrated industry focusing on a small number of hubs. However since the declining cost structure derives from a market interaction rather than large scale economies in production technology, an alternative interpretation is that there is an external economy (a pecuniary externality) between spokes feeding the same hub. Centralization of control of this system in a single firm — a monopoly — can be interpreted either as the direct outcome of the declining average cost or as a means to internalize the externality. Alternatively, many firms may provide service on the spoke routes. Efficient allocation then requires coordinated concentration of the route structure on a high volume hub, in order fully to utilize the low marginal cost of additional passenger volume on each flight and the positive interaction of traffic densities between spokes. There will then be an inter–firm externality to be treated (for efficient allocation) by co–operation or regulation.

2. Modeling Structure, Demand and Cost in a Transportation Network

Let N be a set of n points (cities) in $\mathbb{R}^2$, $n \geq 2$, with typical elements i, j, and k. Let M be an $n \times n$ matrix. The intended interpretation is that M_{ij} is the number of passenger miles that want to travel from i to j. We make the following balance assumption on the demand for air travel, ignoring e.g. the timing issues of business vs.

weekend travelers.

Balanced Demand $M_{ij} = M_{ji} > 0$ for $i \neq j$, and $M_{ii} = 0$ for all i.

Following [9, 10], we describe a route structure as a relation R on N. The intended interpretation is that iRj if there are direct flights connecting i and j while $\neg iRj$ denotes "it is not the case that iRj".[4] By convention, there is no direct flight connecting i with i, that is, $\neg iRi$ for all $i \in N$. Viewing R as a subset of $N \times N$, this can be rewritten as $\Delta \cap R = \emptyset$, where Δ is the diagonal in $N \times N$. When convenient, we shall also regard any relation S on N as a correspondence by defining $S(i) = \{j\colon iRj\}$. Thus, for a route structure R, $R(i)$ is the set of j that can be reached from i by direct flights.

For relations S on N we will use the following notation:

(1) For $m \in \mathbb{N}$, define the relation S^m by $iS^m j$ if there are m points $i = i_1, i_2, \ldots, i_{m-1}, i_m = j$, such that $i_j R i_{j+1}$ for $j = 1, 2, \ldots, m-1$. For a route structure R, $R^m(i)$ is the set of j that can be reached from i in m steps.

(2) For $m \in \mathbb{N}$, let $\bar{S}^m$ denote $\cup_{m' \leq m} S^{m'}$. For a route structure R, $i\bar{R}^m j$ if it is possible to travel from i to j in m or fewer steps. $\bar{S}^n = \bar{S}$ is the transitive closure of S.

(3) For $i \in N$, $S\_(i)$ denotes the set $\{j \in N\colon jSi\}$. For a route structure R, $R\_(i)$ is the set of points from which i can be reached by a single direct flight, $\bar{R}^m\_(i)$ is the set of points from which i can be reached in m or fewer flights, and $R^m\_(i)$ denotes the the set of point from which i can be reached in exactly m flights.

In some of what follows we will restrict attention to symmetric route structures, that is, to route structures where iRj if and only if jRi (there are direct flights from i to j if and only if there are direct flights from j to i).

---

[4] [19, Ch. 6, §III] contains a variety of network models of social phenomena for which this formalism may prove useful.

As an aid to intuition, it is often convenient to have a matrix representation of route structures. For this purpose we will use adjacency matrices ([3, p. 22]). For example, if R is the cyclical route structure $1 \to 2$, $2 \to 3$, $3 \to 4$, and $4 \to 1$, then the following four matrices represent R, R^2, R^3, $\bar{R}^3$, and $\bar{R}^4$. In this case, $\bar{R}^3$ is equal to E, the route structure where Every city is connected by a direct flight to every other city.

$$
\begin{array}{c|cccc|}
i \backslash j & 1 & 2 & 3 & 4 \\
\hline
1 & & o & & \\
2 & & & o & \\
3 & & & & o \\
4 & o & & & \\
\hline
\multicolumn{5}{c}{R}
\end{array},
\quad
\begin{array}{c|cccc|}
 & 1 & 2 & 3 & 4 \\
\hline
1 & & & o & \\
2 & & & & o \\
3 & o & & & \\
4 & & o & & \\
\hline
\multicolumn{5}{c}{R^2}
\end{array},
\quad
\begin{array}{c|cccc|}
 & 1 & 2 & 3 & 4 \\
\hline
1 & & & & o \\
2 & o & & & \\
3 & & o & & \\
4 & & & o & \\
\hline
\multicolumn{5}{c}{R^3}
\end{array},
\quad
\begin{array}{c|cccc|}
 & 1 & 2 & 3 & 4 \\
\hline
1 & & o & o & o \\
2 & o & & o & o \\
3 & o & o & & o \\
4 & o & o & o & \\
\hline
\multicolumn{5}{c}{\bar{R}^3 = E}
\end{array},
\quad
\begin{array}{c|cccc|}
 & 1 & 2 & 3 & 4 \\
\hline
1 & o & o & o & o \\
2 & o & o & o & o \\
3 & o & o & o & o \\
4 & o & o & o & o \\
\hline
\multicolumn{5}{c}{\bar{R}^4 = N \times N}
\end{array}
$$

In a similar fashion, if R is the hub–spoke route structure with the hub at $k = 1$, the corresponding graphs are

$$
\begin{array}{c|cccc|}
i \backslash j & 1 & 2 & 3 & 4 \\
\hline
1 & & o & o & o \\
2 & o & & & \\
3 & o & & & \\
4 & o & & & \\
\hline
\multicolumn{5}{c}{R}
\end{array},
\qquad
\begin{array}{c|cccc|}
 & 1 & 2 & 3 & 4 \\
\hline
1 & o & o & o & o \\
2 & o & o & o & o \\
3 & o & o & o & o \\
4 & o & o & o & o \\
\hline
\multicolumn{5}{c}{R^2 = \bar{R}^2 \supset E}
\end{array}.
$$

We say that a route structure R is connected if it is possible to go from each city i to any other city j using the flights in the route structure. Formally, this is the requirement that $\bar{R}^n = N \times N$, or equivalently, $E \subset \bar{R}^n$.

Let C(R,M) denote some measure of the cost of running the route structure R to serve the demand M, and let B(R,M) denote the benefit (eg. revenue or social benefit) to running the airline R to service the demand M. Implicit in C(R,M) and B(R,M) is the demand for flights chosen by the passengers in response to the route structure R. Note that we have not included prices of flights in the analysis. We wish to describe the efficient route structures for the following two classes of problems, Symmetric and Unrestricted:

Symmetric(m) $\max_R B(R,M) - C(R,M)$ s.t. $E \subset \overline{R}^m$, R symmetric,

Unrestricted(m) $\max_R B(R,M) - C(R,M)$ s.t. $E \subset \overline{R}^m$.[5]

Because $m \geq n$ implies that $\overline{R}^m = \overline{R}$, there is no point in considering these problems for $m > n$.

It is important to note that we do not consider the question of which set of cities in N should be connected, a question is at the center of [2] and [17] in the related context of communications service. That work examines the optimal "equilibrium user sets" ([2]) and the effects of different adjustment processes on the final set of users ([17]). These correspond to the set of connected cities in this context. The reason we avoid analyzing this problem is straightforward: one of the key observations here is that in solving the pair of problems just given, fixed costs tend to be linear in the size of N while benefits are quadratic. When this is true, it seems sensible that all cities will be connected.

We assume that $B(R,M)$ is defined as the sum over all i,j pairs of the benefits to serving the demand M_{ij}, and that this is linear, $b_{ij} \cdot M_{ij}$, $b_{ij} > 0$. Benefit accrues only to flying people between their initial points and their final destination. In particular, no benefit accrues to any intermediate flying needed to go from i to j, though we expect extra cost. Thus, for all connected R, $B(R,M)$ is the constant $\Sigma_{i,j} b_{ij} \cdot M_{ij}$. This implies that we need only characterize solutions to cost minimization problems. We turn now to the cost functions.

For all i,j pairs with $i \neq j$, let α_{ij} and β_{ij} be cost parameters. Let $\tilde{\alpha}$ and $\tilde{\beta}$ be vectors in $\mathbb{R}^{n(n-1)}$ of the α_{ij}'s and β_{ij}'s. We are going to study cost functions of the form $C(R,M;\tilde{\alpha},\tilde{\beta}) = \Sigma_{ij} c_{ij}(R,M;\tilde{\alpha},\tilde{\beta})$ where $c_{ij}(R,M;\tilde{\alpha},\tilde{\beta}) = \alpha_{ij} + \beta_{ij} \cdot m_{ij}(R,M)$ if iRj, and $c_{ij}(R,M;\tilde{\alpha},\tilde{\beta}) = 0$ if $\neg$iRj. Here $m_{ij}(R,M)$ denotes passenger miles moved from i to j when passengers choose their favorite routes in response to R.

To calculate $m_{ij}(R,M)$, we will invoke our General Assumption about Passengers:

[5] This is the restriction that the graphs be m–connected [8].

(GAP) Passengers plan their routes in a fashion that depends on R but not on $\tilde{\alpha}$ or $\tilde{\beta}$; passengers will choose some one of the possible ways of going from i to j.

In specific cases we will also assume one of the following two demand patterns for the passengers: they plan so as to minimize the number of flights that they take; or so as to minimize the total distance they have to fly.[6]

LEMMA 1: Under GAP, for all R and M, C(R,M) is linear in $\tilde{\alpha}$ and $\tilde{\beta}$, and for $(\tilde{\alpha},\tilde{\beta}) \geq (0,0)$ (0 being the origin in $\mathbb{R}^{n(n-1)}$) the mapping $(\tilde{\alpha},\tilde{\beta}) \mapsto \min_R C(R,M;\tilde{\alpha},\tilde{\beta})$ is homogeneous of degree 1, concave, and its graph is a finite polygon.

PROOF: The linearity is clear, and given the linearity, the second claim is immediate. □

Thus, we can give a general characterization of regions of the $(\tilde{\alpha},\tilde{\beta})$ space in which different R's are chosen — they are polygonal cones with their only vertex at the origin. We turn now to analyses of the two classes of problems under various configurations of $\tilde{\alpha}$ and $\tilde{\beta}$.

3. Small Fixed Costs

In this section we consider cost functions for which virtually all of the costs of the transportation network vary directly with passenger mileage flown, the fixed costs are negligible. In the following two sections, we will consider the opposite case. Here, total costs vary linearly with the number of passenger–miles. The cost minimizing route structure is E, direct point–to–point flights between all cities. Because some passengers would be obliged to travel along indirect routes, increasing passenger miles travelled and costs, any system other than point–to–point is inefficient in this case. In particular, no

[6] See [8] for an examination of the qualitative comparative statics of the minimal cost flow problem with respect to cost parameters. Here we subsume the solution to this complicated problem in the m_{ij} functions.

variant of hub spoke can be optimal.

At various points in the analysis it will be convenient to talk about phenomena that are true for sufficiently small, that is, negligible $\mathbf{x}$, $\mathbf{x}$ a vector in $\mathbb{R}^{\ell}$, $\ell \geq 1$. To this end, we will use the notation "$\mathbf{x} \simeq 0$", which is read as "$\mathbf{x}$ is negligible". Formally, the statement "if $\mathbf{x} \simeq 0$ then $S(\mathbf{x})$ is true" means "if $\mathbf{x}^n$ is a sequence with $|\mathbf{x}^n| \to 0$, then for all sufficiently large n, $S(\mathbf{x}^n)$ is true". Thus, negligible vectors or numbers represent the behavior of numbers and vectors very close to 0 or $\mathbf{0}$. Note that the finite sum of negligible vectors or numbers are again negligible. The notation "$\mathbf{y} \simeq \mathbf{x}$" is short–hand for "$\mathbf{y} - \mathbf{x} \simeq 0$."

Let $\tilde{1}$ denote the vector of 1's in $\mathbb{R}^{n(n-1)}$. Recall that costs are given by $C(R,M;\tilde{\alpha},\tilde{\beta}) = \Sigma_{ij}\, c_{ij}(R,M;\tilde{\alpha},\tilde{\beta})$ where $c_{ij}(R,M;\tilde{\alpha},\tilde{\beta}) = (\alpha_{ij} + \beta_{ij}\cdot m_{ij}(R,M))\cdot 1_{\{iRj\}}$ (where 1_A is equal to 1 if A holds and 0 otherwise) and $m_{ij}(R,M)$ denotes passenger miles moved from i to j when passengers choose their favorite routes in response to R.

Theorem A: For all $m \in \mathbb{N}$, if $\tilde{\alpha} \simeq \mathbf{0}$ and $\tilde{\beta} \simeq \beta\cdot\tilde{1}$, $\beta > 0$, then E is the unique solution to the problem Unrestricted(m) and Symmetric(m) when passengers choose their flights either to minimize the distance traveled or to minimize the number of flights. Further, if $\tilde{\alpha} \simeq \mathbf{0}$ and $\tilde{\beta} > \mathbf{0}$, then the cost of the optimal network is bounded below by $\underline{\beta}\cdot\Sigma_{i,j} M_{ij}$ plus a negligible number where $\underline{\beta} = \min\{\beta_{ij}: i \neq j\}$.

Thus, if marginal costs per passenger mile dominate the cost function, and if costs per passenger miles are approximately uniform across routes, then the route structure E, direct flights between every pair of points, dominates. Together, Theorem A and Lemma 1 imply that for negligible $\tilde{\alpha}$, the set of $\tilde{\beta}$ such that E is the solution to Unrestricted(m) and Symmetric(m) is a polygon with non–empty interior containing the ray from the origin generated by $\tilde{1}$.

The last part of Theorem A implies that as n, the number of cities in the network, increases, optimal costs increase quadratically in n. More specifically, if $\underline{M} \leq M_{ij}$ and $\underline{\beta} \leq \beta_{ij}$ for all i,j as n increases, then the cost of the optimal network is at least

$\beta\cdot\underline{M}\cdot n(n-1)$ plus a sum of negligible numbers.

Proof of Theorem A: Because E is symmetric, if we show that it solves Unrestricted(m), then it must also solve Symmetric(m). Let R be the solution. If m = 1, then $E \subset \overline{R}^m$ implies that $R = E$. Suppose $m \geq 2$. Because R is connected, B(R,M) is constant and we need only minimize costs. If $R \neq E$, then some passengers must take more than one flight or fly further to reach their destination. This increases the cost from $C(E,M;\tilde{\alpha},\tilde{\beta})$ by a non–negligible amount, proving the first statement.

The last statement is clear, a total of $\Sigma_{i,j}M_{ij}$ passenger miles must be moved, all of them at a cost of at least β.[7] □

As noted above, for $m \geq n$, $\overline{R}^m = \overline{R}$. If $\tilde{\alpha} = 0$, the study of the problem $\min_R C(R,M;0,\tilde{\beta})$ subject to $\overline{R} = N \times N$ is called network synthesis or network design and can (in principle) be solved by linear programming. When both $\tilde{\alpha} \neq 0$ and $\tilde{\beta} \neq 0$ the problem is highly non–linear.

4. Small Marginal Costs

In this section we begin our treatment of the case where the marginal cost per passenger mile is negligible. Here we will first see that the costs to the optimal networks are at most linear in n. In particular, if marginal costs are negligible and the fixed costs to connecting any pair of cities is approximately equal, then the minimal cost is achieved by connecting all cities with the minimal number of routes. This is n, with routing in the form of a cycle, $i_1 \to i_2 \to \ldots \to i_n \to i_1$ for some permutation of the cities in the network. In general, when the fixed costs are not approximately equal, we show that the maximal number of routes in an optimal network is bounded above by $2(n-1)$. Formally, negligible marginal costs correspond to $\tilde{\beta} \simeq 0$.

[7] As crude as this bound seems, it is tight — consider the case $\tilde{\alpha} \simeq 0$ and $\tilde{\beta} \simeq \beta\cdot\tilde{1}$, $\beta > 0$ in light of the first part of Theorem A.

Definition: A relation R is a cycle that covers N if for all $i \in N$, $\#R(i) = 1$ and $\bar{R}^n = N \times N$ (where $\#T$ denotes the cardinality of a set T).

Thus, R is a cycle if there is only one direct flight out of each point, and if by following the direct flights you will cover every possible destination and return to your starting point in exactly n steps. Note that if R is a cycle then $\#R = n$.

THEOREM B: (i) Under GAP, if $\tilde{\alpha} \simeq \alpha \cdot \tilde{1}$, $\alpha > 0$ and $\tilde{\beta} \simeq 0$, then any solution to Unrestricted(n) is a cycle. (ii) In general, if $\tilde{\alpha} > 0$ (but not necessarily $\simeq \alpha \cdot \tilde{1}$) and $\tilde{\beta} \simeq 0$, then the maximum number of routes in the optimal route structure is less than or equal to $2(n-1)$.

PROOF: (i) Let R be a solution when $\tilde{\alpha} \simeq \alpha \cdot \tilde{1}$, $\alpha > 0$ and $\tilde{\beta} \simeq 0$. Because $\tilde{\beta} \simeq 0$ and cycles use n routes, $\#R \leq n$. If $\#R \leq n-1$ then for some $i \in N$, $\#R(i) = 0$, which implies that it is not possible to leave city i, violating the constraint. Hence $\#R = n$. Because $\#R = n$, if for some i, $\#R(i) \geq 2$, then for some $j \neq i$, $\#R(j) = 0$. But this implies that it is not possible to leave city j, violating the constraint. Hence for all i, $\#R(i) = 1$ so that R is a cycle.

(ii) Now let R be a solution to the general case $\tilde{\alpha} > 0$ and $\tilde{\beta} \simeq 0$. The proof consists in pruning "trunk lines" from R and cities from N in a fashion that removes 2 connecting flights for every city removed from N. We will then note that the remaining flights between cities must be in cycles, and this leads to the $2(n-1)$ bound.

Let N' be a subset of N. We say that $i \in N'$ is in a 2–cycle with $j \in N'$ in R restricted to N' if the only way into and out of i using R and connecting only to points in N' is through j. Formally, this is true if $R(i) \cap N' = R_{-}(i) \cap N' = \{j\}$. We now describe an iterative procedure for pruning N of "trunk lines".

Given $N^t \subset N$, $t = 0, 1, 2, \ldots$, set I^t equal to the set of i such that i is in a 2–cycle with some $j \in N^t$ in R restricted to N^t. If $I^t \neq \emptyset$, pick an arbitrary element, $i^t \in I^t$, and set $N^{t+1} = N^t \backslash \{i^t\}$. Begin with $N^0 = N$ and prune N as far as possible. Denote the resulting set N^∞. Either $\#N^\infty = 1$ or $\#N^\infty \geq 3$ (if $\#N^\infty = 2$, the pruning is clearly not complete). If $\#N^\infty = 1$, then we are done and $\#R = 2(n-1)$. Because only 2–cycles have been deleted, R restricted to the pruned N^∞ is a connected route structure with no 2–cycles. Note that the number of routes removed is $2 \cdot \#(N \backslash N^\infty)$. Thus, to prove the theorem it is sufficient to show that the number of routes in R restricted to N^∞, $\#R_{|N^\infty}$, is less than or equal to $2(\#N^\infty - 1)$.

For $x \geq 3$, we say that a subset $\{i_1, \ldots, i_x\}$ of distinct elements of N^∞ is in an x–cycle if $i_1 R i_2, \ldots, i_{x-1} R i_x$, and $i_x R i_1$. Because $\tilde{\alpha} > 0$ and R is optimal, any x–cycle must not be reversible, that is, $\neg i_2 R i_1, \ldots, \neg i_x R i_{x-1}$, and $\neg i_1 R i_x$. (If any of these city–pairs were in R, their deletion would save a strictly positive amount and cost at most a finite sum of negligible numbers.) Now, N^∞ must be either an x–cycle with $x \geq 3$, or a set of x–cycles with $x \geq 3$ and some overlap. In either case, $\#R_{|N^\infty} \leq 2(\#N^\infty - 1)$. □

Note that once again the costs are at most linear in n. For any $\tilde{\beta}$ let $A(\tilde{\beta})$ be the set of $\tilde{\alpha}$ for which the solution is a cycle. Together, Theorem B and Lemma 1 imply that for any $\tilde{\beta} \simeq 0$, $A(\tilde{\beta})$ is a polygon with non–empty interior containing the ray from the origin determined by the vector $\tilde{1}$. The problem of finding the minimum cost cycle when $\tilde{\beta} = 0$ is the famous Traveling Salesman problem. We therefore suspect that the general version of Unrestricted(n) is at least as hard.

A cycle has an obvious drawback from the point of view of the passengers, their trips may involve as many as n–1 distinct flights. Our proxy for this cost is the restriction that passengers need take no more than m different flights in order to complete their travel plans. If $m = 1$, then the only feasible route structure is a system of direct flights. When $m = 2$, variable costs are low and the fixed costs approximately uniform, we are led to hub–spoke networks with a single hub. When the fixed costs are not uniform, m is unrestricted, but we consider only symmetric route structures, we are led to multi–hub systems.

5. Hub and Spoke Systems

Suppose that we are trying to connect 4 or more cities and that the fixed costs of connecting pairs of cities are approximately equal. If we require that the transportation system economize on fixed costs and not require more than one change of vehicle per trip, then Theorem C shows that hub–spoke systems with one hub are the solution. (At least 4 cities are needed for this result to avoid the cycle $1 \to 2 \to 3 \to 1$.) Example 1 and the subsequent discussion show that the assumptions of Theorem C cannot be substantially loosened. Theorem D shows how hub–spoke systems with multiple hubs may arise when fixed costs are minimized. In both cases, costs are linear in n.

<u>Definition</u>: A two–step hub–spoke system hubbing at $k \in N$ is a relation R such that iRk and kRi for all $i \neq k$, and $\neg$iRj if neither i nor j is equal to k.

Note that two–step hub–spoke systems are connected, indeed, $\bar{R}^2 = N \times N$. Further, two–step hub–spoke systems have $2(n-1)$ routes in them, so we are again led to the linearity of costs in n.

Theorem C: Under GAP, if $\#N \geq 4$, $\tilde{\beta} \simeq 0$, $\tilde{\alpha} \simeq \alpha \cdot \tilde{1}$, $\alpha > 0$, then the only solutions to Unrestricted(2) or to Symmetric(2) are two–step hub–spoke systems.

Proof: Because two–step hub–spoke systems are symmetric, it is sufficient to show that they are the only solutions to Unrestricted(2). As before, it is sufficient to minimize costs. Note that for $\#N = 3$, a cycle has strictly lower cost than a hub and spoke system, so the size of N matters as will be seen more directly below.

Because $\tilde{\beta} \simeq 0$, $\tilde{\alpha} \simeq \alpha \cdot \tilde{1}$, $\alpha > 0$, it is sufficient to show that two–step hub and spoke systems are the only solutions to the problem

$$(**) \qquad \min_R \#R \text{ s.t. } E \subset \bar{R}^2.$$

Two–step hub and spoke systems satisfy the constraint and have $2(n-1)$ routes in them. Hence, if R is a solution to (**) then $\#R \leq 2(n-1)$.

Let $\underline{m}$ denote $\min_{i \in N} \#R(i)$. Because R is connected, we know that $\underline{m} > 0$. If

$\underline{m} \geq 2$, then $\#R \geq 2n$, a contradiction. Hence $\underline{m} = 1$. Let $\bar{m} = \max_{i \in N} \#R(i)$. As an intermediate step in the proof, we will show that if $n = \#N \geq 4$, then $\bar{m} = n-1$.

Let i' be a point in N such that $\#R(i') = 1$. If $\bar{m} \leq n-3$, then it is impossible to get from i' to some point $j \in N$ in 2 or fewer steps, violating the constraint. Hence, either $\bar{m} = n-2$ or $\bar{m} = n-1$.

Suppose that $\bar{m} = n-2$ and let k be a point in N with $\#R(k) = n-2$. Because $\#R \leq 2(n-1)$ and $\underline{m} = 1$, there is at most 1 point $j \in N \backslash \{k\}$ such that $\#R(j) = 2$. Because N has at least 4 points in it, $\#R(i') = \#R(i'') = 1$ for some pair $i' \neq i''$ in N. But because $\bar{m} = n - 2$, from either i' or i'' (or both), it is impossible to reach some point $j \in N$. This violates the constraint, hence $\bar{m} = n-1$.

To finish the proof, let k be that point in N such that $\#R(k) = n - 1$ (there clearly cannot be two k with this property). Because $\bar{m} = n-1$, $\underline{m} = 1$, and $\#R \leq 2(n - 1)$, we know that for all $i \neq k$, $\#R(i) = 1$. If for some $i \neq k$, $R(i) \neq \{k\}$, then there is some $j \in N$ that cannot be reached from i in two steps, violating the constraint. Hence S is a two–step hub and spoke system hubbing at k. □

If the fixed costs of connecting cities are not approximately equal, that is, if $\tilde{\alpha}$ is a long way from $\alpha \cdot \tilde{1}$, then the two–step hub–spoke system may not be the solution with $m = 2$, even if we require that the solution be symmetric. There are four cities in the following example. Using any one of them as the hub would require some flights going over the middle, and this may be too costly.

Example 1 (The Bermuda Triangle): Let $\#N = 4$, $\tilde{\beta} \simeq 0$. Suppose that $\alpha_{ij} = \alpha_{ji}$, $\alpha_{12} = \alpha_{23} = \alpha_{34} = \alpha_{41} = 1$ while $\alpha_{13} = \alpha_{24} = x$. Let S denote a hub and spoke system hubbing at any point in N. The cost of S is $4 + 2x$ plus a finite sum of negligibles. Let R denote the symmetric route structure connecting all pairs except $\{1,3\}$ and $\{2,4\}$. The cost of R is 8 plus a finite sum of negligibles. It is easy to check that $\overline{R}^2 = E$. For $x > 2$, R is strictly better than S. (Draw a box and number the corners 1–4 clockwise. Put a Bermuda triangle in the middle. When x is high, no–one is willing to fly over the triangle.) □

For large x, the optimal route structure in Example 1 is symmetric and has $\#R = \frac{1}{2}\cdot 4^2 = \frac{1}{2}\cdot n^2$. This is a general phenomenon. There are cost structures, essentially no more than complicated versions of Example 1, for all even $n \geq 4$, with the property that when R is the solution to Unrestricted(2), then R is symmetric and $\#R = \frac{1}{2}\cdot n^2$. Thus, the restriction that passengers need make at most one flight change does not by itself imply that costs are linear in the number of cities to be connected. Indeed, this quadratic phenomenon appears for all <u>fixed</u> m as n becomes sufficiently large.

However, as we shall see, imposing symmetry alone without any restriction on the maximum number of flights a passenger need take does lead to optimal networks having linear costs. For this purpose it is useful to introduce some terminology from graph theory (see e.g. [13]).

<u>Definition</u>: A graph G on N is a pair (N,A) where N is a finite set and A is a collection of unordered pairs, arcs, of elements of N. A path from i to j in a graph G is a sequence of nodes in N, $(i,i_2), \ldots, (i_k,j)$ where each unordered pair is in A. G is connected if there is a path from each i to each j in N. A cycle is a path from i to i containing at least one arc, in which no node except i is repeated.[8] If G is connected and contains no cycles, it is called a tree.

<u>Definition</u>: For symmetric R, the graph associated with R is (N,A) where A is the set of unordered pairs (i,j) such that iRj. Note that $\#R = 2\cdot\#A$.

Theorem D: Under GAP, if $\tilde{\beta} \simeq 0$ and $\tilde{\alpha} > 0$, then the graph associated with any solution, R, to Symmetric(n) is a tree. In particular, $\#R = 2(n-1)$.

Proof: Let R be a solution to Symmetric(n) and let G be the graph associated with R. Because $\tilde{\alpha} > 0$, R has no x–cycles, $x \geq 3$. Because R is connected and has no x–cycles with $x \geq 3$, G is a tree (connected and contains no cycles). But this is

[8] Compare this with the definition of x–cycles in the proof of Theorem B.

equivalent to G being connected and having n–1 arcs ([13, Prop. 5.2, p. 27]). □

The two–step hub–spoke systems that arise in Theorem C are trees where only one city, the hub, is directly connected to many other cities. The following two–hub example indicates some of the range of tree structures that may arise as solutions under the conditions of Theorem D.

Example 2 (A Bicoastal Perspective): Suppose that $n = n_1 + n_2$, and that n_1 of the points in N are on the West coast and n_2 on the East coast. Pick a distinguished node on the West coast and call it L.A., pick a distinguished node on the East coast and call it N.Y. Consider the symmetric relation connecting all West (respectively East) coast points to L.A. (respectively N.Y.), as well as connecting L.A. and N.Y. This symmetric relation is connected, $\overline{R}^3 = N \times N$, and its graph is a tree so that $\#R = 2(n-1)$. □

In Example 2, the relation was a pair of two–step hub and spoke systems joined together by a two–step hub and spoke system joining the distinguished nodes L.A. and N.Y. This is but a special case of a more general construction. The restrictions $E \subset \overline{R}$ and $\#R = 2(n-1)$ are always satisfied by such route structures that are built out of smaller two–step hub and spoke systems: Begin with a distinguished point, i_0; Connect it to $n(i_0) \geq 1$ other points using a two–step hub and spoke system hubbing through i_0; For each point i_1 in the resulting system, connect it to $n(i_1) \geq 0$ other points using a two–step hub and spoke system hubbing through i_1. Such a process can be continued indefinitely, having $n(i_t)$ large gives a very "bushy" route structure with a small number of very busy hubs, having $n(i_t) = 1$ gives a "sprawling" route structure with many hubs.[9]

Note that these hub structures arise without any economies of concentration in the cost function — no economies to centralized maintenance and repair for example. Intuitively, it seems clear that adding this complication to the cost structure would tend to

---

[9] To see that Example 2 is of this form, let i_0 be L.A., let $n(i_0) = 1$ and call that one new city N.Y. Connect L.A. to N.Y. using a two–step hub and spoke system hubbing through L.A. Connect each city, L.A. and N.Y., in the resulting network, to the n_1 and n_2 cities on the West and East coasts respectively.

concentrate the route structures into a smaller number of busy hubs.

6. Natural Monopoly and Entry

In the presence of high fixed costs and low marginal costs on each line, optimal airline networks typically display a system–wide scale economy. A small scale economy is implicit in the statement of the problem, the declining average cost in each airline route. One might reasonably expect scale economies to be fully realized once each single route achieves efficient passenger loads on each flight. On the contrary, adding new cities to an existing network introduces new scale economies to the system. Thus, scale economies appear to be consistently present — a many–fold growth in a hub–spoke or related system can result in a more than proportionate increase in measures of welfare (e.g. consumer plus producer surplus).

The system–wide scale economy comes from the interaction of travel demand, route structure, and the route–level declining average cost. There are strong spillover effects between routes to a given hub. Each route generates traffic for the others since from each origin there are passengers for each destination through the hub. Hence the organizational structure of scheduling and pricing will be complicated by the interaction between spokes. In a hub–spoke system each spoke terminus generates travel demand from that terminus to each of the others in the system. Adding spokes increases traffic density and thereby reduces average cost on each of the other spokes. Hence the system is characterized by diminishing average cost over a wide range of activity levels. The scale economy is systemic and large, as was explicitly realized in the airline industry's corporate plans of the early 1980's.

An unlimited diminishing marginal cost structure describes a natural monopoly. A principal implication of the natural monopoly structure is the difficulty faced by start–up competitors. A new firm flying from a new location to a hub generates external benefit to the other line(s) serving the hub. It adds traffic at low marginal cost to each of the other routes. This external benefit can be internalized by a monopoly at the hub. Hence, and additional line to an incremental destination will typically be more profitable for a monopoly centered on the hub than for a new entrant. This excess profitability is true even if the firms have identical pricing.

We have noted that if $C(R,M)$ is the sum of the fixed costs of the lines, then under a variety of different assumptions, when costs are minimized and profits maximized, $\#R$ is linear in n. For illustrative purposes,[10] we will examine the questions, "When does such a system give a profit?" and, "When will it be profitable for an entrant?" for the case of two–step hub–spoke systems under conditions where they must arise, those of Theorem C, $\tilde{\beta} \simeq 0$, $\tilde{\alpha} \simeq \alpha \cdot \tilde{1}$, $\alpha > 0$.

As to the first question recall that the benefits are $B(R,M) = \Sigma_{i \neq j} b_{ij} \cdot M_{ij}$ and the (fixed) costs are $2\alpha \cdot (n-1)$. Supposing that $b_{ij} = b$, there are positive profits if $b \cdot \Sigma_{i \neq j} M_{ij} > 2\alpha \cdot (n-1)$. To get a sense of the implications of this solution, suppose that all the M_{ij}, $i \neq j$, are equal to some M. Then $b \cdot \Sigma_{i \neq j} M_{ij} = bM \cdot n(n-1)$ and $bM \cdot n(n-1) > 2\alpha \cdot (n-1)$ if and only if

$$nM > 2\alpha/b.$$

Now, it is profitable to run a single line between i and j if and only if $b \cdot M > 2\alpha$, that is, if and only if

$$M > 2\alpha/b.$$

Clearly, when set up costs are the dominant factor, it is much more profitable (per line) to run a network than it is to run single lines. Thus, a monopoly centered on the hub finds incremental routes profitable under much weaker conditions than will a start–up firm. It does so since the monopoly internalizes the benefits of traffic spillovers to other lines from the incremental route. We expect to find similar effects whenever fixed costs are

[10] See also [3], where this question is analyzed for an airline serving three cities.

non–negligible.

The analysis above suggests that monopoly is the predictable outcome of an unregulated route structure, inasmuch as a hub–spoke system is characterized by unlimited economies of scale (Theorems B, C, and D). But actual airline route networks do not completely fulfill the implications of this model. There is more variety (and less concentration on the hub) in practical settings than can be characterized fully by the simple model presented here. The essential point is the interaction of scale economies at the level of individual routes and the network structure generating large systemic scale economies. There are three principal discrepancies between the predictions of the model and actual experience: direct flights between non–hub cities, multiple airlines in service, and multiple hubs serving a single airline. These discrepancies all reflect the excessively strong predictions of a single powerful assumption in the analysis, strictly declining average cost per passenger on each flight. Declining average cost per passenger faces a limit in practice not modeled here: the capacity of aircraft economically available given the state of technology and the existing stock of aircraft. Between two high volume destinations, if passenger loads on direct flights can fill the capacity of the largest aircraft, then there may be no need for the flights to be routed through the hub. Similarly, the largest economic capacity of aircraft determines the maximum load on flights to and from a hub. If actual demand to or from most (non–hub) destinations substantially exceeds this capacity, then the scale economy is exhausted at a scale less than the market as a whole. If passenger flows to or from most non–hub destinations exceed several times the economic capacity of a single aircraft, then the scale economy can be fully exploited by each of several firms rather than rely on a monopolistic structure. There is then room in the market for several airlines and several hubs.

7. Conclusion

Airline route structure is considered here as an optimizing decision subject to cost and demand considerations. Route–level scale economies are shown to generate a large and pervasive scale economy in the transportation system under a variety of cost and demand configurations. These systems allow higher traffic densities, that is, more passengers per plane, than would be possible in a system of direct flights. This higher density is achieved by concentrating passengers to each destination on fewer aircraft. As the airlines' marginal cost of an extra passenger on a given plane is small, increasing traffic density reduces average cost per passenger–trip.

Monopoly will lead to efficient allocation with regard to the external economy across routes (subject to the usual concern about inefficient monopoly pricing). Monopoly internalizes the external effects. An alternative policy treatment is to allow many carriers to fly into the various hubs, and to coordinate the flights in order to achieve an efficient allocation in the presence of externalities. Coordination (and perhaps cross subsidies) may be required because of the positive externalities among routes.

References

[1] Bailey, E.E., D. R. Graham, and D.P. Kaplan (1985), *Deregulating the Airlines*, Cambridge, Mass.: MIT Press.

[2] Bental, B. and M. Spiegel (1990), "Consumption Externalities in Telecommunications Services," in A. De Fontenay, M.H. Shugard, and D.S. Sibley, eds., *Telecommunications Demand Modeling: An Integrated View*, New York: North Holland.

[3] Bittlingmayer, G. (1990), "Efficiency and Entry in a Simple Airline Network," *International Journal of Industrial Organization*, v. 8, pp. 245–257.

[4] Brueckner, J.K., N.J. Dyer, and P.T. Spiller, (1990) "Fare Determination in Airline Hub–and–Spoke Networks," photocopy, University of Illinois, Economics Department.

[5] Brueckner, J.K. and P.T. Spiller (1991a), "Competition and Mergers in Airline Networks," *International Journal of Industrial Organization*, v. 9.

[6] Brueckner, J.K. and P.T. Spiller (1991b), "Fares and Network 'Feed': Estimating Economies of Traffic Density in Airline Hub–and–Spoke Systems," photocopy, University of Illinois, Department of Economics.

[7] DeVany, A.S., and E. H. Garges, (1972) "A Forecast of Air Travel and Airport Use in 1980," *Transportation Research*, v. 6, pp. 1– 18.

[8] Granot, F. and A.F. Veinott, Jr. (1985), "Substitutes, Complements and Ripples in Network Flows," *Mathematics of Operations Research*, v. 10, n. 3, pp. 471–497.

[9] Haller, H. (1989a), "Topologies as Trade Infrastructures," Virginia Polytechnic Institute and State University, Department of Economics Working Paper, November.

[10] Haller, H. (1989b), "Large Random Graphs in Pseudo–Metric Spaces," Virginia Polytechnic Institute and State University, Department of Economics Working Paper, November.

[11] Hendricks, K., M. Piccione, and G. Tan (1992), "The Economics of Hubs: The Case of Monopoly," photocopy, University of British Columbia, Department of Economics.

[12] Horwitz, Robert (1989), *The Irony of Regulatory Reform.* New York: Oxford University Press.

[13] Lawler, Eugene (1976), *Combinatorial Optimization, Networks and Matroids.* New York: Holt, Rinehart, and Winston.

[14] Littlechild, S.C.,(1975) "Two Part Tariffs and Consumption Externalities," *Bell Journal of Economics*, v. 6, pp. 661–670.

[15] MacNeille, S., (1992) "Redrawing the International Air Map," *New York Times*, March 1, section 5, p. 23.

[16] McShan, S. and R. Windle, (1989) "The Implications of Hub–and–Spoke Routing for Airline Costs and Competitiveness," *Logistics and Transportation Review*, v. 25, n. 3, pp. 209–230.

[17] Rohlfs, J.H., (1974), "A Theory of Interdependent Demands for a Communications Service," *Bell Journal of Economics and Management Science*, v.5, n. 1, pp. 16–37.

[18] Spiller, P.T., (1989), "A Note on Pricing of Hub–and–Spoke Networks," *Economics Letters*, August, v. 30, n. 2, pp. 165–169.

[19] Stinchcombe, A.L., (1968), *Constructing Social Theories.* New York: Harcourt, Brace, and World.

[20] Uchitelle, L. (1991), "Off Course: What Happened to America's Airlines?" *New York Times Magazine*, September 1, pp. 12 ff.

[21] Vietor, R.H.K., (1990), "Contrived Competition: Airline Regulation and Deregulation, 1925–1988," *Business History Review*, v. 64, pp. 61–108.