

Large-Scale Transit Network Optimization by Minimizing User Cost and Transfers

Fang Zhao, Florida International University

Abstract

This article proposes a methodology for developing optimal transit networks (route structures and headways) that minimizes transit transfers and total user cost while maximizing service coverage, given information on transit demand, transit fleet size, and the street network of the transit service area. The research provides an effective mathematical computational tool with minimal reliance on heuristics. The methodology includes representation of transit route networks and solution search spaces, objective functions representing total user cost and unwillingness of users to make transfers, and a global search scheme based on simulated annealing. The methodology has been tested with published solutions to benchmark problems and has been applied to a large-scale realistic network optimization problem in Miami-Dade County, Florida.

Introduction

As congestion in large urban areas continues to worsen and gas prices began to rise in the recent years, the attractiveness of public transit as an alternative to private cars has also been growing. However, for a public transit system to help meet the growing travel demand and alleviate the congestion problem, it must be able to provide reasonable travel time and convenience relative to private vehicles. Travel time and convenience are affected directly by the configuration of a transit net-

work (TN) and service frequency, although other service and traffic characteristics and pedestrian environment will also have an impact on the willingness of the public to use transit. The quality of a TN may be evaluated in terms of a number of parameters including route directness, service coverage, operator cost, transit user cost (including waiting, in-vehicle, and transfer times), and the average number of transfers required to accomplish a trip. Route directness may be measured by the additional travel time incurred to a transit user when a bus does not follow the most direct route between the user's origin and destination. Service coverage refers to the percentage of total estimated demand (i.e., transit trips) that may be potentially satisfied by the transit services provided, based on a given transit route network. Operator cost is the cost to a transit property to provide transit services within a given network. Transfers are a result of not being able to provide direct services between all pairs of origins and destinations.

Transfers are known to discourage transit use. According to a survey conducted by Stern (1996) of various transit agencies in the United States, about 58 percent of the respondents believed that transit riders were willing to transfer only once per trip. Reducing transfers, therefore, has great potential in increasing the attractiveness of public transit and ridership. Transfers may be reduced by optimizing transit network configuration, or optimally laying out transit routes such that the services are as direct as possible and transfers are minimized. Improvements of TN configuration may also lead to lower transit operating cost and more services provided, which, in turn, help increase transit use.

In TN optimization, route network layouts and route headways are sought that minimize the overall cost of providing transit services, which is generally considered to have two components: user cost and operator cost. Unfortunately, TN design optimization processes that attempt to find global optimal solutions from a search space with reasonable completeness suffer from combinatorial intractability. Newell (1979) observes the difficulty in developing efficient TN optimization methods with traditional mathematical programming techniques and points out that TN design optimization "is generally a nonconvex (even concave) optimization problem for which no simple procedure exists short of direct comparisons of the various local minima." Furthermore, the resultant system for a TN problem is usually a NP-hard, mixed combinatorial optimization problem that is unlikely to be solved with traditional mathematical optimization techniques. The NP-hard problem (the hard problem in nondeterministic polynomial problem/algorithm class) refers to a problem for which the number of elementary numeri-

cal operations is not likely to be expressed or bounded by a function of polynomial form where the variable(s) of the function reflect(s) the size of the problem. The NP-hard intractability is due to the need to search for optimal solutions from a large search space made up by all possible solutions. A mixed problem refers to a problem that involves both continuous and discrete variables; a combinatorial problem usually refers to an integer optimization problem where the unknown variable set (called combinatorial set) consists of all feasible integer subsets of a larger base integer set. In TN optimization, the base set is the set of all street nodes that are suitable to serve as transit stops, and the combinatorial set consists of all street paths (subsets or integer vectors of the base street node set) in the street network that are suitable for transit vehicle operations. Even for a small street node set, the corresponding combinatorial set (i.e., the set that includes all possible paths) may be huge. Baaj and Mahmassani (1991) observe that large-scale TN optimization problems tend to suffer from several forms of difficulties with traditional mathematical approaches, such as nonlinearity, nonconvexity, multiobjectives, and combinatorial intractability due to the discrete nature of the problems. Similar observations are also made by Ceder and Wilson (1986), Charkroborty and Dwivedi (2002), and Zhao and Ubaka (2004), among others. These seem to be why the solutions to most TN optimization problems in practice are either relying on certain heuristic assumptions or are limited to relatively small or idealized networks. To date, the solutions to large-scale transit network problems that include both route network and headway as design components have been mostly limited to the use of various heuristic approaches where the solution search schemes are based on a collection of design guidelines, criteria established from past experiences, and cost and feasibility constraints.

In recent years, genetic algorithm (GA) has been applied to various TN optimization problems. GA is a stochastic algorithm based on natural evolution principle (i.e., genetic inheritance and the Darwinian strife for survival process). Mathematically, GAs may be categorized as weak solution search schemes that make few assumptions about problem domains and function properties, such as the smoothness, uniqueness, or compatibility of the objective functions, design parameters, and constraints. While this makes GAs attractive and popular for complex problems, it also causes GAs to suffer from combinatorial explosive solution costs due to huge solution search spaces often associated with large-scale problems. In the current TN literature, most GA applications are limited to small- or medium-sized network problems. Recently, Agrawal and Mathew (2004) applied a GA approach to a large-scale transit network. However, the travel demand (about 900 origin-desti-

nation pairs) was relatively small, and the search method required multiprocessor parallel processing due to intensive computation involved.

Table 1 summarizes the main features of some of the approaches reported in the literature. In the table, H&M indicates a combination of both mathematical programming methods and heuristic search schemes; MATH stands for mathematical optimization; H&M/AI means a combination of H&M and artificial intelligence techniques; and multiconstraints indicates use of multiple constraints such as maximum/minimum route length, maximum number of routes, minimum frequency, etc. Due to space limitations, the merits, solution strategies, and applicability to practical problems of the individual approaches are not discussed. Detailed information about various optimization approaches may be found in Fan and Machemehl (2004), which provides an extensive review and comparison of various optimization methods for TN design, and Zhao and Gan (2003), among others.

Table 1. Main Features of Some Approaches Used in Transit Network Design

<i>Year</i>	<i>Author</i>	<i>Optimization Objectives</i>	<i>Design Variables</i>	<i>Solution Approaches</i>	<i>Constraints</i>
1979	Dubois et al.	Gen. user cost	Route & frequency	H&M	Operating cost
1979	Mandl	Gen. time	Route	H&M	Coverage & directness
1986	Ceder & Wilson	Gen. user & operator cost	Route, frequency & vehicle scheduling	H&M	Multiconstraints & fleet size
1988	Van Nes et al.	No. of direct trips	Route & frequency	H&M	Operating cost & fleet size
1991	Baaj & Mahmassani	Multi-objects	Route & frequency	H&M/AI	Multiconstraints
1992	Bookbinder et al.	Disutility function-transfer inconvenience	Timetable/headway (offset time)	MATH	Heuristic guidelines
1994	Shih & Mahmassani	Multi-objects	Route & frequency	H&M	Multiconstraints
1998	Pattnaik et al.	Gen. user & operator time	Route network	GA	Frequency & load factor
2002	Chakraborty & Dwivedi	Transfer directness	Route network	GA	Coverage & user cost
2003	Chien et al.	Total operator & user cost	Route shape & headway	MATH	Route length, waiting time, load factors, etc.
2003	Ngamchai & Lovell	Total operator & user cost	Route network & headway	GA	Heuristic guidelines
2004	Agrawal and Mathew	Total operator & user cost	Route network & headway	GA	Multiconstraints
2004	Zhao & Ubaka	Transfer directness	Route network	MATH	Multiconstraints & directness

The development of the combined simulated annealing and fast descent (SAFD) method in this study has been motivated by the lack of optimization procedures that are capable of tackling large-scale TN problems and finding global optimal solutions in terms of both user and operator costs. Unlike other search algorithms such as various heuristic methods and genetic algorithms, which do not theoretically guarantee good performance to ensure a global optimum, simulated annealing is supported by a solid theory. Under fairly general conditions, it has been shown that a global optimal will be obtained with probability 1 (Hajek 1988). The simulated annealing search scheme used in this study is based on the integrated simulated annealing, tabu, and greedy search method developed by Zhao and Gan (2003), originally designed for finding optimal TN route layouts to minimize passenger transfers.

Solution Methodology

For simplicity, the following assumptions were made in this study:

1. The demand pattern, expressed in a transit origin-destination (OD) matrix, remains the same during the period of study.
2. Passengers' choices of routes are based on the shortest travel time. Terminal times are not included, although may be added easily.
3. Transit vehicles have the same seating capacity.
4. Passengers arrive at transit stops randomly (uniform distribution); therefore, the average waiting time to board a vehicle (t_{wait}) is one half of the headway (h), i.e.,

$$t_{wait} = h/2 \quad (1)$$

The following simple (yet widely used) relationships between TN parameters are employed:

$$L \equiv \frac{h \cdot q_{max}}{V_{Seat}} \leq L_{max} \quad (2)$$

$$h = \frac{2 \cdot R_L}{R_{Fleet}} \tag{3}$$

where:

- L is vehicle load factor
- q_{max} represents the critical link passenger flow of a given route
- V_{Seat} indicates vehicle seating capacity
- L_{max} signifies a user-defined maximum allowable load factor
- $2R_L$ is the round-trip in-vehicle travel time
- R_{Fleet} represents the route vehicle fleet size

More complex relationships between various TN parameters may be found in Ngamchai and Lovell (2003), Shih and Mahmassani (1994), Bookbinder and Désilets (1992), among others. The above simplifications do not prevent the proposed SAFD method from solving problems with complex TN parameter relationships such as nonlinear, nonconvex, or stochastic function relationships. Like genetic algorithms, the SAFD search method relies only on the evaluation of objective or constraint functions themselves. Therefore, difficult issues in traditional nonlinear search methods, such as function smoothness, convexity, uniqueness, etc., are not of concern. Theoretically, the proposed SAFD approach should be able to solve transit network optimization problems with dynamic demand iteratively as long as the transit demand (OD matrix) may be obtained after each transit route network and headway update. The challenge of solving dynamic demand problems is to have effective and reliable models to generate a new OD matrix after a new route layout is produced.

Representation of Transit Service Area, Route Network, Transit Demand, and Headways

A transit service area is represented by a set of street nodes, denoted as $N^{(n)} \{n_1, n_2, \dots, n_n\}$, that are connected to each other by a set of street segments, $A^{(m)} = \{a_1, a_2, \dots, a_m\}$. Together, these street node sets and street segment sets are referred to as the *street network* of the transit service area. A street segment a_i in $A^{(m)}$ may

be defined by its two end-street nodes n_{i1}, n_{i2} , *i.e.*, $\mathbf{a}_i = \mathbf{a}_i(n_{i1}, n_{i2})$ and $n_{i1}, n_{i2} \in N^{(n)}$, ($i = 1, 2, \dots, m$). A street segment length is measured by in-vehicle travel time between its two end nodes. A path or a route between any two street nodes is defined as a sequence of nodes, $\mathbf{p} = \mathbf{p}(p_1, p_2, \dots, p_k)$, and there is one street segment connecting two neighboring nodes in the path. In this study, only undirected networks are considered, but the methodology may be easily extended to directed networks. It is also assumed that the street network is connected, meaning that any two nodes in the street network are connected by at least one path. A TN \mathbf{T} consisting of l routes may be represented by a set of route/path vectors

$$\mathbf{T}^{(l)} = \mathbf{T}^{(l)}\{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_l\}, \mathbf{r}_j = \mathbf{r}(n_{j1}, n_{j2}, \dots, n_{js(j)}), (j = 1, 2, \dots, l) \quad (4)$$

where:

n_{jk} is the k -th node ($k = 1, 2, \dots, s(j)$)

$s(j)$ represents the number of transit stops on transit route \mathbf{r}_j

The above TN vector set $\mathbf{T}^{(l)}$ may also be expressed as a TN matrix

$$\mathbf{T} = \mathbf{T}[t_{ij}], t_{ij} = \begin{cases} 1, & \text{if node } j \text{ is on route } i, & i = 1, 2, \dots, l \\ 0, & \text{if node } j \text{ is not on route } i, & j = 1, 2, \dots, n \end{cases} \quad (5)$$

In this study, for the purpose of representation uniqueness, it is assumed that transit stops coincide with street nodes. Vehicle headways of a TN system may be expressed in a vector form

$$\mathbf{h} = (h_1, h_2, \dots, h_l)$$

where:

h_q ($q = 1, 2, \dots, l$) is the vehicle headway of route q

Using relationship (3), a headway vector may be derived from vehicle assignment vector

$$v = (v_1, v_2, \dots, v_l)$$

where:

v_q ($q = 1, 2, \dots, l$) is the number of vehicles operating on route q

Transit demand is given by an OD matrix

$$O = O [o_{ij}]$$

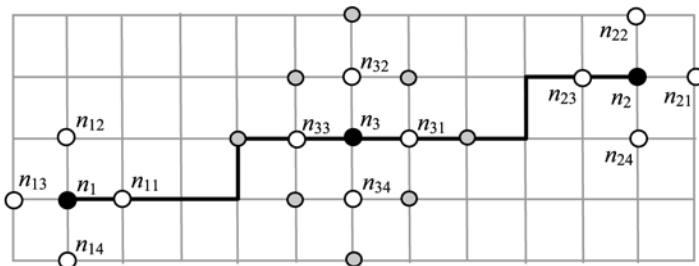
where:

o_{ij} is the number of trips originating from node i and destined for node j

Representation of Search Spaces for Transit Routes and Transit Network

The solution search spaces are locally and iteratively defined, and the size of a local search space may be flexible depending on available computing resources. A local path space is defined by a master path, a key-node representation of the master path, and a set of paths that are in the neighborhood of the master path. A master path is a path from which a local path space is generated. Key-nodes are a set of nodes on a master path based on which paths in the local path space are generated. Figure 1 illustrates a master path (solid line) and its three key-nodes n_1 , n_2 , and n_3 .

Figure 1. Three-Key-Node Representation of a Transit Route



Based on these key-nodes, local node spaces may be defined and a local path space is derived from the local node spaces. An i th order local node space of a master node is defined as the set of nodes that may be connected to the master node with i or fewer street segments. As an example, in Figure 1, the first order node space of the key-node n_1 is comprised of the master node n_1 itself (black circle) and its immediately adjacent nodes (n_{11} , n_{12} , n_{13} , and n_{14}). The second order node space of the key-node n_3 includes all the nodes that may be connected to the master node n_3 with two or fewer street segments (i.e., all the black, white, and gray circles around node n_3). Clearly, a local node space is a subspace of the street node space $N^{(n)}$. As the order i increases, a local node space will approach the original street node space $N^{(n)}$. Therefore, the order of a local node space provides a measurement of the degree of localization.

The procedure to generate a local path space from a master path has three steps:

1. Select s nodes from the node set of the master path $p = p(n_1, n_2, \dots, n_r)$ as the key-nodes.
2. Generate a sequence of s local node spaces of a given order from these s key-nodes.
3. Define the local path space as the set of paths consisting of piecewise shortest path segments that start from each node in the first local node space, sequentially pass through one node in each of the intermediate local node spaces, and end at each node in the last local node space.

The local network search spaces of a transit network $T^{(l)} = \{r_1, r_2, \dots, r_j\}$ is defined as the set of all local path spaces on the local node spaces (of a given order) of all the routes. In general, a route derived from a smaller number of key-nodes will result in better route directness and a smaller local path search space, but its flexibility is also limited. A route with a larger number of key-nodes is relatively more flexible to reach more neighboring nodes and therefore may cover more trips. However, it is also associated with a larger local path search space thus requiring more computing resources.

Simple Constraints for Transit Route Network

It is well known that for a discrete system, identifying and incorporating as many appropriate constraints as possible will significantly reduce the size of the search space. The following constraints are applied in this study:

- Maximum in-vehicle travel time (or route length) constraints for individual transit routes. It is known that a lengthy transit route not only results in difficulty in maintaining schedule, but also presents a safety hazard due to possible driver fatigue.
- Constraint on total transit vehicle fleet size. Since total vehicle fleet size of a transit network is closely related to operator cost, this may be considered an operator cost constraint. The optimization problem may be stated as finding a route network and route headways that result in the optimal service coverage and minimum user cost for a given fixed-operator budget (reflected by a given fleet size).
- Minimum and maximum headway constraints on individual routes. Route headways should be neither too small for operational reasons nor too large to result in long waiting times; the latter adversely affect ridership. Determination of headway is also constrained by vehicle load factor, which limits the number of passengers in a transit vehicle to ensure passenger comfort.

Route Directness Constraints

Route directness is defined as

$$d(\mathbf{r}) = \sum_{i=1}^{q-1} \sum_{j=i+1}^q w_{ij} (u_{ij}^{(S)} / u_{ij}^{(R)}) \tag{6}$$

where:

q is the number of nodes on route $\mathbf{r} = \mathbf{r}(n_1, n_2, \dots, n_q)$

$u_{ij}^{(R)}$ signifies the user travel cost between nodes n_i and n_j along the route

$u_{ij}^{(S)}$ is the user travel cost along the shortest path of the street network between these two nodes

w_{ij} is a weighting factor

For geometry-based route directness, $w_{ij} = 2/(q^2-1)$, and for ridership-based route directness

$$w_{ij} = (o_{ij} + o_{ji}) / \sum_{i=1}^{q-1} \sum_{j=i+1}^q (o_{ij} + o_{ji})$$

where:

o_{ij} and o_{ji} are coefficients of the demand matrix \mathbf{O}

In general, a larger route directness value implies better route directness but may result in higher transit operating cost, while a smaller route directness value may mean possible loss of ridership and, in turn, higher operating cost. A description of the physical meaning of route directness is found in Zhao and Gan (2003).

Network Directness Constraints

The meaning of network directness is similar to that of the route directness except that the directness measurement is based on the geometry or ridership characteristics of the entire network instead of individual routes.

Optimization Objective Functions

The objective function in this study is the total user travel cost, which is the summation of the user travel times of all the trips between the OD pairs in the transit service area

$$U(\mathbf{T}, \mathbf{O}, \mathbf{h}) = \sum_{i,j=1, i \neq j}^n [o_{ij} \cdot U_{ij}(\mathbf{T}, \mathbf{h})] \quad (7)$$

where:

o_{ij} is the number of trips originating from node i and destined for node j

\mathbf{h} represents a headway vector

$U_{ij}(\mathbf{T}, \mathbf{h})$ is the user travel time of one trip between nodes i and j in TN \mathbf{T}

For zero-, one-, and two-transfer trips, user travel time may be expressed as

$$u_{ij}^{(0)} = t_{wait}^{(r)} + t_{invh}^{(r)} \tag{8}$$

$$u_{ij}^{(1)} = t_{wait}^{(r1)} + t_{invh}^{(r1)} + t_{wait}^{(r2)} + t_{invh}^{(r2)} + t_{Tpenl}^{(r1r2)} \tag{9}$$

$$u_{ij}^{(2)} = t_{wait}^{(r1)} + t_{invh}^{(r1)} + t_{wait}^{(r2)} + t_{invh}^{(r2)} + t_{Tpenl}^{(r1r2)} + t_{wait}^{(r3)} + t_{invh}^{(r3)} + t_{Tpenl}^{(r2r3)} \tag{10}$$

where:

$u_{ij}^{(k)}$ ($k = 0, 1, 2$) represents the travel time of a k -transfer trip between a demand node pair i and j

$t_{wait}^{(q)}$ and $t_{invh}^{(q)}$ are, respectively, waiting time and in-vehicle travel time on the transit route q ($q = r, r1, r2, r3$)

In equations (9) and (10), $t_{Tpenl}^{(r1r2)}$ is the penalty for transfers between routes $r1$ and $r2$ expressed in equivalent in-vehicle travel time, while $t_{Tpenl}^{(r1r2)}$ in equation (10) is the penalty for transfers between routes $r2$ and $r3$. All the travel time components in expressions (8), (9), and (10) are functions of TN matrix T and/or route headway vector h . A detailed description of the above travel time components can be found in, for example, Shih and Mahmassani (1994). For uncovered trips, including those that require too many transfers and thus are unlikely to occur, the corresponding travel times are represented by a fictitious travel cost penalty value. The penalty cost associated with trips involving k transfers may be chosen from the following:

$$u_p^{(0)} = t_{wait}^{(max)} + t_{invh}^{(max)} \tag{11}$$

$$u_p^{(1)} = 2 t_{wait}^{(max)} + 2 t_{invh}^{(max)} + t_{Tpenl}^{(max)} \tag{12}$$

$$u_p^{(2)} = 3 t_{wait}^{(max)} + 3 t_{invh}^{(max)} + 2 t_{Tpenl}^{(max)} \tag{13}$$

$$u_p^{(3)} = 4 t_{wait}^{(max)} + 4 t_{invh}^{(max)} + 3 t_{Tpenl}^{(max)} \quad (14)$$

where:

$u_p^{(k)}$ ($k = 0, 1, 2, 3$) represents the maximum possible cost of a k -transfer trip
 $t_{wait}^{(max)}$, $t_{invh}^{(max)}$, and $t_{Tpenl}^{(max)}$ represent, respectively, the maximum possible waiting, in-vehicle travel, and transfer penalty times for all demand trips in the service area

Based on the above definitions and notations, the total user-cost objective function may be expressed as

$$U^{(k)}(\mathbf{T}, \mathbf{O}, \mathbf{h}) = \sum_{i,j=1, i \neq j}^n [o_{ij} \cdot U_{ij}^{(k)}(\mathbf{T}, \mathbf{h})] \quad (15)$$

where:

$U^{(k)}$ ($k = 0, 1, 2$) is the total user-cost function based on k -or-less transfer trips, and the corresponding travel time is defined as

$$U_{ij}^{(0)}(\mathbf{T}, \mathbf{h}) = \begin{cases} u_{ij}^{(0)} & \text{if demand } o_{ij} \text{ is satisfied with a zero-transfer trip} \\ u_p^{(q)} & \text{otherwise } (q = 0, 1, 2, \text{ or } 3) \end{cases} \quad (16)$$

$$U_{ij}^{(1)}(\mathbf{T}, \mathbf{h}) = \begin{cases} u_{ij}^{(0)} & \text{if demand } o_{ij} \text{ is satisfied with a zero-transfer trip} \\ u_{ij}^{(1)} & \text{if demand } o_{ij} \text{ is satisfied with a one-transfer trip} \\ u_p^{(q)} & \text{otherwise } (q = 0, 1, 2, \text{ or } 3) \end{cases} \quad (17)$$

$$U_{ij}^{(2)}(\mathbf{T}, \mathbf{h}) = \begin{cases} u_{ij}^{(0)} & \text{if demand } o_{ij} \text{ is satisfied with a zero-transfer trip} \\ u_{ij}^{(1)} & \text{if demand } o_{ij} \text{ is satisfied with a one-transfer trip} \\ u_{ij}^{(2)} & \text{if demand } o_{ij} \text{ is satisfied with a two-transfer trip} \\ u_p^{(q)} & \text{otherwise } (q = 0, 1, 2, \text{ or } 3) \end{cases} \quad (18)$$

It may be seen from the structure of the total user-cost function $U^{(k)}$ defined in (15) through (18) that minimization of this function has two effects: minimizing the total user cost based on k -or-less transfer trip coverage, and maximizing k -or-less transfer trip coverage. The penalty term $u_p^{(q)}$ regulates the balance between these two effects. In general, the larger the penalty value, the greater relative importance is given to service coverage. Based on the above descriptions, a TN design optimization problem may be stated as follows:

Minimize:

$$U^{(k)}(\mathbf{T}, \mathbf{O}, \mathbf{h}) = \sum_{i,j=1, i \neq j}^n [o_{ij} \cdot U_{ij}^{(k)}(\mathbf{T}, \mathbf{h})] \tag{19}$$

Subject to:

- Route length constraints: $R_{\min}^{(q)} \leq R_L^{(q)} \leq R_{\max}^{(q)}, (q = 1, 2, \dots, l)$ (20)

- Route headway constraints: $h_{\min}^{(q)} \leq h_q \leq h_{\max}^{(q)}, (q = 1, 2, \dots, l)$ (21)

- Route load factor constraints: $\frac{h_q \cdot q_{\max}^{(q)}}{V_{Seat}^{(q)}} \leq L_{\max}^{(q)}, (q = 1, 2, \dots, l)$ (22)

- Route directness constraints: $d_q \geq d_{\min}^{(q)}, (q = 1, 2, \dots, l)$ (23)

- Total TN fleet size constraints: $\sum_{q=1}^l R_{Fleet}^{(q)} \leq N_{Fleet}^{(T)}$ (24)

where:

$R_{\min}^{(q)}$ and $R_{\max}^{(q)}$ represent the minimum and maximum route length constraints for route q

$h_{\min}^{(q)}$ and $h_{\max}^{(q)}$ are the minimum and maximum route headway constraints for route q

$d_{\min}^{(q)}$ is the minimum route directness constraint for route q

$N_{Fleet}^{(T)}$ is the system fleet size

Simulated Annealing Algorithm Search for Optimal Transit Network

The simulated annealing (SA) search scheme is a stochastic process designed to avoid being trapped into poor local optima. Under fairly general conditions and for large problems, it may always be expected to find a global solution faster than a random search method. In an SA search, a solution and its associated local space will replace the incumbents with probability 1 if it has a better goal value or with some probability between 0 and 1 otherwise. The probability to accept a worse solution is proportional to the difference between its goal value and the current best goal value. A slightly worse solution has a higher probability of being accepted than a much worse solution. In the long run, as the number of search iterations becomes sufficiently large, the search process may escape from any local optimum, and eventually should visit a global optimal solution (Hajek 1988). For a TN design optimization problem, the simulated annealing procedure involves the selection of a solution candidate TN T from a local network space based on a given initial TN T_0 . The network T is accepted if the associated objective function $U(T, \mathbf{O}, \mathbf{h}) < U(T_0, \mathbf{O}, \mathbf{h})$. Otherwise, T is accepted with a probability

$$p = \exp(-\Delta/t), \Delta = [U(T, \mathbf{O}, \mathbf{h}) - U(T_0, \mathbf{O}, \mathbf{h})] \quad (25)$$

where:

t ($t > 0$) is the temperature of the annealing process

U is the total user-cost function defined in (15)

The term “temperature” is borrowed from annealing in solids, and has no physical meanings. The value of a “temperature” t in equation (25) merely reflects the fact that the likelihood of accepting solutions with worse goal values is regulated by t . A larger t results in a larger probability of accepting worse solutions, while a smaller t reduces such chances. In practice, a large initial t is often chosen to increase the chance of escaping from a poor local minimum, which is gradually reduced in the search process by a factor τ to enhance the selectivity of the search toward improved solutions.

One difficulty with the SA search method described above is that the search process may repeat the same solutions or a sequence of solutions many times before moving to other search regions. To alleviate this problem, a tabu list is established to keep track of solutions evaluated recently to prevent them from entering the solution search process again. A detailed description of the simulated annealing search process may be found in Zhao and Gan (2003).

Fast Descent Method Search for Optimal Transit Headways

In the SA search process, the TN headway vector h in equation (25) remains a passive network parameter (i.e., h does not play an active role in the SA search process). The TN headway vector h will be modified only if its vector components violate their associated constraints, such as (21), (22), etc. This is due to the fact that for large-scale TN problems, simultaneous search for both optimal route networks and headways may be computationally intractable due to the huge solution population size to expose any meaningful characteristics (e.g., promising search directions, local minima, etc.). The search for a better headway h is performed in a separate process called the fast descent (FD) search. The basic idea behind this method is to find the vector components (or directions) of a cost function such that appropriate adjustments of these components will lead to the fast descent of the cost function. The procedure for applying a FD search method to search for optimal route headways is outlined as follows:

1. If the maximum fleet size constraint is not violated, find the route (the FD component) in the network for which a decrease in its headway (as a result of increasing the number of vehicles operating on this route) will result in the FD of the total user-cost function. Update existing route headway with the new headway. Repeat this step until the maximum fleet size constraint is violated. Go to step 2.
2. Find two routes in the route network for which decreasing the headway for one route (by increasing the number of vehicles on the route) and increasing the headway for the other route (by reducing an equal number of vehicles on this route) will result in the FD of the total user-cost function. Update existing route headways with the new headways. Repeat this step until a (local) minimal cost value is reached. Start a new round of SA search iteration for a better transit route network layout.

In this study, the SA and FD search processes are integrated together in an iterative manner (i.e., an FD search process for better route headways will be executed after one or more SA search iterations for better route network layouts).

Numerical Experiments

The first experiment was based on a real network in Switzerland (Mandl 1979). This problem was also used by Shih and Mahmassani (1994) and Baaj and Mahmassani (1991) as a benchmark to test their approaches to TN design optimization. Mandl's problem consisted of a street network of 15 nodes with a total demand of 15,570 trips per day. The length of a street segment was defined in terms of in-vehicle travel time in minutes. The maximum route length was constrained to 40 minutes. In Table 2, the first row identifies the source of the solutions to the benchmark problem, which include Mandl's problem as well as its variations constructed by Baaj and Mahmassani (1991) and Shih and Mahmassani (1994). The second row identifies solutions to the benchmark problem. The problems/solutions differ in their number of routes, total fleet size, and/or the search method used. Mandl (1979), Baaj and Mahmassani (1991), and Shih and Mahmassani (1994) all used a transfer penalty of 5 minutes, which was also assumed in the proposed SAFD method. The methods used to obtain the results are indicated in the third row. For each solution, the unshaded column provides the statistics for the layouts produced in the original studies, and the shaded column gives the statistics for the results produced from the SAFD method developed in this study. All the SAFD results in this table were generated with the total user-cost objective function $U^{(1)}$ defined in equation (15). The operator cost is reflected by the TN fleet sizes shown in the 10th row. It may be seen that the percentages of zero-transfer trips are higher for all solutions produced from this study. Except for Mandl's original results, all solutions provided 100 percent trip coverage with trips requiring zero or one (one-or-less) transfer. The savings in total user travel time in hours from the SAFD method are shown in the 8th row.

The second experiment involved a large-scale TN optimization problem based on the service area of the Miami-Dade Transit Agency (MDTA), which encompasses a region of about 300 square miles with a population of about 2.3 million. MDTA ranks as the 16th largest transit agency in the United States. At the time of this research, MDTA operated 83 transit routes, including a rail rapid transit system of 22.5 route miles (Metrorail), a 4.5-mile downtown automated circulation system (Metromover), and 81 bus routes with about 4,500 transit stops. The street net-

Table 2. Comparison of Results from Different Methods

Problem Source	Mandl ¹		Baaj & Mahmassani						Shih & Mahmassani			
	1		1		2		3		1		2	
Search method	Mandl	SAFD ²	B&M ³	SAFD	B&M	SAFD	B&M	SAFD	S&M ⁴	SAFD	S&M	SAFD
0-transfer trips (%)	69.94	95.31	78.61	95.18	79.96	95.44	80.99	92.49	82.59	94.03	87.73	95.12
1-transfer trips (%)	29.93	4.69	21.39	4.82	20.04	4.56	19.01	7.51	17.41	5.97	12.27	4.88
2-transfer trips (%)	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Total user travel cost (min)	219094	185158	205646	190998	209318	195466	217954	190478	203936	189460	204028	196956
Savings in total user travel cost (hr)	-	565.60	-	244.13	-	230.87	-	457.93	-	241.27	-	117.87
Travel directness	1.335	1.128	1.253	1.115	1.276	1.141	1.328	1.144	1.243	1.092	1.243	1.125
Total fleet sizes	99	99	89	89	77	77	82	82	84	84	68	68
Number of routes	4	4	6	6	8	8	7	7	6	6	8	8
Average transfers	1.302	1.047	1.214	1.048	1.200	1.046	1.190	1.075	1.174	1.060	1.123	1.049
Network directness	0.811	0.970	0.912	0.963	0.959	0.954	0.856	0.908	0.835	0.979	0.940	0.979

¹Mandl's method

²Simulated annealing and FD method

³Baaj and Mahmassani's method

⁴Shih and Mahmassani's method

Table 3. Comparison of SAFD Search Results with the Existing Network

Network Parameters	Existing Network		User-Cost Objective Function $U^{(0)}$				User-Cost Objective Function $U^{(1)}$			
	$h_{max} = 20$		$h_{max} = 20$		$h_{max} = 30$		$h_{max} = 20$		$h_{max} = 30$	
Transfer penalty time (min)	5	10	5	10	5	10	5	10	5	10
Zero-transfer trips (%)	14.38	14.38	32.37	32.37	31.99	31.99	32.34	31.73	31.08	32.39
One-or-less transfer trips (%)	55.17	55.17	82.35	82.35	82.48	82.48	86.41	86.24	85.89	86.35
Two-or-less transfer trips (%)	65.23	65.23	86.48	86.48	86.60	86.60	89.84	90.03	89.97	90.92
Total covered trips (%)	65.69	65.69	86.51	86.51	86.64	86.64	89.88	90.10	90.00	90.97
Average user cost (minutes)	42.8	47.1	39.7	42.5	39.8	42.6	37.4	40.1	37.8	39.9
Total fleet sizes	600	600	600	600	600	600	600	600	600	600
Average transfers (two-or-less)	1.934	1.934	1.673	1.673	1.678	1.678	1.678	1.690	1.700	1.694
CPU time (hours)	-	-	0.3270	0.266	0.264	0.223	5.178	7.182	4.806	5.477

work used in this experiment consisted of 4,300 street segments and 2,804 street nodes, and the longest bus routes were about 32 miles. Total length of the transit system was about 1,300 route miles, not including some small loops at the ends of some routes or in shopping centers.

The OD matrix was generated from the 1999 validated Miami-Dade travel demand model, which provided the daily number of passenger trips between each pair of traffic analysis zone centroids. These trips were manually distributed to the surrounding street network nodes with considerations given to land-use patterns, proximity, and street network connectivity. Total demand was 161,944 daily transit trips. They were distributed, sparsely and unevenly, between about 120,000 demand (OD) pairs.

Operator cost is reflected by the network fleet size of 600 transit vehicles, which is about the same as that operated by MDTA. The total number of transit lines in the example remains the same as the existing system, and Metrorail and Metromover alignments are fixed in the optimization process. Other constraints and data used in this example include: for bus routes, the minimum and maximum headways are $h_{\min} = 4$ min and $h_{\max} = 20$ min; for Metrorail, $h_{\min} = 6$ min and $h_{\max} = 10$ min; and for Metromover, $h_{\min} = 1$ min and $h_{\max} = 6$ min. The minimum and maximum route lengths (in-vehicle time) are $R_{\min} = 10$ min and $R_{\max} = 90$ min; average in-vehicle travel speeds are 21 mph for bus and 31 mph for Metrorail. Since no data were available regarding the appropriate value of transfer penalty time in Miami-Dade County, results obtained from two sets of transfer penalty values are presented. One is $t_{Tpenl} = 5$ minutes, which is the same as that used in the first example. The other is $t_{Tpenl} = 10$ minutes. Ideally, the value of transfer penalty value should be determined through transit user surveys since transfer penalty reflects transit users' tolerance to experiences of unpleasantness or inconvenience during vehicle transfers. Penalty values are likely to vary across different geographic areas and change with demographics, socioeconomics, topography, climate, quality of transfer facilities, etc.

All the numerical results were obtained on a personal computer with a 2.8GHz CPU and 1GB RAM memory. Table 3 presents the results from the proposed SAFD method. There are two sets of results, one based on the zero-transfer total user-cost function $U^{(0)}$ and the other based on the one-or-less transfer total user-cost function $U^{(1)}$. For references, TN parameters for the existing network are also included. The first row in the table identifies the objective functions used to generate the corresponding numerical results. The second row provides the maximum

route headway constraints for different test cases, while the third row indicates the transfer penalty values for various test cases. For various objective functions, headway constraint, and transfer penalty combinations, statistics for the TNs obtained from the optimization process are provided. Results in Table 3 show that, user cost appears to be more sensitive to the value of transfer penalty time than to other TN parameters. This implies that an accurate transfer penalty value is needed to obtain a good estimate of user cost and improvement in transfer facilities and that transfer quality will help reduce the transfer penalty.

From Table 3, it may be seen that results obtained from this study have significant improvement over the existing one. The zero-transfer trips increased from 14.38 percent based on the existing network to 32.37 percent with objective function $U^{(0)}$ and $h_{\max} = 20$, an improvement of about 125 percent. The one-or-less transfer trip coverage increased from 55.17 percent to 86.41 percent with objective function $U^{(1)}$, $h_{\max} = 30$, and $t_{\text{pen}} = 5\text{min}$, an improvement of about 57 percent. Assuming most transit riders are only willing to transfer once per trip, the one-or-less trip coverage shown in the fifth row would be the actual total service coverage of the corresponding TN. The remaining trips either require two or more transfers or are not satisfied. The second from the last row in Table 3 presents the average transfers per trip for the two-or-less transfer trips that involve the use of Metrorail and Metromover. The high level of service of the rail lines is more likely to encourage people to use transit even if the trips may require two transfers. The eighth row in Table 3 shows the per user cost based on two-or-less transfer trips. As expected, larger transfer penalty time values result in higher user costs.

The modeled network is not a perfect description of the actual network even though care has been taken to prepare the input data as accurately and completely as possible. Consequently, some of the differences in the statistics from the existing network and the results generated from the SAFD may be attributed to the inaccuracy in the modeled network. However, the main purpose of the second example is not to show the superiority of the SAFD results over the existing network, but rather to demonstrate that for a large-scale transit network with a given transit demand pattern as well as a constraint set, the proposed method is able to improve the initial network configuration in a reasonable amount of time.

Conclusions

A mathematical stochastic method for large-scale TN optimization has been described. A stochastic local search method based on simulated annealing and fast descent search has been developed and has been shown to be capable of tackling large-scale transit network design optimization problems and producing results in a reasonable amount of time. The solution methodology is generally applicable to a wide range of practical TN problems, does not favor any particular transit network configurations, and gives reasonably good results in a reasonable amount of time. The methodology also allows results to improve and approach the global optimum as the computer resource or power increases.

Future improvements to the solution method may include, for example, the development of time-dependent TN optimization methods to optimize a TN by taking into consideration changes in network conditions and OD distribution during different time periods of a day, inclusion of terminal access times in the user costs, and removal of simplifying assumptions regarding transit fare and fixed demand.

Acknowledgements

The author would like to thank the reviewers for their careful reading of this paper and for their helpful comments and suggestions.

References

- Agrawal, J., and T. V. Mathew. 2004. Transit route network design using parallel genetic algorithm. *Journal of Computing in Civil Engineering ASCE* 18: 248–256.
- Baaj, M. H., and H. S. Mahmassani. 1991. An AI-based approach for transit route system planning and design. *Journal of Advanced Transportation* 25, 2: 187–210.
- Bookbinder, H. J., and A. Désilets. 1992. Transfer optimization in a transit network. *Transportation Science* 26, 2: 106–118.
- Ceder, A., and N.H.M. Wilson. 1986. Bus network design. *Transportation Research—Part (20B)*4: 331–344.
- Charkroborty, P., and T. Dwivedi. 2002. Optimal route network design for transit systems using genetic algorithms. *Engineering Optimization* 34, 1: 83–100.

- Chien, S., B. V. Dimitrijevic, and L. N. Spasovic. 2003. Optimization of bus route planning in urban commuter networks. *Journal of Public Transportation* 6, 1: 53–80.
- Fan, W., and R. B. Machemehl. 2004. Optimal transit route network design problem: algorithms, implementations, and numerical results, Report No. SWUTC/04/167244-1, Austin, TX: Center for Transportation Research, University of Texas at Austin.
- Hajek, B. 1988. Cooling schedules for optimal annealing. *Math. of Operations Research* 13: 311–329.
- Mandl, C. E. 1979. Evaluation and optimization of urban public transportation networks. *Third European Congress on Operational Research*, Amsterdam, Netherlands.
- Newell, G. F. 1979. Some issues related to the optimal design of bus routes, *Transportation Science* 13: 20–35.
- Ngamchai, S., and D. J. Lovell. 2003. Optimal time transfer in bus transit route network design using a genetic algorithm. *Journal of Transportation Engineering* 129, 5: 510–521.
- Shih, M. C., and H. S. Mahmassani. 1994. *A design methodology for bus transit networks with coordinated operations*. SWUTC/94/60016-1. Austin, TX: Center for Transportation, University of Texas at Austin.
- Stern, R. 1996. Passenger transfer system review. *Synthesis of Transit Practice* 19, Transportation Research Board.
- Van Nes, R. R. Hamerslag, and B.H. Immers. 1988. Design of public transport networks. *Transportation Research Record* 1202, 74–83.
- Zhao, F., and A. Gan. 2003. *Optimization of transit network to minimize transfers: methodologies for route network optimization*. Technical Report for the Florida Department of Transportation, Department of Civil and Environmental Engineering, Florida International University, Miami, FL.
- Zhao, F., and I. Ubaka. 2004. Transit network optimization—minimizing transfers and optimizing route directness. *Journal of Public Transportation* 7, 1: 67–82.

About the Author

FANG ZHAO (zhaof@fiu.edu) received her Ph.D. in civil engineering from Carnegie Mellon University in 1991. Since 1992 she has been teaching in the Civil and Environmental Engineering Department at Florida International University. She is currently an associate professor and deputy director of the Lehman Center for Transportation Research. Her main research interests include public transportation, travel demand modeling, geographic information systems, and transportation data modeling. Dr. Zhao is a registered professional engineer in the State of Florida and serves on both the TRB Transit Planning and Development Committee and the New Technologies and Systems Committee.