Public Transp (2009) 1: 233–251
DOI 10.1007/s12469-009-0015-8

ORIGINAL PAPER

Congested multimodal transit network design

Quentin K. Wan · Hong K. Lo

Published online: 5 November 2009

Abstract The planning of transit services is vital to transit-oriented metropolises. It is a complex, multi-objective decision process, especially for services operated by the private sector. Traveler's desire for direct, affordable, and quality services often conflicts with the profit-making objective of private operators. In a multi-modal network, partly collaborative and partly competitive interactions among transit modes further complicate the problem. To simplify the planning problem, existing studies generally consider transit network design from the perspective of a single mode while neglecting the modal interactions. The lack of a comprehensive approach across transit modes may result in an unbalanced supply of transit services, weakening the financial viability of the services and, more importantly, adding unnecessarily to congestion, especially in already congested districts. This study explicitly considers these interactions in a multi-modal network framework. We develop a systematic phase-wise methodology for multi-modal network design, considering both the effect of congestion and integration of modal transfers. Inter-route and inter-modal transfers are modeled through the State Augmented Multi-modal (SAM) network approach developed in earlier studies. An illustrative example is included to demonstrate the design procedure and its salient features.

1 Introduction

Transit systems play an important role in the planning and development of transit-oriented metropolises, as in many Asian and European cities. Transit network design

Q.K. Wan
Transport Research Institute, Napier University, Edinburgh, UK

H.K. Lo (✉)
Department of Civil and Environmental Engineering, Hong Kong University of Science and Technology, Hong Kong, China
e-mail: cehklo@ust.hk

(TND) identifies a set of optimal routes to serve the travel demands between specific origin–destination (OD) pairs, subject to a set of feasibility conditions. These feasibility conditions cater to the specific requirements of the local jurisdiction, such as covering geographic constraints and satisfying local regulations. TND is a complex, multi-objective decision process, especially for services operated by the private sector. TND must attend to the perspectives of users, operators, as well as the overall system performance.

Travelers' desire for quality and direct but inexpensive services often conflicts with the profit-making objective of private operators. In particular, indirect services, wherein transfers are necessary, could lower operating costs; yet too many transfers would make the services unattractive from the users' perspective. On the other hand, too many direct routes could result in low passenger loads and service redundancy that adds to road congestion, especially in downtown areas. In planning a transit network, therefore, it is important to fully consider the interactions between all transport modes so as to ensure that the TND is sustainable under competition and regulation.

Previous studies mainly considered networks of a single transit mode, not capturing the interactions between co-existing modes. Existing TND approaches focus primarily on maximizing the coverage to potential demand or on minimizing travelers' generalized cost. For example, Dufourd et al. (1996) proposed heuristics to align a rapid transit line under a population coverage constraint. Baaj and Mahmassani (1995) developed a transit route generation heuristic that accounts for user and operator costs. On the other hand, Ceder and Wilson (1986), and Pattnaik et al. (1998) studied the problem of bus network design to minimize generalized costs. Michaelis and Schöbel (2009) describe a recent approach where they first design vehicle routes, then split them to lines and finally calculate a (periodic) timetable. While the objective in all three steps is customer-oriented, costs can be controlled during the whole process.

In the reality of a multi-modal network, the simplification of studying a single mode often leads to either an over-design (as the contribution from the other modes is ignored) or that the characteristics of multi-modal demands are not captured. In a multi-modal transit market, the patronage on a specific mode and route depends heavily on the competition and/or cooperation of the co-existing transit services. Moreover, alternative modes like auto or taxi should be explicitly captured in the choice set. Our previous studies on competitive transit services found that private operators would adopt different competitive or cooperative strategies to increase their profits, which would result in drastic changes in passenger demands on their services (e.g., Lo et al. 2003b). This result illustrates the significance of considering the interactions among different modes for TND. Bruno et al. (1998) pioneered the design of a multi-modal network by studying a hybrid pedestrian transit system. In that study, however, the model is limited in its ability to work with multi-modal, concurrent public transport lines.

In Lo et al. (2003a), we develop an approach, referred to as the State-Augmented Multi-modal (SAM) network, by combining discrete choice modeling with network equilibrium methods to capture travelers' combined-mode choices in a network of multi-modal services. The salient features of the SAM approach include the considerations of: (i) passengers' transfer behavior and preference for the combined-mode connections; (ii) transfer congestion and penalty; (iii) non-linear fare structure and inter-modal transfer fare discount; (iv) onboard passenger volume of the arriving transit vehicle prior to a stop or terminal (that affects passengers' ability and comfort to get onboard the arriving vehicle); (v) in-vehicle crowdedness or discomfort function; and finally (vi) road congestion for the transit modes that travel on the shared roadway network. In other words, the SAM network considers a range of factors that are important for modeling travelers' combined-mode choices.

In this paper, drawing upon the SAM network approach, we develop a formulation for TND that explicitly considers travelers' multimodal choices and preferences. In addition, to cater to the privately operated transit market, the formulation considers the financial viability of the services to be offered. This issue has received little attention in previous TND studies. Moreover, as evident in Hong Kong, in order to compete, private operators often increase their bus frequencies along major (and profitable) corridors, thus adding significantly to road congestion. Incorporating congestion consideration in TND, therefore, is essential to maintain the overall mobility of a region, especially for already congested corridors.

Summarizing, this paper develops a modeling framework for TND that captures travelers' multimodal choices and transfers, financial viability of the services, and overall system congestion. As is typical, network design problems involve integer formulations, making them difficult to solve. This study develops a two-phase heuristic to solve the formulation. Some numerical examples are provided to illustrate the properties of the formulation and the quality of the heuristic solutions.

2 Model formulation

The congested multimodal transit network design (CMTND) problem can be formulated as a bi-level mathematical program (e.g., Bard 1998; Colson et al. 2005). The upper-level optimizes the network configuration and service frequency of financially viable services so as to maximize the social welfare of the system; whereas the lower level models travelers' combined mode, route, and transfer choices in response to the prevailing design. In the previous section, we discussed briefly the important factors to be captured in a multi-modal network model. To encapsulate these factors, the SAM approach provides an apt alternative to the conventional way of modeling the physical network. On the other hand, the attributes of the physical network can be recovered from its SAM counterpart. In the model development depicted herein, we formulate the model in terms of the SAM notations. Nevertheless, for ease of comprehension, we provide notation cross-references between the physical and SAM networks below. One can refer to Lo et al. (2003a, 2004) for details of the SAM modeling approach.

Consider a SAM network $G = (\mathbf{N}, \mathbf{A})$ with node set $\mathbf{N}$ and link set $\mathbf{A}$. Each SAM node $n \in \mathbf{N}$ contains information on both the location of a point in the spatial domain and the transport modes associated with a trip arriving at or departing from that location. The link set $\mathbf{A}$ is a collection of two types of links, denoted as direct in-vehicle link a_d and transfer link a_t, i.e. $\mathbf{A} = \mathbf{A}_\mathrm{d} \cup \mathbf{A}_\mathrm{t}$. Direct in-vehicle links are connections between boarding and alighting stops on the same mode without any transfers, and the flow on each direct in-vehicle link represents the passenger volume between the associated stops of the specific transport mode. One thing to note is that direct links are abstract node to node connections on the SAM network. They only specify the connectivity between the concerned nodes (i.e., 1 when there is; 0 otherwise) but do not specify the exact routing between the connected nodes on the actual network. Transfer links, on the other hand, are connections between transport modes. Embedded implicitly in the construction of the transfer link set $\mathbf{A}_\mathrm{t}$, are transfer rules that capture specific types of transfers deemed as probable or those that would actually be used by travelers. An example of the probable transfer state diagram can be found in Lo et al. (2003a). In the CMTND problem, we denote $\mathbf{x}$ as the design scheme, representing a set of transit lines and their associated service frequencies. For example, the transit line configuration can be expressed as a set of binary variables, whose value is equal to 1 when a link is served by a transit line, or zero otherwise. Similarly, on the SAM network, the transit line configuration can be expressed in terms of binary variables with respect to direct in-vehicle links. As the design scheme varies, both the network structure and its associated traffic performance change. Therefore, we can express the path (passenger) flow vector $\mathbf{h}(\mathbf{x})$ and the corresponding vector of path utility $\boldsymbol{\xi}(\mathbf{x})$ as functions of the design scheme $\mathbf{x}$. As is typical, the utility of a path is a function of its total in-vehicle travel time, total waiting time, transfer penalty, and fare.

Mathematically, the upper level program of the CMTND problem is formulated as:

[CMTND]

$$\max_{\mathbf{x}} \ W(\mathbf{h}(\mathbf{x}), \boldsymbol{\xi}(\mathbf{x})) \tag{1}$$

$$\text{subject to} \quad \Pi(\mathbf{x}, \mathbf{h}(\mathbf{x})) \geq 0, \tag{2}$$

$$\mathbf{x} \in \mathbf{X} \tag{3}$$

where $\mathbf{h}(\mathbf{x})$ and $\boldsymbol{\xi}(\mathbf{x})$ are solutions of the lower level program—i.e., the combined modal split and stochastic user equilibrium (SUE) assignment problem—evaluated with the design scheme $\mathbf{x}$. In the upper-level program, the social welfare function W is maximized. In this study, we adopt the Bergson-Samuelson social welfare function, which measures social welfare in terms of individual utilities. We consider a special case of the Bergson-Samuelson social welfare function, whose functional form is linear (based on the perspective of purely utilitarian) and the same weight is assigned to all individuals (Mas-Colell et al. 1995). Thus, we have $W = \mathbf{h}^\mathrm{T}\boldsymbol{\xi}$ for the current context.

The objective of social welfare maximization is subject to the financial viability constraint (2) and design feasibility constraint (3). In (2), Π is the profit function for each of the transit lines under design, defined to be the difference between its revenue and operating cost. In this study, we consider that the transit fares are regulated by the government (as is typically the case) without the flexibility to be changed drastically. Transit fares, therefore, are not a design variable, but are set relative to those of the existing transport modes. Under this context, the revenue of each transit line is a sole function of the path flow vector $\mathbf{h}$. On the other hand, the operating cost depends on the transit line configuration and service frequencies; hence a function of $\mathbf{x}$. A transit service is said to be financially viable if it can self-sustain in financial terms, i.e. $\Pi \geq 0$. This constraint can be relaxed according to the specific situations of the local jurisdiction, wherein perhaps subsidy or partial cost recovery is acceptable. In (3), $\mathbf{X}$ is the feasible set of design schemes, which ensures the connectivity of the transit line through the 0-1 variables, non-negativity of the service frequency, and requirements on the upper and lower bounds, etc.

The lower level program is formulated as a variational inequality (VI) problem: Find $\mathbf{h}^* \in \Omega(\mathbf{x})$ such that

$$C(\mathbf{h}^*)^{\mathrm{T}} \cdot (\mathbf{h}(\mathbf{x}) - \mathbf{h}^*) \geq 0, \quad \forall \mathbf{h}(\mathbf{x}) \in \Omega(\mathbf{x}) \tag{4}$$

where C is the vector of travel cost functions. Problem (4) is a conventional user equilibrium model, where Ω is the set of feasible flows defined by the conservation of flow equations, non-negativity of flows, etc. Given $\mathbf{x}$, problem (4) solves for the SUE assignment pattern on the combined mode, route, and transfer problem. More detailed specifications of (4), including the type of utility functions used, are depicted in Lo et al. (2003a, 2004).

In solving problem (4), as inherited from the SAM modeling structure, the effect of congestion is captured. Briefly, the SAM network structure divides transport modes into three sub-classes based on their effect and interactions on network congestion (Lo et al. 2003a, 2003b). Class-1 transport mode has exclusive right-of-ways without congestion interactions with other transport modes, such as subway. Classes-2 and -3 transport modes share the same roadway segments and hence the congestion together. Specifically, we distinguish the Class-2 modes as transit services with fixed routes and frequencies, such as buses, whereas the Class-3 modes are those without, such as taxi or auto. Subsequently, by accounting for the volume of each of these classes of transport modes and combining their respective traffic volumes on the shared physical links, we explicitly capture the effect of congestion on the travel time of the corresponding modes. In the end, using the SAM network as a modeling platform, we develop asymmetric path utility functions for the SUE assignment procedure, which is solved by Methods of Successive Averages (MSA). Further details can be found in Lo et al. (2003a, 2004).

3 Phase-wise design procedure

CMTND is intrinsically combinatorial as the problem involves discrete choices of alternatives. This type of problem is generally NP-hard in the absence of solution algorithms that solve the problem to guaranteed optimality in polynomial time. Magnanti and Wong (1984) provided a classical review of the use of (mixed) integer programming models for this type of transportation network design problem, which is rarely solved to optimality for larger problem instances.

The practically unattainable optimality suggests one to sacrifice optimality for efficiency in obtaining good solutions. Previous studies first identified a preferable network configuration and then assigned the corresponding service frequencies. This greatly reduces the problem complexity as the discrete variables are processed in one sub-problem, followed by a separate optimization involving the continuous variables. Additionally, this technique of specifying a different objective in each sub-problem offers a way to address the multi-objective nature of the combined problem. In this study, a similar procedure is developed to consider passengers' choices as modeled with a SAM network. For ease and clarity of presentation, hereafter we consider that it is the bus service that is under planning, while the other existing transit services are considered as fixed in a multi-modal transit system, with the interactions among the different modes captured. The synopsis of each phase is, respectively:

1. Phase 1 determines the demands arising for the mode under planning (i.e. bus service in the illustrative example) while considering multi-modal choices for travelers in the presence of co-existing modes,
2. Phase 2 determines the bus routes and their associated frequencies to carry the demands found in Phase 1.

Financial viability of the bus services is to be maintained in Phase 1. The Phase 1 procedure is repeated until all the bus service connections pass the test. Then Phase 2 determines the number, routing, and frequency of bus lines needed to carry the demands found in Phase 1. By separating the problem into two phases, with each phase achieving certain objectives, they collectively solve the combined problem in an approximate manner. The approach resembles the common solution strategy of restructuring the combined problem as a series of decisions with ascending levels of detail.

3.1 Phase 1 problem

Phase 1 estimates the amount of passengers, as part of the total demand, which would use the concerned mode (bus in this case) in a congested network. For a given set of service configuration, we determine the matrix of point-to-point demands on the concerned mode under planning. In the end, a profile of the resulting bus network, expressed as the point-to-point connectivity, is attained. The CMTND problem in Phase 1 is expressed in a modified form, as in (5) and (6), which includes the lower level problem as in (4):

[CMTND-1]

$$\max_{\bar{\mathbf{x}}} \ W = \mathbf{h}(\bar{\mathbf{x}})^{\mathrm{T}} \boldsymbol{\xi}(\bar{\mathbf{x}}) \tag{5}$$

$$\text{subject to} \quad \frac{1}{\hat{f}_b \cdot c_b}\left(\sum_{\phi}\sum_{v} h_v^{\phi}(\bar{\mathbf{x}}) \cdot \delta_{a_{\mathrm{d}}}^{v}\right) \geq \mu_{\mathrm{critical}}, \quad \forall a_{\mathrm{d}} \in \bar{\mathbf{x}} \tag{6}$$

where $\mathbf{h}(\bar{\mathbf{x}})$ and $\boldsymbol{\xi}(\bar{\mathbf{x}})$, solved in problem (4), are the solution of the combined modal split and SUE assignment problem at the lower-level evaluated at design scheme $\bar{\mathbf{x}}$.

In this modified problem, we express the bus service $\bar{\mathbf{x}}$ as the point-to-point connectivity. That is, we are indicating whether there are direct bus services connecting the concerned locations, or in the SAM network terminology, whether there are direct (bus) links between the concerned nodes. As a result, the design scheme $\bar{\mathbf{x}}$ for this modified CMTND problem is composed of binary variables on a set of bus direct in-vehicle links only. The exact routing and service frequency are to be determined in

Phase 2 and are not known a priori in Phase 1. In the objective function (5), the purely utilitarian social welfare function is maximized by providing bus services to link selected pairs of boarding and alighting points. The financial viability condition (2) is expressed in a simpler form in (6), which basically checks for the load factors of the services to be offered. The frequency $\hat{f}_b$ in Phase 1 is tentatively fixed to fulfill the minimum requirement of the regulation. In (6), c_b is the vehicle capacity; $\delta^v_{a_\text{d}}$ is an indicator parameter which is 1 if the transit path v travels on the bus direct link a_d, or zero otherwise. The value μ_critical sets the minimum average bus-load such that the concerned bus service can be financially viable. The numerical value of μ_critical takes into account a variety of factors like vehicle running and maintenance cost, wages of drivers and back-up staff, bus fare, etc.

3.2 Phase 1 heuristic

CMTND-1 is a bi-level discrete network design problem. Previous heuristics for the discrete network design problem generally fall into three different approaches: (i) the *add* heuristic; (ii) the *delete* heuristic; and (iii) a hybridized *add-delete* heuristic. The *add* heuristic, starting with an empty network, adds the next most favorable element to the network one by one (e.g. based on demand load, marginal cost, etc.) till some defined criterion is met. The *delete* heuristic is similar but it starts with a large number of elements in the network, and in some cases the fully connected network, and gradually removes elements till termination. The last type is a mix of the previous two, in which the algorithm chooses to add or delete in the next step based on the current condition.

This study develops a *delete* heuristic. The heuristic gradually prunes away service connections that do not meet the financial viability or load factor constraint. According to the logit model in the SUE assignment model, the satisfaction function is monotonic with respect to the size of the choice set. That is, the perceived utility of the system is higher when there are more alternatives in the choice set. Under certain conditions, this is equivalent to saying that as the choice set is larger, the objective function (5) is higher. Generally, from travelers' perspective, it is desirable to have more direct point-to-point bus services. And by starting with a fully-connected network, generally the system has the highest objective function value. However, more direct services will result in low passenger volume per service, which may not meet the financial viability or load factor condition. By gradually pruning services from the network, the remaining transit services enjoy the "concentration of flow" benefit (Rea 1971). Firstly, their load factors go up. Secondly, higher demand per service can afford more frequent services, leading to reductions in headway and hence shorter transit journey time.

The problem size, however, would be prohibitive if we start off with a fully-connected SAM network for a large physical multi-modal network. In the more likely applications to expand or improve an existing (bus) system, prior knowledge of the usage pattern and preference is helpful in identifying a limited number of services to be added, hence without the need to consider all possible node-to-node connections. To design for a brand new network without any prior understanding of the transfer and usage pattern, it may be necessary to run the entire problem once, albeit a large effort, in order to gain insights about the system performance.

To summarize, the heuristic procedure is as follows:

Step 0: Network transformation. Transform the existing multi-modal city network (physical network) into a SAM network $G^0 = (\mathbf{N}, \mathbf{A})$ in accordance with a set of appropriate transition rules (for details, see Lo et al. 2003a, 2003b).

Step 1: Initialization. Connect all the bus direct in-vehicle links $\mathbf{A}_\mathrm{d}^1$ between all node pairs of interest. Set $n := 1$.

Step 2: Network update and Assignment. Establish the set of transfer links $\mathbf{A}_\mathrm{t}^n$ in conjunction with $\mathbf{A}_\mathrm{d}^n$. Solve the SUE assignment problem (4) on the SAM network $G^n = (\mathbf{N}, \mathbf{A}^n)$, where $\mathbf{A}^n = \mathbf{A} \cup \mathbf{A}_\mathrm{d}^n \cup \mathbf{A}_\mathrm{t}^n$.

Step 3: Financial viability check. Evaluate the average bus-load for each bus direct in-vehicle link $a_\mathrm{d} \in \mathbf{A}_\mathrm{d}^n$. If the load factor of each direct in-vehicle link satisfies the minimum requirement in (6), set $\mathbf{A}_\mathrm{d}^* = \mathbf{A}_\mathrm{d}^n$, go to Step 5; otherwise, go to Step 4.

Step 4: Updating direct links. Remove the bus direct link with the lowest load factor from $\mathbf{A}_\mathrm{d}^n$ to form $\mathbf{A}_\mathrm{d}^{n+1}$. Set $n = n + 1$ and go to Step 2.

Step 5: Termination. The point-to-point bus demand matrix is obtained from the SUE assignment solution $\mathbf{h}$ on the network $G^* = (\mathbf{N}, \mathbf{A}^*)$, where $\mathbf{A}^* = \mathbf{A}^n$.

The above heuristic starts with forming the SAM network from the background network without bus services (Step 0), then all direct bus services are added (Step 1) to obtain the complete SAM network on which the SUE transit assignment is performed (Step 2). Subsequently, the resultant bus service load is checked for financial viability (Step 3). If the financial viability criterion is not met, an updated SAM network is formed by deleting direct bus services with low load factors (Step 4). With the updated network structure, the procedure of SUE assignment is re-run. As expected, the overall flows would change drastically from the SAM network obtained in the last iteration. The combined mode-transfer-route splits of all travel demands (which include the demand reallocated from the deleted bus direct links) are re-estimated by the SUE assignment on the newly updated SAM network. The process is reiterated until all bus services are financially viable.

Upon termination, the heuristic identifies a set of direct bus links and the SUE path flows. The flows on the direct bus links can be viewed as a set of bus OD demands. Though the optimality condition is not guaranteed, the heuristic aims to produce the best objective function value by gradually removing links from the network.

3.3 Phase 2 problem

The objective of Phase 2 is to design a set of cost-minimizing transit routes to serve the bus OD demand determined in Phase 1, which is fixed in this phase. Given that the demand is fixed, and that the fare is set independently or given, cost minimizing is equivalent to profit maximizing. To speed up the computation, existing transit network design procedures typically assume certain simplifying conditions. For example, Ceder and Wilson (1986) excluded all the paths that are some percentage longer than the shortest path, whereas Pattnaik et al. (1998) only considered those transit routes which are combinations of some specified shorter routes. In this study, we assume the following conditions:

(C1) All origin and destination nodes of the demands are covered.
(C2) Sufficient bus capacities are introduced to carry all the demands.
(C3) Every route is acyclic.
(C4) Every route serves both directions with the same line frequency.

Conditions (C1) and (C2) ensure that there is no residual or un-served demand. Condition (C3) excludes any cyclic routes. Finally, condition (C4) retains the conservation of transit vehicles along a bus route, and is commonly adopted in practice. In Phase 2, the design scheme $\tilde{\mathbf{x}}$ is determined under this set of assumptions.

The problem can be formulated as a multi-route transit network design problem (MRTNDP). Wan and Lo (2003) depicted a mixed integer formulation for the MRTNDP. In the following, in the interest of space, we only provide a brief summary. Unlike Phase 1, where the design variables are the point-to-point connectivity indicators or the direct in-vehicle link a_{d} in SAM network, the MRTNDP expresses the design variables in terms of links (a) in the physical network, and each transit route (r) as an itinerary of nodes to be visited on the physical network. Thus, the design scheme $\tilde{\mathbf{x}}$ is now a collection of binary variables $x_a^r = 1$ if route r traverses link a; zero otherwise. The design scheme also includes the frequency vector whose component f_r is associated with a particular bus route r. The formulation is:

[MRTNDP]

$$\min_{\tilde{\mathbf{x}},\mathbf{f},\mathbf{q}} \sum_r \sum_a \kappa_a f_r x_a^r \tag{7}$$

$$\text{subject to} \quad \sum_r q_{st}^r = q_{st}, \qquad \forall (s,t) \in \mathbf{\Theta} \tag{8}$$

$$c_b f_r - \sum_{(s,t)\in\mathbf{\Theta}} d_{su}^r d_{ut}^r q_{st}^r \geq \sum_{(u,v)\in\mathbf{\Theta}} d_{uv}^r q_{uv}^r, \qquad \forall u \in \bar{\mathbf{\Theta}} \tag{9}$$

$$c_b f_r - \sum_{(s,t)\in\mathbf{\Theta}} d_{tu}^r d_{us}^r q_{st}^r \geq \sum_{(u,v)\in\mathbf{\Theta}} d_{vu}^r q_{uv}^r, \qquad \forall u \in \bar{\mathbf{\Theta}} \tag{10}$$

$$0 \leq q_{st}^r \leq M(d_{st}^r + d_{ts}^r), \qquad \forall (s,t) \in \mathbf{\Theta},\ \forall r \tag{11}$$

$$f_{\min} \leq f_r \leq f_{\max}, \qquad \forall r \tag{12}$$

$$d_{st}^r = 0 \text{ or } 1, \qquad \forall (s,t) \in \mathbf{\Theta},\ \forall r \tag{13}$$

$$\tilde{\mathbf{x}} \in \tilde{\mathbf{X}} \tag{14}$$

where $\mathbf{q} = (\ldots, q_{st}^r, \ldots)$ is a vector of decision variables; q_{st} is the OD demand from s to t; κ_a is the marginal operating cost of bus service on link a; d_{ij}^r is the direct in-vehicle link indicator from i to j and equals 1 if route r serves between the respective boarding and alight stops; M is a big number; $\mathbf{\Theta}$ is the set of OD pairs; and $\bar{\mathbf{\Theta}}$ is the set of bus alighting points. In this formulation, the optimal allocation of flows between each OD pair between routes are set as decision variable $\mathbf{q}$, which is determined together with the route configuration $\tilde{\mathbf{x}}$.

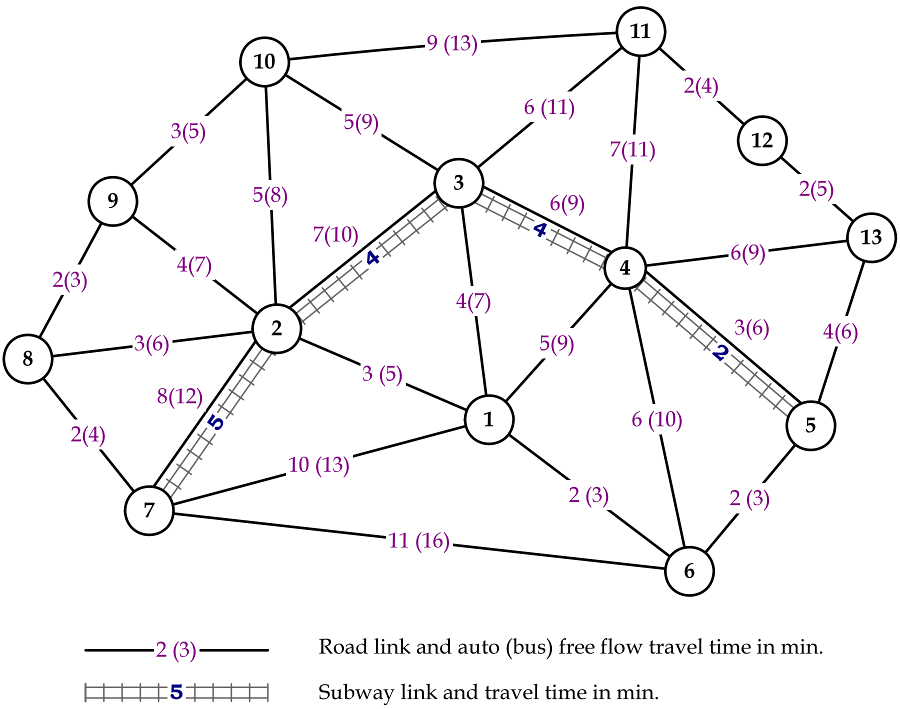
The objective function (7) is to minimize the sum of operating costs of all the transit lines. Though the operating cost appears to be defined at the link level, constraint (14) on line configuration ensures that the operating cost of a route is the sum of marginal costs over all links of the entire route. Conditions (C1) and (C2) are expressed in (8), (9) and (10). Constraint (8) ensures that all the OD demands are satisfied by the bus services. Constraints (9) and (10) depict the capacity requirements for the boarding passenger volume at each node along each transit route in both directions. For example, the left-hand side of (9) is the available passenger space on route r prior to node u in the forward direction, which is required to be higher than the boarding volume at that node. Similarly, constraint (10) considers the available passenger space in the backward direction. Constraint (11) assigns demand q_{st}^{r} on route r if it provides a direct service for the OD movement, in either the forward or backward direction. Constraint (12) sets the bounds of service frequencies. Constraints (13) and (14) are the feasibility constraints for the transit routes. Wan and Lo (2003) introduced node labels which facilitate the representation of the route structure and it is capable of expressing both (13) and (14) in a number of linear constraints in terms of these node labels. Further details are provided in Wan and Lo (2003), which also discussed the linearization procedures to obtain an equivalent mixed integer liner program (MILP) of the MRTNDP. One can then use existing MILP tools to solve the problem. The solution of the MRTNDP is a set of transit routes and their corresponding frequencies.

4 An illustrative example

To illustrate the proposed approach, we consider the case of bus route design in a multi-modal network. To simplify the illustration, we only consider (i) auto and taxi, (ii) subway, and (iii) bus. Figure 1 depicts the network topology, link travel time and subway fare structure. The network has 13 nodes and 26 links. A subway operates along the node alignment $7 \leftrightarrow 2 \leftrightarrow 3 \leftrightarrow 4 \leftrightarrow 5$ in a segregated track with an exclusive right-of-way. Only auto, taxi, and bus share the congestion on the roadway network. Auto and taxi have the same free flow travel time in minutes (shown as the number on each link in Fig. 1), while bus, due to its lower speed and the need for passenger boarding and alighting, has a longer free flow travel time (shown as the bracketed number on each link in Fig. 1). The travel time of the subway is shown next to its track in Fig. 1. Also shown in Fig. 1 is the subway fare table, demonstrating its nonlinear structure, as is typical in reality. In particular, this fare structure was set up to illustrate the effect of the probable transfer rules of the SAM network approach. Even though a traveler can reduce his fare slightly via certain subway-to-subway transfers (e.g., the direct fare from node 7 to node 4 is \$12; a passenger can save \$2 by taking the subway from node 7 to node 2 first, exiting the station at node 2 and reentering the station at node 2, then taking the subway again from node 2 to node 4), most people would not choose to do that, as is captured in the probable transfer rules (Lo et al. 2003a, 2003b). In fact, instances of such a fare structure can be found in Hong Kong.

Road congestion is captured by the Bureau of Public Roads (BPR) performance function with a uniform link capacity of 1,500 pcu/hr, expressed as:

$$t_a = t_a^0 \cdot \left[1 + 0.15\left(\frac{v_a}{c_a}\right)^4\right] \tag{15}$$



Subway Fare ($)	Destination Station				
Origin Station	2	3	4	5	7
2	-	5	7	12	3
3	5	-	5	7	12
4	7	5	-	7	12
5	12	7	7	-	12
7	3	12	12	12	-

Fig. 1 The illustrative multi-modal network

where t_a and t_a^0 are, respectively, the travel time and free flow travel time of link a, v_a the traffic volume in pcu (passenger car units), c_a the practical link capacity, taken as $c_a = 1500 \times 0.75 = 1125$ pcu/hr.

For the disutility associated with a crowded transit vehicle, we follow Nielson (2000) and Lo et al. (2003a) and adopt the following discomfort function:

$$\zeta_b^a = t_b^a \cdot \left[1 + 0.1 \left(\frac{v_b^a}{f_b \cdot c_b} \right)^{1.3} \right] \tag{16}$$

where ζ_b^a, t_b^a, v_b^a are, respectively, the modified travel time, the actual in-vehicle time, passenger flow on transit vehicle of mode b at link a; f_b and c_b are the service frequency and transit capacity as introduced earlier. For consistency, we define $\zeta_{\text{car}}^a = t_a$ for auto/taxi. With reference to the case in Hong Kong, the vehicle capacity of a bus is taken as 100 passengers while that of subway is taken as 3,000 passengers.

Table 1 Total travel demand [prs/hr]

Origin node	Destination node							
	3[a]	4[a]	5[a]	6	7[a]	9	10	12
3[a]	–	805	448	743	638	826	698	211
4[a]	1040	–	798	648	885	798	1077	130
5[a]	438	599	–	430	417	462	368	106
6	600	1162	470	–	549	639	680	158
7[a]	302	337	403	399	–	335	422	60
9	667	790	569	499	531	–	862	116
10	1043	870	681	481	722	913	–	177
12	109	131	82	118	89	106	85	–

[a]Locations of subway stations

Mode choice in this multi-modal transit network is modeled with the logit model, including in-vehicle time, waiting time, and fare as attributes. The utility of any chosen path is simply the sum of its associated direct link(s) or transfer link(s) (Lo et al. 2003a). We use the popular linear-in-parameter utility function (e.g. Ben-Akiva and Lerman 1985). The utility attribute coefficients are: $-0.047/\text{min}$ for in-vehicle time, $-0.1087/\text{min}$ for waiting time, $-0.0387/\$$ for fare, and mode specific constants relative to bus are, respectively, -1.3 for auto and 0.7 for subway. The transfer penalties are, respectively, 0.1 for transfer with the subway, 0.5 between bus and auto/taxi, and 5 between buses. Finally, the overall OD demand matrix in passengers per hour (prs/hr) is shown in Table 1.

4.1 Phase 1 procedure and solution

As discussed earlier, the fares are assumed to be exogenous to this problem. In this illustrative example, we consider the bus fare ρ to be proportional to the free flow travel time, as: $\rho(i, j) = 0.5t_0^{ij}$, where $\rho(i, j)$ and t_0^{ij} are the bus fare (in dollars) and auto free flow travel time from i to j. In case of several auto paths available between the OD pair, it is taken as the weighted in-vehicle travel time of these paths by the modal split. Meanwhile, the bus travel time is taken as:$t_{\text{bus}}^{ij} = 2t_{\text{auto}}^{ij}$, where t_{bus}^{ij} and t_{auto}^{ij} are travel time of bus and auto on the same link, respectively. The proportionality constant addresses the delay of bus due to its generally lower speed and passenger boarding and alighting at stops.

The procedure starts with the fully connected SAM network, i.e. point-to-point bus services are available between all location pairs. For the network shown in Fig. 1, a total of 156 bus direct links is initially generated. We consider that each direct link has a uniform frequency of 6 buses per hour. After each SUE assignment on the SAM network, these bus direct links are recursively removed if their average bus loads fall below $\mu_{\text{critical}} = 0.2$, or an equivalent demand of 120 prs/hr, which is set as the criterion for financial viability. The result from Phase 1 is summarized in Table 2. There are totally 9 direct bus services, or bus direct in-vehicle links, at the termination of the Phase 1 heuristic.

The result demonstrates three key characteristics. Firstly, the result from Phase 1 shows that direct bus service links are not limited to the original origins or destina-

Table 2 Bus direct service links after Phase 1

Direct link	Patronage [prs/hr]	Direct link	Patronage [prs/hr]	Direct link	Patronage [prs/hr]
$2 \rightarrow 10$	1432	$4 \rightarrow 3$	121	$4 \rightarrow 6$	615
$5 \rightarrow 4$	287	$5 \rightarrow 6$	967	$7 \rightarrow 3$	145
$7 \rightarrow 4$	153	$7 \rightarrow 6$	880	$7 \rightarrow 10$	146

tions of the total demand table as shown in Table 1. For example, direct link 2→ 10 is introduced (see Table 2) even though the total demand table (Table 1) indicates no such direct demand. This indicates that this direct bus link mainly acts as a feeder service for the demand leaving the subway at node 2. It illustrates that this modeling platform takes into account inter-modal transfers in planning the bus network. Secondly, the direct bus links determined in Phase 1 are not limited to neighboring locations: for example, direct bus links are provided between nodes 7 and 4 and between nodes 7 and 10 (Table 2). The third feature is illustrated by not providing bus service to nodes 9 and 12, though there are travel demands going to and from these nodes. One may note that the subway also does not serve these nodes. The result indicates that passengers to and from these nodes generally take taxi or drive to their destinations or to other nodes for transfers. The main reason for taking out services from these nodes is due to the financial viability condition or the low demand associated with these nodes. The model reveals this. If for welfare reasons, such connections are to be added, at least the planner is aware of the implication—subsidy is needed for their service provision.

4.2 Phase 2 procedure and solution

As discussed earlier, this study considers only acyclic routes. In addition, all bus routes serve in both directions along the same alignment with the same frequency. In this case study, we consider at most 4 routes in the solution, i.e. $r = 1, 2, 3$ or 4. The allowable frequency ranges from $f_{\min} = 3$ bus/hr to $f_{\max} = 25$ bus/hr. The marginal operating cost of each link is taken to be proportional to the corresponding bus free flow travel time, i.e. $\kappa_a = 200 t^0_{a,\text{bus}}$. The MILP was solved using the commercial optimizer CPLEX®. For practical applications use of the state-of-the-art computing technology invariably brings an edge in solving the inherently difficult problem. In solving our example problem, the solution consists of 3 bus routes and has an objective function value of 15,498. The route alignment is as shown in Fig. 2. Route 1 (i.e. $2 \leftrightarrow 10$) has a frequency $f_1 = 24$ bus/hr, while those of Route 2 (i.e. $7 \leftrightarrow 8 \leftrightarrow 9 \leftrightarrow 10$) and Route 3 (i.e. $3 \leftrightarrow 4 \leftrightarrow 5 \leftrightarrow 6 \leftrightarrow 7$) are $f_2 = 3$ bus/hr and $f_3 = 16$ bus/hr.

The bus network configuration obtained aims to satisfy the direct bus OD demand (determined in Phase 1) without any transfers. However, the resultant bus routes typically would traverse other nodes as well. As a result, the obtained bus network usually provides more direct services than required. For example, in Route 2, other than serving the direct demand of $7 \rightarrow 10$, it also brings about other direct services $7 \leftrightarrow 8$, $7 \leftrightarrow 9$ and $8 \leftrightarrow 9$. Thus, the resulting network should always be better than what is specified at the end of Phase 1.

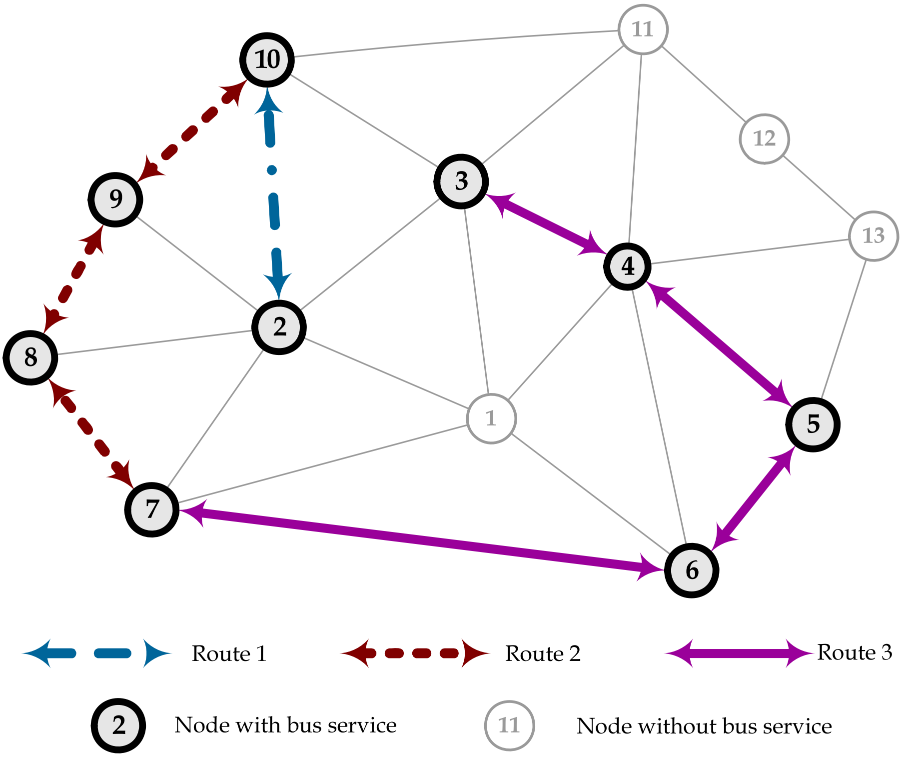


Fig. 2 Resulting bus routes

It is interesting to note that although we specify the maximum routes to be generated in Phase 2 to be 4, only 3 routes are generated. So the formulation is able to automatically select the optimal number of routes as well. The results also demonstrate the ability of the formulation to design transit services that consider the interactions with other transport modes. According to the results, Routes 1 and 2 are feeder services to work with the subway via nodes 2 and 7. On the other hand, Route 3 overlaps with the subway substantially. As Phase 1 incorporates travelers' preferences for the combined mode choices while figuring out the direct bus demands, the final network configuration produced in Phase 2 reflects what is best from the travelers' perspectives.

4.3 Performance of the resulting network

The two-phase heuristic starts with a fully-connected SAM network, which is gradually pruned by removing links with low load factors. At the end of Phase 1, the pruned SAM network and resultant bus demands are obtained (Table 2). These bus demands are then fed into Phase 2 to produce the bus routes for their accommodation (Fig. 2). In order to examine the performance of the multi-modal network determined at the end of Phase 2, a complete transit assignment on the final network is performed again with the SAM network approach.

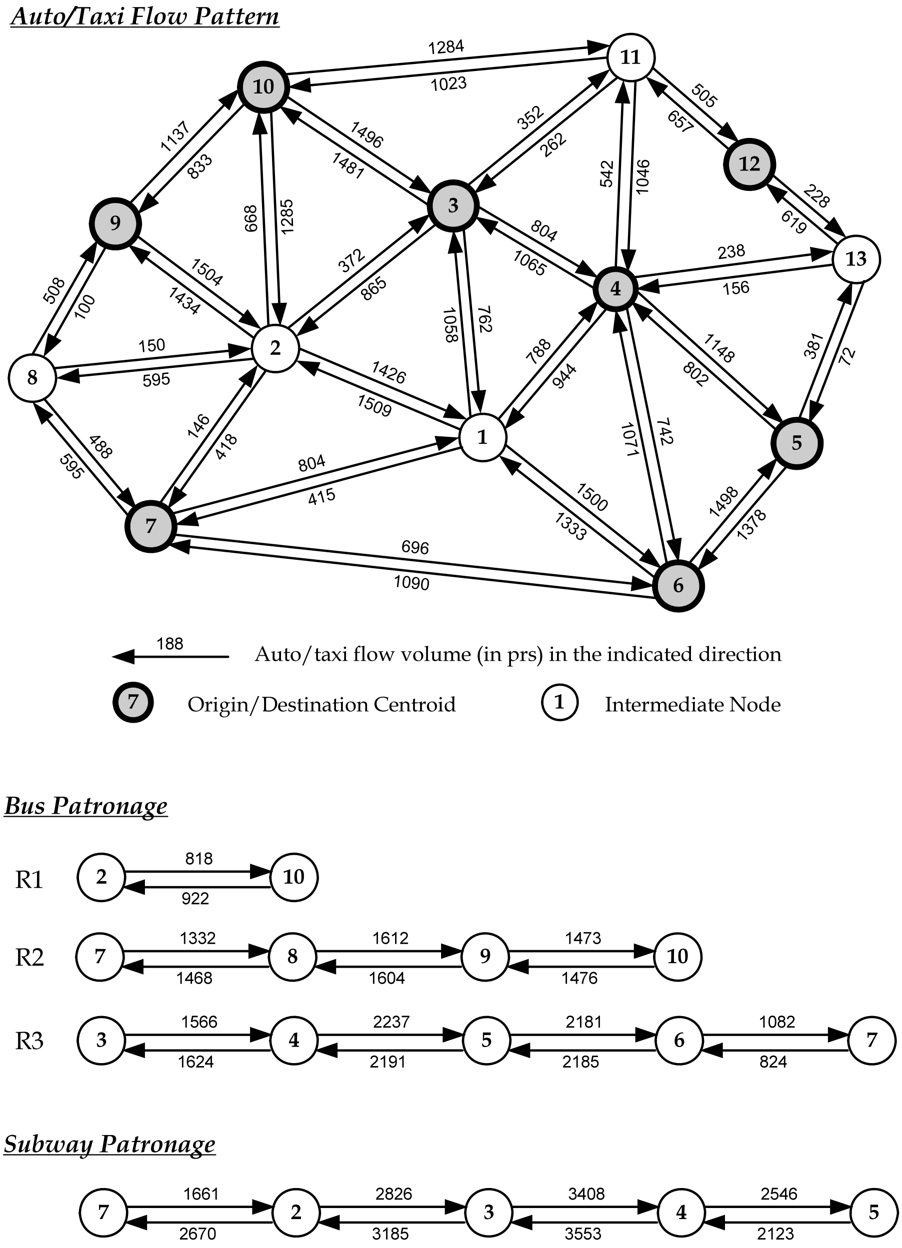


Fig. 3 Resulting auto/taxi flow and transit patronage pattern on the network

The resulting traffic flow pattern is shown in Fig. 3. The number next to each link shows the passenger volume by auto/taxi; whereas the passenger volumes on the different transit lines are shown along with the transit route diagram. In addition,

Table 3 Summary and statistics on passengers' choice

Choice	Volume [prs/hr]	Percentage [%]
(1) Auto/taxi	6,953	24.2
(2) Subway	1,873	6.5
(3) Single bus route	5,456	19.0
(4) Auto/taxi & subway	5,034	17.5
(5) Subway & bus	3,471	12.1
(6) Auto/taxi & bus	4,140	14.4
(7) Auto/taxi, subway & bus	1,789	6.2
Total	28,722	100.0

Table 4 Summary and statistics on inter-modal transfers

Number of transfers	Volume [prs/hr]	Percentage [%]
(1) Nil, i.e. direct service	14,282	49.7
(2) 1 transfer	12,011	41.8
(3) 2 transfers	2,429	8.5
(4) More	0	0.0
Total	28,722	100.0

Tables 3 and 4 depict summary and statistics on passengers' modal choices and the needs of transfer, respectively.

According to the result, we find that auto/taxi alone is the most popular transport mode which attracts around a quarter of the travel demand. This high usage may reflect the high reliance on auto/taxi at nodes 11, 12 and 13 in the absence of transit services there and/or a preference to auto/taxi over the transit services. The former is related to the financial viability in Phase 1 and the parameter μ_{critical}, while the latter concerns with the utility parameters in the mode choice model. Another observation is the relatively low usage on the subway alone. However, a detailed examination reveals that a total of 42.3% of the travel demand travels on the subway for at least part of their journeys. The three bus routes serve 19.0% of the total demand without transfers; whereas including those with transfers, they serve 32.7% of the total demand. Table 4 further demonstrates the important role of transfer in a multi-modal transit network. While 49.7% of the demand travels with transfers, the other 50.3% do involve at least one transfer. In particular, around 80% of the trips with transfers complete their trips with a single transfer. In general, with the SAM network modeling tool, one can derive many measures to gauge the performance of a multi-modal network.

One limitation of this approach, however, is that currently the two phases of the heuristic scheme are implemented separately. Phase 1 produces the initial SAM network and bus demands; the demands are then used in Phase 2 to determine the bus routes. The SAM network associated with the bus routes generated in Phase 2 may be different from the SAM network determined in Phase 1. This inconsistency is manifested in the result of bus route 2, which serves mainly the demand from node 7 to node 10 and carries a relatively light volume of 146 passengers per hour according to the demand calculated in Phase 1 (Table 2). Accordingly, in Phase 2, the frequency for Route 2 is optimized to be 3 per hour, which is sensible per the light demand between nodes 7 and 10 found in Phase I. To determine the eventual performance of the multi-modal network obtained in Phase 2, a final traffic assignment on the Phase 2 network is performed again. The loadings for bus route 2 are determined to be around 1332 to 1612 passengers per hour (Fig. 3), which are different from the loading estimated in Phase 1. This discrepancy points to the need of fully integrating the two phases, such that the loadings determined from the two phases are consistent. Conceptually, this extension can be accomplished by iterating between the two phases until the loadings converge. Given that Phase 2 involves a mixed-integer program, computational efficiency may be an issue. Therefore, how to integrate the two phases so as to minimize the problem size of Phase 2 and the number of iterations required is not trivial. We will leave this extension to a future study.

5 Concluding remarks

In a multi-modal network where transfers are common, the connections with the other modes are crucial to be incorporated in the planning process. The framework developed designs transit services while taking into consideration the other existing transit services, travelers' preferences for the combined-mode choices, as well as the financial viability requirement. In particular, it extends the existing approaches by explicitly incorporating the interactions with co-existing transit services.

In this study, the importance of inter-route and inter-modal transfers is treated with the SAM network, and the congestion effect incorporated in the SUE assignment procedure. The two-phase methodology can be considered as a sequence of two relaxed optimizations. The first phase seeks to maximize the social welfare by providing the largest set of bus service connections, subject to the financial viability constraints. The second phase seeks to minimize the operation cost by detailing the bus routings and frequencies to fulfill the service connections determined in the first phase.

We have presented the two phases separately. In either phase, certain assumptions have been made to simplify the formulation and make this framework tractable. These assumptions are generally not restrictive and can be relaxed in future studies. For example, in Phase 1, the demand is expressed as a function of the transit service frequency, which is not exactly known prior to Phase 2. On the other hand, the bus routes determined in Phase 2 can be used to refine the connection of point-to-point direct services in Phase 1. Hence, there may be inconsistency between the two phases. Also, different principles of passenger allocation among transit routes are assumed in the two phases to reflect the different interests of travelers and operators. In Phase 1, the passenger flows are modeled according to the principle of stochastic user equilibrium; whereas in Phase 2, operating cost minimization governs the transit service availability. This is a typical example of demand-supply interaction to achieve equilibrium via a sequence of iterative action-reaction processes. To improve this formulation, a feedback procedure can be carried out to improve the consistency between the two phases. A possible iterative scheme would work like this: for any OD pair with point-to-point service as derived from the line configuration of Phase 2, update the corresponding direct in-vehicle links with the parameters (such as service frequency) in the fully-connected SAM network as the starting network in Phase 1. The logic behind is that the lines from Phase 2 should be more financially viable than other possible direct services. At the same time, as illustrated in our example, some direct services with lesser demand can be made financially viable via demand transferred from other services. As a result, the feedback loop of updating the results between the two phases should better reflect the actual passenger movements or choices in the physical network. The iterative design process should be stopped if a specified consistency criterion between the two phases is met, from which the final design can be chosen by comparison of the network performance evaluation.

The proposed framework can be improved in several aspects. Firstly, the passenger modal choice models should be calibrated with real data, as they have substantial implications on what they will actually use and therefore on the service connections to be provided. Secondly, the SUE assignment procedure is computationally demanding, especially for congested networks and large passenger perception variations. Similarly, the computational burden of solving the linearized MRTNDP of considerate size is formidable. However, as extensive transit network design is not frequently performed, for incremental service planning of adding a small number of potential routes, the computation time is acceptable in many cases. In practice, one may also divide the whole city into zones of reasonable scope and work on each district sequentially. In this way, planners should find this design framework helpful.

Acknowledgements The authors would like to thank the anonymous referees and Professor Stefan Voss for comments and helpful suggestions. This study is sponsored by the Competitive Earmarked Research Grants, HKUST6161/02E and 617105, of the Hong Kong Research Grant Council.

References

Baaj MH, Mahmassani HS (1995) Hybrid route generation heuristic algorithm for the design of transit networks. Transp Res Part C 3:31–50

Bard JF (1998) Practical bilevel optimization: algorithms and applications. Kluwer Academic, Dordrecht

Ben-Akiva M, Lerman SR (1985) Discrete choice analysis: theory and application to travel demand. MIT Press, Cambridge

Bruno G, Ghiani G, Improt G (1998) A multi-modal approach to the location of a rapid transit line. Eur J Oper Res 104:321–332

Ceder A, Wilson NHM (1986) Bus network design. Transp Res Part B 20:331–344

Colson B, Marcotte P, Savard G (2005) Bilevel programming: a survey. Q J Oper Res 3:87–107

Dufourd H, Gendreau M, Laporte G (1996) Locating a transit line using tabu search. Locat Sci 4:1–19

Lo HK, Yip CW, Wan KH (2003a) Modeling transfer and nonlinear fare structure in multi-modal network. Transp Res Part B 37:149–170

Lo HK, Yip CW, Wan QK (2003b) Modeling competitive multi-modal transit services. In: Lam WHK, Bell MGH (eds) Advanced modeling for transit operations and service planning. Pergamon, Amsterdam, pp 231–256

Lo HK, Yip CW, Wan QK (2004) Modeling competitive multi-modal transit services: a nested logit approach. Transp Res Part C 12:251–272

Magnanti TL, Wong RT (1984) Network design and transportation planning: models and algorithms. Transp Sci 18:1–55

Mas-Colell A, Whinston MD, Green JR (1995) Microeconomic theory. Oxford University Press, New York

Michaelis M, Schöbel A (2009, to appear) Integrating line planning, timetabling, and vehicle scheduling: a customer-oriented heuristic. Public Transp. doi:10.1007/s12469-009-0014-9

Nielsen OA (2000) A stochastic transit assignment model considering differences in passengers utility functions. Transp Res Part B 34:377–402

Pattnaik SB, Mohan S, Tom VM (1998) Urban bus transit route network design using genetic algorithm. J Transp Eng 124:368–375

Rea JC (1971) Designing urban transit systems: An approach to the route-technology selection problem. Report no. UMTA-URT-49(70)-71-6 by the Urban Transportation Program, University of Washington, Seattle, WA

Wan QK, Lo HK (2003) A mixed integer formulation for multiple-route transit network design. J Math Model Algorithms 2:299–308