

Multimodal Mobility Systems: Joint Optimization of Transit Network Design and Pricing

Qi Luo

Clemson University
Clemson, SC, USA
qluo2@clemson.edu

Samitha Samaranayake

Cornell University
Ithaca, NY, USA
samitha@cornell.edu

Siddhartha Banerjee

Cornell University
Ithaca, NY, USA
sbanerjee@cornell.edu

Abstract

The performance of multimodal mobility systems relies on the seamless integration of conventional mass transit services and the advent of Mobility-on-Demand (MoD) services. Prior work is limited to individually improving various transport networks' operations or linking a new mode to an existing system. In this work, we attempt to solve transit network design and pricing problems of multimodal mobility systems en masse. An operator (public transit agency or private transit operator) determines frequency settings of the mass transit system, flows of the MoD service, and prices for each trip to optimize the overall welfare. A primal-dual approach, inspired by the market design literature, yields a compact mixed integer linear programming (MILP) formulation. However, a key computational challenge remains in allocating an exponential number of hybrid modes accessible to travelers. We provide a tractable solution approach through a decomposition scheme and approximation algorithm that accelerates the computation and enables optimization of large-scale problem instances. Using a case study in Nashville, Tennessee, we demonstrate the value of the proposed model. We also show that our algorithm reduces the average runtime by 60% compared to advanced MILP solvers. This result seeks to establish a generic and simple-to-implement way of revamping and redesigning regional mobility systems in order to meet the increase in travel demand and integrate traditional fixed-line mass transit systems with new demand-responsive services.

CCS Concepts: • Applied computing → Transportation;
• Theory of computation → Market equilibria; • Mathematics of computing → Combinatorial optimization.

Keywords: Multimodal mobility systems, Market design, Mixed Integer Programming

ACM Reference Format:

Qi Luo, Samitha Samaranayake, and Siddhartha Banerjee. 2021. Multimodal Mobility Systems: Joint Optimization of Transit Network Design and Pricing. In *Nashville '21: ACM/IEEE International Conference on Cyber-Physical Systems, May 19–21, 2021, Nashville, TN*. ACM, New York, NY, USA, 11 pages. <https://doi.org/>

1 Introduction

Emerging Mobility-on-Demand (MoD) solutions, from ride-hailing platforms (e.g., Uber, Lyft) to on-demand bus services [1], provide responsive and reliable mobility options to urban commuters. Nevertheless, there are growing concerns that MoD would compete with or substitute for conventional transit, increase traffic congestion, or adopt discriminatory price mechanisms [5, 7, 9]. This work proposes a new framework for building a multimodal mobility system as a first step in the full integration of Mobility-on-Demand (MoD) services and regional mass transport networks. Mass transit (MT) operates as an affordable top-down system where users have to adapt to available routes and service times. On the other hand, MoD platforms provide flexible door-to-door service but are not affordable for every traveler. This work aims to explore the opportunity to get the best of both systems by tightly integrating the two systems at the system design level, thereby understanding the true potential of a fully integrated multimodal mobility system.

A primary motive for establishing a multimodal mobility system is to ensure that MT and MoD services are seamlessly integrated into an efficient transportation network. The system's success hinges upon the confluence of two modes and meeting the latent travel demand given user preferences. MT provides low-cost services along fixed routes to a prearranged timetable. In contrast, MoD can provide access to underserved communities by MT (due to various reasons) at a higher expense. On a network design level, integrating MoD helps extend MT's reach through first- or last-mile connections; On an operational level, redirecting travelers by MoD can harness underused MT routes and reduce the overall operational costs. This work aims to provide a tractable optimization framework that simultaneously design networks and determines service prices in favor of the public good.

1.1 Backgrounds

Multimodal mobility systems are an emerging service that cuts across traditional transit boundaries and emerging MoD applications. Transportation agencies are playing a prominent role in shaping such services by experimenting with new service models and rules for governing integration with on-demand mobility service providers [13]. For example, the Chicago Regional Transportation Authority and Lyft

launched an incentive program that connects commuters from nearby metro-stations to office buildings [14]. The local transportation agency is responsible for allocating resources between improving MT services and subsidizing MoD services as first- or last-mile connections. MoD, either owned by the public sector or a private company, can benefit from reaching a broader basis of customers [2].

A vast stream of literature has studied the *centralized* planning, tactical, and operational decisions in transit or MoD systems. Planning decisions include choosing service regions or bus stop locations. Tactical decisions include setting bus routes' frequency. Operational decisions include timetabling of MT and routing of MoD. Interested readers are referred to the reviews on MT [8] and this stream of work on MoD [16]. More recently, joint decision-making on different levels has been studied thoroughly taking both short-term and long-term travelers' behavior into consideration. Zhang et al. [18] investigated the joint optimization of MT service frequency and the fare to maximize a weighted sum of profit and consumers' surplus. Sun and Szeto [15] evaluated the effectiveness of sectional fares in MT by solving a bilevel program that jointly determines the fares and frequency setting. These approaches do not apply to a multimodal setting due to the exponential blowup in computation time when including multiple modes.

Modeling travelers' mode choice is a critical technical challenge of optimizing multimodal networks. The latent travel demand is realized only when the prices and the quality of service (QoS) are both satisfying. In general, this choice model is treated as a lower-level optimization in centralized decisions. Beheshtian et al. [3] studied the multimodal marketplace in which service providers bids for the use of road segments and travelers pay the roadway tolls by the clearing prices. In addition, transportation agencies may incentivize travelers to use one mode of transport over the others. Brands et al. [4] designed a tradable permit scheme to regulate road transport externalities and empirically tested the existence of dynamic equilibrium on the mobility market. Characterizing those user equilibria is a computationally heavy task, regardless of the recourse under various tactical and operational plans. A simplified framework is proposed by Wischik [17]. It studied the multimodal mobility market's knock-on effect on travelers' route and mode choices by combining a discrete choice model with a multipath resource allocation model. By solving a single optimization problem, a transportation agency can infer the choice models and then control traffic flows externalized by travelers' prices. Leveraging this idea, this work focuses on solving a joint pricing and network design problem for multimodal mobility systems.

1.2 Contributions

To the best of our knowledge, this work is the first attempt to propose a scalable framework for solving the joint pricing and network design problem in multimodal mobility

networks using exact methods. This problem is notoriously difficult to solve even in single-mode transportation networks, and exacerbated by the complexities introduced in the multimodal setting. Two critical challenges in the multimodal setting are the growth of feasible combinations of modes and modeling user preferences. Our work computes the optimal prices for hybrid trips by a primal-dual approach developed in the market design literature. The computed prices ensure that each traveler is willing to pay for their utility-maximizing option. The optimal design of the marketplace achieves maximal social welfare. Besides, the multimodal mobility network needs to determine the frequency of MT routes and relocate MoD vehicle fleets to meet travel demand. Since the number of possible modes is enormous, we combine approximation algorithms, Benders decomposition, and cutting-plane approaches into a unified solver to obtain implementable algorithm for optimal pricing and operational planning. This approach's performance in facilitating the convergence of the two modes is experimentally evaluated using a real-world case study in Nashville, Tennessee.

The outline of this work is as follows. Section 2 describes the formulation of the joint pricing and network design problem. Section 3 presents a decomposition scheme that takes advantage of the special structure of the subproblems. Section 4 conducts a case study applying the developed approach to the city of Nashville. Finally, we draw some conclusions in Section 5.

2 Model

We study a joint pricing and network design problem in a multimodal mobility system. With a welfare maximization goal in its mind, a transportation agency is responsible for pricing multimodal mobility services and determining the routes and frequency setting of MT to meet demand for affordable and efficient mobility services.

2.1 A motivating example for multimodal systems

The following example (Figure 1) shows how to coordinate the operations of MT and MoD subject to travelers' behavior to maximize overall welfare.

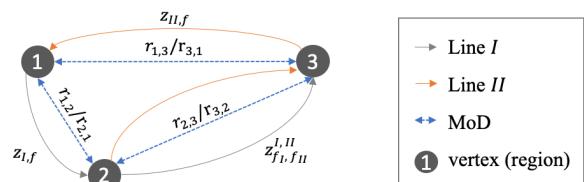


Figure 1. Pricing and network design problem in a simple multimodal network

The network has two one-way MT lines, *I* and *II*, and MoD is available on each edge. As MoD serves trip requests in

an on-demand fashion, it needs to rebalance idle vehicles to high demand locations [12]. The rebalancing flows between each pair of vertices i, j at steady-state are r_{ij} . There are two types of travelers θ_1 and θ_2 with different utility functions. Travel demand from vertex 1 to vertex 3 is split to $\lambda_{1,3}(\theta_1)$ and $\lambda_{1,3}(\theta_2)$, respectively. Travelers of type θ_1 are price-sensitive and travelers of type θ_2 are QoS-sensitive.

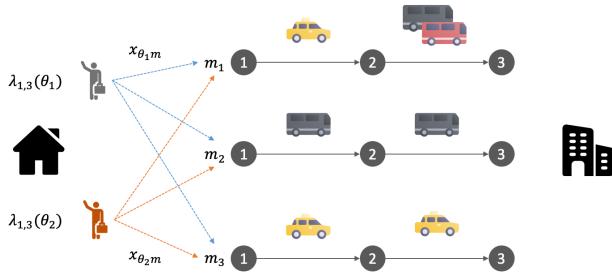


Figure 2. Travelers’ choice model in a multimodal network

Travelers have unified access to use a mix of two mobility services through a trip planner. For example, each traveler from vertex 1 to vertex 3 are provided with three options (called “mode” throughout this work): (a) taking low-frequency MT Line I, (b) taking MoD, or (c) a multimodal mode, i.e., taking MoD to vertex 2 and transfer to MT Line I or Line II, which collectively arrive at a high frequency (Figure 2). We define the time between MT vehicles past a given point as the headway, and the MT service frequency in unit of the average number of vehicles per unit time is the inverse of the headway. Assume that Line I and Line II operate at frequencies of f_I and f_{II} , respectively. If a traveler chooses mode (c), she will take the first arriving MT vehicle at the vertex 2. We model this option by creating a bundle of two lines from which each traveler chooses Line $i \in \{I, II\}$ by a probability of $f_i/(f_I + f_{II})$. Since mode (b) is expensive and mode (a) may not meet travelers’ QoS requirements, improving the competitiveness of mode (c) when appropriate is the primary goal of this work.

After the system presents the aforementioned modes and their corresponding prices and QoS, each traveler chooses the mode m that maximizes the $(v_m - p_m)$, $m \in \{m_1, m_2, m_3\}$, where v_m is the valuation and p_m is the charged price (including both MT fares and MoD fares). Without integrated multimodal access, θ_1 travelers tend to take Line I and θ_2 travelers tend to choose MoD. As a result, Line I is overloaded, and the MoD vehicle fleet stagnates at vertex 3.

From the transportation agency’s perspective, the goal is to create a multimodal market in which the market clearance prices for all modes, including these hybrid modes, lead to maximal social welfare. This work focuses on frequency setting and rebalancing decisions in addition to pricing is because frequency setting is the most fundamental tactical decision in MT and rebalancing is the unique and central

task in MoD operations. In this market design procedure, an obvious trade-off is the resource allocation between MT and MoD. Providing high-frequency MT services on Line I enhances the QoS while increasing its operating costs. In this example, it is potentially beneficial to subsidize some travelers to use the multimodal mode (c) to gain access to the low-cost Line II. On the other hand, they should not be over-subsidized to choose the MoD mode (b). Since the system’s welfare is measured by the difference between utilities received from serving travel demand and the incurred operating costs, this approach can redistribute realized traffic streams to improve MT’s utilization while reducing the cost of operating the MoD vehicle fleet.

2.2 Setting

In a general setting, the multimodal mobility network is represented by a directed graph $G = (V, E)$ consisting of n vertices corresponding to pickup/dropoff locations. There are L transit lines where each line is a sequence of edges $\{e_{ij}\}$, which joins a sequence of distinct vertices. MoD is assumed to be available on each edge. A traveler from the starting vertex s to the terminal vertex t is provided with a menu of available hybrid modes consisting of MT and MoD denoted as \mathcal{M}_{st} . She employs a mixed strategy, i.e., the trip planner provides her a randomized option from the chosen modes to maximize her expected utility. Let $[K] := \{1, 2, \dots, K\}$ throughout this work and $[K]^2$ denote the Cartesian product of sets. The notation used throughout this work is summarized in the nomenclature Table 4 in Appendix A.

2.2.1 Frequency setting in bundled service. The service frequency on each line is a vector $\mathbf{f} \in \mathbb{Z}^{|L|}$. There exists a natural lower bound such that, when setting the frequency at this bound, MT becomes noncompetitive for all travelers compared with other options. We rescale the frequency such $f_\ell \in [F]$ for all $\ell \in [L]$. The transit network’s configuration is fixed throughout the frequency setting decision. Let $z_{\ell,f}$ be a binary variable corresponding to line ℓ being set to frequency f , i.e., $z_{\ell,f} = 1$ if the frequency f is turned on and $z_{\ell,f} = 0$ if the frequency is off. Line ℓ ’s vehicle capacity is \mathcal{V}_ℓ . The unit setup cost incurred for raising the frequency of a transit line is o_ℓ so the total setup cost of MT is $\sum_{\ell \in [L], f \in [F]} o_\ell z_{\ell,f}$.

When a set of MT lines L_{ij} passes through the same edge e_{ij} , the system creates a set of bundled services $\mathcal{B}_{ij} := \text{comb}(L_{ij})$. Each bundled service $b_{ij} \in \mathcal{B}_{ij}$ includes one or more MT lines. The probability of a traveler who chooses b_{ij} taking line $\ell \in b_{ij}$ is proportional to its frequency $f_\ell / \sum_{\ell' \in b_{ij}} f_{\ell'}$, i.e., on a first-come-first-serve (FCFS) basis. The on/off decision for each bundle is denoted a binary variable $z_\psi^{b_{ij}}$ where ψ_ℓ is the corresponding frequency in ψ at its $\ell \in b_{ij}$ entry. In example of Figure 1, $b_{23} = \{I, II\}$ and $\psi = \{f_I, f_{II}\}$. $z_\psi^{b_{ij}} = 1$ means that the system provides bundled service b_{ij} to all modes

passing the edge e_{ij} . As each traveler is allowed to transfer at most once, a preprocess truncates the set of modes \mathcal{M}_{st} when it is expanded to include bundled services at varying frequencies.

2.2.2 Demand model. We consider a large market, fluid scaling of the system. Each vertex s has exogenous, non-atomic flows $\{\lambda_{st}, t \in V/\{s\}\}$. Heterogeneous travelers of type $\theta \in \Theta$ have valuations drawn from a discrete distribution q . Thus the flows are split as $\lambda_{st}(\theta) := \lambda_{st} q_{st}(\theta)$. Travelers from a starting vertex s to a terminal vertex t plan their trips from options (called “hybrid modes”) $m \in \mathcal{M}_{st}$ defined by a sequence of edges, each specifying the mode of transport as $\{(s, i_1), (i_1, i_2), \dots, (i_k, t)\}$.

Observing existing services and their corresponding QoS and prices, travelers make their trip plans (possibly using a multimodal option). A traveler from vertex s to vertex t chooses the mode $m \in \mathcal{M}_{st}$ that maximizes their utility $v_{\theta m} - p_{\theta m}$. The valuation for trip (s, t) is type- and mode-dependent, and denoted as $\mathbf{v}(\theta) \in \mathbb{R}^{|\mathcal{M}_{st}|}$. $x_{\theta m} \in [0, 1]$ denotes the probability that a traveler of type θ traveling from s to t chooses mode $m \in \mathcal{M}_{st}$ and $\sum_{m \in \mathcal{M}_{st}} x_{\theta m} = 1$.

Since MoD and MT have different pricing practices, their operational costs are calculated separately in each hybrid mode $m \in \mathcal{M}_{st}$. The total operational cost of hybrid mode m is given by $c_m = \sum_{e_{ij} \in E_m} c_{ij} + \sum_{\ell \in L: \ell \sim m} c_\ell$, where $\ell \sim m$ means that mode m uses Line ℓ . Payments for MoD are distance-based, and we let c_{ij} denotes the operating cost of carrying passengers by MoD from i and j . Payments for MT are made per-use. The chosen Line ℓ 's fares depend on the od pair $s - t$ and the operational cost of Line ℓ is c_ℓ .

2.2.3 MILP formulation. This section proposes a generic MILP formulation for the joint pricing and network design problem in (1). The transit network design problem determines the MT frequency f for all $\ell \in [L]$ (\mathbf{z}) and rebalancing flows of MoD (\mathbf{r}). The objective is maximizing the overall welfare collected from operating the multimodal mobility system. The welfare measures the difference between the utility of serving all travel demand, operational costs of hybrid modes, MoD's rebalancing costs, and MT's setup costs.

The constraints are categorized as follows:

1. (1a) are the capacity constraints on each edge. This constraint guarantees that the total MT flow passing through edge e_{ij} (left-hand side) is no larger than the total capacity carried by MT lines (right-hand side). The flow distribution for each MT line is satisfied by enforcing a FCFS rule in trip assignment.
2. (1b) - (1c) describe the relationship between MT's line setups and bundled service setups. A bundle b_{ij} is accessible if and only if all lines included $\ell \in b_{ij}$ are operated at the frequency levels of ψ .
3. (1d) - (1e) are the ranking constraints for MT frequency setting. When the system determines Line ℓ 's frequency

to be at level ℓ , low-frequency options are turned on as travelers tend to choose a high-QoS service when paying the same price.

4. (1f) is the budget constraint for MT. The cycle time of Line ℓ is T_ℓ and the total number of MT vehicles is \mathcal{R} .
5. (1g) ensures the flow-balance (via rebalancing) of the MoD fleet.
6. (1h) is a linkage constraint in mode generation. A mode m is available if each $\ell \sim m$ is open. It can be relaxed to $Cx_{\theta m} \leq z_{\ell, f}$ with arbitrary $C > 0$.
7. (1i) are the demand conservation constraints.

$$\begin{aligned} \max_{\mathbf{z}, \mathbf{x}, \mathbf{r}} \quad & \sum_{\substack{(s,t) \in [n]^2 \\ \theta \in \Theta_{st}}} \lambda_{st}(\theta) \sum_{m \in \mathcal{M}_{st}} (v_{\theta m} - c_m) x_{\theta m} - \sum_{i,j} c_{ij} r_{ij} \\ & - \sum_{\substack{\ell \in [L] \\ f \in [F]}} o_\ell z_{\ell, f} \\ \text{s.t.} \quad & \sum_{\substack{(s,t): e_{ij} \in (s,t) \\ \theta \in \Theta, m \in \mathcal{M}_{st}}} \lambda_{st}(\theta) x_{\theta m} \leq \sum_{\substack{(\ell,f) \in (b_{ij}, \psi), \\ \forall \psi \in [F]^{|\mathcal{B}_{ij}|}, \\ \forall b_{ij} \in \mathcal{B}_{ij}}} \mathcal{V}_\ell z_{\ell, f}, \forall e_{ij} \in E \\ (1a) \end{aligned}$$

$$z_{\psi}^{b_{ij}} \leq z_{\ell, f}, \quad \forall (\ell, f) \in (b_{ij}, \psi), \forall \psi \in [F]^{|\mathcal{B}_{ij}|}, \\ \forall b_{ij} \in \mathcal{B}_{ij}, \forall e_{ij} \in E \quad (1b)$$

$$z_{\psi}^{b_{ij}} \geq \sum_{(\ell, f) \in (b_{ij}, \psi)} z_{\ell, f} - |\mathcal{B}_{ij}| + 1, \quad \forall \psi \in [F]^{|\mathcal{B}_{ij}|}, \\ \forall b_{ij} \in \mathcal{B}_{ij}, \forall e_{ij} \in E \quad (1c)$$

$$z_{\ell, f} \leq z_{\ell, f'}, \quad \forall f \geq f', \forall \ell \in [L] \quad (1d)$$

$$z_{\ell, f} \geq z_{\ell, f''}, \quad \forall f \leq f'', \forall \ell \in [L] \quad (1e)$$

$$\sum_{\ell \in [L]} \sum_{f \in [F]} T_\ell z_{\ell, f} \leq \mathcal{R} \quad (1f)$$

$$\begin{aligned} \sum_{j \in [n]} \left(r_{ji} + \sum_{\substack{\theta \in \Theta \\ m: (j,i) \in E_m}} \lambda_{st}(\theta) x_{\theta m} \right) = \\ \sum_{k \in [n]} \left(r_{ik} + \sum_{\substack{\theta \in \Theta \\ m: (i,k) \in E_m}} \lambda_{st}(\theta) x_{\theta m} \right), \quad \forall i \in [n] \quad (1g) \end{aligned}$$

$$x_{\theta m} \leq z_{\ell, f}, \quad \forall (\ell, f) \sim m, \forall \theta \in \Theta, \forall \ell \in L, \forall f \in [F] \quad (1h)$$

$$\sum_{m \in \mathcal{M}_{st}^M} x_{\theta m} \leq q_{st}(\theta), \quad \forall \theta \in \Theta, \forall (s, t) \in [n]^2 \quad (1i)$$

$$z_{\psi}^{b_{ij}} \in \{0, 1\}, \quad \forall \psi \in [F]^{|\mathcal{B}_{ij}|}, \forall b_{ij} \in \mathcal{B}_{ij}, \forall e_{ij} \in E \quad (1j)$$

$$z_{\ell, f} \in \{0, 1\}, \quad \forall f \in [F], \forall \ell \in [L] \quad (1k)$$

$$0 \leq x_{\theta m} \leq 1, \quad \forall \theta \in \Theta, m \in \mathcal{M}_{st}, (s, t) \in [n]^2 \quad (1l)$$

$$r_{ij} \geq 0, \quad \forall e_{ij} \in E \quad (1m)$$

This formulation overcomes two main technical challenges in prior work. First, the transit network's design subject to

travelers' mode choice is modeled by a primal-dual approach. This approach conserves the user equilibrium and avoids the computational tractability issue in the conventional bilevel programming approach [15]. Second, considering p_m as primal variables under the demand elasticity leads to a nonlinear mathematical program. We convert the pricing decisions into a dual problem (explained in Section 2.2.4) and simplify the formulation by linearization tricks. This model can be extended to including the MoD platform as a separate decision-maker [2]. The privately-owned MoD company is better off by participating in this multimodal network to gain a broader basis of customers.

2.2.4 Dual prices for hybrid modes. The LP-relaxation of (1) computes the dual prices $p_{\theta m}$ for each mode $m \in \mathcal{M}_{st}$. The dual variables for the MILP formulations are as follows:

1. Edge-base price: p_{ij} for each $e_{ij} \in E$ corresponding to (1a).
2. Bundled price: $\beta_{\psi}^{b_{ij}}$ and $\eta_{\psi}^{b_{ij}}$ for $(b_{ij}, \psi) \in (\mathcal{B}_{ij}, |F|^{b_{ij}})$ and $e_{ij} \in E$ corresponding to (1b) and (1c).
3. Line price: $w_{\ell, f, f'}$ with $f \leq f'$, $f, f' \in [F]$ corresponding to (1d) and (1e).
4. MT vehicle setup price: u corresponding to (1f).
5. MoD connection price: γ_i for $i \in [n]$ corresponding to (1g).
6. Line setup cost: $\zeta_{\theta m \ell f}$ for all $(\ell, f) \sim m, \theta \in \Theta, \ell \in L, f \in [F]$ corresponding to (1h).
7. Path-base utility: $v_{\theta, st}$ for all $\theta \in \Theta, m \in \mathcal{M}_{st}, (s, t) \in [n]^2$ corresponding to (1i).

Traveler of type θ traveling from s to t is charged a price $p_{\theta m}$ computed x as follows:

$$p_{\theta m} = \sum_{e_{ij} \in E_m} \left[p_{ij} + \beta_{\psi}^{b_{ij}} + \eta_{\psi}^{b_{ij}} + \gamma_i \right] + w_{\ell, f, f'} + u + \zeta_{\theta m \ell f}, \\ m \sim (b_{ij}, \psi), \quad (2)$$

where $m \sim (b_{ij}, \psi)$ means that the bundled service on edge e_{ij} is included in the hybrid mode m ; E_m and V_m are edges and vertices trespassed by this mode.

It is worth mentioning that the dual prices are not well-defined in the original MILP (1). Section 3 justifies the definition of dual prices with a decomposition scheme. After fixing the frequency setting of MT, the travelers' mode choices are formulated as an LP [17] and the pricing problem regarding the realized network flow is precisely defined. Solving these subproblems will return part of the dual price and feasibility/optimality cuts. The algorithm then resolves the master problem to approach a feasible frequency-setting plan for MT such that the total welfare is maximized. In the final step, extra setup costs for MTs are added evenly for all users, reflecting the equal-sharing practice in public transportation sectors. This decomposition is also computationally efficient as the subproblems are solvable in polynomial time.

Proposition 2.1. Under the optimal dual prices p_{θ} , travelers of θ from vertex s to vertex t in graph G will adopt a mixed strategy $x_{\theta m}^*$ over available modes $m \in \mathcal{M}_{st}$ such that

$$x_{\theta m} \in \arg \max \left\{ \sum_{\substack{\theta \in \Theta, \\ m \in \mathcal{M}_{st}}} (v_{\theta m} - p_{\theta m}) x_{\theta m} \right\}, \forall (s, t) \in [n]^2.$$

This mixed strategy for choosing modes can be implemented by a randomized assignment policy as follows. Observing that each optimal mode m gives type θ travelers the same utility, the system can randomly provide mode m with a probability of $x_{\theta m}^*$ and keep track of the past assignment. In the following assignment, the system matches the empirical distribution of mode assignments with these $x_{\theta m}^*$.

3 Pricing services with decomposition

The joint optimization of transit network design and pricing formulated above is notoriously hard to solve directly regarding the large-scale urban networks. This section starts with a negative result for the integrality gap of the original MILP formulation. To overcome the computational challenge, we design a bilevel decomposition scheme such that the dual prices of modes are computed in recourse. Finally, we shed light on the computational complexity of the proposed algorithm.

3.1 Overview of decomposition framework

The dual prices can be computed from the constraints (1a)–(1f) in the LP-relaxation of (1). Nevertheless, since both the frequency setting z and mode choice x are integer variables, we show that the integrality gap of the original formulation is arbitrarily large.

In a simple network of two vertices, $L = 1$, and F equals a constant \hat{F} , we consider only MT in a round trip. Without loss of generality, we assume the optimal frequency is f^* such that $\sum_{f \in [f^*]} T_I z_{I,f} \leq \mathcal{R}$ and $\sum_{f \in [f^*+1]} T_I z_{I,f} > \mathcal{R}$. Hence, $z_{I,1} = \dots = z_{I,f^*} = 1$ and $z_{I,f^*+1} = \dots = z_{I,F} = 0$. We can denote the mode choice variable x_f in this case as the modes only differ in frequency. The setting up cost is $f^* o_\ell$ and the total capacity is $f^* V_I$. The constraint (1a) determines x_f as $x_f > 0$ for the maximal $v_{\theta f} - c_f$ for $f \in [f^*]$ and equals 0 for others. We denote the objective value as v^* . We can find fractional solutions for its LP-relaxation counterpart as follows. We set $z_{I,f} = f^*/\hat{F}$ for all $f \in [\hat{F}]$ which is feasible for the constraints in (1). The total capacity and the setup costs are equal to the original problem while all modes are open. x_f is expanded to $f \in [\hat{F}]$ and the integrality gap $\max\{v_{\theta f} - c_f\}_{f \in [f^*+1, \hat{F}]} / v^*$ is unbounded.

The large integrality gap motivates the development of a decomposition framework in Figure 3. The original MILP is computationally intractable as the dimension of binary

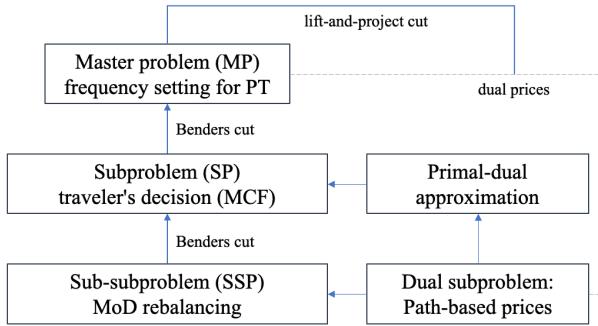


Figure 3. Decomposition framework for joint pricing and transit network design problem

variables z and x are in the scale of $2^{L \times F \times |E|}$ and the resulting number of constraints explodes.

We decompose the MILP formulation to two levels of subproblems that solve z , x , and r sequentially.

3.1.1 Subproblem: traveler choice. We start with subproblems for any transit network design, which are solved through fully polynomial time approximation schemes. Let $y_{\theta m} = \lambda_{st}(\theta)x_{\theta m}$ and the values from the subproblem $\hat{y}_{\theta m} = \lambda_{st}(\theta)\hat{x}_{\theta m}$. Given a MT network design \mathbf{z} and travelers' choice \mathbf{y} , the sub-subproblem $SUB(\mathbf{z}, \mathbf{y})$ (the MoD rebalancing problem) is a min-cost flow problem:

$$SUB(\mathbf{z}, \mathbf{y}) = \min_{\mathbf{r}} \sum_{i,j} c_{ij} r_{ij} \quad (3)$$

$$\text{s.t. } \sum_{j \in [n]} \left(r_{ji} + \sum_{\substack{\theta \in \Theta \\ m: (j,i) \in E_m}} \hat{y}_{\theta m} \right) = \sum_{k \in [n]} \left(r_{ik} + \sum_{\substack{\theta \in \Theta \\ m: (i,k) \in E_m}} \hat{y}_{\theta m} \right), \quad \forall i \in [n] \quad (3a)$$

$$r_{ij} \geq 0, \quad \forall e_{ij} \in E \quad (3b)$$

The dual of the MoD rebalancing problem is

$$\max_{\boldsymbol{\gamma}} \sum_{i \in [n]} \left(\sum_{\substack{\theta \in \Theta \\ m: (i,k) \in E_m}} \hat{y}_{\theta m} - \sum_{\substack{\theta \in \Theta \\ m: (j,i) \in E_m}} \hat{y}_{\theta m} \right) \gamma_i \quad (4)$$

s.t. $\gamma_i - \gamma_j \leq c_{ij}, \quad \forall e_{ij} \in E$

γ_i unrestricted, $\forall i \in [n]$

The feasible set of the dual problem is independent of \mathbf{x} . Solving the subsubproblem adds either feasible cuts or

optimal cuts to the subproblem $P(\mathbf{z})$ in (5):

$$P(\mathbf{z}) = \max_{\mathbf{y}} \sum_{\substack{(s,t) \in [n]^2 \\ \theta \in \Theta_{st} \\ m \in \mathcal{M}_{st}}} (v_{\theta m} - c_m) y_{\theta m} + \phi \quad (5)$$

$$\text{s.t. } (1a), (1h), (1i) \quad (5a)$$

$$\sum_{i \in [n]^e} \left(\sum_{\substack{\theta \in \Theta \\ m: (i,k) \in E_m}} y_{\theta m} - \sum_{\substack{\theta \in \Theta \\ m: (j,i) \in E_m}} y_{\theta m} \right) \gamma_i^e \geq \phi \quad (5b)$$

$$\sum_{i \in [n]^u} \left(\sum_{\substack{\theta \in \Theta \\ m: (i,k) \in E_m}} y_{\theta m} - \sum_{\substack{\theta \in \Theta \\ m: (j,i) \in E_m}} y_{\theta m} \right) \gamma_i^u \geq 0 \quad (5c)$$

$$0 \leq y_{\theta m} \leq \lambda_{st}, \quad \forall \theta \in \Theta, m \in \mathcal{M}_{st}, (s,t) \in [n]^2 \quad (5d)$$

where γ^e are extreme points of (4) and γ^u are unbounded directions of (4).

The traveler choice subproblem (5) is a variation of the weighted multicommodity flow problem (WMFP). If each traveler adopts a deterministic mode choice strategy, it is reduced to an integral MFP problem, which is a well-known NP-complete problem (MAX SNP-hard). Compared to a direct LP approach, the subproblem computation's main bottleneck is the number of commodities $|\Theta| \cdot |\mathcal{M}_{st}|$ in multimodal mobility networks. A primal-dual approximation algorithm (Algorithm 1) is used to simultaneously compute travelers' routes and their corresponding prices. The essence of the algorithm is repeated solve shortest-path problems on the dual solutions to WMFP such that the approximation ratio is independent of the number of commodities [6].

The dual problem of the fractional subproblem (5) is as follows:

$$D(\mathbf{z}) = \min \sum_{\substack{(\ell,f) \in (b_{ij}, \psi), \\ \forall \psi \in [F]^{[b_{ij}]}, \\ \forall b_{ij} \in \mathcal{B}_{ij}}} V_{\ell} \hat{z}_{\ell,f} p_{ij} + \sum_{\substack{\theta \in \Theta \\ m \in \mathcal{M}_{st} \\ (\ell,f) \sim m}} \hat{z}_{\ell,f} \zeta_{\theta m \ell f} \sum_{\substack{\theta \in \Theta \\ (s,t) \in [n]^2}} \lambda_{st}(\theta) (\lambda_{st}(\theta) v_{\theta,st} + \alpha_{\theta m}) \quad (6)$$

$$\text{s.t. } \sum_{e_{ij} \in E} p_{ij} + \sum_{(\ell,f) \sim m} \zeta_{\theta m \ell f} + \alpha_{\theta m} + v_{\theta,st} \geq v_{\theta m} - c_m, \quad \forall \theta \in \Theta, \forall m \in \mathcal{M}_{st}, \forall (s,t) \in [n]^2 \quad (6a)$$

$$A_{\phi}^T \boldsymbol{\beta} \geq 0 \quad (6b)$$

$$\boldsymbol{p}, \boldsymbol{\zeta}, \boldsymbol{v}, \boldsymbol{\alpha}, \boldsymbol{\beta} \geq 0 \quad (6c)$$

where $\boldsymbol{p}, \boldsymbol{\zeta}, \boldsymbol{v}, \boldsymbol{\alpha}, \boldsymbol{\beta}$ are dual variables corresponding to constraints (1a), (1h), (1i), (5d), (5b) - (5c), respectively. A_{ϕ} is the coefficient matrix of (5b) and (5c). We denote the objective function of the dual problem as $g(\boldsymbol{p}, \boldsymbol{\zeta}, \boldsymbol{v}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

The approximation algorithm for WMCF (Algorithm 1) provides a $(1 - \epsilon)$ approximation, where ϵ is the error tolerance. The number of constraints in (5) is upper-bounded by

Algorithm 1: Approximation subroutine for traveler choice problem

Result: Dual prices related to traveler choice and benders cuts (7b), (7c).

Initialization: frequency setting \hat{z} in a network;
 $\mathbf{y} = 0, \mathbf{p}, \boldsymbol{\zeta}, \mathbf{v}, \boldsymbol{\alpha}, \boldsymbol{\beta} = \delta$;

while new cuts are generated from (4) **do**

- | **for** $(s, t) \in [n]^2, i \in [\log_{1+\epsilon} \frac{1+\epsilon}{\delta}], j \in [K]$ **do**
- | | $P \leftarrow \min_{D(\hat{z})} \mathcal{P}_j$;
- | | **while** $p(\mathcal{P}_i^{st}) < \min\{1, \delta(1 + \epsilon)^i\}$ **do**
- | | | $V^* \leftarrow \min_{e_{ij} \in P} V_\ell \hat{z}_{\ell,f}$;
- | | | $y_{\theta m} \leftarrow y_{\theta m} + V^*$;
- | | | $\forall e_{ij} \in P, p_{ij} \leftarrow p_{ij}(1 + \frac{\epsilon V^*}{V_\ell \hat{z}_{\ell,f}})$ Update P .
- | | **end**
- | **end**
- Solving subsubproblem $SUB(\mathbf{z}, \mathbf{y})$ with $\hat{\mathbf{y}}$ and add either feasibility or optimality cuts.

end

Algorithm 2: Algorithm for joint pricing and transit network design problem

Result: Frequency setting for MT, rebalancing flows for MoD, and prices.

Initialization: $\mathbf{z} = 0$ for all $z_{\ell,f}$ and z_ψ^{bij} ;

for $\ell \in [L]$ **do**

- | Update $Z_\ell = [z_{\ell,f}]_{f \in [F]}$ such that, if $z_{\ell,f} > 0$, set $z_{\ell,f'} = 1$ for all $f' < f$ and $z_{\ell,f'} = 0$ for all $f' > f$.
- | **for** $f \in [F]$ **do**
- | | Solve the LP-relaxation of (7) to obtain Z^* ;
- | | Generate lift-and-project cuts with Z^* for $Z \in \mathbb{R}^{L \times F}$ for (7) (Appendix A);
- | | Update all bundled services z_ψ^{bij} .
- | **end**
- Solve subproblem (5) by Algorithm 1 and generate either feasibility or optimality cuts;

end

$\bar{m} := n^2 |\Theta| LF$. Let $\delta = \epsilon / \bar{m}^2$ be the initial value of dual variables in (6). The number of commodities is $K = |\mathcal{M}_{st}| \cdot |\Theta|$. The paths of commodity i for each (s, t) is denoted as \mathcal{P}_i^{st} . We assume that, using the dual values in (6), there exists a shortest-path oracle that finds the path with minimum cost for \mathcal{P}_i^{st} in polynomial time as $\min_{D(\hat{z})} \mathcal{P}_i^{st}$. The shortest path computed using the dual values is $p(\mathcal{P}_i^{st})$. Dual prices for each commodity are derivatives from the subroutine in Algorithm 1.

3.1.2 Master problem: frequency setting. The top level of the original formulation (1) is a frequency setting problem with bundled services:

$$\max_{\mathbf{z}} \tau - \sum_{\substack{\ell \in [L] \\ f \in [F]}} o_\ell z_{\ell,f} \quad (7)$$

$$\text{s.t. (1b)} - \text{(1f)}, \text{(1j)} - \text{(1k)} \quad (7a)$$

$$g(\mathbf{p}^e, \boldsymbol{\zeta}^e, \mathbf{v}^e, \boldsymbol{\alpha}^e, \boldsymbol{\beta}^e) \geq \tau \quad (7b)$$

$$g(\mathbf{p}^u, \boldsymbol{\zeta}^u, \mathbf{v}^u, \boldsymbol{\alpha}^u, \boldsymbol{\beta}^u) \geq 0 \quad (7c)$$

$(\mathbf{p}^e, \boldsymbol{\zeta}^e, \mathbf{v}^e, \boldsymbol{\alpha}^e, \boldsymbol{\beta}^e)$ are extreme points of the dual subproblem (6) and $(\mathbf{p}^u, \boldsymbol{\zeta}^u, \mathbf{v}^u, \boldsymbol{\alpha}^u, \boldsymbol{\beta}^u)$ represent the unbounded directions.

We use the lift-and-project cutting-plane method for solving the binary optimization in (7). For completeness, we describe this method in Appendix B. The final algorithm is summarized in Algorithm 2.

4 Numerical experiment

4.1 Experiment setup and data description

We conduct a numerical experiment that retrofits the current MTA bus system in Nashville, Tennessee with MoD as a first or last-mile connection. Residents in Nashville, like

most metropolitan areas in the United States, have unequal access to transit systems. Therefore, determining the efficacy of extending the MT structure with the assistance of MoD connections is a meaningful activity.

Table 1 summarizes the data used in the experiment. The trip demand data estimates the origin-destination matrix at the census tract level of daily commuting trips, which is used for λ_{st} . There are two types of travelers $\theta \in \{1, 2\}$ corresponding to price-sensitive and QoS-sensitive travelers. We can include more types of travelers by expanding set Θ in other applications. For all o-d pairs, $q_{st}(1) = q_{st}(2) = 0.5$. For each kind of traveler, the utility functions follow

$$v_{\theta m} = \begin{cases} p_{\theta m}^2 + \sum_{e_{ij} \in E_m} (c_f \cdot f_e - c_l \cdot l_e) & \theta = 1 \\ p_{\theta m} + \sum_{e_{ij} \in E_m} (c_f \cdot f_e^2 - c_l \cdot l_e) & \theta = 2 \end{cases}$$

Each candidate mode set \mathcal{M}_{st} includes all combinations of bus routes with no more than two transfers and MoD as first- or last-mile connections. We truncated \mathcal{M}_{st} in long-distance trips if there exist more than five hybrid combinations of MoD and MT, because travelers are not capable of evaluating a large size of alternative options. For each combination, travelers consider all possible frequencies of MT as in Proposition 2.1.

The MTA bus routes and trip demand data are shown in Figure 4. The size of optimization in (1) depends on fixed network properties such as the scale of MT network and the dimension of trip demand od matrix, as well as user-specific design parameters such as the MT frequency levels F or the total number of MT vehicles \mathcal{R} . Table 2 summarizes the benchmark model used throughout the remaining analysis.

Table 1. Summary of numerical experiment data sources

Item	Datasource	Description
Trip demand	Access Nashville [10]	Origin-destination movement matrix
Bus route	Nashville MTA [11]	GIS and GTFS
Cost parameters	Operational cost [2]	MT cycling time, setup cost, Rebalancing cost of MoD

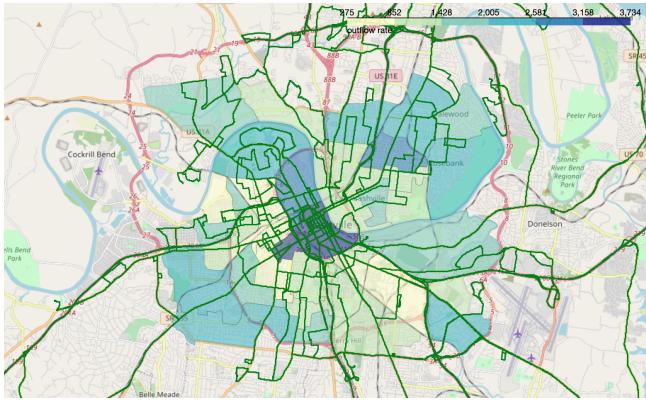


Figure 4. Nashville MTA bus system (green lines) and origin-destination flow (outflow) data (chloropleth map)

Table 2. Summary of parameters in Nashville MTA case study

Item	Notation	Value
Bus lines	L	48
Bus frequency level	F	5
Number of vertices (census tracts)	$ V $	47
Number of o-d pairs with $\lambda_{st} > 0$	(s, t)	836
Number of modes	$ \mathcal{M}_{st} $	1 to 5
Number of MT frequency setting variables	$z_{t,f}$	240
Number of bundled variable	z_{ψ}^{bij}	320
Number of choice variables	$x_{\theta m}$	139, 744
Number of MoD rebalancing variables	r_{ij}	664
Total number of buses	\mathcal{R}	20
Vehicle capacity	\mathcal{V}_t	75

4.2 Optimal prices and transit network design in Nashville, TN

The optimal frequency setting of the MTA bus system is shown in Figure 5. The lines connecting suburban regions

to the downtown area are expected to run more frequently than the downtown area. MoD carries travelers to these lines and aggregate flow via these main corridors. This result, on the other hand, has limitations since o-d trip demand data predicts everyday commuting trips predominantly. Furthermore, the proportion of the population able to engage in this multimodal mobility service is not specified.

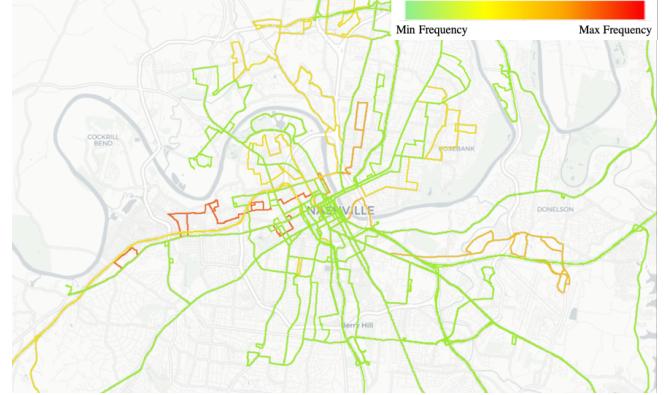


Figure 5. Optimal frequency setting for Nashville MTA bus system

The computational efficiency of the proposed algorithm is shown in Table 3. Two benchmarks that we compared with are state-of-the-art MIP solvers (Gurobi 9.0 on a 1.1 GHz Dual-Core processor). The runtime of the proposed algorithm is significantly shorter for both cases. When the size of the network or the frequency level increases, our approach outperforms the MIP solver as the proposed algorithm only solves a sequence of LP. Besides, the MIP solvers cannot return the dual prices directly because it is not well-defined. In contrast, as the subproblems are multicommodity flow problems with rebalancing constraints, dual prices are well-defined for each (s, t) path in the multimodal network.

4.3 Sensitivity analysis

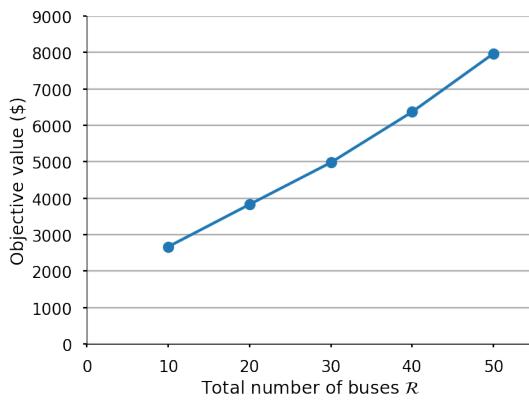
Since several operational parameters of the status-quo system are unknown, we conduct a sensitivity analysis in Figure 6. The most critical design parameters for MT are the total number of buses \mathcal{R} and vehicular capacities V_t . We test the optimal design's sensitivity with regard to these two parameters.

As the MT system's available resources increase, the overall welfare increases significantly as more travelers can save money by using MT facilities. Note that we do not consider the purchase costs of additional MT buses, but assume the current system's bus fleet size may vary from Table 2. On the other hand, it is not inherently a profitably practice to use large-capacity vehicles as we can raise the frequencies on specific routes to offset the reduced per-vehicle capacity. This coincides with the recent development in operating small-size flexible MT.

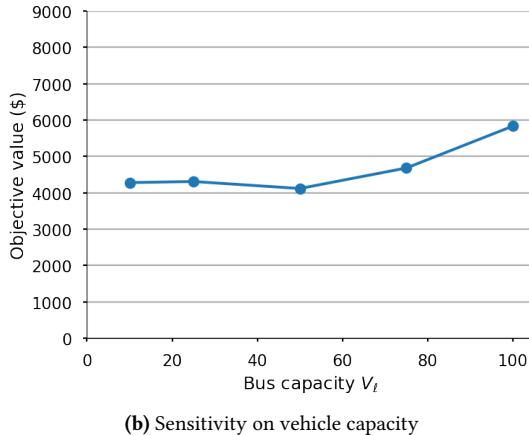
Table 3. Computational results for the Nashville MTA case study

Method	Runtime (second)		UB-LB gap (%)		Objective	
	Inst. 1	Inst. 2	Inst. 1	Inst. 2	Inst. 1	Inst. 2
MIP solver with conservative cut generation	3600*	2832	20.74%	0.0%	4246	5932
MIP solver with automatic cut generation	3600*	2422	21.93%	0.0%	4050	5569
Benders decomposition and approximation algorithm	1232	921	-	-	4815	5208

- Instance 1 is from Table 2.
- Instance 2 changes the frequency level $F = 3$ and types of customers $|\Theta| = 3$.
- Runtime with * means the solver hits the time limits and stops early.
- $UB - LB$ gap is the difference between the upper bound and lower bound of the solutions.



(a) Sensitivity on total vehicle number



(b) Sensitivity on vehicle capacity

Figure 6. Sensitivity analysis of the optimal value regarding MT parameters

5 Conclusion

As a building block for any multimodal mobility ecosystem, this work studies a joint pricing and transit network design problem. A unified MILP framework solves the frequency setting of MT in the master problem and solves the price and MoD rebalancing in subproblems sequentially. This primal-dual approach is substantially easier to implement and more efficient in computation compared to the traditional bilevel program with equilibrium constraints. With the advent of a shared, on-demand, and diverse mobility economy, a case study in Nashville, TN, highlights the ability of using this approach in solving real-world challenges.

The limitations of this work include the type of choice model for travelers that can be accommodated by this linear model, even though it admits the MNL, which is the most common model of discrete choice in transportation. Furthermore, we presume that information in the system is publicized such that each traveler is completely rational in making their travel plans, which might not be the case in reality. We leave these questions for future work.

Acknowledgement

This material is based upon work partially supported by the National Science Foundation under Grant No. CNS-1952011 and CMS-1839346.

References

- [1] Javier Alonso-Mora, Samitha Samaranayake, Alex Wallar, Emilio Frazoli, and Daniela Rus. 2017. On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. *Proceedings of the National Academy of Sciences* 114, 3 (2017), 462–467.
- [2] Siddhartha Banerjee, Chamsi Hssaine, Samitha Samaranayake, and Ragavendran Gopalakrishnan. 2020. Welfare-revenue tradeoffs for a mobility marketplace. *Working paper, Cornell Univiersity, Ithaca, NY* (2020).
- [3] Arash Beheshtian, R Richard Geddes, Omid M Rouhani, Kara M Kockelman, Axel Ockenfels, Peter Cramton, and Wooseok Do. 2020. Bringing the efficiency of electricity market mechanisms to multimodal mobility across congested transportation systems. *Transportation Research Part A: Policy and Practice* 131 (2020), 58–69.
- [4] Devi K Brands, Erik T Verhoef, Jasper Knockaert, and Paul R Koster. 2020. Tradable permits to manage urban mobility: market design and experimental implementation. *Transportation Research Part A: Policy and Practice* 137 (2020), 34–46.
- [5] Benjamin G Edelman and Damien Geradin. 2015. Efficiencies and regulatory shortcuts: How should we regulate companies like Airbnb and Uber. *Stanford Technology Law Review* 19 (2015), 293.
- [6] Lisa K Fleischer. 2000. Approximating fractional multicommodity flow independent of the number of commodities. *SIAM Journal on Discrete Mathematics* 13, 4 (2000), 505–520.
- [7] Alejandro Henao and Wesley E Marshall. 2019. The impact of ride-hailing on vehicle miles traveled. *Transportation* 46, 6 (2019), 2173–2194.
- [8] Omar J Ibarra-Rojas, Felipe Delgado, Ricardo Giesen, and Juan Carlos Muñoz. 2015. Planning, operation, and control of bus transport systems: A literature review. *Transportation Research Part B: Methodological* 77 (2015), 38–75.

- [9] Qi Luo, Zhiyuan Huang, and Henry Lam. 2019. Dynamic Congestion Pricing for Ridesourcing Traffic: a Simulation Optimization Approach. In *Proceedings of Winter Simulation Conference*.
- [10] Tennessee Metro Government of Nashville & Davidson County. 2020. Nashville Transportation Demand Management. <https://www.nashville.gov/Planning-Department/Long-Range-Planning/Transportation-Planning.aspx>
- [11] Nashville MTA. 2020. Metropolitan Transit Authority Bus Routes (GIS). <https://data.nashville.gov/Transportation/Metropolitan-Transit-Authority-Bus-Routes-GIS-/rbzs-b4vk>
- [12] Marco Pavone, Stephen L Smith, Emilio Frazzoli, and Daniela Rus. 2012. Robotic load balancing for mobility-on-demand systems. *The International Journal of Robotics Research* 31, 7 (2012), 839–854.
- [13] Helen KRF Pinto, Michael F Hyland, Hani S Mahmassani, and I Ömer Verbas. 2019. Joint design of multimodal transit networks and shared autonomous mobility fleets. *Transportation Research Part C: Emerging Technologies* (2019).
- [14] RTA. 2019. RTA, Bannockburn, Glenstar and Lyft announce first mile/last mile transit pilot. <https://rideonrtachicago.com/2019/04/09/rt-a-bannockburn-glenstar-and-lift-announce-first-mile-last-mile-transit-pilot/>
- [15] Shangjue Sun and WY Szeto. 2019. Optimal sectional fare and frequency settings for transit networks with elastic demand. *Transportation Research Part B: Methodological* 127 (2019), 147–177.
- [16] Hai Wang and Hai Yang. 2019. Ridesourcing systems: A framework and review. *Transportation Research Part B: Methodological* 129 (2019), 122–155.
- [17] Damon Wischik. 2018. The price of choice: models, paradoxes, and inference for ‘mobility as a service’. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 604–610.
- [18] Junlin Zhang, Robin Lindsey, and Hai Yang. 2018. Public transit service frequency and fares with heterogeneous users under monopoly and alternative regulatory policies. *Transportation Research Part B: Methodological* 117 (2018), 190–208.

A Summary of notation

Table 4. Summary of notation

Notation	Definition
Decision variables:	
$x_{\theta m}$	Type θ traveler’s mode choice strategy for mode m
$z_{\ell, f}$	Binary variable for setting MT route ℓ ’s frequency at level f
$z_{\psi}^{b_{ij}}$	Binary variable for setting bundled MT services b_{ij} ’s frequency at level ψ
r_{ij}	MoD’s rebalancing flows on edge e_{ij}
$y_{\theta m}$	Intermediate variable $y_{\theta m} = \lambda_{st}(\theta)x_{\theta m}$
$p_{\theta m}$	Prices for mode m charged for type θ travelers
$\boldsymbol{p}, \boldsymbol{\zeta}, \boldsymbol{v}, \boldsymbol{\alpha}, \boldsymbol{\beta}$	Dual variables of subproblem
τ, ϕ	Intermediate variables for benders’ decomposition
Parameters:	
$G = (V, E)$	Graph of multimodal mobility network consists of vertices V and edges E
n	Cardinality of vertices V
Θ	Type space of travelers
λ_{st}	Travel demand from vertex s to vertex t
$q_{st}(\theta)$	Probability density of type θ for $s - t$ trip
L	Total number of MT lines
F	Maximum frequency level for each line
\mathcal{M}_{st}	Set of hybrid modes available for $s - t$ trip
$v_{\theta m}$	Valuation of mode m to type θ travelers
E_m	Edges trespassed by mode m
V_m	Vertices trespassed by mode m
\mathcal{B}_{ij}	Set of bundled MT services on edge e_{ij}
o_ℓ	Unit setup cost for line ℓ
c_m	Operational cost of mode m
\mathcal{R}	Total number of available MT vehicles
T_ℓ	Cycling time of line ℓ
\mathcal{P}_i^{st}	Set of paths for i^{th} commodity’s $s - t$ trip in subproblem’s approximation algorithm
K	Number of commodities in subproblem
ϵ, δ	Error tolerance in approximation algorithm
\mathcal{V}_ℓ	Vehicle capacity on line ℓ
Acronyms:	
MoD	Mobility-on-Demand
MNL	Multinomial Logit model

B Lift-and-project cuts for master problem

We first apply the convexification procedure to all constraints in (7). For example, we can denote the left-hand coefficient matrix of (1b) as A_2 . Since $z_{\ell,f} \geq 0$, we can rewrite the constraint as

$$\begin{aligned} z_{\ell,f}(Z^{b_{ij}} - A_2^{b_{ij}} Z_{\ell'}) &\leq 0, & \forall b_{ij} \in \mathcal{B}_{ij}, \forall e_{ij} \in E \\ (1 - z_{\ell,f})(Z^{b_{ij}} - A_2^{b_{ij}} Z_{\ell'}) &\leq 0, & \forall b_{ij} \in \mathcal{B}_{ij}, \forall e_{ij} \in E \end{aligned}$$

and linearize the constraints as

$$\begin{cases} z_{\ell,f}^2 \leftarrow z_{\ell,f} \\ z_{\ell,f} z_{\ell',f'} \leftarrow w_{\ell,f}, & \ell \neq \ell' \text{ or } f' \neq f \end{cases}.$$

Denote the generic form of constraints in the LP-relaxation as $AZ \leq b$. Given the fractional solution Z^* , the lift-and-project cuts are generated by the following cut generating

LP:

$$\begin{aligned} &\text{maximize } \kappa Z^* - \kappa_0 && (8) \\ \text{s.t. } &\kappa \leq uA + u_0 \mathbf{e}_{\ell,f} \\ &\kappa \leq vA - v_0 \mathbf{e}_{\ell,f} \\ &\kappa_0 \geq ub \\ &\kappa_0 \geq vb - v_0 \\ &\mathbf{1}^\top u + u_0 + \mathbf{1}^\top v + v_0 = 1 \\ &u, v, u_0, v_0 \geq 0 \end{aligned}$$

where $\mathbf{e}_{\ell,f}$ is an all-zero vector but one at the ℓ, f entry.

Let $A_{\ell,f}$ be the column index of A . We add the strengthened lift-and-project cuts for $z_{\ell',f'}$ as follows:

$$\sum_{\ell \in [L], f \in [F]} \alpha_{\ell,f} z_{\ell,f} \leq \beta,$$

where $\alpha_{\ell,f} =$

$$\begin{cases} \max\{u^\top A_{\ell,f} + u_0 \lceil m_{\ell,f} \rceil, v^\top A_{\ell,f} - v_0 \lfloor m_{\ell,f} \rfloor\}, (\ell', f') \neq (\ell, f) \\ \max\{u^\top A_{\ell,f} - u_0, v^\top A_{\ell,f} + v_0\}, (\ell', f') = (\ell, f) \\ \max\{u^\top A_{\ell,f}, v^\top A_{\ell,f}\}, \ell = L, f = F \end{cases}$$

$$\beta = \min\{u^\top b, v^\top b + u_0\}$$

$$m_{\ell,f} = \frac{v^\top A_{\ell,f} - u^\top A_{\ell,f}}{u_0 + v_0}.$$