Transportation Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Plan Your System and Price for Free: Fast Algorithms for Multimodal Transit Operations

Siddhartha Banerjee; , Chamsi Hssaine; , Qi Luo; , Samitha Samaranayake

informs
https://pubsonline.informs.org/journal/trsc

TRANSPORTATION SCIENCE
Vol. 59, No. 1, January–February 2025, pp. 13–27
ISSN 0041-1655 (print), ISSN 1526-5447 (online)

Plan Your System and Price for Free: Fast Algorithms for Multimodal Transit Operations

Siddhartha Banerjee,[a] Chamsi Hssaine,[b,]* Qi Luo,[c] Samitha Samaranayake[d]

[a] Department of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14853; [b] Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, California 90089; [c] Department of Business Analytics, Tippie College of Business, University of Iowa, Iowa City, Iowa 52242; [d] School of Civil and Environmental Engineering, Cornell University, Ithaca, New York 14853

*Corresponding author

Contact: sbanerjee@cornell.edu, https://orcid.org/0000-0002-8954-4578 (SB); hssaine@usc.edu, https://orcid.org/0000-0003-3104-3575 (CH); qi-luo-1@uiowa.edu, https://orcid.org/0000-0002-4103-7112 (QL); samitha@cornell.edu, https://orcid.org/0000-0002-5459-3898 (SS)

Received: December 30, 2022
Revised: October 26, 2023; July 24, 2024
Accepted: September 16, 2024
Published Online in Articles in Advance: October 25, 2024

https://doi.org/10.1287/trsc.2022.0452

Abstract. We study the problem of jointly pricing and designing a smart transit system, where a transit agency (the *platform*) controls a fleet of demand-responsive vehicles (cars) and a fixed line service (buses). The platform offers commuters a menu of options (*modes*) to travel between origin and destination (e.g., direct car trip, a bus ride, or a combination of the two), and commuters make a utility-maximizing choice within this menu, given the price of each mode. The goal of the platform is to determine an optimal set of modes to display to commuters, prices for these modes, and the design of the transit network in order to maximize the social welfare of the system. In this work, we tackle the *commuter choice* aspect of this problem, traditionally approached via computationally intensive bilevel programming techniques. In particular, we develop a framework that efficiently decouples the pricing and network design problem: Given an efficient (approximation) algorithm for centralized network design *without prices*, there exists an efficient (approximation) algorithm for decentralized network design *with prices and commuter choice*. We demonstrate the practicality of our framework via extensive numerical experiments on a real-world data set. We moreover explore the dependence of metrics such as welfare, revenue, and mode usage on (i) transfer costs and (ii) cost of contracting with on-demand service providers and exhibit the welfare gains of a fully integrated mobility system.

Funding: This work was supported by the National Science Foundation [Awards CMMI-2308750, CNS-1952011, and CMMI-2144127].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/trsc.2022.0452.

Keywords: multimodal transportation • pricing • network design

1. Introduction

Providing efficient, comprehensive, and cost-effective transportation solutions has long been a challenge for public transit agencies. In particular, a central problem faced by these agencies is the fact that they are resource constrained and often face a *ridership versus coverage* problem (Walker 2018). On the one hand, a key objective of ridership-oriented transit is efficiency: By maximizing the number of commuters served, not only do transit agencies benefit from lower public subsidies per rider, but they also reduce traffic congestion and the environmental impact of gas emissions by targeting cities' highest-demand corridors. On the other hand, public transit ideally should provide access to mobility for the entire community and be an equitable service, which in many cases comes at the cost of reduced service in high ridership areas.

In the past decade, however, mobility-on-demand (MoD) companies such as Lyft and Uber have entered into the conversation, dramatically altering the mobility landscape. These flexible and adaptive transportation services are increasingly viewed as a means of (partially) reconciling the aforementioned tension between ridership and coverage. Indeed, by using ride-hailing services to connect previously underserved communities to transit hubs, agencies can expand access to more affordable and sustainable transportation solutions, all while being able to concentrate resources on high-density corridors. Examples of pilot projects focusing on enhancing first/last-mile connectivity via multimodal trips (with supporting services including microtransit and ride-sharing options like UberPool) include LA Metro and Via (LA Metro 2019), Dallas Area Rapid Transit and Lyft (DART 2019), Moovel in Germany (Moovel 2024), Whim in Helsinki (Bloomberg 2018), UbiGo in Stockholm (Civitas 2020), and MARTA (MARTA Reach 2022).

Central to the operations of these types of integrated systems is the ability to design a system that is consistent with *commuter choice*. Specifically, transit systems are characterized by their inherently decentralized nature; that is, rather than being assigned to trips, commuters choose the mass transit routes that minimize their travel times. In a similar vein, when faced with hybrid transportation options, commuters will choose the option that maximizes their net utility, a more meaningful metric than travel times given the heterogeneity of options along different dimensions, for example, travel time, comfort, convenience, and, importantly, price. This last aspect is crucial in the successful design of a multimodal system; whereas certain transit systems charge a flat fare for all trips (e.g., $2.9 for a subway ride in New York City; New York City MTA 2024), there exists a tradeoff between simple pricing (such as a flat fare) and trip-specific pricing, which is potentially more efficient. Thus, the question preoccupying transit agencies should no longer solely be how to design such a transportation network, but how to simultaneously *price* and design this integrated system. Indeed, it has been empirically shown that jointly solving these two problems—that have by and large been considered separately in the literature—can lead to substantial gains in system efficiency (Bertsimas, Sian Ng, and Yan 2020). Although joint pricing and network design is considered computationally difficult in general (see Section 2), in this paper, we develop a new framework that demonstrates that pricing and designing a network in the presence of commuter choice is no harder than network design under nonstrategic behavior for a large family of problem instances.

1.1. Summary of Our Contributions

We consider a model in which a transit agency (henceforth, the platform) controls a fixed-line service and has access to a fleet of demand-responsive vehicles (e.g., via a prenegotiated contract with a ride-hailing service or an in-house fleet of taxis). The platform is faced with a set of nonatomic passenger flows and offers commuters the choice of a number of ways of traveling between origin and destination nodes: A commuter can travel by bus for the entirety of the trip (walking to and from the bus stations closest to their origin and destination); they can use a ride-hailing service; or they can combine these two travel options by using the ride-hailing service for the first and last legs of their trip, and traveling by bus in between. We refer to these different options as travel *modes*.

A commuter has a valuation for each mode, drawn from a known distribution. Given the prices set by the platform, they choose the mode that maximizes their net utility. The platform, on the other hand, incurs an operating cost for each mode, as well as a cost to design the transit network (e.g., a fixed cost for each bus line). The focus of the platform is the *long-term* planning and design of the multimodal marketplace. Concretely, the goal is to determine the optimal set of modes to display to commuters, prices for these modes, and the design of the transit network in order to maximize the total welfare of the system, that is, the sum of its profit and commuter utilities. We refer to this problem as the welfare maximization problem.

One of the key hurdles of transit planning is accounting for commuters' strategic behavior: doing so requires computing equilibria of an underlying game, a task known to be Polynomial Parity Arguments on Directed graphs (PPAD) complete in general (Daskalakis, Goldberg, and Papadimitriou 2009). Indeed, the vast majority of existing techniques use computationally intensive iterative methods based on a bilevel programming formulation to compute the optimal set of planning decisions that are consistent with commuter choice (Parbo, Nielsen, and Prato 2014, Verbas and Mahmassani 2015, Yu et al. 2015, Pinto et al. 2019). Our main methodological contribution in this respect is to show that, when the transit planner can use pricing as a lever to coordinate commuter choice, joint pricing and line planning is no harder than vanilla line planning in the presence of *nonstrategic* behavior. More specifically, we propose a methodological framework that disentangles the two sources of complexity in the welfare maximization problem: (i) designing the transportation network and (ii) pricing the trips offered by the platform. The framework tackles the problem in two steps:

1. (Approximately) solve a *centralized* welfare maximization problem, that is, a single-level assignment problem that relaxes the commuter choice constraints of the original problem.
2. Compute the prices that induce the flows corresponding to this (approximate) solution.

A priori, not only is it unclear *how* to compute the prices described in Step 2, but it is also not evident that such prices even exist. Our work answers these two questions in the affirmative for a broad class of commuter valuation distributions. As a warmup, we first consider the widely used multinomial logit (MNL) model of commuter choice, showing that the tractable closed-form expression for commuters' choice probabilities can be inverted to yield the appropriate prices. We then show that such an approach efficiently computes prices for a much broader class of continuous valuation distributions, under mild regularity conditions.

A natural next question is whether such an approach—that is, solving a centralized problem and computing prices that induce at least as high a welfare—can be leveraged for the space of *discrete* valuation distributions. This space of distributions is of particular interest because, in the case of transportation networks, supply costs and constraints (e.g., budget, capacity, and circulation constraints) are often linear in the decision variables. Modeling commuter valuations by a discrete distribution (e.g., by bucketing commuters into different types), in such cases, allows the platform to leverage the computational sophistication of mixed integer linear programming solvers compared with mixed integer convex programming solvers, which one would need under continuous valuation distributions. We show that, for arbitrary discrete valuation distributions, the welfare-maximizing prices arise from the dual of an appropriately chosen linear program. This framework thus has far-reaching computational implications for pricing and network design in decentralized settings: *Given an efficient (approximation) algorithm for the centralized welfare maximization problem, there exists an efficient (approximation) algorithm for the welfare maximization problem in the presence of commuter choice*.

Although the approach of solving pricing problems in the assignment space has been explored in simple settings within the revenue management and mechanism design communities (Gallego and Van Ryzin 1994), our contribution here is demonstrating that these insights can be leveraged in much more complicated settings arising in the transportation literature, wherein one frequently only has access to an approximate solution. This powerful idea then vastly simplifies the computational complexity of many problems considered in recent work.

Finally, we demonstrate the practicality of our framework via extensive numerical experiments on a real-world data set. In particular, we show that in large-scale settings, the time required to solve the mixed integer linear program (MILP) associated with the centralized welfare maximization problem dwarfs that of the linear program that computes the welfare-maximizing prices. In other words, you can *plan your system and price for "free."* Our framework enables us to explore the dependence of key platform metrics such as welfare, revenue, profit, travel times, and mode usage on the quality of bus lines considered by the planner, transfer costs incurred by commuters because of first- and last-mile connections, as well as the per-mile cost a ride-hailing company would charge the platform for its services. Our results highlight that the major gains from introducing hybrid trips come from efficiently partitioning trip types: In such a system, MoD-only options are reserved for short trips, with longer trips reserved for transit-only options. Hybrid modes, finally, are most useful for the longest trips that are poorly connected to existing transit lines. These experiments also emphasize the importance of reducing transfer frictions for hybrid trips, as well as low MoD operating costs.

1.2. Paper Organization

In Section 2, we survey relevant literature. We then present the basic model and define the welfare maximization problem in Section 3. In Section 4, we develop intuition for our main approach by solving the pricing problem under MNL commuter choice and extend this result to continuous valuation distributions under mild conditions. We then build on this basic approach in Section 5 and present our pricing framework for the setting of discrete valuation distributions. Finally, in Section 6 and Online Appendix C, we demonstrate its applicability via numerical experiments on a real-world data set. All proofs are relegated to the Online Appendix.

2. Related Work

We review the most closely related lines of work in this section.

2.1. Line Planning and Commuter Choice

Much of the work on finding an optimal set of lines that minimizes some function of passenger waiting times and transit operator costs, subject to commuter choice, has relied on *bilevel* programming formulations (Constantin and Florian 1995). Although bilevel optimization is known to be strongly NP-hard in general (Hansen, Jaumard, and Savard 1992), Fontaine and Minner (2014) show that, under the assumption that the principal is allowed to choose among agents' optimal decisions, the bilevel network design problem with the objective of minimizing passenger waiting times can be cast as a nonlinear single-level problem via the Karush–Kuhn–Tucker (KKT) conditions, and then approximately solved via linearization tricks. From a technical perspective, our work is most similar in spirit to this latter paper in its reliance on linear programming duality to reduce the decentralized pricing problem to a centralized pricing problem.

From a modeling perspective, a typical objective considered in prior work is that of commuter travel times (Borndörfer, Grötschel, and Pfetsch 2007, Bertsimas, Sian Ng, and Yan 2020) rather than system welfare. The problem we consider subsumes this objective under the assumption that commuter valuations are a nonincreasing function travel times. When commuters all take the same mode (e.g., mass transit), travel times are a natural objective to minimize; however, we argue that this ceases to be the case once hybrid modes—and, as a result, heterogeneity in factors such as comfort and convenience—are introduced. System welfare has more commonly been considered as an objective in the line of work on pricing and matching for ride-hailing platforms (Cashore, Frazier, and Tardos 2023). Moreover, although our work does not consider the important problem of incentivizing ride-hailing services to lend their services to a welfare-maximizing platform (e.g., a transit agency), the fact that welfare includes the platform's profit (and as a result, operating costs incurred from the ride-hailing legs of each mode) captures the true cost to the ride-hailing firm for serving those rides.

2.2. Design of Multimodal Mobility Systems

Although the bulk of the work on network design has focused exclusively on bus routing for mass transit systems, a recent line of work which considers the integration of public transportation and ride-hailing services has emerged. Auad-Perez and Van Hentenryck (2022) and Périvier et al. (2021), respectively, consider shuttle fleet-sizing and line planning optimization for multimodal systems; these works, however, do not incorporate commuter choice in their models. Basciftci and Van Hentenryck (2023) address this gap, using a bilevel exact decomposition method to capture riders' mode preferences. Pinto et al. (2019) also proposed a bilevel heuristic for the problem of joint transit network redesign and shared-use autonomous vehicle fleet sizing. Finally, closely related to our paper is recent work by Lanzetti et al. (2023), who study the interactions of a public transit provider with profit-maximizing operators of autonomous fleets. In contrast to us, their focus is on analyzing game theoretic equilibria that arise from commuter choice in a fragmented system; they do not consider the transit operator's network design problem. Motivated by a recent push from cities to create integrated mobility marketplaces (U.S. Department of Transportation 2015), our focus is not on profit maximization for the MoD provider but solely on the transit operator's welfare maximization problem.

2.3. Pricing and Commuter Choice

The most active line of work with respect to pricing applied to transportation has been on toll setting, in which the goal is to either maximize the revenue raised from tolls located on edges of a network or to induce desirable equilibrium flows (Colson, Marcotte, and Savard 2007). Although bilevel schemes have been used to solve variants of this problem (Huang 2002), various heuristics have been proposed to reduce the bilevel complexity of the problem (Brotcorne et al. 2001, Wu, Yin, and Lawphongpanich 2011, Wu et al. 2012).

Our work is technically closely related to that of Wischik (2018), who considers a similar problem from the perspective of a ride-hailing service offering multiple modes. The key insight of this work is that, when the multinomial logit model is used to model passenger choice, the problem can efficiently be solved by formulating it as an equivalent resource allocation problem. We generalize this work with respect to the MNL model by (i) considering the line planning and assortment optimization problem and showing that one can still obtain (approximately) optimal prices, assuming oracle access to a feasible solution of the associated centralized problem, and (ii) showing that these insights extend to a broader class of continuous valuation distributions. We moreover differ from this work by considering arbitrary discrete valuation distributions and showing that dual-based pricing is optimal for this setting.

2.4. Joint Pricing and Frequency Setting of Transportation Networks

Our paper joins a small line of work that considers the problem of joint pricing and frequency setting of transportation networks. Sun and Szeto (2019) propose a bilevel programming model to find the set of profit-maximizing fares and frequencies, although they do not consider multimodal options. Also in the single mode setting, Bertsimas, Sian Ng, and Yan (2020) consider the problem of pricing and frequency setting in order to minimize system wait time in mass transit networks under an MNL choice model. In order to handle the nonconvexity induced by the passengers' demand function, they propose a first-order method that solves a series of locally linear approximations. Our work shows that, in the case of the MNL model, this nonconvexity in the prices is a red herring; as long as the problem is concave in the *quantile* (i.e., the assignment) space—which we show it is in the case of the welfare objective—the pricing aspect of the problem can be solved exactly.

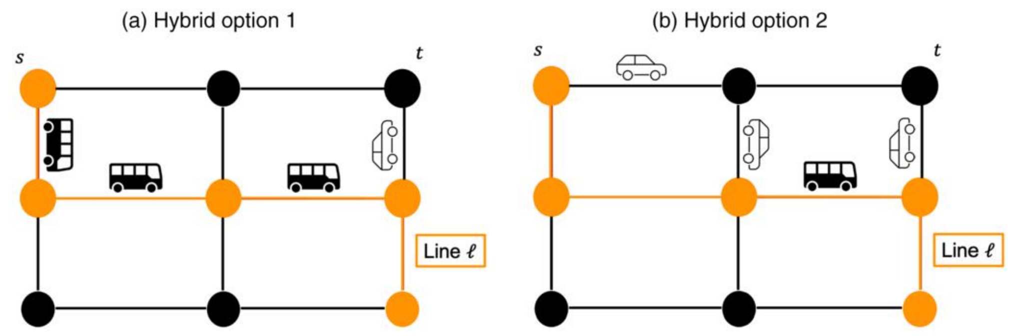
3. Preliminaries

We model the transportation network as a directed weighted graph $G = (V, E)$, with $|V| = n$ nodes corresponding to pickup and dropoff locations and edges representing roads between nodes.

3.1. Supply Model

A single mobility provider (henceforth, the platform) operates the network and controls a fleet of demand-responsive vehicles, as well as a fixed-line, mass-transit service. The platform makes a set of network design-related (or *supply*) decisions, such as the set of routes its mass transit service operates, and the routes' corresponding frequencies. Given the network design, the platform presents commuters with a menu of possible ways to travel between their origin and destination nodes: A commuter can complete their trip entirely by transit (with potential walking for first/last mile) entirely via the demand-responsive service or via a hybrid combination of demand-responsive and fixed-line legs. Formally, given origin-destination pair (s, t), a hybrid trip option m (henceforth referred to as a mode) is defined by a sequence of trip segments $((s, i_1), (i_2, i_3), \ldots, (i_k, t) | i_1, \ldots, i_k \in V^k)$, and the service (e.g., transit operating at a certain frequency, or demand-responsive option) associated with each segment. Let $\mathcal{M}_{st}$ denote the set of all possible modes that can feasibly complete an (s, t) trip, with $\mathcal{M} = \{\mathcal{M}_{st} | (s, t) \in V^2\}$. Figure 1 provides a simple illustration of two modes available to a commuter traveling from s to t. We assume that commuters sharing the same origin and destination are shown the same set of modes and that the platform has an upper bound $k \in \mathbb{N}$ on the number of modes it wishes to display to commuters. We use $\mathbf{y} \in \{0, 1\}^{|\mathcal{M}|}$ to denote the indicator vector representing the set of modes displayed to commuters. Finally, the platform sets a price p_m for each mode m, uniform across commuters, with $\mathbf{p} = (p_m)_{m \in \mathcal{M}}$. We specify later on how

Figure 1. (Color online) Example of a Transit Network with a Single Bus Line ℓ and a Single (s, t) Commuter



Note. The commuter can use one of the two hybrid modes comprising a bus segment, and first- and last-mile car segments.

the set of supply decisions, as well as $\mathbf{p}$ and $\mathbf{y}$ will be chosen.

We use $C(\mathbf{z})$ to denote the total cost incurred by the platform for a vector of supply decisions $\mathbf{z}$ and assume C is convex in $\mathbf{z}$. We let c_m denote the cost incurred by the platform for a trip completed via mode m. This cost may include, for instance, the cost of compensating the ride-hailing service for the trip (prenegotiated as part of a contract) and the associated mass transit operating costs.

3.2. Commuter Choice Model

We consider a large-market, fluid scaling of the demand side of the system. That is, each pair of nodes $(s, t) \in V^2$ is associated with an exogenous nonatomic mass (or *flow*) of commuters seeking to travel from s to t (henceforth referred to as (s, t) commuters), denoted by $\lambda_{st} \in \mathbb{R}_+$. An (s, t) commuter has valuation V_m for mode $m \in \mathcal{M}_{st}$, drawn from a known, exogenous distribution F_{st} with support $\mathcal{V} \subseteq \mathbb{R}^{|\mathcal{M}_{st}|}$; let $\mathbf{V}_{st} = (V_m)_{m\in\mathcal{M}_{st}}$ (the assumption that valuations are drawn from a known distribution is common in much of the pricing literature; Gallego and Van Ryzin 1994, Banerjee, Freund, and Lykouris 2022).

Example 1 illustrates the generality of the valuation abstraction, and how it allows the platform to model broad heterogeneity in commuters' travel preferences.

Example 1. Consider an (s, t) commuter, whose valuation for mode m is given by

$$V_m = \alpha_{st} + \beta_{st} t_m^{\text{car}} + \gamma_{st} t_m^{\text{transit}} + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, 1)$, t_m^{car} denotes the duration of the trip from s to t completed by car, and t_m^{transit} denotes the duration completed by mass transit. Here, $\alpha_{st} \in \mathbb{R}_+$ represents the commuter's base valuation for completing the trip, and $\beta_{st} \in \mathbb{R}, \gamma_{st} \in \mathbb{R}$, respectively, represent the value the commuter places on the relative convenience of a car and the sustainability of transit.

Given a set of displayed modes and their corresponding prices, commuters make a randomized utility-maximizing decision about their chosen mode. If all modes generate negative utility, commuters opt out of the marketplace. (The assumption that the opt-out option produces zero utility is without loss of generality.) Let $x_m(\mathbf{V}_{st}, \mathbf{p}, \mathbf{y})$ denote the probability that an (s, t) commuter with valuation vector $\mathbf{V}_{st}$ chooses m given price menu $\mathbf{p}$ and displayed modes $\mathbf{y}$, with $\mathbf{x}_{st}(\mathbf{V}_{st}, \mathbf{p}, \mathbf{y}) = (x_m(\mathbf{V}_{st}, \mathbf{p}, \mathbf{y}))_{m\in\mathcal{M}_{st}}$. Formally,

$$\mathbf{x}_{st}(\mathbf{V}_{st}, \mathbf{p}, \mathbf{y}) \in \arg\max_{\mathbf{x}'} \left\{ \sum_{m\in\mathcal{M}:y_m=1} (V_m - p_m)x'_m \,\middle|\, \sum_{m\in\mathcal{M}} x'_m \le 1, x'_m \le y_m, x'_m \ge 0 \;\; \forall m \in \mathcal{M} \right\}. \tag{1}$$

In case of nonuniqueness, commuters break ties in favor of the platform, a commonly made assumption in the literature known as *partial cooperation* (Bialas and Karwan 1984, Dempe 2002, Fontaine and Minner 2014). Let $\mathbf{x}(\mathbf{V}, \mathbf{p}, \mathbf{y}) = (\mathbf{x}_{st}(\mathbf{V}_{st}, \mathbf{p}, \mathbf{y}), (s, t) \in V^2)$. For ease of notation, we at times omit dependence of $\mathbf{x}$ on its arguments.

Remark 1. In theory, commuters could deviate from the mode they have chosen (e.g., by getting on one bus line instead of another, if both bus lines get them to the same destination). We preclude such deviations (also referred to as "self-constructed" modes) from the model, an assumption that could be practically implementable via, for example, trip-specific tickets.

Remark 2. We briefly remark on the choice of the fluid model. In addition to flow models being commonly used in the line planning literature (Borndörfer, Grötschel, and Pfetsch 2007), it is well known that they provide strong approximations to many stochastic systems in large-market regimes and have been used widely for the study of these systems. Notable examples include dynamic pricing and revenue management (Gallego and Van Ryzin 1994, Gallego and Topaloglu 2019). Fluid models have also found extensive use in the study of ride-hailing systems (Banerjee, Freund, and Lykouris 2022), where there are theoretical guarantees of convergence between these systems and the corresponding fluid model. Moreover, though our model assumes stationary arrival demand patterns, our framework can be used to solve for optimal flows and prices in nonstationary settings by solving separately for "stationary periods" (e.g., peak versus off-peak hours). One can also interpret the demand flows as *forecasts* of future demand. The use of such forecasts to compute long-term plans in advance of online decision making given realized demand is used extensively in industry, wherein a central planning system uses forecasts to generate a high-level, aggregate network plan. This is done even if all components of the system are flexible; in our setting, doing so is even more important because one component of the system—the bus lines—must be fixed in advance.

3.3. Welfare Maximization Problem

We now introduce the platform's optimization problem. We consider the problem of *long-term* planning and pricing decisions for the multimodal marketplace, with the goal of maximizing *platform welfare*, which we formally define below.

Definition 1 (Welfare). Given a set of displayed modes $\mathbf{y}$, a price vector $\mathbf{p}$ for these modes, and supply decisions $\mathbf{z}$, the *welfare* $W(\mathbf{p},\mathbf{y},\mathbf{z})$ of the platform is the sum of the expected commuter utilities and platform profit (where the expectation is taken with respect to the valuation distributions). Formally,

$$W(\mathbf{p},\mathbf{y},\mathbf{z}) = \sum_{(s,t)\in V^2} \lambda_{st}\mathbb{E}\left[\sum_{m\in\mathcal{M}_{st}} (V_m - c_m)x_m(\mathbf{V}_{st},\mathbf{p},\mathbf{y})\right] - C(\mathbf{z}). \tag{2}$$

The platform's optimization problem is to determine the network design, set of modes, and associated prices to maximizes the induced welfare, subject to a set of network design constraints (also referred to as the *feasible region*), denoted by $\mathcal{N}$ (to be specified at the end of the section). This is given by the following bilevel program:

$$\max_{\mathbf{p},\mathbf{y},\mathbf{z}} \quad \sum_{(s,t)\in V^2} \lambda_{st}\mathbb{E}\left[\sum_{m\in\mathcal{M}_{st}} (V_m - c_m)x_m(\mathbf{V}_{st},\mathbf{p},\mathbf{y})\right] - C(\mathbf{z}) \tag{P}$$

$$\text{s.t.} \quad \mathbf{x}_{\text{st}}(\mathbf{V}_{\text{st}},\mathbf{p},\mathbf{y}) \in \arg\max_{\mathbf{x}'} \left\{ \sum_{m\in\mathcal{M}:y_m=1} (V_m - p_m)x'_m \,\middle|\, \sum_{m\in\mathcal{M}} x'_m \le 1, x'_m \le y_m, x'_m \ge 0 \ \forall m \right\} \quad \forall (s,t)\in V^2, \mathbf{V}_{st} \in \mathcal{V}^{|\mathcal{M}_{st}|}, \tag{3}$$

$$\sum_{m\in\mathcal{M}_{st}} y_m \le k \quad \forall (s,t)\in V^2$$

$$(\mathbb{E}[\mathbf{x}(\mathbf{V}_{st},\mathbf{p},\mathbf{y})],\mathbf{y},\mathbf{z}) \in \mathcal{N}, \quad y_m \in \{0,1\} \ \forall m\in\mathcal{M}, \quad p_m \in \mathbb{R} \ \forall m \in \mathcal{M}. \tag{4}$$

Here, Constraint (3) encodes commuters' utility maximization problem, and Constraint (4) ensures that at most k modes are displayed to commuters for each origin-destination pair. We use OPT to denote the optimal value of (P).

The platform's network design constraints $\mathcal{N}$ couple supply and flow of demand. We first present two examples of possible constraint sets that our model encompasses, before presenting them in complete generality.

Example 2. Our first example models a platform that operates a fleet of buses and contracts with a ride-hailing company to provide first- and last-mile car rides, as in Périvier et al. (2021). Let $\kappa \in \mathbb{N}$ denote the fixed capacity of a bus. A bus *route* is a fixed sequence of consecutive edges of G, said to be served at *frequency* $f \in \{0,1,\dots,\overline{F}\}$ if f buses are operated on the route throughout the time window of interest, with $\overline{F} \in \mathbb{N}$ (assuming an upper bound on the set of possible frequencies is without loss of generality for a system with a finite number of passengers). We define a bus *line* ℓ to be a combination of a bus route r_ℓ and an associated frequency f_ℓ, and let $\mathcal{L} = \{(r,f) | (r,f) \in \mathcal{R} \times [\overline{F}]\}$, with $L = |\mathcal{L}|$. Let $\mathcal{L}_m$ denote the set of lines used by m, and $\mathcal{E}_m$ the set of edges of m traversed by the lines in $\mathcal{L}_m$. Finally, we let $c_\ell \in \mathbb{R}_+$ denote the fixed cost of opening line ℓ, assumed to be increasing in f_ℓ. Because a mode is partially defined by a set of bus lines in this example, each of which has an associated frequency, a commuter's valuation here will be both route *and* frequency dependent.

In this case, the platform's supply decision $\mathbf{z}$ is the set of bus lines to operate, and the constraint it seeks to enforce is that the bus capacity is not exceeded on any given edge of its route, in expectation. Formally, $C(\mathbf{z}) = \sum_{\ell\in\mathcal{L}} c_\ell z_\ell$, and the network design constraint is

$$\sum_{(s,t)\in V^2} \lambda_{st} \sum_{\substack{m\in\mathcal{M}_{st}: \\ \ell\in\mathcal{L}_m, \\ e\in\mathcal{E}_m}} \mathbb{E}[x_m(\mathbf{V}_{st},\mathbf{p},\mathbf{y})] \le \kappa f_\ell z_\ell \quad \forall \ell \in \mathcal{L}, e \in r_\ell.$$

We include the welfare maximization problem associated with this example in Online Appendix B.1.

Although Example 2 models a situation in which the platform contracts out ride-hailing trips to create hybrid modes, our general model also subsumes a setting in which the platform operates both a fleet of buses and a fleet of unit-capacity vehicles (cars), as shown in Example 3.

Example 3. In this setting, as in Example 2, the platform must (i) decide the set of lines to operate, denoted by $\mathbf{z}' \in \{0,1\}^L$, and (ii) make a set of empty-vehicle rebalancing decisions to satisfy the demand for car trips. We use $\eta_{ij} \in \mathbb{R}_+$ to denote the rebalancing rate between nodes i and j and let $c_{ij} \in \mathbb{R}_+$ be the cost per unit of rebalancing. Moreover, for $m \in \mathcal{M}$ let $\mathcal{T}_m$ denote the set of source and destination nodes (or *terminal* nodes) for the car segments of mode m.

In this case, $C(\mathbf{z}) = \sum_{\ell \in \mathcal{L}} c_\ell z'_\ell + \sum_{(i,j)\in V^2} c_{ij}\eta_{ij}$, and the constraint set $\mathcal{N}$ is the same as that of Example 2, with the additional constraint that the flow of unit-capacity vehicles must form a valid *circulation*, that is, for all $i \in [n]$,

$$\sum_{j\in[n]}\left[\eta_{ij} + \sum_{(s,t)\in V^2}\lambda_{st}\sum_{\substack{m\in\mathcal{M}_{st}:\\ (i,j)\in\mathcal{T}_m}}\mathbb{E}[x_m(\mathbf{V}_{st},\mathbf{p},\mathbf{y})]\right]$$

$$=\sum_{j\in[n]}\left[\eta_{ji} + \sum_{(s,t)\in V^2}\lambda_{st}\sum_{\substack{m\in\mathcal{M}_{st}:\\ (j,i)\in\mathcal{T}_m}}\mathbb{E}[x_m(\mathbf{V}_{st},\mathbf{p},\mathbf{y})]\right].$$

In addition to the basic capacity and circulation constraints presented in the examples above, the platform may be interested in incorporating the following constraints in $\mathcal{N}$:

- A budget constraint for the fixed costs for opening lines and operating each mode:

$$\sum_{\ell\in\mathcal{L}} c_\ell z'_\ell + \sum_{(s,t)\in V^2}\lambda_{st}\sum_{m\in\mathcal{M}_{st}} c_m \mathbb{E}[x_m(\mathbf{V}_{st},\mathbf{p},\mathbf{y})] \le B, \quad B\in\mathbb{R}_{>0}.$$

- An upper bound $N \in \mathbb{N}$ on the number of buses:

$$\sum_{\ell\in\mathcal{L}} \tau_\ell f_\ell z'_\ell \le N,$$

where $\tau_\ell \in \mathbb{R}_{>0}$ represents the time required for a bus to complete route r_ℓ.

Having presented these motivating examples for the network design constraints, we now define the problem's feasible region that our results encompass, in complete generality. In particular, we assume that $\mathcal{N}$ is defined by a collection of convex functions $g_i, i\in[N_1], h_j, j\in[N_2]$, with $N_1\in\mathbb{N}, N_2\in\mathbb{N}$, and constants $\alpha_{mi}, m\in\mathcal{M}, i\in[N_1], \beta_{mj}, m\in\mathcal{M}, j\in[N_2]$, as follows:

$$\mathcal{N} = \Bigg\{(\mathbb{E}[\mathbf{x}(\mathbf{V}_{st},\mathbf{p},\mathbf{y})],\mathbf{y},\mathbf{z}) \,\Big|\, \sum_{(s,t)\in V^2}\lambda_{st}\sum_{m\in\mathcal{M}_{st}}\alpha_{mi}\mathbb{E}[x_m(\mathbf{p},\mathbf{y},\mathbf{z})] + g_i(\mathbf{y},\mathbf{z}) \le 0\ \forall i\in[N_1],$$
$$\sum_{(s,t)\in V^2}\lambda_{st}\sum_{m\in\mathcal{M}_{st}}\beta_{mj}\mathbb{E}[x_m(\mathbf{p},\mathbf{y},\mathbf{z})] + h_j(\mathbf{y},\mathbf{z}) = 0\ \forall j\in[N_2]\Bigg\}.$$

In words, our results capture settings in which the functions defining the network design constraints are linear in the expected demand and additively separable across $\mathbb{E}[\mathbf{x}]$ and $(\mathbf{y},\mathbf{z})$, as is the case in the practical examples above.

4. Warmup: MNL Commuter Choice

As motivation for our main result, we consider the setting in which commuter choices are governed by a discrete-choice model typically used in the transportation and revenue management literature: the *MNL model* (McFadden 1973). Under this model, an (s,t) commuter has valuation $V_m = v_m + \epsilon$ for mode m, where v_m is a deterministic base valuation, and ϵ is a Gumbel-distributed random variable with location 0 and scale 1, drawn independent and identically distributed (i.i.d.) across modes. Commuters' value for the outside option is simply $\epsilon \sim \text{Gumbel}(0,1)$. Let $q_m(\mathbf{p},\mathbf{y})$ denote the ex ante probability an (s,t) commuter chooses mode m given prices $\mathbf{p}$, that is, $q_m(\mathbf{p},\mathbf{y}) = \mathbb{E}[x_m(\mathbf{V}_{st},\mathbf{p},\mathbf{y})]$, and $\mathbf{q}(\mathbf{p},\mathbf{y}) = (q_m(\mathbf{p},\mathbf{y}))_{m\in\mathcal{M}}$. We will equivalently refer to these probabilities $\mathbf{q}(\mathbf{p},\mathbf{y})$ as *quantiles*. Under the MNL model, we have

$$q_m(\mathbf{p},\mathbf{y}) = \frac{e^{v_m - p_m} y_m}{1+\sum_{m'\in\mathcal{M}_{st}} e^{v_{m'}-p_{m'}} y_{m'}}. \tag{5}$$

Proposition 1 leverages (5) to establish that, under the MNL model, our original bilevel problem reduces to a single-level problem. We defer its proof to Online Appendix A.1.

Proposition 1. *Under MNL choice, the welfare maximization problem (P) reduces to the following single-level optimization problem:*

$$\max_{\mathbf{p},\mathbf{y},\mathbf{z}} \sum_{(s,t)\in V^2}\lambda_{st}\left(\sum_{m\in\mathcal{M}_{st}}(v_m - c_m)q_m(\mathbf{p},\mathbf{y})\right) - \sum_{(s,t)\in V^2}\lambda_{st}\left(\sum_{m\in\mathcal{M}_{st}} q_m(\mathbf{p},\mathbf{y})\log q_m(\mathbf{p},\mathbf{y})\right)$$
$$- \sum_{(s,t)\in V^2}\lambda_{st}\left(1-\sum_{m\in\mathcal{M}_{st}} q_m(\mathbf{p},\mathbf{y})\right)\log\left(1-\sum_{m\in\mathcal{M}_{st}} q_m(\mathbf{p},\mathbf{y})\right) - C(\mathbf{z})$$
$$\text{s.t.}\ \ q_m(\mathbf{p},\mathbf{y}) = \frac{e^{v_m-p_m}y_m}{1+\sum_{m'\in\mathcal{M}_{st}} e^{v_{m'}-p_{m'}}y_{m'}}\ \forall m\in\mathcal{M}_{st}, (s,t)\in V^2$$
$$\sum_{m\in\mathcal{M}_{st}} y_m \le k\quad \forall (s,t)\in V^2$$
$$(\mathbf{q}(\mathbf{p},\mathbf{y}),\mathbf{y},\mathbf{z})\in\mathcal{N},\quad y_m\in\{0,1\}\ \forall m\in\mathcal{M},\quad \mathbf{p}\in\mathbb{R}. \tag{MNL-P}$$

Note that the objective of (MNL-P) depends only on the prices to the extent that the prices determine the quantiles $\mathbf{q}$. In addition to this, the objective—although nonconvex in the space of prices—is concave in the quantile space (Cover 1999). In light of these two facts, in what follows, it will be useful to instead think of welfare as a function of the quantiles $\mathbf{q}$ rather than $\mathbf{p}$.

4.1. Solving a Decentralized Problem via a Centralized Assignment Program

As a first step toward building a solution for the welfare maximization problem, we consider a relaxation of this problem: the *centralized welfare maximization problem*. In the centralized welfare maximization problem, the goal is to determine the network design, set of modes, and an *assignment of demand to modes* that maximizes the induced welfare. Under the MNL model, this can be formulated via following mixed integer convex program (MICP):

$$
\begin{aligned}
\max_{\mathbf{q},\mathbf{y},\mathbf{z}} \quad & \sum_{(s,t)\in V^2} \lambda_{st}\left(\sum_{m\in\mathcal{M}_{st}} (v_m - c_m) q_m\right) \\
& - \sum_{(s,t)\in V^2} \lambda_{st}\left(\sum_{m\in\mathcal{M}_{st}} q_m \log q_m\right) \\
& - \sum_{(s,t)\in V^2} \lambda_{st}\left(1 - \sum_{m\in\mathcal{M}_{st}} q_m\right)\log\left(1 - \sum_{m\in\mathcal{M}_{st}} q_m\right) \\
& - C(\mathbf{z}) \\
\text{s.t.} \quad & \sum_{m\in\mathcal{M}_{st}} q_m \le 1 \quad \forall (s,t)\in V^2 \\
& q_m \le y_m \quad \forall m\in\mathcal{M} \\
& \sum_{m\in\mathcal{M}_{st}} y_m \le k \quad \forall (s,t)\in V^2 \\
& (\mathbf{q},\mathbf{y},\mathbf{z})\in\mathcal{N}, \qquad q_m \ge 0 \quad \forall m\in\mathcal{M}, \\
& \qquad\qquad y_m\in\{0,1\} \quad \forall m\in\mathcal{M}.
\end{aligned}
\tag{MNL-CP}
$$

We emphasize that the centralized welfare maximization problem *is not a pricing problem*, and thus $\mathbf{q}$ is a *decision variable* rather than a function of the prices $\mathbf{p}$. We call this problem "centralized" because it considers a world in which commuters have no choice and can be dictated by the transit operator to travel via a specific mode.

Our key insight is that, given any feasible solution to (MNL-CP), there exists a set of prices that implement a *decentralized* solution *without any loss to the objective*. Otherwise stated, pricing and designing a network in the presence of strategic behavior is no harder than simply designing the network with nonstrategic commuters. We defer the proof of Proposition 2 to Online Appendix A.2.

Proposition 2. *Suppose the platform has access to an oracle $\mathcal{O}$ that returns a feasible solution $(\mathbf{q}^{\mathcal{O}}, \mathbf{y}^{\mathcal{O}}, \mathbf{z}^{\mathcal{O}})$ to (MNL-CP) and let $W^{\mathcal{O}}$ be the objective value of (MNL-CP) corresponding to this solution. Define prices $\mathbf{p}^{\mathcal{A}}$ as follows:*

$$
p_m^{\mathcal{A}} = v_m - \log\left(\frac{q_m^{\mathcal{O}}}{1 - \sum_{m'\in\mathcal{M}_{st}} q_{m'}^{\mathcal{O}}}\right) \qquad \forall (s,t)\in V^2, m\in\mathcal{M}_{st} \text{ s.t. } y_m^{\mathcal{O}} = 1.
$$

Finally, let $W^{\mathcal{A}}$ denote the system welfare induced by $(\mathbf{p}^{\mathcal{A}}, \mathbf{y}^{\mathcal{O}}, \mathbf{z}^{\mathcal{O}})$. Then, the following holds:

1. *The commuter choice probabilities induced by $\mathbf{p}^{\mathcal{A}}$ are feasible for* (MNL-P), *and*
2. *The objectives satisfy $W^{\mathcal{A}} = W^{\mathcal{O}}$.*

In practice, the oracle $\mathcal{O}$ can be any heuristic for solving this mixed integer convex program. For instance, one may feed this problem to a state-of-the-art solver. Depending on the complexity of the constraint set $\mathcal{N}$, the solver need not output an optimal solution within a desired time limit but the best solution after a fixed number of iterations.

Observe that, for any feasible solution $\mathbf{p}$ to (MNL-P), the *induced expected choice* $\mathbf{q}(\mathbf{p}, \mathbf{y})$ is feasible to (MNL-CP). Thus, (MNL-CP) is an upper bound on (MNL-P), and we obtain the following corollary, which states that under our framework, any approximation scheme for (MNL-CP) can be leveraged to obtain an approximation scheme for (MNL-P) with the same performance guarantee.

Corollary 1. *Let OPT^{CP} and OPT respectively denote the optimal values of* (MNL-CP) *and* (MNL-P). *Suppose moreover that $W^{\mathcal{O}} \ge \alpha\, OPT^{CP}$, for some $\alpha > 0$. Then, $W^{\mathcal{A}} \ge \alpha\, OPT$.*

Proposition 2 shows that joint pricing and network design under the MNL model is no harder than network design in a centralized setting. This insight, however, does not simply hold for MNL choice. Proposition 3 establishes that, under mild regularity conditions, it extends to the entire space of continuous valuation distributions.

Proposition 3. *Suppose that the valuation distributions $\{F_{st}\}_{(s,t)\in V^2}$ are such that the following conditions hold:*

1. *$\mathbf{q}(\mathbf{p}, \mathbf{y})$ is efficiently invertible, and*
2. *The welfare function $W(\mathbf{q}, \mathbf{y}, \mathbf{z})$ is concave in the quantiles $\mathbf{q}$.*

Then, given an oracle $\mathcal{O}$ that produces a feasible solution $(\mathbf{q}^{\mathcal{O}}, \mathbf{y}^{\mathcal{O}}, \mathbf{z}^{\mathcal{O}})$ to (MNL-CP), *prices $\mathbf{p}^{\mathcal{A}} = \mathbf{p}^{-1}(\mathbf{q}^{\mathcal{O}})$ induce a feasible flow, and moreover $W^{\mathcal{A}} = W^{\mathcal{O}}$.*

We omit the proof of this fact, as it is identical to that of Proposition 2.

Having established minimal conditions for the use of this centralized pricing framework within the space of continuous valuations, a natural next step is to see whether it can be extended to the space of discrete valuation distributions. Discrete valuation distributions are particularly attractive from a computational perspective because in real-world settings the welfare maximization problem (P) tends to be *linear* in the supply decisions $\mathbf{z}$ (see Examples 2 and 3). Thus, modeling commuters' valuations via discrete distributions would allow the platform to formulate the centralized welfare maximization problem as an MILP, for which existing solvers outperform MICP solvers (Lubin et al. 2018).

Remark 3. The technique of solving for optimal quantiles and inverting these to obtain prices is frequently used in revenue management (Gallego and Van Ryzin 1994). This idea is presented here within the context of the lesser considered problem of *welfare maximization* in order to motivate our main result in the following section, in particular, in settings where the quantile function is not easily invertible.

5. Main Result

In this section, we show how to leverage a similar approach—that is, obtaining prices via a single-level centralized problem in the assignment space—for the space of arbitrary *discrete* valuation distributions. We first introduce some notation.

5.1. Notation

A commuter is associated with a *discrete type* θ defined by their origin-destination pair, as well as a valuation profile $\mathbf{v}_\theta = (v_{\theta m})_{m \in \mathcal{M}_{st}}$ for the available modes, with $\mathbf{v}_\theta \in \mathcal{V}$. Abusing notation, we let λ_θ denote the total flow of type θ commuters and define Θ to be the set of all commuter types. We use $x_{\theta m}(\mathbf{p}, \mathbf{y})$ to denote the probability that a type θ commuter chooses mode m given price menu $\mathbf{p}$ and displayed modes $\mathbf{y}$, with $\mathbf{x}_\theta(\mathbf{p}, \mathbf{y}) = (x_{\theta m}(\mathbf{p}, \mathbf{y}))_{m \in \mathcal{M}_{st}}$. For ease of notation, in the remainder of the paper we often omit the dependence of the set of modes on the origin-destination pair (s, t), with it being clear from context that $v_{\theta m} = 0$ for $m \notin \mathcal{M}_{st}$ for (s, t) commuters of type θ. Under this model, welfare is given by

$$W(\mathbf{p}, \mathbf{y}, \mathbf{z}) = \left[\sum_{\theta \in \Theta} \lambda_\theta \sum_{m \in \mathcal{M}} (v_{\theta m} - c_m) x_{\theta m}(\mathbf{p}, \mathbf{y})\right] - C(\mathbf{z}).$$

Recall, in the MNL setting, given a feasible flow to (MNL-CP), we obtained the prices inducing at least as high a welfare by inverting commuters' choice probabilities. Unfortunately, for arbitrary discrete valuation distributions, existence of such an inverse is not guaranteed. We next show that invertibility is in fact not necessary to obtain equivalent prices. As in the MNL setting, consider the centralized welfare maximization problem for discrete valuations, which, instead of optimizing over prices, optimizes over $\phi_{\theta m} = \lambda_\theta x_{\theta m}$, the *flow* of type θ commuters assigned to mode m:

$$\begin{aligned}
\max_{\boldsymbol{\phi}, \mathbf{y}, \mathbf{z}} \quad & \sum_{\theta \in \Theta} \sum_{m \in \mathcal{M}} (v_{\theta m} - c_m) \phi_{\theta m} - C(\mathbf{z}) && \text{(CP)} \\
\text{s.t.} \quad & \sum_{m \in \mathcal{M}} \phi_{\theta m} \le \lambda_\theta \quad \forall \theta \in \Theta, \\
& \sum_{\theta \in \Theta_m} \phi_{\theta m} \le y_m \left(\sum_{\theta \in \Theta_m} \lambda_\theta\right) \quad \forall m \in \mathcal{M} \\
& \sum_{m \in \mathcal{M}_{st}} y_m \le k \quad \forall (s,t) \in V^2 && (6) \\
& (\boldsymbol{\phi}, \mathbf{y}, \mathbf{z}) \in \mathcal{N}, \quad \phi_{\theta m} \ge 0 \ \ \forall \theta \in \Theta, m \in \mathcal{M}, \\
& y_m \in \{0,1\} \ \ \forall m \in \mathcal{M},
\end{aligned}$$

where Θ_m denotes the set of types for whom mode m is available, and $\boldsymbol{\phi} = (\phi_{\theta m})_{\theta \in \Theta, m \in \mathcal{M}}$. Here, Constraint (6) enforces that the set of commuters Θ_m who can feasibly take mode m to complete their trip cannot be assigned to m unless it is displayed.

Given the modes displayed $\mathbf{y}$ and the supply decisions $\mathbf{z}$, what remains of (CP) is a linear program in $\boldsymbol{\phi}$, solvable in polynomial time. We formally define this subproblem, denoted by SP($\mathbf{z}, \mathbf{y}$):

$$\begin{aligned}
\max_{\boldsymbol{\phi}} \quad & \sum_{\theta \in \Theta} \sum_{m \in \mathcal{M}} (v_{\theta m} - c_m) \phi_{\theta m} && (\mathrm{SP}(\mathbf{z}, \mathbf{y})) \\
\text{s.t.} \quad & \sum_{m \in \mathcal{M}} \phi_{\theta m} \le \lambda_\theta \quad \forall \theta \in \Theta, && (7) \\
& \sum_{\theta \in \Theta_m} \phi_{\theta m} \le y_m \left(\sum_{\theta \in \Theta_m} \lambda_\theta\right) \quad \forall m \in \mathcal{M}, && (8) \\
& \sum_{m \in \mathcal{M}} \alpha_{mi} \left(\sum_{\theta \in \Theta_m} \phi_{\theta m}\right) + g_i(\mathbf{y}, \mathbf{z}) \le 0 \quad \forall i \in [N_1], && (9) \\
& \sum_{m \in \mathcal{M}} \beta_{mj} \left(\sum_{\theta \in \Theta_m} \phi_{\theta m}\right) + h_j(\mathbf{y}, \mathbf{z}) = 0 \quad \forall j \in [N_2]. \\
& && (10) \\
& \phi_{\theta m} \ge 0 \quad \forall \ell \in \mathcal{L}
\end{aligned}$$

Our algorithm makes use of the dual of SP($\mathbf{z}, \mathbf{y}$), given by

$$\begin{aligned}
\min_{\mathbf{u}, \zeta, \mu, \nu} \quad & \sum_{\theta \in \Theta} \lambda_\theta u_\theta + \sum_{m \in \mathcal{M}} \zeta_m y_m \left(\sum_{\theta \in \Theta_m} \lambda_\theta\right) - \sum_{i \in [N_1]} \mu_i g_i(\mathbf{y}, \mathbf{z}) \\
& - \sum_{j \in [N_2]} \nu_j h_j(\mathbf{y}, \mathbf{z}) \\
& && (\text{D-SP}(\mathbf{z}, \mathbf{y})) \\
\text{s.t.} \quad & u_\theta \ge v_{\theta m} - c_m - \zeta_m - \sum_{i \in [N_1]} \alpha_{mi} \mu_i - \sum_{j \in [N_2]} \beta_{mj} \nu_j \\
& \qquad \forall m \in \mathcal{M}, \ \theta \in \Theta_m \\
& u_\theta \ge 0 \quad \forall \theta \in \Theta, \quad \zeta_m \ge 0 \quad \forall m \in \mathcal{M}, \quad \mu_i \ge 0 \\
& \qquad \forall i \in [N_1] \\
& && (11)
\end{aligned}$$

Here, the dual variables $\mathbf{u}, \zeta, \mu$, and ν, respectively, correspond to primal Constraints (7), (8), (9), and (10) and dual Constraints (11) to primal variables $\boldsymbol{\phi}$. For concreteness, we present the centralized welfare maximization problem for Example 2 and its corresponding dual subproblem in Online Appendix B.1.

We now present our algorithm. As in the MNL setting, Algorithm 1 assumes access to an oracle that produces a feasible solution to (CP). Algorithm 1 is particularly attractive because of the interpretability of its outputs: The price of mode m is composed of its operating cost c_m, the cost ζ_m of displaying the mode (i.e., using up one unit of the budget k for modes), andthe costs $\sum_{i\in[N_1]}\alpha_{mi}\mu_i$ and $\sum_{j\in[N_2]}\beta_{mj}\nu_j$ related to the network design constraints.

Algorithm 1 (Multimodal Pricing via LP Duality)
Input: oracle $\mathcal{O}$ for (CP)
Output: prices $\mathbf{p}$, displayed modes $\mathbf{y}$ and supply decisions $\mathbf{z}$
Run $\mathcal{O}$. Let $\mathbf{z}$ denote the set of supply decisions returned by $\mathcal{O}$.
Solve linear program D-SP($\mathbf{z}, \mathbf{y}$). Let $(\mathbf{u}, \zeta, \mu, \nu)$ denote an optimal solution to D-SP($\mathbf{z}, \mathbf{y}$).
Set $p_m = c_m + \zeta_m + \sum_{i\in[N_1]}\alpha_{mi}\mu_i + \sum_{j\in[N_2]}\beta_{mj}\nu_j, \quad \forall m \in \mathcal{M}$.
Make supply decisions $\mathbf{z}$, display modes $\mathbf{y}$, and set prices $\mathbf{p}$.

Let $W^{\mathcal{A}}$ denote the welfare induced by Algorithm 1. Theorem 1 establishes that pricing the decentralized system is no harder than planning and assignment for a centralized system. We defer its proof to Online Appendix B.2.

Theorem 1. *Suppose the platform has access to an oracle $\mathcal{O}$ that returns a feasible solution $(\boldsymbol{\phi}^{\mathcal{O}}, \mathbf{y}^{\mathcal{O}}, \mathbf{z}^{\mathcal{O}})$ to (CP), and let $W^{\mathcal{O}}$ be the objective value of (CP) corresponding to this solution. Let $W^{\mathcal{A}}$ denote the system welfare induced by $\mathbf{p}^{\mathcal{A}}$, the prices output by Algorithm 1. Then, the following holds:*

1. *The commuter choice probabilities induced by $\mathbf{p}^{\mathcal{A}}$ are feasible for (P), and*
2. *The objectives satisfy $W^{\mathcal{A}} \geq W^{\mathcal{O}}$.*

As in the setting with MNL choice, for any feasible solution $\mathbf{p}$ to the original expected problem, the induced flow is feasible to (CP). Thus, (CP) is an upper bound on the optimal welfare for the original problem, and we obtain the following corollary of Theorem 1, which establishes that our algorithm inherits any approximation guarantee that the oracle has with respect to (CP).

Corollary 2. *Let OPT^{CP} and OPT, respectively, denote the optimal values of (CP) and (P). Suppose moreover that $W^{\mathcal{O}} \geq \alpha\, OPT^{CP}$, for some $\alpha > 0$. Then, $W^{\mathcal{A}} \geq \alpha\, OPT$.*

When (CP) can be solved exactly (i.e., $\alpha = 1$), Corollary 2 implies that $W^{\mathcal{A}} = OPT$. That is, $\mathbf{p}^{\mathcal{A}}$ is optimal for the original problem. Because $W^{\mathcal{A}} = W^{\mathcal{O}} \geq OPT^{CP} \geq OPT$ and $W^{\mathcal{A}} = OPT$, this immediately implies that $OPT^{CP} = OPT$ as well.

Thus, we have shown that, given an efficient oracle for the nonstrategic problem, our framework efficiently computes prices for arbitrary discrete valuation distributions by leveraging the power of linear programming duality.

6. Numerical Experiments: Case Study on the Manhattan Network

Finally, we demonstrate the practicality of our framework by deploying it on the Manhattan road network, using real data from the OpenStreetMap (OSM) database (Boeing 2017) and historical records of for-hire vehicle trips in New York City (NYC) (NYC Open Data 2024).

Although optimizing for welfare, we are also interested in the platform's revenue and profit, fraction of demand served (i.e., *throughput*), the distribution of trips across commuter types (which we partition based on value of time), and travel times, as we vary key inputs to the model. All experiments were run on a workstation with an eight-core, 3.6-GHz processor and 16 GB RAM, using a state-of-the-art solver (Gurobi 10.0).

6.1. Experimental Setup

We consider the setting described in Example 2, in which a platform operates a fleet of shuttles (also referred to as buses) and contracts with a ride-hailing company for first- and last-mile car rides (see Online Appendix B.1 for the mathematical formulation of the corresponding centralized welfare maximization problem and dual subproblem).

6.1.1. Line Inputs. We assume access to a candidate set of lines $\mathcal{L}$, constructed by Périvier et al. (2021). (Although our numerical experiments make use of a candidate set, our framework can be deployed without access to such a set. Such an assumption is however standard in the transit planning literature (Ceder and Wilson 1986, Chakroborty and Wivedi 2002, Fan and Machemehl 2006, Auad-Perez and Van Hentenryck 2022) and is necessary to develop constant-factor approximations to the line planning problem (Périvier et al. 2021).) We let $c_\ell = \$50\, d_\ell$, where d_ℓ denotes the distance (in miles) traveled by line ℓ. We moreover set the bus capacity $\kappa = 160$ and frequency $f_\ell = 1$ for all lines ℓ. (Note that this is equivalent to running a 40-person shuttle every 15 minutes.) In Online Appendix C.4, we perform a sensitivity analysis on κ to determine the impact of bus capacity on key system metrics.

6.1.2. Travel Modes. A mode consists of at most one trip segment completed by bus and at most two car trips. This design decision stems from the fact that mixed trips force commuters to incur at the minimum first- or last-mile car-to-bus transfers; any additional trip segments could be deemed excessive. (We study the impact of allowing for more than one bus segment in Online Appendix C.3.) We refer to any mode that is composed of a transit segment and at least one MoD segment as a *hybrid* mode.

For each (s, t) pair, we assume that the MoD-only mode consists of the shortest path between s and t. Moreover, for each line in $\mathcal{L}$, we construct either a transit-only or a hybrid mode. Specifically, we consider the nearest bus stops to s and t, respectively, and calculate the associated first- and last-mile distances, d_m^{FM} and d_m^{LM}. We fix a maximum walking radius $d_{\max} = 0.25$ miles (Yang and Diez-Roux 2012); if the first- and last-mile distances are both lower than $d_{\max}$, then this mode is transit only. Otherwise, it is hybrid, with either the first or last mile (or both) served by MoD. We omit hybrid modes for which the total first- and last-mile distances exceed the length of the direct MoD trip. We assume average speeds of 3 miles per hour (mph) for walking, 7 mph for transit (New York City Department of Transportation 2021), and 8.5 mph for ride-hailing (Bertsimas et al. 2019), respectively, and let τ_m denote the duration of the (s, t) trip completed by m.

The platform incurs no operating cost for the transit segment of mode m (i.e., all transit costs are subsumed in the line costs c_ℓ). The operating cost c_m for a mode with a MoD leg is composed of a fixed initial cost of \$3, as well as a constant cost per-MoD mile, denoted by c^{MoD}. Formally, we let $c_m = 3 + c^{\text{MoD}} \cdot (d_m^{\text{FM}} \cdot \mathbb{1}(d_m^{\text{FM}} > d_{\max}) + \text{d}_m^{\text{LM}} \cdot \mathbb{1}(d_m^{\text{LM}} > d_{\max}))$. We initially assume $c^{\text{MoD}} = \$3.5$ per mile (New York City Taxi & Limousine Commission 2022) and later on perform a sensitivity analysis to determine the effect of c^{MoD} on marketplace outcomes. Finally, we let $k = 5$ for the display constraint.

6.1.3. Demand Inputs. We consider one hour's worth of for-hire vehicle trips in Manhattan on February 8, 2018, using the NYC Open Data platform (NYC Open Data 2024). For computational efficiency, we cluster the trips into 4,700 origin-destination pairs, imposing an upper bound of $d_{\max}$ on cluster diameter.

Commuters have a base valuation γ_m for mode m, with $\gamma_m = \$10$ for a direct trip by car, $\gamma_m = \$7$ for a hybrid mode, and $\gamma_m = \$5$ if m exclusively uses transit. The average value of time across the entire commuter population is taken to be $\alpha_\tau = \$18.6/\text{hour}$ (Liu et al. 2019). We further partition the population based on their value of time (different time sensitivities across the commuter population may, for instance, depend on the income level (Börjesson, Fosgerau, and Algers 2012)): 75% of commuters are of "low" type and have an average value of time of $\beta_\theta \alpha_\tau = 0.75\alpha_\tau = \$13.95$ per hour; 25% of commuters are of "high" type and have a value of time of $1.75\alpha_\tau = \$32.55$ per hour. (These multipliers were chosen to ensure the aggregate average value of time of α_τ. We perform a sensitivity analysis on the type multipliers in Online Appendix C.5 to study the impact of type heterogeneity on system outcomes.) Commuters incur a transfer disutility $c^{\text{transfer}} > 0$ for each transfer. We initially assume $c^{\text{transfer}} = \$2$, although in subsequent experiments, we vary c^{transfer} to understand its impact on the efficiency benefits of an integrated marketplace.

Putting this together, the valuation of a type θ commuter for mode m is given by

$$v_{\theta m} = \left(\gamma_m + \beta_\theta \alpha_\tau (\tau_m^{\max} - \tau_m) - c^{\text{transfer}} n_m^{\text{transfer}}\right)^+, \quad (12)$$

where $\tau_m^{\max}$ is the walking time from origin to destination, and n_m^{transfer} is the number of transfers associated with mode m.

6.2. Results

We run Algorithm 1 on the above setup, where the solution oracle $\mathcal{O}$ is an exact solution to CP.

6.2.1. Impact of Size of Candidate Set of Lines. We first investigate how the quality of the returned solution trades off with the runtime associated with the two optimization problems solved by our algorithm, as the size of the candidate set of lines, denoted by L, increases. Intuitively, by increasing the size of the candidate set, the quality of the welfare-optimal solution improves, as potentially higher-quality lines are included into the set. This, however, comes at the cost of a significantly larger decision space.

Tables 1 and 2 illustrate this tradeoff. We observe that the number of modes, and the corresponding variables and constraints in the MILP, increase linearly in L.

Table 1. Runtime Dependence on the Size of the Candidate Set of Lines

		Number of variables				Runtime (s)	
L	$\lvert\mathcal{M}\rvert$	Binary	Continuous	Number of constraints	MIP gap (%)	CP	D-SP
50	152,402	152,452	304,804	169,503	0.55%	30.9	2.5
100	298,981	299,081	597,962	320,121	0.45%	123.1	5.0
200	604,866	605,066	1,209,732	634,654	0.86%	1,115.2	16.6
400	1,221,573	1,221,973	2,443,146	1,268,104	0.79%	9,765.5	35.9
800	2,434,172	2,434,972	4,868,344	2,513,553	0.66%	42,582.2	57.2

Note. The CP and D-SP columns respectively correspond to the runtimes associated with solving the Centralized Welfare Maximization MILP and the smaller welfare maximization subproblem, given the set of lines returned by the MILP.

Table 2. Dependence of System Metrics on Size of the Candidate Set of Lines

L	Welfare	Revenue	Profit	MoD costs	Transit costs	Throughput	No. of open lines
50	276,942	179,453	26,565	136,315	16,573	96.58%	40
100	288,160	121,026	1,058	97,822	22,147	96.42%	48
200	295,359	95,279	−10,678	82,868	23,088	96.55%	43
400	298,867	81,554	−18,464	76,342	23,676	96.30%	44
800	301,099	78,903	−18,125	73,274	23,754	96.61%	45

Although the MILP solver is able to return a near-optimal solution for all instances, the runtime required to solve (CP) exhibits a supra-linear increase, with more than 11 hours required when $L = 800$. This dwarfs that of the linear program whose solution we use to compute the optimal set of prices. These results highlight the methodological contribution of our algorithmic framework: Despite the intractability of the centralized problem, computing welfare-optimal prices, *given a solution*, remains a tractable task.

In regard to system metrics, as L increases, so does the total welfare of the system, with a more than 8% increase between $L = 50$ and $L = 800$. Although throughput remains approximately constant at 96%, revenue and profit steeply decrease, with the system being unprofitable for $L \geq 200$. Table 3 helps to understand this behavior. As the size (and as a result, the quality) of the candidate set of lines increases, there is a significant increase in transit usage across both types of commuters. This increase results in a drop in revenue; although total costs also decrease, this net decrease is not sufficient to recoup the revenue loss.

Table 3 also displays how mode usage varies by type under welfare-optimal solutions. As noted above, both high- and low-type commuters migrate toward transit options as line quality increases; the way in which they do this, however, differs. The fraction of high-type commuters using hybrid options (a little over a third) drops 3 points between $L = 50$ and $L = 800$, whereas there is a more than 20-point decrease in MoD usage. Low-type commuters, on the other hand, see a more than 25-point drop in hybrid usage and a 13-point drop in MoD usage. Still, across both types of commuters, hybrid options are almost (if not more) attractive than MoD-only options from a welfare perspective, thus highlighting the gains from introducing these mixed trips. Table 4 also illustrates that the different modes serve different "types" of trips: MoD-only trips are heavily biased toward short trips, with the median MoD-only trip lasting 10–15 minutes; the median transit-only trip is significantly longer, lasting 45–50 minutes; finally, hybrid trips are the longest, with a median trip length of 50–55 minutes. This precisely highlights the type of trip for which hybrid options are most advantageous: long trips that are poorly connected to existing transit lines.

6.2.2. Impact of Transfer Costs.

We next study how the benefits of such an integrated system vary with transfer costs, for a fixed candidate set of lines of size $L = 100$. Our results are shown in Table 5, which illustrate the welfare gains of a fully integrated platform. We observe a 14% decrease in welfare between $c^{\text{transfer}} = \$0$ and $c^{\text{transfer}} = \$8$ and a 50-point decrease in hybrid usage across both types of commuters. As transfer costs increase, MoD-only options become significantly more attractive to commuters; this then explains the corresponding increase in revenue and profit. Still, more than 20% of served commuters choose the hybrid option when $c^{\text{transfer}} = \$8$.

Figure 2 (see Online Appendix C.1) further illustrates the change in selected bus lines as the transfer cost increases. Under zero transfer costs, more than 70% of served demand uses options involving transit, which results in a dense set of open bus lines. As transfer costs increase, this set becomes more sparse; still, we see that selected lines serve the region evenly.

These results highlight the existence of settings in which hybrid trips improve on nonmixed options from

Table 3. Mode Usage by Commuter Type

	High-type commuters			Low-type commuters		
L	Hybrid	Transit	MoD	Hybrid	Transit	MoD
50	38.56%	1.57%	59.86%	53.33%	6.21%	40.47%
100	48.52%	8.72%	42.76%	51.10%	16.84%	32.06%
200	42.74%	17.99%	39.27%	41.53%	27.96%	30.51%
400	38.37%	23.83%	37.79%	32.84%	37.62%	29.54%
800	35.23%	28.43%	36.33%	27.53%	45.19%	27.28%

Table 4. Trip Time Quantiles (in Hours) by Mode Type

	Hybrid			Transit			MoD		
L	25%	50%	75%	25%	50%	75%	25%	50%	75%
50	0.72	0.92	1.18	0.56	0.76	1.01	0.15	0.25	0.38
100	0.67	0.87	1.15	0.53	0.70	0.90	0.14	0.20	0.28
200	0.64	0.85	1.16	0.58	0.81	1.03	0.13	0.19	0.25
400	0.62	0.84	1.15	0.60	0.80	1.06	0.13	0.18	0.24
800	0.62	0.84	1.11	0.57	0.78	1.06	0.12	0.18	0.23

Table 5. Dependence of System Metrics on c^{transfer}

				High-type commuters			Low-type commuters			
c^{transfer}	Welfare	Revenue	Profit	Hybrid	Transit	MoD	Hybrid	Transit	MoD	Throughput
0	306,033	108,614	−2,068	68.02%	1.91%	30.07%	74.98%	3.85%	21.17%	95.96%
2	288,160	121,026	1,056	48.52%	8.72%	42.76%	51.10%	16.84%	32.06%	96.42%
4	276,624	141,317	6,569	33.21%	11.78%	55.00%	37.70%	22.79%	39.51%	96.46%
6	268,844	159,927	14,318	24.74%	13.45%	61.81%	29.67%	22.71%	47.61%	96.38%
8	262,335	178,304	18,194	16.10%	14.17%	69.73%	24.71%	24.01%	51.28%	96.26%

a welfare standpoint, even under high transfer costs; they moreover emphasize the importance of frictionless MoD-to-transit transfers within these settings.

6.2.3. Impact of MoD Cost. We next study the impact of the per-mile MoD cost on system metrics. (Recall, we assume that these costs are prenegotiated and enforced via a contract between the platform and a ride-hailing company.) Our results are shown in Table 6. As the per-mile cost of the MoD service increases from $2 to $10 per mile, we observe a more than 30% decrease in welfare. Revenue and profit, on the other hand, are nonmonotonic: They exhibit a decrease from $2 to $4 per mile, followed by an increase from $5 to $10 per mile. The reason for this behavior is as follows: When c^{MoD} is low, commuters can take high-valuation, MoD-only trips at low cost, resulting in high welfare. As the per-mile cost of these trips increases, commuters move away from MoD-only trip to hybrid trips, which are associated with lower valuations; because these also include transit legs, these cannot be priced as high, which results in lower revenue and profit. However, past a certain point, the high cost per mile of the first- and last-mile legs, their maintained high usage, and decreased costs because of the MoD-only trips result in both higher revenue and profit.

Overall, these results highlight the welfare optimality of hybrid options, even at high MoD per-mile costs, as well as the importance of the inclusion of MoD-only options in an integrated marketplace.

7. Conclusion

Although ride-hailing services have been viewed as competitors to cities' public transit operations in recent years, in this paper, we investigated the extent to which, to the contrary, they can be leveraged as *complements*, because of the potentially massive gains from combining their on-demand capabilities with the sustainability of mass transit options. Specifically, we approached the question of designing an integrated mobility marketplace from a central planner's perspective via a market design lens, that is, by tackling the joint problem of pricing and network design of such an system. In our main methodological contribution, we leveraged linear programming duality to show that the pricing and optimization of these systems can be decoupled by solving a closely related centralized assignment problem that ignores commuter choice entirely, and deployed this framework to a real-world data set to obtain insights into the welfare impacts of such an integration relative to the status quo fragmented system.

This paper lends itself to a number of natural directions for future work. First, our work was concerned with problem of long-term planning and pricing of the multimodal system, before commuter flows are even realized. As noted above, the ability to solve the fluid problem we consider is often a necessary precursor to the design of real-time algorithms in online settings; the question of how to optimally dispatch an on-demand fleet to match the target flows computed by the platform is an important direction for future work.

Table 6. Dependence of System Metrics on c^{MoD}

				High-type commuters			Low-type commuters			
c^{MoD}	Welfare	Revenue	Profit	Hybrid	Transit	MoD	Hybrid	Transit	MoD	Throughput
2	329,823	250,057	34,459	5.29%	0.58%	94.14%	14.26%	2.84%	82.90%	96.36%
3	296,704	166,943	24,113	35.16%	4.47%	60.37%	39.75%	8.92%	52.33%	95.99%
4	279,284	114,719	208	52.39%	10.56%	37.06%	53.15%	21.65%	25.20%	96.33%
5	266,711	107,939	2,745	58.10%	16.09%	25.80%	56.33%	26.78%	16.88%	96.24%
6	256,176	110,134	10,100	60.59%	21.27%	18.14%	57.71%	31.19%	11.11%	96.11%
8	239,423	113,463	21,148	61.03%	29.20%	9.77%	54.46%	39.18%	6.36%	94.63%
10	225,688	121,717	38,492	59.64%	34.54%	5.82%	48.95%	47.44%	3.62%	90.36%

Additionally, our work does not treat the question of stakeholder *incentives*. An important question would be how to design contracts with ride-hailing services, which we assume in this work to be prenegotiated.

Acknowledgments

The authors thank the editor, associate editor, and two anonymous reviewers whose valuable suggestions led to significant improvements. The authors also thank Lu Liu at Clemson University for assisting with numerical experiments and Noémie Périvier for assisting with data processing.

References

Auad-Perez R, Van Hentenryck P (2022) Ridesharing and fleet sizing for on-demand multimodal transit systems. *Transportation Res. Part C Emerging Tech.* 138:103594.

Banerjee S, Freund D, Lykouris T (2022) Pricing and optimization in shared vehicle systems: An approximation framework. *Oper. Res.* 70(3):1783–1805.

Basciftci B, Van Hentenryck P (2023) Capturing travel mode adoption in designing on-demand multimodal transit systems. *Transportation Sci.* 57(2):351–375.

Bertsimas D, Sian Ng Y, Yan J (2020) Joint frequency-setting and pricing optimization on multimodal transit networks at scale. *Transportation Sci.* 54(3):839–853.

Bertsimas D, Delarue A, Jaillet P, Martin S (2019) Travel time estimation in the age of big data. *Oper. Res.* 67(2):498–515.

Bialas WF, Karwan MH (1984) Two-level linear programming. *Management Sci.* 30(8):1004–1020.

Bloomberg (2018) Helsinki's MaaS app, whim: Is it really mobility's great hope? Accessed October 14, 2024, https://www.bloomberg.com/news/articles/2018-10-25/is-helsinki-s-maas-app-whim-the-future.

Boeing G (2017) OSMnx: A python package to work with graph-theoretic OpenStreetMap street networks. *J. Open Source Software* 2(12):215.

Börjesson M, Fosgerau M, Algers S (2012) On the income elasticity of the value of travel time. *Transportation Res. Part A Policy Practice* 46(2):368–377.

Borndörfer R, Grötschel M, Pfetsch ME (2007) A column-generation approach to line planning in public transport. *Transportation Sci.* 41(1):123–132.

Brotcorne L, Labbé M, Marcotte P, Savard G (2001) A bilevel model for toll optimization on a multicommodity transportation network. *Transportation Sci.* 35(4):345–358.

Cashore JM, Frazier PI, Tardos É (2023) Dynamic pricing provides robust equilibria in stochastic ridesharing networks. *Math. Oper. Res.* 49(3):1303–2047.

Ceder A, Wilson NH (1986) Bus network design. *Transportation Res. Part B Methodological* 20(4):331–344.

Chakroborty P, Wivedi T (2002) Optimal route network design for transit systems using genetic algorithms. *Engrg. Optim.* 34(1):83–100.

Civitas (2020) Evaluation of UbiGo Stockholm. Accessed October 14, 2024, https://civitas.eu/sites/default/files/eccentric_m3.5._evaluation_of_ubigo_stockholm.pdf.

Colson B, Marcotte P, Savard G (2007) An overview of bilevel optimization. *Ann. Oper. Res.* 153(1):235–256.

Constantin I, Florian M (1995) Optimizing frequencies in a transit network: A nonlinear bi-level programming approach. *Internat. Trans. Oper. Res.* 2(2):149–164.

Cover TM (1999) *Elements of Information Theory* (John Wiley & Sons, Hoboken, NJ).

DART (2019) DART, Lyft creating new transit choices. Accessed October 14, 2024, https://www.dart.org/about/news-and-events/newsreleases/newsrelease-detail/dart–lyft-creating-new-transit-choices-1213.

Daskalakis C, Goldberg PW, Papadimitriou CH (2009) The complexity of computing a Nash equilibrium. *SIAM J. Comput.* 39(1):195–259.

Dempe S (2002) *Foundations of Bilevel Programming* (Springer Science & Business Media, Boston).

Fan W, Machemehl RB (2006) Using a simulated annealing algorithm to solve the transit route network design problem. *J. Transportation Engrg.* 132(2):122–132.

Fontaine P, Minner S (2014) Benders decomposition for discrete–continuous linear bilevel problems with application to traffic network design. *Transportation Res. Part B Methodological* 70:163–172.

Gallego G, Topaloglu H (2019) *Revenue Management and Pricing Analytics,* International Series in Operations Research & Management Science (Springer, New York).

Gallego G, Van Ryzin G (1994) Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Sci.* 40(8):999–1020.

Hansen P, Jaumard B, Savard G (1992) New branch-and-bound rules for linear bilevel programming. *SIAM J. Sci. Statist. Comput.* 13(5):1194–1217.

Huang HJ (2002) Pricing and logit-based mode choice models of a transit and highway system with elastic demand. *Eur. J. Oper. Res.* 140(3):562–570.

LA Metro (2019) LA Metro – Ride with Via. Accessed October 14, 2024, https://ridewithvia.com/news/via-and-la-metro-launch-pilot-program-to-evolve-public-transit-in-los-angeles.

Lanzetti N, Schiffer M, Ostrovsky M, Pavone M (2023) On the interplay between self-driving cars and public transportation. *IEEE Trans. Control Network. Systems* 11(3):1478–1490.

Liu Y, Bansal P, Daziano R, Samaranayake S (2019) A framework to integrate mode choice in the design of mobility-on-demand systems. *Transportation Res. Part C Emerging Tech.* 105:648–665.

Lubin M, Yamangil E, Bent R, Vielma JP (2018) Polyhedral approximation in mixed-integer convex optimization. *Math. Programming* 172(1):139–168.

MARTA Reach (2022) On-demand rideshare pilot program. Accessed October 14, 2024, https://www.itsmarta.com/reach.aspx.

McFadden D (1973) Conditional logit analysis of qualitative choice behavior. Zarembka P, ed. *Frontiers in Econometrics* (Academic Press, New York), 105–142.

Moovel (2024) Introducing the world's first transit engagement platform. Accessed October 14, 2024, https://www.moovelus.com/.

New York City Department of Transportation (2021) New York City mobility report. Accessed October 14, 2024, https://www.nyc.gov/html/dot/html/about/mobilityreport.shtml.

New York City MTA (2024) Accessed October 14, 2024, https://new.mta.info/fares.

New York City Taxi & Limousine Commission (2022) Taxi fare. Accessed October 14, 2024, https://www.nyc.gov/site/tlc/passengers/taxi-fare.page.

NYC Open Data (2024) NYC open data. Accessed October 14, 2024, https://opendata.cityofnewyork.us/.

Parbo J, Nielsen OA, Prato CG (2014) User perspectives in public transport timetable optimisation. *Transportation Res. Part C Emerging Tech.* 48:269–284.

Périvier N, Hssaine C, Samaranayake S, Banerjee S (2021) Real-time approximate routing for smart transit systems. Chaintreau A, Golubchik L, Zhang Z-L, eds. *Proc. ACM Measurement Analysis Comput. Systems*, vol. 5, no. 2 (Association for Computing Machinery, New York), 1–30.

Pinto HK, Hyland MF, Mahmassani HS, Verbas İÖ (2019) Joint design of multimodal transit networks and shared autonomous mobility fleets. *Transportation Res. Proc.* 38:98–118.

Sun S, Szeto W (2019) Optimal sectional fare and frequency settings for transit networks with elastic demand. *Transportation Res. Part B Methodological* 127:147–177.

US Department of Transportation (2015) Smart city challenge. Accessed October 14, 2024, https://www.transportation.gov/smartcity.

Verbas İÖ, Mahmassani HS (2015) Integrated frequency allocation and user assignment in multimodal transit networks: Methodology and application to large-scale urban systems. *Transportation Res. Rec.* 2498(1):37–45.

Walker J (2018) The transit ridership recipe. Accessed October 14, 2024, https://humantransit.org/basics/the-transit-ridership-recipe.

Wischik D (2018) The price of choice: Models, paradoxes, and inference for 'mobility as a service'. *Proc. 56th Annual Allerton Conf. Comm. Control Comput.* (IEEE, New York), 604–610.

Wu D, Yin Y, Lawphongpanich S (2011) Pareto-improving congestion pricing on multimodal transportation networks. *Eur. J. Oper. Res.* 210(3):660–669.

Wu D, Yin Y, Lawphongpanich S, Yang H (2012) Design of more equitable congestion pricing and tradable credit schemes for multimodal transportation networks. *Transportation Res. Part B Methodological* 46(9):1273–1287.

Yang Y, Diez-Roux AV (2012) Walking distance by trip purpose and population subgroups. *Amer. J. Preventive Medicine* 43(1): 11–19.

Yu B, Kong L, Sun Y, Yao B, Gao Z (2015) A bi-level programming for bus lane network design. *Transportation Res. Part C Emerging Tech.* 55:310–327.