Transportation Research Part C 169 (2024) 104864

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Pricing for multi-modal pickup and delivery problems with heterogeneous users

Mark Beliaev [a,*], Negar Mehr [b], Ramtin Pedarsani [a]

[a] Department of Electrical and Computer Engineering, University of California Santa Barbara, Santa Barbara, CA 93106, USA
[b] Department of Mechanical Engineering, University of California Berkeley, Berkeley, CA 94720, USA

ARTICLE INFO

Keywords:
Transportation networks
Congestion games
Optimization

ABSTRACT

In this paper, we study the pickup and delivery problem with multiple transportation modalities, and address the challenge of efficiently allocating transportation resources while price matching users with their desired delivery modes. More precisely, we consider that orders are demanded by a heterogeneous population of users with varying trade-offs between price and latency. To capture how prices affect the behavior of heterogeneous selfish users choosing between multiple delivery modes, we construct a congestion game taking place over a form of star network, where each source–sink pair is composed of parallel links connecting users with their preferred delivery method. Using the unique geometry of this network, we prove that one can set prices explicitly to induce any desired network flow, i.e, given a desired allocation strategy, we have a closed-form solution for the delivery prices. We conclude by performing a case study on a meal delivery problem with multiple courier modalities using data from real world instances.

1. Introduction

As the world continues to integrate with digital technology, we become more reliant on e-commerce services such as food delivery and ride-hailing. The global food delivery market has seen exponential growth, with the most mature markets becoming four to seven times larger from 2018 to 2021 (Ahuja et al., 2021). In 2022, Uber reported a 19% year-over-year increase in online bookings, marking a daily average of 23 million trips on their platform (Uber, 2023). Despite this growth, many pickup and delivery services operate under low profit margins due to high driver wages (Shetty et al., 2022).

To fulfill market demands and mitigate these costs, recent efforts have been made to introduce autonomous transportation methods for food delivery and ride hailing, such as delivery drones, electric vertical takeoff and landing (eVTOL) aircrafts, and sidewalk autonomous delivery robots (SADRs) (Moshref-Javadi and Winkenbach, 2021; Starship, 2023). For example, Archer Aviation Inc. has recently revealed their plans to provide passengers with an eVTOL aircraft travel option between O'Hare International Airport and Vertiport Chicago that takes roughly 10 min, a trip which can take upwards of an hour or more during rush hour traffic (Gump, 2023).

As these modalities are introduced, it is important for service providers to develop new resource allocation strategies that efficiently utilize emergent transportation modalities, while coinciding with customer preferences. With the advent of urban air mobility, recent research has studied demand modeling, operations, and integration with existing infrastructure (Garrow et al., 2021). For example, surveys examining commuter preferences regarding transportation have found significant heterogeneity in individuals' value of time, and that the median value of time for air taxis is larger compared to other modalities (Binder et al.,

* Corresponding author.
E-mail addresses: mbeliaev@ucsb.edu (M. Beliaev), negar@illinois.edu (N. Mehr), ramtin@ucsb.edu (R. Pedarsani).

https://doi.org/10.1016/j.trc.2024.104864
Received 20 August 2023; Received in revised form 16 June 2024; Accepted 16 September 2024
Available online 2 November 2024
0968-090X/

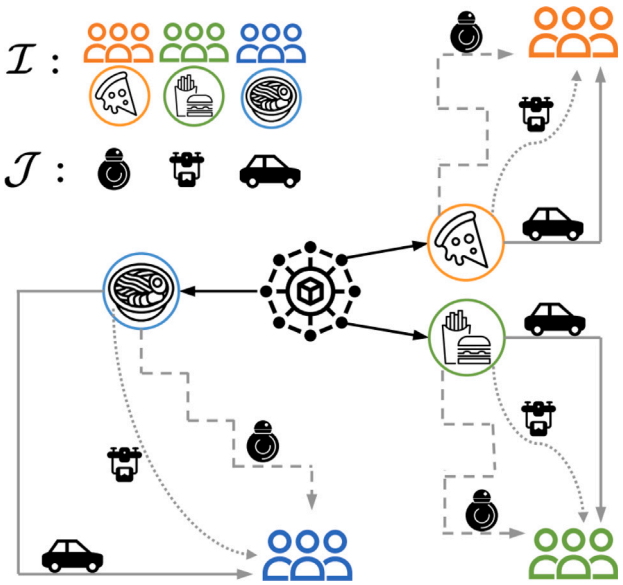

Fig. 1. We represent the pickup and delivery problem as a congestion game played over a star network. Each source-sink pair is denoted by $i \in \mathcal{I}$, which can be viewed as a population of users at some location demanding a particular order at a certain rate. Each source-sink pair is connected by a set of parallel edges $j \in \mathcal{J}$, which can be viewed as the set of delivery modes the users choose from. Note that we are not concerned with how the couriers are routed to the pickup or delivery location, and instead focus on how we allocate the different delivery modes for each order. Specifically, our goal is to induce an optimal allocation of transportation modalities by appropriately setting prices for each order-modality pair.

2018; Garrow et al., 2019). As the inclusion of urban air mobility is predicted to disrupt urban transportation, it will be crucial for existing rideshare providers to adapt their pricing and allocation strategies. In light of these developments, we address the challenge of using customer preferences to set prices that efficiently allocate transportation resources amongst them.

This paper examines the pickup and delivery problem with multiple transportation modalities, and demonstrates how one can achieve a desired allocation strategy for a set of orders by appropriately setting prices for each modality. Specifically, we consider orders demanded by a heterogeneous population of users with varying trade-offs between price and latency. This problem is analogous to a congestion game taking place over a form of star network: one central *source* node is used to connect a set of *sink* nodes, where for each source–sink pair, there is a directed graph composed of parallel links between the source and sink node. This way, we can use the star network to represent a central delivery system responsible for connecting users placing some order, with their preferred transportation modality. Note that by using parallel links to represent the modalities, we are not concerned with routing individual vehicles, and instead focus on allocating transportation resources. We illustrate this congestion game in Fig. 1, depicting a meal delivery problem where customers at different locations place orders for food via the available delivery modes. This unique network structure enables us to show that we can explicitly define prices to induce any desired network flow, i.e, given a desired allocation strategy, we have a closed-form solution for the delivery prices.

The main contributions of this work are:

- We construct a congestion game that captures how prices affect the behavior of heterogeneous selfish users choosing between multiple delivery modes.
- Building on results from prior works, we prove that in these settings, the set of prices can be explicitly defined for any desired network flow.
- We demonstrate our results with a case study on a meal delivery service with multiple courier modalities, using real world instances provided by Grubhub (Reyes et al., 2018).
- Under additional assumptions, we extend our results by allowing users to have varying trade-offs for each modality. We demonstrate this with a case study on a taxi service with urban air transportation to and from the O'Hare International Airport, using data provided by the city of Chicago (Transportation, 2023).

2. Related work

The application of emerging transportation modalities such as unmanned aerial vehicles or drones has drawn a lot of attention. Many works look at how drones can be utilized in logistic operations such as delivery systems (Beliaev et al., 2023), urban air taxi (Ale-Ahmad and Mahmassani, 2023), on-demand meal delivery (Liu, 2019), as well as many other applications (Moshref-Javadi and Winkenbach, 2021). Other works look at safety verification for dynamical systems utilizing drones to account for factors such as collision avoidance (Llanes et al., 2022) and schedule feasibility (Wei et al., 2021). The pickup and delivery vehicle routing problem

with drones has also been considered by some, where mixed integer linear programming models are used to find routing solutions for optimizing various objectives (Gacal et al., 2020; Lu et al., 2022). Unlike these works, our research lies in the broader field of congestion games, specifically building on previous works that consider pricing in non-atomic congestion games.

Congestion games aim to allocate traffic over transportation networks represented by graphs, where each road corresponds to an edge with a latency function representing the travel time experienced by users on that edge (Dafermos and Sparrow, 1969). In these settings, one aims to find the optimal network flow that minimizes a social cost, such as the aggregate latency experienced by all users. However, if we assume that users are self-interested and choose their routes selfishly by minimizing their individual latency, the resulting flow follows a network equilibrium (Wardrop, 1952; Sheffi, 1985). One area of research is focused on categorizing the trade-off in social cost between the optimal network flow and the equilibrium network flow (Roughgarden, 2005; Lazar et al., 2021, 2017). Many works specifically look at how tolling can be used to price network edges such that the equilibrium network flow corresponds to the desired optimal flow (Dafermos, 1973; Cole et al., 2003a; Brown and Marden, 2017). In our work, we make the distinction that users are heterogeneous in their trade-off between price and time.

While in the homogeneous case it has been long known that marginal cost pricing can guarantee that the equilibrium flow equals the optimal flow (Beckmann et al., 1956), this strategy does not hold for heterogeneous populations. More recent research has demonstrated that for directed graphs with one source–sink pair, optimal tolls exist and can be found by solving a polynomial size set of linear inequalities, given that the number of users in the heterogeneous population is finite (Cole et al., 2003b). In this seminal paper, it was assumed that the model was nonatomic, meaning that each user corresponded to an infinitesimal unit of flow, and inelastic, meaning that the demand could not change as a function of the road parameters. Following this work, others have improved the result by considering multicommodity networks (Karakostas and Kolliopoulos, 2004; Fleischer et al., 2004), allowing user demand to be elastic (Karakostas and Kolliopoulos, 2006), and addressing the atomic setting (Fotakis et al., 2010). In this paper, we keep the assumption of a nonatomic model with inelastic demands, but consider a graph structure which is unique to the pickup and delivery problem considered. By exploiting this graph structure, we can define prices explicitly to induce any desired network flow without limiting it to an optimal flow. Whereas prior works directly use Linear Program (LP) formulations to find edge prices in general directed graphs, our theoretical results imply that one can first find path prices combinatorially to simplify the LP formulation.

The rest of the paper is organized as follows. In the subsequent Section 3, we formally introduce the problem setting and show how it is analogous to a congestion game. Following this, in Section 4 we describe our theoretical result in the general framework of the aforementioned congestion game, and in Section 5, we describe the specific framework that is used to model the pickup and delivery problem and find the optimal allocation strategy. We go on to apply our theoretical results on this framework in Section 6, which consists of our two case studies using the public Grubhub dataset (Reyes et al., 2018) for meal delivery and the Chicago Transportation Network Providers dataset (Transportation, 2023) for taxi services. Lastly, we conclude our work in Section 7, listing potential avenues for improvement and further research.

3. Problem formulation

We model our pickup and delivery problem using a static system, where during a given time interval,[1] there is a set of orders $\mathcal{I}$ demanded by a population of users. We consider that each order originates from a unique neighborhood composed of a heterogeneous population represented by the interval $[0,1]$, where each point $a \in [0,1]$ is a non-cooperative and infinitesimal unit referred to as a user. We sort these users by money sensitivity, where in general, we can view $\alpha_i : [0,1] \to (0,\infty)$ as an unbounded, non-decreasing function representing the trade-off between price and time for users in the population corresponding to order i. Thus, when placing an order $i \in \mathcal{I}$, each user chooses one of the J delivery modes $j \in \mathcal{J} : \{1,\dots,J\}$ based on latency $\ell_{i,j}$, dollar price $\tau_{i,j}$, and their money/time valuation $\alpha_i(a)$. We assume users placing order $i \in \mathcal{I}$ have inelastic demands, i.e., they will not switch their demand to a different order and will always choose one of the J delivery modes. Our goal is then to find the set of delivery prices which would induce some desired allocation of users between the delivery modes $j \in \mathcal{J}$ for each order $i \in \mathcal{I}$ (see Table 1).

This problem is analogous to a congestion game played over a star network as portrayed in Fig. 1, where each order $i \in \mathcal{I}$ corresponds to a source–sink pair connected by a set of parallel edges $\mathcal{J}$ representing the different delivery options. Each source–sink pair $i \in \mathcal{I}$ has an associated demand of traffic flow at the sink which represents the population of users $a \in [0,1]$ requesting deliveries. Although we model this flow demand with the unit interval to simplify notation, we can allow for an arbitrary demand r_i at each source–sink pair i. The edge corresponding to modality $j \in \mathcal{J}$ for source–sink pair $i \in \mathcal{I}$ has a congestion dependent latency $\ell_{i,j}$, which represents the time needed to complete the order, and a price issued to control congestion $\tau_{i,j}$, which represents the dollar price paid by the user, both of which are assumed to be nonnegative. Note that when we drop index j from the notation of terms like $\ell_{i,j}$ by writing ℓ_i, we refer to the set of latencies $\{\ell_{i,j}\}_{j\in\mathcal{J}}$ over all the edges $\mathcal{J}$ for a given source–sink pair i.

With this approach, we can view network flow as an allocation of users over the delivery modes. To represent such allocation strategies, we define $0 \le x_{i,j} \le 1$ as the flow of users on edge $j \in \mathcal{J}$ corresponding to source–sink pair $i \in \mathcal{I}$, where $\sum_{j\in\mathcal{J}} x_{i,j} = 1$ must be satisfied. More precisely, for each source–sink pair i we view this flow as a Lebesgue-measurable function $x_i : [0,1] \to \mathcal{J}$ which corresponds to a flow over the edges $\{x_{i,j}\}_{j\in\mathcal{J}}$. We use notations $x = \{x_{i,j}\}_{i\in\mathcal{I},j\in\mathcal{J}}$ and $\tau = \{\tau_{i,j}\}_{i\in\mathcal{I},j\in\mathcal{J}}$ to denote the entire set of edge flows and edge prices, respectively. As we will later show in Section 5 when defining the specific optimization problem, we can use x as a decision variable to find an optimal allocation strategy for a given objective, and explicitly define prices τ that induce this desired strategy. For now, we continue to detail how latency and user equilibrium are considered in our framework.

[1] Without loss of generality, we can define the time interval during which the orders $\mathcal{I}$ are demanded as one hour, using the same unit of time for all variables and constants throughout our formulation.

Table 1
Notations.

$\mathcal{I} \mid i \in \mathcal{I}$	$\triangleq$	Set of orders | individual order (source–sink pair)
$\mathcal{J} \mid j \in \mathcal{J}$	$\triangleq$	Set of modes | individual mode (edge)
J	$\triangleq$	Number of modes, i.e., the cardinality of $\mathcal{J}$
$a \in [0,1]$	$\triangleq$	A non-cooperative and infinitesimal unit referred to as a user
$\alpha_i : [0,1] \to (0,\infty)$	$\triangleq$	Function representing the trade-off between price and time for users placing order i.
$\ell_{i,j}$	$\triangleq$	Latency: total time required to complete order i using mode j
$\tau_{i,j}$	$\triangleq$	Dollar price of placing order i using mode j
$x_{i,j}$	$\triangleq$	Flow of users on edge j corresponding to source–sink pair i
$x_i : [0,1] \to \mathcal{J}$	$\triangleq$	Function representing flow over the edges $\mathcal{J}$ for source–sink pair i
$p^a_{i,j}$	$\triangleq$	User cost (hours) $a \in [0,1]$ assigns to edge j for source–sink pair i
$s_{i,j} \mid t_{i,j} \mid u_{i,j}$	$\triangleq$	Service time | travel time | pickup time (order i, mode j)
$r_i \mid d_i$	$\triangleq$	Pickup location | drop-off location (for order i)
N_j	$\triangleq$	the total number of vehicles for mode j
μ_j	$\triangleq$	Order completion rate for mode j in units of orders per hour
$\rho_j \mid \bar{\rho}_j$	$\triangleq$	Utilization of mode j | upper bound on utilization of mode j
$\beta_{i,j}$	$\triangleq$	Portion of available couriers distributed around pickup location r_i such that their travel times are uniform in $[0, k_j]$
k_j	$\triangleq$	Constant unit of time used for computing $\beta_{i,j}$
c_j	$\triangleq$	Cost in dollars for completing one order with mode j
$\bar{c}_j$	$\triangleq$	Cost in dollars per hour for operating mode j
C	$\triangleq$	Cost in dollars per hour for operating the entire system

3.1. Congestion

We first describe the congestion element of our framework, namely, the latency function defined for each edge. Specifically, we assume that each edge $j \in \mathcal{J}$ corresponding to source–sink pair $i \in \mathcal{I}$ has a nonnegative and continuous latency $\ell_{i,j}$ as a function of the entire network flow x. Each latency function $\ell_{i,j}$ describes the time it takes for an order i delivered by modality j to arrive at the customer's location from the moment it was placed. We note that in order to claim our main theoretical result, we do not need any further restrictions on the latency functions $\ell_{i,j}$. We leave further discussion regarding latency to Section 5, where we model latency using concepts from queuing theory for our application. Until then, we stick with the aforementioned assumptions and simply use notation $\ell_{i,j}(x)$ when defining edge latency.

3.2. User equilibrium

We are now ready to discuss how users choose between the different delivery modes. When confronted with a set of prices τ_i and latencies ℓ_i for the varying edge options $j \in \mathcal{J}$, user $a \in [0,1]$ will choose the edge with the smallest cost $\ell_{i,j}(x) + \alpha_i(a)\tau_{i,j}$. Essentially, every source–sink pair $i \in \mathcal{I}$ corresponds to its own nonatomic game in which users $a \in [0,1]$ choose between the $j \in \mathcal{J}$ pure strategies available. The non-cooperative behavior of users results in a Nash equilibrium, which is a stable point where no user has an incentive to unilaterally alter their chosen strategy. Specifically, we let $p^a_{i,j}(x, \tau_i) = \ell_{i,j}(x) + \alpha_i(a)\tau_{i,j}$ represent the evaluation user $a \in [0,1]$ assigns to edge j for source–sink pair i.

Definition 1. For a given source–sink pair $i \in \mathcal{I}$, we call the flow $x_i : [0,1] \to \mathcal{J}$ an equilibrium or Nash flow for instance $(\alpha_i, \ell_i, \tau_i)$ if for any user $a \in [0,1]$ and edge $j \in \mathcal{J}$:

$$p^a_{i,x_i(a)}(x, \tau_i) \le p^a_{i,j}(x, \tau_i). \tag{1}$$

The existence of such Nash flows is a well known and a general result (Schmeidler, 1973).

Proposition 1. *For a given source–sink pair $i \in \mathcal{I}$, any instance $(\alpha_i, \ell_i, \tau_i)$ admits a Nash flow $x_i : [0,1] \to \mathcal{J}$ satisfying Eq. (1).*

We point out that the above Proposition requires not only the cost function $p^a_{i,j}$ to be nonnegative and continuous, but also the set of possible flows to be a nonempty, convex, and compact. This is trivially satisfied when considering only the demand, as done when presenting our theoretical results in Section 4. However, we will need to make sure the set is nonempty when including a supply constraint, as done when formulating our optimization problem in Section 5.

Note that in the above results pertaining to Nash equilibria, for each source sink-pair $i \in \mathcal{I}$, we consider the flow x_i independently from the entire network flow x, keeping the remaining flows constant. Since the latency $\ell_{i,j}(x)$ is assumed to be a function of the entire network flow x, one may require a network flow $x : \{x_i\}_{i \in \mathcal{I}}$ for which all source sink pairs $i \in \mathcal{I}$ exhibit Nash equilibrium under their corresponding flow $x_i : [0,1] \to \mathcal{J}$. We use the term *stable allocation strategy* to encompass this notion, formally defining it below.

Definition 2. For a given star network defined by the source–sink pairs $i \in \mathcal{I}$ and edges $j \in \mathcal{J}$, we call the network flow $x : \{x_i\}_{i \in \mathcal{I}}$ a stable allocation strategy for instance (α, ℓ, τ) if for all source–sink pairs $i \in \mathcal{I}$, the corresponding flow $x_i : [0,1] \to \mathcal{J}$ is an equilibrium flow satisfying Eq. (1).

As we show in the subsequent section, any network flow x is a stable allocation strategy for some set of prices τ.

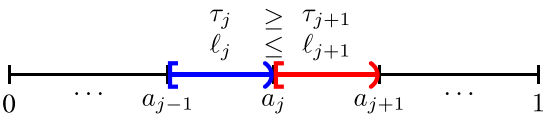

Fig. 2. A sketch depicting how a canonical Nash flow splits the population $a \in [a_0, a_J]$ into subintervals $[a_{j-1}, a_j) : x(a) = j$, where $a_0 = 0$, $a_J = 1$, and $j \in \mathcal{J}$. Note that order i is left out from notation.

4. Theoretical results

Before stating our main results, we need to elaborate on one more property of equilibrium flows that applies to individual source–sink pairs. Intuitively, we expect Nash flows to exhibit a structure where users $a \in [0, 1]$ close to 0, who value time more than money, will choose an option with small latency but large price. Similarly, users further away from 0 will choose an option with a relatively larger latency but a smaller price. Finally, users close to 1 will choose an option with very large latency in order to pay a very small price. We encapsulate this notion below.

Definition 3. For a given source–sink pair $i \in \mathcal{I}$, a flow x_i at Nash equilibrium is *canonical* if:

- For any edge $j \in \mathcal{J}$, the users assigned to j form a possibly empty or degenerate subinterval of $[0, 1]$.
- If $a_1 < a_2$, then $\ell_{i,x_i(a_1)}(x) \leq \ell_{i,x_i(a_2)}(x)$.
- If $a_1 < a_2$, then $\tau_{i,x_i(a_1)} \geq \tau_{i,x_i(a_2)}$.

In other words, a canonical Nash flow x_i splits $[0, 1]$ into at most J potentially degenerate sub intervals, inducing an ordering over the edges to which x_i assigns users that is nondecreasing in latency and nonincreasing in prices. We portray this in Fig. 2. Using results from prior work which proposed this definition (Cole et al., 2003b), we can state the following existence property, providing an independent proof of this proposition in Appendix A for completeness.

Proposition 2. *For a given source–sink pair $i \in \mathcal{I}$, every instance $(\alpha_i, \ell_i, \tau_i)$ admits a canonical Nash flow.*

With these properties, we can say that for a given source–sink pair $i \in \mathcal{I}$ and instance $(\alpha_i, \ell_i, \tau_i)$, there exists a canonical Nash flow $\tilde{x}_i : [0, 1] \to \mathcal{J}$. This canonical Nash flow represents the flow $\{\tilde{x}_{i,j}\}_{j \in \mathcal{J}}$, where users in interval $[a_{j-1}, a_j] \in [0, 1]$ are routed on edge j for some corresponding set $a_0 \leq a_1 \leq \cdots \leq a_J$, with $a_0 = 0$ and $a_J = 1$. In the pickup and delivery setting, we can assume that the delivery provider already has a set of flows $\{x_{i,j}\}_{j \in \mathcal{J}}$ representing the desired allocation strategy for order i, and wants to find a corresponding set of prices $\{\tau_{i,j}\}_{j \in \mathcal{J}}$ such that the induced equilibrium flow $\{\tilde{x}_{i,j}\}_{j \in \mathcal{J}}$ is equal to the desired flow. Building on top of the aforementioned results, we find a closed-form solution to this problem.

Theorem 1. *For a given source–sink pair $i \in \mathcal{I}$, any desired flow $\{x_{i,j}\}_{j \in \mathcal{J}}$ is an equilibrium flow for instance $(\alpha_i, \ell_i, \tau_i)$, where the set $\mathcal{J} : \{1, \ldots, J\}$ orders the edges by non-decreasing latency, $\alpha_i : [0, 1] \to (0, \infty)$ is a non-decreasing distribution function, ℓ_i is the set of corresponding edge latencies, and τ_i is the set of prices defined by:*

$$\tau_{i,j} = \tau_{i,J} + \sum_{k=j}^{J-1} \frac{\ell_{i,k+1} - \ell_{i,k}}{\alpha_i(a_k)} \qquad \forall j \in \mathcal{J}, \tag{2}$$

where $\tau_{i,J}$ is any predefined price for the cheapest option.

Proof. The proof strategy is as follows: given the result of Proposition 2 which states that every instance $(\alpha_i, \ell_i, \tau_i)$ admits a canonical Nash flow, we use the properties of canonical Nash flows along with a subset of the inequalities defined for Nash equilibrium in Eq. (1) to show that for some desired flow $\{x_{i,j}\}_{j \in \mathcal{J}}$ to be at equilibrium, there is only one set of valid prices τ_i. We complete the proof by showing that the corresponding set of prices τ_i does indeed satisfy all of the inequalities defined in Eq. (1). The full proof is provided in Appendix B. □

It follows directly that given any network flow $x : \{x_i\}_{i \in \mathcal{I}}$ representing a desired allocation strategy over all orders, one can independently set prices $\tau : \{\tau_i\}_{i \in \mathcal{I}}$ for each source–sink pair to make x a stable allocation strategy.

Corollary 1. *For a given star network defined by source–sink pairs $i \in \mathcal{I}$ and edges $j \in \mathcal{J}$, any network flow $x : \{x_i\}_{i \in \mathcal{I}}$ is a stable allocation strategy for instance (α, ℓ, τ) when the set of prices τ is defined according to Eq. (2).*

We note that under additional assumptions, the results of Theorem 1 and Corollary 1 can be extended to the setting where users have different trade-offs between price and latency $\alpha_{i,j}$ for different modes of transportation j as well as orders i.

Corollary 2. *For a given source–sink pair $i \in \mathcal{I}$ and instance $(\alpha_i, \ell_i, \tau_i)$, if a desired flow $\{x_{i,j}\}_{j \in \mathcal{J}}$ is inducible under some equilibrium flow $\tilde{x}_i : [0, 1] \to \mathcal{J}$ that routes users in interval $[a_{j-1}, a_j] \in [0, 1]$ on edge j for some corresponding set $a_0 \leq a_1 \leq \cdots \leq a_J$, then the set of prices τ_i must be defined by:*

$$\tau_{i,j} = \frac{\alpha_{i,j+1}(a_j)}{\alpha_{i,j}(a_j)} \tau_{i,j+1} + \frac{\ell_{i,j+1} - \ell_{i,j}}{\alpha_{i,j}(a_j)} \qquad \forall j \in \{1, \ldots, J-1\}, \tag{3}$$

where $a_0 = 0$ *and* $a_J = 1$, α_i *is a set of non decreasing functions* $\alpha_{i,j} : [0,1] \to (0,\infty)$ *for* $j \in \{1,\dots,J\}$ *such that given* $a < a'$, $\frac{\alpha_{i,j+1}(a)}{\alpha_{i,j}(a)} \geq \frac{\alpha_{i,j+1}(a')}{\alpha_{i,j}(a')}$ *for all* $j \in \{1,\dots,J-1\}$, ℓ_i *is the set of corresponding edge latencies, and* $\tau_{i,J}$ *is any predefined price for the cheapest option. It follows directly that if the network flow* $x : \{x_i\}_{i\in\mathcal{I}}$ *is a stable allocation strategy for instance* (α,ℓ,τ), *then the set of prices* τ *is defined by Eq. (3).*

The proof is provided in Appendix C, where we can no longer rely on Proposition 2 that allows us to utilize the properties of canonical Nash flows. Due to this, we only consider well behaved equilibrium flows which divide the users into J potentially degenerate sub intervals, inducing an ordering over the edges which preserves the assumption placed on the distribution functions $\alpha_{i,j}$. Intuitively, this assumption requires the modalities $\{1,\dots,J\}$ ordered by their relative luxury, since users $a \in [0,1]$ closer to 0, who have a greater value of time overall, will have a proportionally higher value of time for more luxurious modes. Hence, Corollary 2 tells us that if the desired flow $\{x_{i,j}\}_{j\in\mathcal{J}}$ can be induced by such an equilibrium flow, then the prices $\{\tau_{i,j}\}_{j\in\mathcal{J}}$ must follow Eq. (3). Although this is a weaker result compared to Theorem 1, it can be strengthened and used in applicable settings as we will demonstrate in our second case study provided in Section 6.

Lastly, we point out that so far we have not made any claims regarding the uniqueness of equilibrium flows. While the uniqueness of Nash flows for networks composed of parallel links is a known result (see Theorem 1 in Orda et al. (1993) or Proposition 3.3 in Milchtaich (2000)), it requires additional assumptions: convexity and strict monotonicity of cost functions $p^a_{i,j}$ with respect to flow $x_{i,j}$. To keep our results in Theorem 1, as well as Corollaries 1 and 2, general for any arbitrary latency functions $\ell_{i,j}$, we forego making such restrictions. However, we note that the latency function $\ell_{i,j}$ used in our case studies satisfies the aforementioned requirements. Furthermore, due to the results of Proposition 2, any two allocation strategies x and x' that are stable for instance (α,ℓ,τ), must induce the same ordering over the edges $j \in \mathcal{J}$ that is nondecreasing in latency $\ell_{i,j}$, and nonincreasing in prices $\tau_{i,j}$, for all source–sink pairs $i \in \mathcal{I}$.

5. Pickup and delivery problem

To show the usability of our model, we apply our theoretical framework to the pickup and delivery problem with multiple courier types. Our goal is to find the optimal allocation strategy with respect to some objective, where we will use Theorem 1 to set the prices which induce this desired strategy. Our objective will be to find the optimal values of x which minimize the expected latency over all orders:

$$L(x) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \ell_{i,j}(x)x_{i,j}. \tag{4}$$

Using the model and optimization problem developed in this section, we will perform case studies on both a meal delivery service and a taxi service in the following Section 6. Before we set up and solve this optimization problem, we first specify how latency is measured, and how cost is accounted for.

5.1. Latency model

We begin by characterizing each order $i \in \mathcal{I}$ by a 2–tuple $\langle r_i, d_i\rangle$, consisting of a pick-up and drop-off location, respectively. We would like our system to model the time it takes to complete a customer's order from the moment it was placed. We refer to this as the latency $\ell_{i,j}$ for order $i \in \mathcal{I}$ and modality $j \in \mathcal{J}$, computing it as:

$$\ell_{i,j}(x) = s_{i,j} + t_{i,j} + u_{i,j}(x). \tag{5}$$

Essentially, the above Eq. (5) splits the latency $\ell_{i,j}$ into three components: service time $s_{i,j}$, travel time $t_{i,j}$, and pickup time $u_{i,j}$ for modality j of order i.

In total, to compute the latency $\ell_{i,j}$ of an order, we account for how long it takes a courier to arrive at the designated pickup location $u_{i,j}$, the travel time between the pickup and drop-off locations $t_{i,j}$, and the service time required $s_{i,j}$. We view the service time $s_{i,j}$ as a constant representing the time spent at the pickup and drop-off locations when completing order i using modality j. Some examples of this include parking for vehicle couriers, landing for aerial couriers, loading, and unloading. Similarly, we define the travel time $t_{i,j}$ as the time it takes to physically travel between pickup r_i and drop-off d_i locations using modality j. The travel time $t_{i,j}$ between locations can be pre-computed separately for each modality j and order i using some known functions. Lastly, we view the pick-up time $u_{i,j}$ as the time it takes for a courier of modality j to arrive at pick-up location r_i. Unlike the other two components, the time required for pickup $u_{i,j}$ should depend on the availability of couriers captured by our decision variable x, as well as the expected travel time between the pickup location and nearest available courier.

To account for the availability of couriers, we use the concept of server utilization from queuing theory. Specifically, we use the $M/M/c$ queue as an approximate model for the availability of couriers since we can obtain closed form formulas for the average order arrival and order completion rates. For a given modality j, we set c to be the total number of couriers N_j, approximate the rate at which users are placing orders as $\sum_{i\in\mathcal{I}} x_{i,j}$, and define the rate at which an order is completed by these types of couriers as μ_j. Note that we can define the order completion rate μ_j as a constant provided by historical data, or estimate it using the parameters of our problem instance as we will do in the case studies following. Drawing these analogies allows us to define the utilization ρ_j of our queuing system for couriers of modality j as:

$$\rho_j = \frac{\sum_{i\in\mathcal{I}} x_{i,j}}{N_j\mu_j}. \tag{6}$$

In our regime of interest, the rate of order arrivals is magnitudes larger than the rate of order completions, and hence the number of available couriers c needs to be large. Using the $M/M/c$ latency function, one can show that in this regime of interest, the time spent waiting for an available server is negligible unless we are close to the capacity limit (Kelly-Bootle and Lutek, 1990). For example, given a system with $c = 50$ servers and a demand of 100 requests per hour, when the server utilization is high at $\rho = 0.99$, the average time spent in the system is 84 min, with 55 min in the queue. Once we lower the utilization to $\rho = 0.9$, the average time spent in the system is 29 min, with only 2 min spent in the queue. This means that the expected waiting time for a courier to be available is relatively small compared to the latency required to complete the order, given that the utilization parameter ρ_j is below a reasonable threshold. Thus, to make sure that customers are not experiencing long wait times for couriers to respond, we can upper-bound the utilization parameter ρ_j for all courier types, and ignore the effect of varying availability.

To model the time a courier must spend traveling to the pick-up location r_i, we take a probabilistic approach by calculating the expected travel time of the nearest available courier. Specifically, we assume that for modality j, some portion $\beta_{i,j} \in (0, 1]$ of available couriers are distributed around the pick-up location r_i such that their travel times are uniform in $[0, k_j]$. Note that we can choose k_j as some constant unit of time from which $\beta_{i,j}$ is estimated based on the pick-up location and modality. Since we know that the expected number of available couriers will be $(1 - \rho_j)N_j$, we can define the pick-up time as the expected travel time of the nearest courier:

$$u_{i,j}(x) = \frac{k_j}{1 + \beta_{i,j} N_j (1 - \rho_j)}, \tag{7}$$

where we used the fact that the expected minimum value of n independent uniform random variables in $[0, 1]$ is $\frac{1}{n+1}$.

Before continuing, we want to note that although our latency function $\ell_{i,j}(x)$ defined in Eq. (5) does not account for traffic congestion on the road, this is not an inherent limitation of our model. Specifically, we choose to set the travel time $t_{i,j}$ as constant in order to focus our latency model on the congestion caused by courier utilization, as captured by the pick-time $u_{i,j}$ defined in Eq. (7). Since we expect couriers to be a small percentage of total traffic, we believe this simplification to be justified. Nonetheless, given that our theoretical results in Section 4 make no restrictions to the latency function $\ell_{i,j}$, and the fact that our current formulation of latency $\ell_{i,j}$ is non-linear and non-convex, one can consider a model for the travel time $t_{i,j}$ which accounts for congestion.

5.2. Cost model

Before setting up our optimization problem, we need to model the cost of operating this system. We define the average dollar cost of completing a single order using a courier of modality j as the delivery cost c_j. This way, we can define the total cost of running our delivery system given the allocation strategy x:

$$C(x) = \sum_{j=1}^{J} c_j \Big(\sum_{i \in \mathcal{I}} x_{i,j} \Big), \tag{8}$$

where $C(x)$ is units of dollars per hour because $x_{i,j}$ is a rate of orders per hour. We use this model in our first case study concerning the meal delivery service. Alternatively, one can define a cost model using fixed hourly wages $\bar{c}_j$ for different courier modalities j, making the total cost of our system independent of the allocation strategy x:

$$C = \sum_{j=1}^{J} \bar{c}_j N_j, \tag{9}$$

where $\bar{c}_j$ is now in units of dollars per hour. Whereas we only account for completed orders in Eq. (8), by modeling couriers using average wages in Eq. (9) we are accounting for the unused couriers. We use this model in our second case study concerning the taxi service as information regarding wages is readily available. In practice, more sophisticated cost models can be utilized by addressing statistics such as profit margins, travel distance for couriers, and other information that is available to the service provider.

5.3. Optimization problem

We are now ready to set up the overall optimization problem.

$$\min_{x} \quad L(x) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \ell_{i,j}(x) x_{i,j} \tag{10}$$

$$\text{subject to} \quad C(x) \le \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \tau_{i,j} x_{i,j}, \tag{11}$$

$$\rho_j(x) \le \bar{\rho} \qquad \forall j \in \mathcal{J}, \tag{12}$$

$$\sum_{j \in \mathcal{J}} x_{i,j} = 1, \qquad \forall i \in \mathcal{I}, \tag{13}$$

$$0 \le x_{i,j} \le 1, \qquad \forall i \in \mathcal{I}, j \in \mathcal{J}. \tag{14}$$

For this case study, we want to find the allocation strategy x which minimizes expected latency L, as shown in Eq. (10). In addition, we constrain the operational cost in Eq. (11) to be less than the total compensation received from all deliveries. Note that because

Table 2
Results for a meal delivery system with 100 car couriers when there is only one modality available. The total operational cost is $1375 per hour.

	Cars
Orders (%)	100
Utilization ρ_j (%)	88
Latency ℓ_j (min)	19
Distance (miles)	1.47
Price τ_j ($)	5.00

Theorem 1 allows us to arbitrarily set price for the cheapest delivery option, one can always set the minimum allotted price to satisfy the constraint after optimization. We also constrain courier utilization in Eq. (12) by choosing an appropriate upper bound $\bar{\rho}$ for all modalities $j \in \mathcal{J}$. Note that with this additional supply constraint, we assume that the number of available couriers N_j is large enough so that the set of possible solutions is nonempty, as required by Proposition 1. We use the constraint in Eq. (13) to satisfy demands for each order $i \in \mathcal{I}$. Finally, we bound our decision variable between the domain of $[0, 1]$ in Eq. (14) so that there are no negative values in the solution. Once we find a desirable allocation strategy by solving this optimization problem, we can set prices using our theoretical results such that this flow is induced at equilibrium.

The optimization problem defined above is non-linear and non-convex, and we use a public implementation of the interior-point filter line-search algorithm (Wächter and Biegler, 2006) to solve it, noting that this method can be used to solve nonlinear programs on the order of a million variables (Biegler and Zavala, 2009). As aforementioned, many choices can be made for the formulation of the latency functions $\ell_{i,j}$, cost constraint C, and the optimization objective L. To efficiently use the interior point method, it is desired for the objective function and constraints to be twice differentiable so that the Hessian can be defined. We include details regarding this implementation in Appendix E, and provide our code online (Beliaev, 2023).

6. Case studies

We can now discuss the setting and results in our case studies. We first model a meal delivery system with three transportation modalities: cars, delivery drones, and sidewalk autonomous delivery robots (SADRs). In this setting, we model each population corresponding to order i with a trade-off function α_i that is independent of the modalities provided. Our second setting considers a taxi service transporting customers to and from the Chicago O'Hare International Airport (ORD) using three transportation modalities: cars, luxury cars, and electric vertical takeoff and landing (eVTOL) aircrafts. Unlike the first study where users are simply ordering food, we now model our populations with different trade-off functions $\alpha_{i,j}$ for each modality. We list how the problem parameters are defined and our results below, providing the full implementation in our code online (Beliaev, 2023).

6.1. Meal delivery: Grubhub instance

Setup. To define the problem parameters for our optimization formulation, we used real world instances from Grubhub (Reyes et al., 2018), which list information about the orders placed and car couriers available throughout a given time interval. Although there is no consideration of other modalities, we use the provided information as a basis and define our remaining parameters to be consistent. For service time $s_{i,j}$, we directly used the given pickup and dropoff times. Similarly for travel time $t_{i,j}$, we used the provided distances between restaurant and customer locations, converting them to time by using constant speeds for all modalities. For cars, we set the speed to 11.93 mph according to the dataset. For drones, we set the speed to 30 mph, using the upper end of the reported range of 13–34 mph (Macrina et al., 2020) since food deliveries are relatively light. Finally for SADRs, we set the speed to 4 mph, as they are expected to operate at the typical speed of pedestrians (Gehrke et al., 2023).

To calculate pickup time $u_{i,j}$, we computed all the parameters required in Eq. (7). The number of couriers N was directly chosen for each instance so that the problem was feasible under the utilization capacities $\bar{\rho}_j$. For cars and SADRs, we set $\bar{\rho}_j$ to 0.9, while for drones, we decreased this value to 0.8 due to the smaller number of vehicles utilized. We then generated courier locations for all three modalities, and computed the portion of available couriers $\beta_{i,j}$ that were at most $k = 10$ min away from the restaurant corresponding to order i. For car couriers, we directly sampled from the provided locations, while for drone couriers, we sampled uniformly from a grid spanning the restaurant locations. To capture SADRs delivering from restaurants closer to downtown, we sampled their locations uniformly from a grid centered in the middle of all restaurant locations, with length and width equal to their coordinate's respective standard deviations. Using these parameters, we estimated the mean rate μ_j of order completions as the inverse of expected latency $\mathbb{E}_i[\ell_{i,j}]^{-1}$ for each modality j, assuming load was equally distributed across them.

To compute the operational costs c_j for each modality, we set the cost per order to $5 for car and drone deliveries as they are expected to be competitive under certain regimes (Cornell et al., 2023), while using a lower cost per order of $1.50 for SADRs as the current cost per order is expected to drop from $2 to $1 in the near future (Jennings and Figliozzi, 2019). For user trade-off between price and time $\alpha(a)$, we used a linear function with the lowest evaluation $\alpha(1)$ set to $10 per hour, and the highest $\alpha(0)$ set to $100 per hour, for all orders $i \in \mathcal{I}$. We go on to discuss the results of our case study for an instance with 505 unique orders, each demanded with an equal rate of approximately 0.54 deliveries per hour.

Table 3
Results for a meal delivery system with 20 car, 24 drone, and 100 SADR couriers available, with a total operational cost of $856.84 per hour.

	Cars	Drones	SADRs	Total
Orders (%)	17	29	54	100
Utilization ρ_j (%)	90	80	88	86
Latency ℓ_j (min)	24	15	26	23
Distance (miles)	1.69	2.42	0.88	1.47
Price τ_j ($)	4.13	7.08	0.65	3.12

Results. We first consider the case when there are only 100 car couriers available, with no other transportation modality. We show the result in Table 2, listing the portion of orders delivered and the courier utilization as percentages, the average latency ℓ_j for modality j in minutes, the distance between customer and restaurant in miles, and finally the delivery price in dollars. For this setting, prices were set so that the money accrued from delivery services was equal to the operational cost. Since car couriers have an operational cost of $5.00 per order, we need an average delivery price of $5.00 per order to satisfy it. Note that although in this setting the minimum delivery price can be set arbitrarily for all orders since users have no choice to make, when we introduce other delivery modalities this is no longer the case as the prices must follow Eq. (2) to satisfy Nash conditions.

Next, we consider the case when there are 20 car, 24 drone, and 100 SADR couriers available displayed in Table 3. With the introduction of low cost SADRs into the system, the average latency has increased from 18 to 23 min, while the average price has decreased from $5.00 to $3.12 per order. Although a high cost and fast delivery service is now available via drones, customers are expected to pay a premium. These trends are expected for two reasons: (1) SADRs are much cheaper to operate making the total operational cost smaller, and (2) the faster speed of drones allows us to charge users who favor shorter delivery times more than users who favor cheaper delivery prices. It is interesting to note that drones are used to complete orders for customers furthest away from their chosen restaurant, while SADRs are used for customers closest to their chosen restaurant. This is due to the travel speeds of the different transportation modalities, as drones can travel efficiently between distant destinations, while robots are expected to operate in a smaller range as they have a pedestrian pace.

Overall, this case study shows that by setting prices according to users' trade-offs between money and time, one can implement a desired allocation strategy over multiple delivery modalities while improving their profit margins. One flaw with the study above is that we assumed users in different neighborhoods had identical distributions for their value of time. Although we had no information that would allow us to model this discrepancy, we expect that this would have an effect on where drones are utilized, as wealthier customers may reside in areas that are closer to downtown locations. Such an effect is considered in the following case study, where we model users' value of time (VOT) based on previous data.

6.2. Taxi service with urban air transportation: Chicago O'hare international airport

Setup. For the second setting, we consider passengers requesting rides to and from the ORD Airport via three available transportation modalities: cars, luxury cars, and eVTOL aircrafts. Using the publicly available data provided by rideshare companies in the city of Chicago (Transportation, 2023), we analyzed all taxi requests between January 2023 and March 2023, collecting travel times, taxi fares, as well as pickup and dropoff locations. The dataset is conveniently divided into 77 city areas, with the vertiport and ORD airport located at area #31 and area #75, respectively. Accordingly, we set the number of unique orders in our problem to 153, representing all trips to and from the airport.[2] We set the total demand roughly equal to the average rate of 1550 orders per hour reported in the dataset, and distributed this proportionally among the orders.

For travel time $t_{i,j}$ by car, we used the mean of all reported travel times between respective city areas. When considering trips via eVTOL aircrafts, 10 min of flight time was added to the travel time required to get to and from the vertiport by car (Gump, 2023). To calculate pickup time $u_{i,j}$, we computed all the parameters required in Eq. (7). The number of couriers N was directly chosen for each instance so that the problem was feasible under the utilization capacities $\bar{\rho}_j$, with the additional consideration that each eVTOL aircrafts would carry 4 passengers (Gump, 2023). For both car modalities we set $\bar{\rho}_j$ to 0.95, while for eVTOL aircrafts we decreased this value to 0.75 due to the smaller number of vehicles utilized. To estimate the portion of available car taxis near each city area $\beta_{i,j}$, we assumed that the higher densities of drivers were located in areas with higher demand. To model this, we exponentially scaled $\beta_{i,j}$ between 0.1 and 0.3 according to the demand at the corresponding city area. We set k to 5 min for both car modalities, and 0 min for eVTOL aircrafts as they do not need to travel to the pickup location. Service time $s_{i,j}$ was set to 0 min for all modalities. Using these parameters, we estimated the mean rate μ_j of order completions as the inverse of expected latency $\mathbb{E}_i[\ell_{i,j}]^{-1}$ for each modality j, assuming load was equally distributed across them. Based on estimated driver wages in Chicago reported by Uber (Rideshare, 2023), we set the hourly wages $\bar{c}_j$ for standard cars and luxury cars to $25 and $40 per hour, respectively.[3] For eVTOL aircrafts we assumed an operational cost of $200 per hour for each vehicle, meaning $50 for each passenger.

[2] We assumed that trips which start and end in area #75 were departing at the airport.
[3] The reported average wage was $28.03 per hour for all drivers.

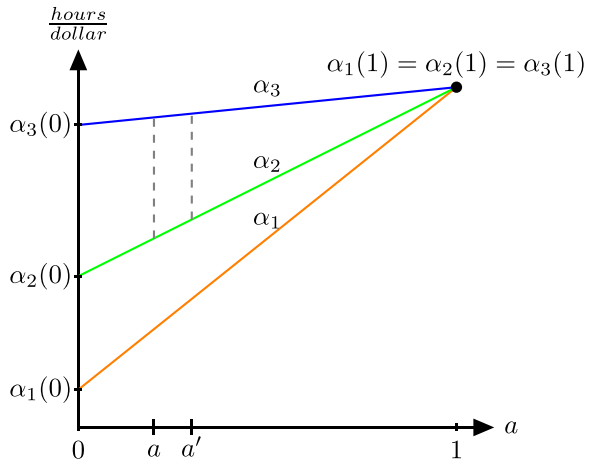

Fig. 3. The above schematic is an example of the user preference functions $\alpha_{i,j}$ for three modes $\mathcal{J} = \{1, 2, 3\}$, where the index corresponding to order $i \in \mathcal{I}$ has been left out for convenience. The x-axis represents users $a \in [0, 1]$ and the y-axis represents hours per dollar, the inverse of the value of time. We can see that the three lines all intersect at $a = 1$, signifying that the most frugal users do not care about the modality and prefer the cheapest option. All three lines have varying slopes, and we sort their indices so that for $a < a'$, $\frac{\alpha_{i,j+1}(a)}{\alpha_{i,j}(a)} \geq \frac{\alpha_{i,j+1}(a')}{\alpha_{i,j}(a')}$ for all $j \in \{1, \ldots, J-1\}$. The dashed gray lines help visualize this property: as we increase a to 1, the ratio $\frac{\alpha_3(a)}{\alpha_2(a)}$ decreases due to α_2 having a larger slope than α_3, reaching a minimum value of 1 when $a = 1$. Intuitively, this orders modalities by their slopes in decreasing order, representing a decrease in relative luxury. For our setting, this order is eVTOL aircrafts, luxury vehicles, and standard vehicles.

Table 4
Results for the airport taxi system with 990 standard vehicles and 110 luxury vehicles, with a total operational cost of $23,100 per hour.

	Standard	Luxury	Total
Orders (%)	91	9	100
Utilization ρ_j (%)	94	85	93
Latency ℓ_j (min)	35	35	35
Price τ_j ($)	41.35	60.63	43.06
Profit (1000$ per hour)	38.60	5.04	43.59

Table 5
Results for the airport taxi system with 900 standard vehicles, 100 luxury vehicles, and 25 eVTOL aircrafts, with a total operational cost of $26,000 per hour.

	Standard	Luxury	eVTOL	Total
Orders (%)	83	8	9	100
Latency ℓ_j (min)	35	30	23	34
Utilization ρ_j (%)	94	87	75	92
Price τ_j ($)	40.96	42.50	112.15	47.26
Profit (1000$ per hour)	34.80	2.38	10.08	47.26

Unlike the previous case study, we rely on Eq. (3) to derive our prices. Hence, in addition to modeling the heterogeneity in individuals' value of time (VOT), we assumed that more luxurious transportation modes are more valuable to the user. Previous studies have shown that users are willing to pay more for air taxi transportation compared to ground transportation (Fu et al., 2019; Binder et al., 2018). To represent this, for each order i we computed the mean VOT based on the dataset, and scaled it by $2, 1.5$, and 1 for eVTOL aircrafts, luxury cars, and standard cars, respectively. Furthermore, we assumed that the most frugal users at $a = 1$ would prefer the cheapest option regardless of the modality, and set $\alpha_{i,j}(1)$ for all modes j equal to one standard deviation below the computed mean VOT for that order i. To connect these two points for each order i and modality j, we used a straight line, making our overall model linear with larger slopes representing more luxurious modalities. We visualize this in Fig. 3. Note that because of the common intercepts at $\alpha_{i,j}(1)$ for all modalities j and the constant slopes, the assumption in Corollary 1 is satisfied. Furthermore, by not offering users travel via eVTOL aircrafts when they are slower than the other two options, we can strengthen Corollary 1 by guaranteeing that any desired flow is indeed inducible under some well behaved equilibrium flow. This result is shown in Appendix D. We make this aforementioned restriction in our formulation, and also explicitly check that the Nash equilibrium conditions of our solutions are indeed satisfied under the derived prices.

Results.

We first consider the case when there are only 1100 car couriers available, with 110 luxury vehicles and 990 standard vehicles. We show the result in Table 4, listing the portion of orders delivered and the courier utilization as percentage, the average latency ℓ_j

for modality j in minutes, the delivery price in dollars, and finally the total profit in 1000's of dollars per hour. For this setting, the minimum price was set so that users taking a standard taxi would pay their expected amount. Recall that setting the minimum price does not alter the Nash equilibrium since demand is inelastic, and hence we set it accordingly to reflect current prices experienced by users. We see this reflected in the results, as customers are expected to pay \$41.35 for a standard cab, and \$60.63 for a luxury cab on average. We expect this since both vehicles have identical travel times, but the mean VOT for luxury vehicles 50% higher compared to standard vehicles. This shows the importance of modeling VOT separately for different transportation modes, as such an effect would not be observed otherwise.

Next we consider the case when 25 eVTOL aircrafts are introduced into the system, displayed in Table 5. As expected, we can see that this premium transportation option comes with a high cost of \$112.15 per order. Although services provided by standard vehicles are mostly unaffected, demand for the luxury vehicle option is severely impacted. Specifically, luxury vehicles have a harder time competing and are required to lower their price from \$60.63 to \$42.50. In addition, we see their average trip latency drop from 35 to 30 min, signifying that commuters located further away are more willing to pay the price for the eVTOL aircraft option. With the higher costs required to operate luxury vehicles, we can expect such competition to greatly affect profits provided by this service, most likely lowering the expected wage for luxury vehicle drivers.

This case study outlines the importance of modeling VOT separately for different transportation modalities, as this can point out potential pitfalls in market strategies as new transportation modalities are introduced into the rideshare ecosystem. As we observed, even though eVTOL aircrafts only took 9% of the total orders, they had a large affect on the profitability of already existing luxury transportation options.

7. Conclusion

We model the pickup and delivery problem with multiple transportation modalities as a congestion game played over a star network, and show that we can explicitly define prices to induce any desired network flow. With this framework, we construct case studies for both a meal delivery service and a taxi service. In the first setting, we show that by utilizing autonomous transportation methods which are more efficient, one can set prices according to users' trade-offs between money and time to induce a desired allocation strategy while improving their profit margins. The second setting considers the additional assumption that users' trade-offs may differ between transportation modalities, and shows that such consideration are crucial to predict trends as new transportation modalities are introduced. We go over some of the implications of our work, pointing out limitations and directions for improvement.

We first note that in the setting of non-atomic congestion games taking place on graphs composed of one source–sink pair, prior works have asked if a feasible solution can be found to compute optimal prices for edges combinatorially, without relying on LP formulations (Cole et al., 2003b). Our main theoretical result states that in these settings, one can define optimal prices for paths combinatorially, implying that the LP formulation used to find prices for edges can be simplified. This points to the possibility that other network structures inherit properties which allow one to find prices efficiently, and we leave this direction for future works.

Further, we point out that our case studies only provide two examples where such a model is useful. Due to the general construction of the congestion game defined, our analysis is practical for any application that utilizes a platform to price match customers with different transportation methods. Since our formulation poses little restriction on the latency function defined, one can construct a model that is suitable for the desired application. Of course our framework gives no guarantees on finding the optimal allocation strategy, and instead provides a method by which prices can be set to induce a desired strategy.

One direction for future work is relaxing our formulation to allow for elastic demand. Such a consideration is interesting as it would permit one to optimize for profits directly, since the minimum price set would now affect the total user demand. The difficulty of such an extension lies in the ordering permutations required to compute the delivery prices. In this setting, additional assumptions may be required as one would need to compute derivative information while keeping track of ordering permutations that depend directly on the decision variable.

Lastly, we want to comment on the ethical implications of our first case study. On the positive side, our results show that by utilizing more autonomous transportation methods one can improve profit margins. However, this is true because our model considers car couriers operated by humans as less cost efficient. While one may have financial incentives to substitute part of their current workforce with autonomous machines, other decisions can be made that improve wages and work conditions for employees. Such a discussion is beyond the scope of our work, and is a topic that should be carefully addressed by policy makers before corporations are allowed to make decision that greedily improve their profits.

CRediT authorship contribution statement

Mark Beliaev: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Negar Mehr:** Writing – review & editing, Resources, Project administration, Funding acquisition, Conceptualization. **Ramtin Pedarsani:** Writing – review & editing, Resources, Project administration, Funding acquisition, Conceptualization.

Acknowledgments

This work was supported by the National Science Foundation, United States [award numbers 1952920 and 2145134].

Appendix A. Proof of Proposition 2

As stated by prior work, the proof of Proposition 2 (Proposition 2.4 in Cole et al. (2003b)) can be arrived at using a rearrangement argument, showing the stronger statement that an arbitrary Nash flow can be reorganized into a canonical one without changing the disutility incurred by any agent or the induced flow on paths. As this proof is omitted from the aforementioned work, we provide an independent result for completeness.

Our proof strategy is as follows: assuming that x is an arbitrary Nash flow, we use the inequalities defined by a Nash flow in Eq. (1) to show that the additional inequalities defined by a canonical Nash flow in Definition 3 must follow. The existence of canonical Nash flows follows directly from the well known result stated in Proposition 1. Note that for sake of notation, we drop the subscript referring to orders $i \in \mathcal{I}$, as it should be clear that the proof applies to an individual source–sink pair.

Formally, x is an equilibrium flow for instance (α, ℓ, τ) if for all edges $j \in \{1, \dots, J\}$, no user a traveling on edge j should want to switch to any other edge $j' \in \{1, \dots, J\}$:

$$\ell_j + \alpha(a)\tau_j \leq \ell_{j'} + \alpha(a)\tau_{j'} \quad \forall a \in \{a : x(a) = j\}, \tag{A.1}$$

where we leave out denoting the flow x in latency $\ell_j(x)$. Given an arbitrary Nash flow x for instance (α, ℓ, τ), we would like to show that for any two users $a_1, a_2 \in [0, 1]$ where $a_1 < a_2$: $\ell_{x(a_1)} \leq \ell_{x(a_2)}$, and $\tau_{x(a_1)} \geq \tau_{x(a_2)}$ must hold.

Clearly, when users a_1 and a_2 are routed on the same edge j, the aforementioned inequalities hold. To show that they hold in general, we assume that user a_1 is routed on edge j, $a_1 \in \{a : x(a) = j\}$, and user a_2 is routed on edge j', $a_2 \in \{a : x(a) = j'\}$, where $j \neq j'$. It follows directly from Eq. (A.1) that:

$$\ell_j - \ell_{j'} \leq \alpha(a_1)(\tau_{j'} - \tau_j), \tag{A.2}$$

and similarly,

$$\ell_j - \ell_{j'} \geq \alpha(a_2)(\tau_{j'} - \tau_j). \tag{A.3}$$

Given $\alpha(a) \geq 0 \ \forall a \in [0, 1]$ and $\alpha(a_1) \leq \alpha(a_2)$ from definition, we can infer from the above two inequalities that $\tau_j \geq \tau_{j'}$ and $\ell_j \leq \ell_{j'}$. This is trivially shown by contradiction: assume that $\tau_{j'} - \tau_j$ is positive, and divide by it on both sides of the inequalities to arrive at the contradiction $\alpha(a_2) \leq \frac{\ell_j - \ell_{j'}}{\tau_{j'} - \tau_j} \leq \alpha(a_1)$, implying that $\tau_{j'} - \tau_j$ and $\ell_j - \ell_{j'}$ are negative. Since $\ell_{x(a_1)} \leq \ell_{x(a_2)}$ and $\tau_{x(a_1)} \geq \tau_{x(a_2)}$ must hold for any two users $a_1, a_2 \in [0, 1]$ where $a_1 < a_2$, it follows directly that for any edge $j \in \mathcal{J}$, the users assigned to edge j by a flow x at Nash equilibrium form a (potentially empty or degenerate) subinterval of $[0, 1]$. This completes the proof.

Appendix B. Proof of Theorem 1

Note that for sake of notation, we will drop the subscript referring to orders $i \in \mathcal{I}$, as it should be clear that the proof applies to an individual source–sink pair. In addition, we assume that indexes $j \in \mathcal{J} : \{1, \dots, J\}$ correspond to the set of edges sorted by non-decreasing latency. Note that throughout our proof, we apply Proposition 2 which allows us to assume that any Nash flow x is a canonical Nash flow, as demonstrated in Appendix A above.

We define two adjacent intervals that are formed by our flow x: users $a \in [a_{j-1}, a_j]$ on the left experience latency ℓ_j and price τ_j, while users $a \in [a_j, a_{j+1}]$ on the right experience latency ℓ_{j+1} and price τ_{j+1}. The two intervals are portrayed in Fig. 2, where we note that this definition holds for $j \in \{1, \dots, J-1\}$. Using the inequalities defined in Eq. (1), we know that for x to be a (canonical) Nash flow for instance (α, ℓ, τ), no user a from the left interval $a \in [a_{j-1}, a_j]$ should want to switch to the delivery option corresponding to the right interval:

$$\ell_j + \alpha(a)\tau_j \leq \ell_{j+1} + \alpha(a)\tau_{j+1} \quad \forall a \in [a_{j-1}, a_j], \tag{B.1}$$

where we leave out denoting the flow x in latency $\ell_j(x)$. It follows:

$$\tau_j - \tau_{j+1} \leq \frac{\ell_{j+1} - \ell_j}{\alpha(a)} \quad \forall a \in [a_{j-1}, a_j], \tag{B.2}$$

$$\tau_j - \tau_{j+1} \leq \min_{a \in [a_{j-1}, a_j]} \left(\frac{\ell_{j+1} - \ell_j}{\alpha(a)} \right). \tag{B.3}$$

The preceding inequality can be simplified further by using the non-decreasing property of function α defining the population's price sensitivity: for any $a_1, a_2 \in [0, 1]$ such that $a_1 \leq a_2$, given user $a \in [a_1, a_2]$, $\max \alpha(a) = \alpha(a_2)$ and $\min \alpha(a) = \alpha(a_1)$. This comparison results in the following condition which must be true for x to be a Nash flow:

$$\tau_j - \tau_{j+1} \leq \frac{\ell_{j+1} - \ell_j}{\alpha(a_j)}. \tag{B.4}$$

We can repeat this process by enforcing that no user a from the right interval $a \in [a_j, a_{j+1}]$ should want to switch to the edge on the left:

$$\ell_{j+1} + \alpha(a)\tau_{j+1} \leq \ell_j + \alpha(a)\tau_j \quad \forall a \in [a_j, a_{j+1}], \tag{B.5}$$

$$\tau_j - \tau_{j+1} \geq \max_{a \in [a_j, a_{j+1}]} \left(\frac{\ell_{j+1} - \ell_j}{\alpha(a)} \right), \tag{B.6}$$

which results in the following:

$$\tau_j - \tau_{j+1} \geq \frac{\ell_{j+1} - \ell_j}{\alpha(a_j)}. \tag{B.7}$$

From (B.4) and (B.7) we can see that the two inequalities force the set of prices $\{\tau_{i,j}\}_{j \in \mathcal{J}}$ to follow:

$$\tau_j - \tau_{j+1} = \frac{\ell_{j+1} - \ell_j}{\alpha(a_j)} \quad \forall j \in \{1, \dots, J-1\}, \tag{B.8}$$

where if τ_J is given, the rest of the prices can be found recursively as defined in Eq (2).

To complete the proof, we must show that for this set of prices τ, the desired x is indeed an equilibrium flow. Formally, x is an equilibrium flow for instance (α, ℓ, τ) if for all edges $j \in \{1, \dots, J\}$ no user a in interval $a \in [a_{j-1}, a_j]$ should want to switch to any other edge $j' \in \{1, \dots, J\}$:

$$\ell_j + \alpha(a)\tau_j \leq \ell_{j'} + \alpha(a)\tau_{j'}. \tag{B.9}$$

Clearly, these inequalities hold when $j = j'$, and hence we show that they hold when $j > j'$ and $j < j'$. Starting with the former, when $j > j'$ we are considering that no user choosing edge j will switch to any edge j' on the left, where by definition $\tau_j \leq \tau_{j'}$ and $\ell_j \geq \ell_{j'}$. Rearranging Eq. (B.9), we have the following for all edges $j > j'$:

$$\begin{aligned}
&\ell_j + \alpha(a)\tau_j \leq \ell_{j'} + \alpha(a)\tau_{j'} \quad \forall a \in [a_{j-1}, a_j], \\
&\tau_{j'} - \tau_j \geq \max_{a \in [a_{j-1}, a_j]} \left(\frac{\ell_j - \ell_{j'}}{\alpha(a)} \right), \\
&\sum_{k=j'}^{J-1} \frac{\ell_{k+1} - \ell_k}{\alpha(a_k)} - \sum_{k=j}^{J-1} \frac{\ell_{k+1} - \ell_k}{\alpha(a_k)} \geq \frac{\ell_j - \ell_{j'}}{\alpha(a_{j-1})}, \\
&\sum_{k=j'}^{j-1} \frac{\ell_{k+1} - \ell_k}{\alpha(a_k)} \geq \sum_{k=j'}^{j-1} \frac{\ell_{k+1} - \ell_k}{\alpha(a_{j-1})}.
\end{aligned}$$

Since $\alpha(a_{j-1}) \geq \alpha(a_k)$ when $j' \leq k \leq j-1$, every summation term on the left hand side is strictly greater than or equal to every summation term on the right hand side, validating the inequalities in Eq. (B.9) for $j > j'$. We can do the same for $j < j'$, where now $\tau_j \geq \tau_{j'}$ and $\ell_j \leq \ell_{j'}$:

$$\begin{aligned}
&\ell_j + \alpha(a)\tau_j \leq \ell_{j'} + \alpha(a)\tau_{j'} \quad \forall a \in [a_{j-1}, a_j], \\
&\tau_j - \tau_{j'} \leq \min_{a \in [a_{j-1}, a_j]} \left(\frac{\ell_{j'} - \ell_j}{\alpha(a)} \right), \\
&\sum_{k=j}^{J-1} \frac{\ell_{k+1} - \ell_k}{\alpha(a_k)} - \sum_{k=j'}^{J-1} \frac{\ell_{k+1} - \ell_k}{\alpha(a_k)} \leq \frac{\ell_j - \ell_{j'}}{\alpha(a_j)}, \\
&\sum_{k=j}^{j'-1} \frac{\ell_{k+1} - \ell_k}{\alpha(a_k)} \leq \sum_{k=j}^{j'-1} \frac{\ell_{k+1} - \ell_k}{\alpha(a_j)}.
\end{aligned}$$

This time, since $\alpha(a_j) \leq \alpha(a_k)$ when $j \leq k \leq j'-1$, every summation term on the left hand side is strictly less than or equal to every summation term on the right hand side. This completes the proof.

Remark 1. Although Eq. (2) in Theorem 1 defines prices for parallel edges, this is equivalent to finding prices for paths in general directed graphs composed of one source–sink pair. Briefly consider a directed graph $G = (V, E)$ with source s and sink t, with edges $e \in E$ and simple $s-t$ paths $P \in \mathcal{P}$. Since the desired flow $x : [0,1] \mapsto \mathcal{P}$ is provided, the path latency can be found using $\ell_P(x) = \sum_{e \in P} \ell_e(x_e)$, where the edge flow is given by $x_e = \sum_{P : e \in P} x_P$. The proofs for Proposition 2 and Theorem 1 stated above follow directly by replacing edges $j \in \mathcal{J}$ with paths $P \in \mathcal{P}$. Note that this does not guarantee that there exists a unique set of additive edge prices τ_e such that $\tau_P = \sum_{e \in P} \tau_e$ is true for all paths $P \in \mathcal{P}$. Since users choosing between delivery modes can be represented by parallel edges, we forego defining paths in our formulation to be concise.

Appendix C. Proof of Corollary 2

The proof strategy is as follows: similar to the proof of Theorem 1, using a subset of the inequalities defined for Nash equilibrium in Eq. (1), we show that for some desired Nash flow $\{x_{i,j}\}_{j \in \mathcal{J}}$ there is only one set of valid prices τ_i that satisfies this subset of inequalities. Note that we drop the subscript referring to orders $i \in \mathcal{I}$, as it should be clear that the proof applies to an individual source–sink pair. In addition, we assume that indexes $j \in \mathcal{J} : \{1, \dots, J\}$ are ordered to satisfy the following assumption:

$$\text{Given } a \leq a' : \frac{\alpha_{j+1}(a)}{\alpha_j(a)} \geq \frac{\alpha_{j+1}(a')}{\alpha_j(a')} \qquad \forall j \in \{1, \dots, J-1\}, \tag{C.1}$$

where α is a set of non decreasing functions $\alpha_j : [0,1] \to (0,\infty)$ for $j \in \{1,\dots,J\}$.

As before, we define two adjacent intervals that are formed by our desired flow x: users $a \in [a_{j-1}, a_j]$ on the left experience latency ℓ_j and price τ_j, while users $a \in [a_j, a_{j+1}]$ on the right experience latency ℓ_{j+1} and price τ_{j+1}. The two intervals are no longer guaranteed to exhibit the properties of a canonical Nash flow portrayed in Fig. 2. Nonetheless, using the inequalities defined in Eq. (1), we know that for x to be a Nash flow for instance (α, ℓ, τ), no user a from the left interval $a \in [a_{j-1}, a_j]$ should want to switch to the delivery option corresponding to the right interval:

$$\ell_j + \alpha_j(a)\tau_j \le \ell_{j+1} + \alpha_{j+1}(a)\tau_{j+1} \quad \forall a \in [a_{j-1}, a_j], \tag{C.2}$$

where we leave out denoting the flow x in latency $\ell_j(x)$. It follows:

$$\tau_j - \frac{\alpha_{j+1}(a)}{\alpha_j(a)}\tau_{j+1} \le \frac{\ell_{j+1} - \ell_j}{\alpha_j(a)} \quad \forall a \in [a_{j-1}, a_j], \tag{C.3}$$

$$\max_{a\in[a_{j-1},a_j]}\left(\tau_j - \frac{\alpha_{j+1}(a)}{\alpha_j(a)}\tau_{j+1}\right) \le \min_{a\in[a_{j-1},a_j]}\left(\frac{\ell_{j+1} - \ell_j}{\alpha_j(a)}\right). \tag{C.4}$$

The LHS of the above inequality can be simplified using the assumption made in Eq. (C.1), whereas the LHS can be simplified as before by using the non-decreasing property of function α_j defining the population's price sensitivity. This results in the following condition which must be true for x to be a Nash flow:

$$\tau_j - \frac{\alpha_{j+1}(a_j)}{\alpha_j(a_j)}\tau_{j+1} \le \frac{\ell_{j+1} - \ell_j}{\alpha_j(a_j)}. \tag{C.5}$$

We can repeat this process by enforcing that no user a from the right interval $a \in [a_j, a_{j+1}]$ should want to switch to the edge on the left:

$$\ell_j + \alpha_j(a)\tau_j \ge \ell_{j+1} + \alpha_{j+1}(a)\tau_{j+1} \quad \forall a \in [a_j, a_{j+1}], \tag{C.6}$$

$$\min_{a\in[a_j,a_{j+1}]}\left(\tau_j - \frac{\alpha_{j+1}(a)}{\alpha_j(a)}\tau_{j+1}\right) \ge \max_{a\in[a_j,a_{j+1}]}\left(\frac{\ell_{j+1} - \ell_j}{\alpha_j(a)}\right), \tag{C.7}$$

which results in the following:

$$\tau_j - \frac{\alpha_{j+1}(a_j)}{\alpha_j(a_j)}\tau_{j+1} \ge \frac{\ell_{j+1} - \ell_j}{\alpha_j(a_j)}. \tag{C.8}$$

From (C.5) and (C.8) we can see that the two inequalities force the set of prices $\{\tau_{i,j}\}_{j\in\mathcal{J}}$ to follow:

$$\tau_j - \frac{\alpha_{j+1}(a_j)}{\alpha_j(a_j)}\tau_{j+1} = \frac{\ell_{j+1} - \ell_j}{\alpha_j(a_j)} \quad \forall j \in \{1,\dots,J-1\}, \tag{C.9}$$

where if τ_J is given, the rest of the prices can be found recursively. This completes the proof.

Remark 2. Note that unlike Theorem 1, we cannot generalize the result of Corollary 2 to apply for any desired flow. Since we cannot rely on Proposition 2 to order the edges as we do for a canonical Nash flow, we cannot show that τ_i does indeed satisfy all of the inequalities defined in Eq. (1). Hence we only prove that for a desired flow $\{x_{i,j}\}_{j\in\mathcal{J}}$ to be an equilibrium flow, the prices $\{\tau_{i,j}\}_{j\in\mathcal{J}}$ must follow Eq. (C.9).

Appendix D. Generalization of Corollary 2

As mentioned in Section 6, by not offering users travel via eVTOL aircrafts when they are slower than the other two options, we can guarantee that any desired flow is indeed inducible by a well behaved equilibrium flow. More precisely, we assume that there are three modalities $J = 3$, and the most premium option $j = 1$ must have the smallest latency $\ell_{i,1} \le \ell_{i,2}, \ell_{i,3}$ for all considered orders i. Given this, we can show that all of the inequalities defined for Nash equilibrium in Eq. (1) can be satisfied under some choice of the cheapest price $\tau_{i,3}$, which is free for us to define. We proceed to show this, dropping the subscript referring to orders $i \in \mathcal{I}$ as it should be clear that the proof applies to an individual source–sink pair.

First recall that the result of Corollary 2 directly satisfies the inequalities defined for Nash equilibrium when adjacent edges are considered. In otherwords, the prices given by Eq. (3) guarantee that users will not switch from luxury cars $j = 2$ to standard cars $j = 3$ or vice-versa, as well as luxury cars $j = 2$ to eVTOL aircrafts $j = 1$ or vice-versa. This result is shown in Appendix C for the general setting. Since we only have three edges in this setting, two inequalities are left to check. No user a from the leftmost interval $a \in [0, a_1]$ corresponding to eVTOL aircrafts $j = 1$ should want to switch the right most interval corresponding to standard cars $j = 3$:

$$\ell_1 + \alpha_1(a)\tau_1 \le \ell_3 + \alpha_3(a)\tau_3 \quad \forall a \in [0, a_1], \tag{D.1}$$

and vice-versa:

$$\ell_1 + \alpha_1(a)\tau_1 \ge \ell_3 + \alpha_3(a)\tau_3 \quad \forall a \in [a_2, 1]. \tag{D.2}$$

As done in Appendix C, the above two inequalities can be simplified to remove the intervals, and combined to form the following:

$$\frac{\alpha_1(a_1)}{\alpha_3(a_1)}\tau_1 - \frac{\ell_3 - \ell_1}{\alpha_3(a_1)} \le \tau_3 \le \frac{\alpha_1(a_2)}{\alpha_3(a_2)}\tau_1 - \frac{\ell_3 - \ell_1}{\alpha_3(a_2)}. \tag{D.3}$$

Note that since $\ell_3 - \ell_1 \ge 0$ and $\alpha_3(a_1) \le \alpha_3(a_2)$, we have $-\frac{\ell_3-\ell_1}{\alpha_3(a_1)} \le -\frac{\ell_3-\ell_1}{\alpha_3(a_2)}$. In addition, we have $\frac{\alpha_1(a_1)}{\alpha_3(a_1)} \le \frac{\alpha_1(a_2)}{\alpha_3(a_2)}$ due to the assumption stated in Corollary 2. This means that the domain of possible values for τ_3 is not degenerate, meaning we can set τ_3 to satisfy both Eqs. (D.1) and (D.2) simultaneously. In practice, we set τ_3 to reflect current prices as mentioned in Section 6, and verify that it is within the permissible range.

Appendix E. Implementation details

We use a public implementation of the interior-point filter line-search algorithm (Wächter and Biegler, 2006), and integrate it with the dataset of Grubhub instances (Reyes et al., 2018) and Chicago taxi services (Transportation, 2023) using Python. We briefly outline the results needed to implement the algorithm, and provide our code online (Beliaev, 2023).

First we define derivative information for the latency, where we use the fact that $\frac{d\rho_j(x)}{dx_{i',j'}} = \frac{1_{[j=j']}}{N_j\mu_j}$.

$$\ell_{i,j} = s_{i,j} + t_{i,j} + k_j[1 + \beta_{i,j}N_j(1-\rho_j(x))]^{-1}, \tag{E.1}$$

$$\frac{d\ell_{i,j}}{dx_{i',j'}} = 1_{[j=j']}\frac{\beta_{i,j}k_j}{\mu_j}[1 + \beta_{i,j}N_j(1-\rho_j(x))]^{-2}, \tag{E.2}$$

$$\frac{d^2\ell_{i,j}}{dx_{i'',j''}dx_{i',j'}} = 1_{[j=j'=j'']}\frac{2\beta_{i,j}^2k_j}{\mu_j^2}[1 + \beta_{i,j}N_j(1-\rho_j(x))]^{-3}. \tag{E.3}$$

Now we can use this to get derivative information for the objective function.

$$L(x) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\ell_{i,j}x_{i,j}, \tag{E.4}$$

$$\frac{dL(x)}{dx_{i',j'}} = \frac{1}{|\mathcal{I}|}\left[\left(\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\frac{d\ell_{i,j}}{dx_{i',j'}}x_{i,j}\right) + \ell_{i',j'}\right], \tag{E.5}$$

$$\frac{d^2L(x)}{dx_{i'',j''}dx_{i',j'}} = \frac{1}{|\mathcal{I}|}\left[\left(\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\frac{d^2\ell_{i,j}}{dx_{i'',j''}dx_{i',j'}}x_{i,j}\right) + \frac{d\ell_{i'',j''}}{dx_{i',j'}} + \frac{d\ell_{i',j'}}{dx_{i'',j''}}\right]. \tag{E.6}$$

Next, we derive derivative information for the two constraints separately, starting with the utility constraints in Eq. (12) which we denote as $g_j(x) \le \bar{\rho} \quad \forall j \in \mathcal{J}$.

$$g_j(x) = \rho_j(x), \tag{E.7}$$

$$\frac{dg_j(x)}{dx_{i',j'}} = \frac{1_{[j=j']}}{N_j\mu_j}, \tag{E.8}$$

$$\frac{d^2g_j(x)}{dx_{i'',j''}dx_{i',j'}} = 0. \tag{E.9}$$

Finally, we denote the flow constraints in Eq. (13) as $h_i(x) = 1 \quad \forall i \in \mathcal{I}$, listing the derivative information below.

$$h_i(x) = \sum_{j\in\mathcal{J}}x_{i,j}, \tag{E.10}$$

$$\frac{dh_i(x)}{dx_{i',j'}} = 1_{[i=i']}, \tag{E.11}$$

$$\frac{d^2h_i(x)}{dx_{i'',j''}dx_{i',j'}} = 0. \tag{E.12}$$

Note that the cost constraint in Eq (11) is not required for optimization as the minimum price can be manually set to satisfy it.

References

Ahuja, K., Chandra, V., Lord, V., Peens, C., 2021. Ordering in: The Rapid Evolution of Food Delivery. Vol. 22, McKinsey & Company.

Ale-Ahmad, H., Mahmassani, H.S., 2023. Factors affecting demand consolidation in urban air taxi operation. Transp. Res. Rec. 2677 (1), 76–92, arXiv:https://doi.org/10.1177/03611981221098396. [Online]. Available: http://dx.doi.org/10.1177/03611981221098396.

Beckmann, M., Mcguire, C., Winsten, C., 1956. Studies in the Economics of Transportation. Yale University Press, New Haven, Connecticut, USA.

Beliaev, M., 2023. Interior point impementation for pickup and delivery. [Online]. Available: https://github.com/mbeliaev1/mdrp.

Beliaev, M., Mehr, N., Pedarsani, R., 2023. Congestion-aware bi-modal delivery systems utilizing drones. Future Transp. 3 (1), 329–348, [Online]. Available: https://www.mdpi.com/2673-7590/3/1/20.

Biegler, L., Zavala, V., 2009. Large-scale nonlinear programming using IPOPT: An integrating framework for enterprise-wide dynamic optimization. Comput. Chem. Eng. 33 (3), 575–582, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0098135408001646. Selected Papers from the 17th European Symposium on Computer Aided Process Engineering held in Bucharest, Romania, May 2007.

Binder, R., Garrow, L.A., German, B., Mokhtarian, P., Daskilewicz, M., Douthat, T.H., 2018. If you fly it, will commuters come? a survey to model demand for eVTOL urban air trips. arXiv:https://arc.aiaa.org/doi/pdf/10.2514/6.2018-2882. [Online]. Available: https://arc.aiaa.org/doi/abs/10.2514/6.2018-2882.

Brown, P.N., Marden, J.R., 2017. The robustness of marginal-cost taxes in affine congestion games. IEEE Trans. Autom. Control 62 (8), 3999–4004.

Cole, R., Dodis, Y., Roughgarden, T., 2003a. How much can taxes help selfish routing?. In: Proceedings of the 4th ACM Conference on Electronic Commerce. EC '03, Association for Computing Machinery, New York, NY, USA, pp. 98–107, [Online]. Available: http://dx.doi.org/10.1145/779928.779941.

Cole, R., Dodis, Y., Roughgarden, T., 2003b. Pricing network edges for heterogeneous selfish users In: Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing. STOC '03, Association for Computing Machinery, New York, NY, USA, pp. 521–530, [Online]. Available: http://dx.doi.org/10.1145/780542.780618.

Cornell, A., Kloss, B., Presser, D., Riedel, R., 2023. Drones take to the sky, potentially disrupting last-mile delivery [Online]. Available: https://www.mckinsey.com/industries/aerospace-and-defense/our-insights/future-air-mobility-blog/drones-take-to-the-sky-potentially-disrupting-last-mile-delivery.

Dafermos, S.C., 1973. Toll patterns for multiclass-user transportation networks Transp. Sci. 7 (3), 211–223, [Online]. Available: http://www.jstor.org/stable/25767702.

Dafermos, S.C., Sparrow, F.T., 1969. The traffic assignment problem for a general network J. Res. Natl. Bur. Stand. B 73 (2), 91–118.

Fleischer, L., Jain, K., Mahdian, M., 2004. Tolls for heterogeneous selfish users in multicommodity networks and generalized congestion games In: 45th Annual IEEE Symposium on Foundations of Computer Science. pp. 277–285.

Fotakis, D., Karakostas, G., Kolliopoulos, S.G., 2010. On the existence of optimal taxes for network congestion games with heterogeneous users In: Algorithmic Game Theory.

Fu, M., Rothfeld, R., Antoniou, C., 2019. Exploring preferences for transportation modes in an urban air mobility environment: Munich case study Transp. Res. Rec. 2673 (10), 427–442, arXiv:https://doi.org/10.1177/0361198119843858. [Online]. Available: http://dx.doi.org/10.1177/0361198119843858.

Gacal, J.B., Urera, M.Q., Cruz, D.E., 2020. Flying sidekick traveling salesman problem with pick-up and delivery and drone energy optimization In: 2020 IEEE International Conference on Industrial Engineering and Engineering Management. IEEM, pp. 1167–1171.

Garrow, L.A., German, B.J., Leonard, C.E., 2021. Urban air mobility: A comprehensive review and comparative analysis with autonomous and electric ground transportation for informing future research Transp. Res. C 132, 103377, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X21003788.

Garrow, L.A., German, B., Mokhtarian, P., Glodek, J., 2019. A survey to model demand for eVTOL urban air trips and competition with autonomous ground vehicles arXiv:https://arc.aiaa.org/doi/pdf/10.2514/6.2019-2871. [Online]. Available: https://arc.aiaa.org/doi/abs/10.2514/6.2019-2871.

Gehrke, S.R., Phair, C.D., Russo, B.J., Smaglik, E.J., 2023. Observed sidewalk autonomous delivery robot interactions with pedestrians and bicyclists Transp. Res. Interdiscip. Perspect. 18, 100789, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590198223000362.

Gump, B., 2023. United airlines and archer announce first commercial electric air taxi Route in Chicago [Online]. Available: https://www.archer.com/news/united-airlines-and-archer-announce-first-commercial-electric-air-taxi-route-in-chicago.

Jennings, D., Figliozzi, M., 2019. Study of sidewalk autonomous delivery robots and their potential impacts on freight efficiency and travel Transp. Res. Rec..

Karakostas, G., Kolliopoulos, S.G., 2004. Edge pricing of multicommodity networks for heterogeneous selfish users In: 45th Annual IEEE Symposium on Foundations of Computer Science. pp. 268–276.

Karakostas, G., Kolliopoulos, S.G., 2006. Edge pricing of multicommodity networks for selfish users with elastic demands Algorithmica 53, 225–249.

Kelly-Bootle, S., Lutek, B.W., 1990. Chapter 5 - queueing theory In: Allen, A.O. (Ed.), Probability, Statistics, and Queuing Theory with Computer Science Applications, second ed. In: Computer Science and Scientific Computing, Academic Press, San Diego, pp. 247–375.

Lazar, D.A., Coogan, S., Pedarsani, R., 2017. Capacity modeling and routing for traffic networks with mixed autonomy In: 2017 IEEE 56th Annual Conference on Decision and Control. CDC, pp. 5678–5683.

Lazar, D.A., Coogan, S., Pedarsani, R., 2021. Routing for traffic networks with mixed autonomy IEEE Trans. Autom. Control 66 (6), 2664–2676.

Liu, Y., 2019. An optimization-driven dynamic vehicle routing algorithm for on-demand meal delivery using drones Comput. Oper. Res. 111, 1–20, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0305054819301431.

Llanes, C., Abate, M., Coogan, S., 2022. Safety from fast, in-the-loop reachability with application to UAVs In: 2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems. ICCPS, pp. 127–136.

Lu, Y., Yang, C., Yang, J., 2022. A multi-objective humanitarian pickup and delivery vehicle routing problem with drones Ann. Oper. Res. 1–63.

Macrina, G., Di Puglia Pugliese, L., Guerriero, F., Laporte, G., 2020. Drone-aided routing: A literature review Transp. Res. C 120, 102762, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X20306744.

Milchtaich, I., 2000. Generic uniqueness of equilibrium in large crowding games Math. Oper. Res. 25 (3), 349–364, [Online]. Available: http://www.jstor.org/stable/3690472.

Moshref-Javadi, M., Winkenbach, M., 2021. Applications and research avenues for drone-based models in logistics: A classification and review Expert Syst. Appl. 177, 114854, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421002955.

Orda, A., Rom, R., Shimkin, N., 1993. Competitive routing in multiuser communication networks IEEE/ACM Trans. Netw. 1 (5), 510–521.

Reyes, D., Erera, A., Savelsbergh, M., Sahasrabudhe, S., O'Neil, R., 2018. The meal delivery routing problem Optim. Online 6571.

2023. Become a rideshare driver in Chicago, IL [Online]. Available: https://www.uber.com/us/en/e/drive/chicago-il-us/.

Roughgarden, T., 2005. Selfish Routing and the Price of Anarchy. MIT press Cambridge.

Schmeidler, D., 1973. Equilibrium Points of Nonatomic Games. LIDAM Reprints CORE 146, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), [Online]. Available: https://EconPapers.repec.org/RePEc:cor:louvrp:146.

Sheffi, Y., 1985. Urban Transportation Networks: Equilibrium Analysis With Mathematical Programming Methods. Prentice Hall.

Shetty, A., Qin, J., Poolla, K., Varaiya, P., 2022. The value of pooling in last-mile delivery In: 2022 IEEE 61st Conference on Decision and Control. CDC, pp. 531–538.

2023. Starship food delivery app [Online]. Available: https://www.starship.xyz/starship-food-delivery-app/.

2023. Transportation network providers trips 2023 [Online]. Available: https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips-2023-/n26f-ihde.

2023. Uber announces results for fourth quarter and full year 2022 [Online]. Available: https://www.businesswire.com/news/home/20230208005139/en/Uber-Announces-Results-for-Fourth-Quarter-and-Full-Year-2022.

Wächter, A., Biegler, L.T., 2006. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming Math. Program. 106, 25–57.

Wardrop, J.G., 1952. Some theoretical aspects of road traffic research Proc. Inst. Civ. Eng. 1 (3), 325–362, arXiv:https://doi.org/10.1680/ipeds.1952.11259. [Online]. Available: http://dx.doi.org/10.1680/ipeds.1952.11259.

Wei, Q., Nilsson, G., Coogan, S., 2021. Scheduling of urban air mobility services with limited landing capacity and uncertain travel times In: 2021 American Control Conference. ACC, pp. 1681–1686.