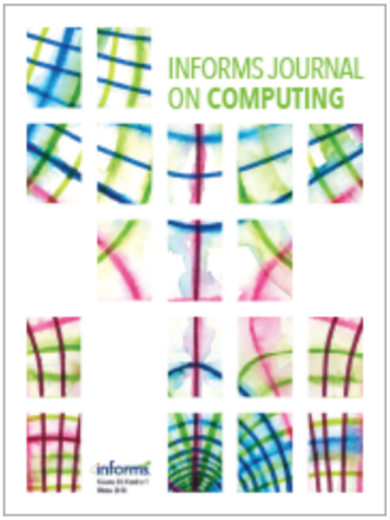


INFORMS Journal on Computing

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

Path-Based Formulations for the Design of On-demand Multimodal Transit Systems with Adoption Awareness

Hongzhao Guan, Beste Basciftci, Pascal Van Hentenryck

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

INFORMS JOURN L ON COMPUTING
rticles in dvance, pp. 1–22
ISSN 1091-9856 (print), ISSN 1526-5528 (online)

Path-Based Formulations for the Design of On-demand Multimodal Transit Systems with doption wareness

Hongzhao Guan,[a] Beste Basciftci,[b,*] Pascal Van Hentenryck[a]

[a] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30345; [b] Department of Business Analytics, Tippie College of Business, University of Iowa, Iowa City, Iowa 52242

*Corresponding author

Contact: hguan7@gatech.edu, https://orcid.org/0000-0002-5006-1165 (HG); beste-basciftci@uiowa.edu, https://orcid.org/0000-0002-3876-2559 (BB); pascal.vanhentenryck@isye.gatech.edu, https://orcid.org/0000-0001-7085-9994 (PVH)

Received: J nu ry 17, 2023
Revised: September 15, 2023; November 26, 2023; J nu ry 19, 2024
ccepted: J nu ry 20, 2024
Published Online in rticles in dvance: M rch 28, 2024

https://doi.org/10.1287/ijoc.2023.0014

bstract. This paper reconsiders the On-Demand Multimodal Transit Systems (ODMTS) Design with Adoptions problem (ODMTS-DA) to capture the latent demand in on-demand multimodal transit systems. The ODMTS-DA is a bilevel optimization problem, for which Basciftci and Van Hentenryck proposed an exact combinatorial Benders decomposition. Unfortunately, their proposed algorithm only finds high-quality solutions for medium-sized cities and is not practical for large metropolitan areas. The main contribution of this paper is to propose a new path-based optimization model, called P-Path, to address these computational difficulties. The key idea underlying P-Path is to enumerate two specific sets of paths which capture the essence of the choice model associated with the adoption behavior of riders. With the help of these path sets, the ODMTS-DA can be formulated as a single-level mixed-integer programming model. In addition, the paper presents preprocessing techniques that can reduce the size of the model significantly. P-Path is evaluated on two comprehensive case studies: the midsize transit system of the Ann Arbor – Ypsilanti region in Michigan (which was studied by Basciftci and Van Hentenryck) and the large-scale transit system for the city of Atlanta. The experimental results show that P-Path solves the Michigan ODMTS-DA instances in a few minutes, bringing more than two orders of magnitude improvements compared with the existing approach. For Atlanta, the results show that P-Path can solve large-scale ODMTS-DA instances (about 17 millions variables and 37 millions constraints) optimally in a few hours or in a few days. These results show the tremendous computational benefits of P-Path which provides a scalable approach to the design of on-demand multimodal transit systems with latent demand.

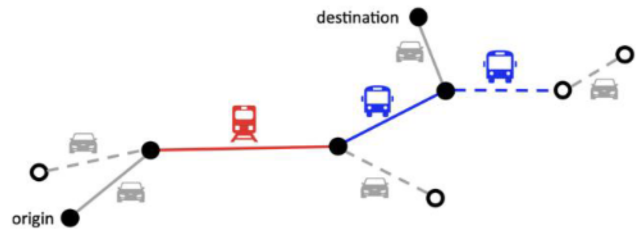
History: Accepted by Andrea Lodi, Design & Analysis of Algorithms—Discrete.
Funding: This work was partially supported by National Science Foundation Leap-HI [Grant 1854684] and the Tier 1 University Transportation Center (UTC): Transit - Serving Communities Optimally, Responsively, and Efficiently (T-SCORE) from the U.S. Department of Transportation [69A3552047141].
Supplemental Material: The software that supports the findings of this study is available within the paper and its Supplemental Information (https://pubsonline.informs.org/doi/suppl/10.1287/ijoc.2023.0014) as well as from the IJOC GitHub software repository (https://github.com/INFORMSJoC/2023.0014). The complete IJOC Software and Data Repository is available at https://informsjoc.github.io/.

Keywords: transit network optimization • bilevel optimization • integer programming • on-demand services • travel mode adoption • latent demand

1. Introduction

On-Demand Multimodal Transit Systems (ODMTSs) tightly integrate on-demand dynamic shuttles with fixed transit services such as rail and high-frequency buses. As illustrated in Figure 1, ODMTSs are operated around a number of hubs that are transit stations for high-frequency buses and urban rails. On-demand shuttles in ODMTSs are primarily utilized as the feeders to/from the hubs and represent an effective solution to the first-mile and last-mile problem faced by the vast majority of transit agencies. A realistic ODMTS pilot conducted in Atlanta, Georgia, USA, has demonstrated the efficacy of ODMTS in providing efficient services and economical sustainability. ODMTSs particularly benefit local communities engaging in short-distance trips or seeking connections to high-frequency routes, such as the rail system in Atlanta (Van Hentenryck et al. 2023). Furthermore, several simulation studies have shown that ODMTSs can provide significant cost and convenience benefits under varied settings

Figure 1. (Color online) An Example of ODMTS with a Sample Rider Path from Their Origin to Destination (Solid Line)



(Mahéo et al. 2019, Dalmeijer and Van Hentenryck 2020, Auad et al. 2021). However, these studies focus on designing an ODMTS for the existing transit users and neglect additional riders who could potentially adopt the system, given its higher convenience.

To address this gap by integrating rider behavior, Basciftci and Van Hentenryck (2023) propose the ODMTS Design with Adoption (ODMTS-DA) problem that captures the latent demand in ODMTS. The ODMTS-DA is a bilevel optimization problem, where the transit agency suggests riders paths from their origins to destinations by utilizing the transit network design and on-demand shuttles. After observing this path, the riders have the choice to adopt or reject the ODMTS. To solve this bilevel problem, an exact combinatorial Benders decomposition was proposed by taking into account the binary nature of the rider choices. Nevertheless, their proposed algorithm only finds high-quality solutions for medium-sized cities and is not practical for large metropolitan cities.

This paper reconsiders the ODMTS-DA and proposes a new path-based optimization model, called P-PATH, to address these computational difficulties. The main contributions of this paper can be summarized as follows:

1. It proposes P-PATH, a new path-based optimization model that replaces the bilevel subproblems by computing a number of specific path sets, that capture the essence of the subproblems and the mode choice models. By leveraging these path sets, the bilevel model can then be reformulated as a single-level mixed-integer programming (MIP) model.

2. It presents a number of preprocessing techniques to reduce the sizes of the path sets and the number of variables and constraints in the MIP model, enhancing computational efficiency.

3. It demonstrates, on two case studies, the computational benefits of P-PATH, which provides a scalable approach to the design of on-demand multimodal transit systems with latent demand. The midsize transit system of the Ann Arbor – Ypsilanti region in Michigan (which was studied by Basciftci and Van Hentenryck 2023) and the large-scale transit system for the city of Atlanta. The computational experiments show that P-PATH solves different Michigan ODMTS-DA instances in a few minutes, demonstrating more than two orders of magnitude improvements in comparison with the existing approach. The results over various Atlanta ODMTS-DA instances (comprising over 50,000 trips) show that P-PATH can obtain optimal solutions for large-scale instances, with up to 17 millions of variables and 37 millions of constraints derived from a graph with more than 2,400 nodes, within a few hours/days.

To ease understanding, P-PATH is presented in two steps. In a first step, another path formulation, called C-PATH, is presented. C-PATH replaces the follower subproblems in the original bilevel ODMTS-DA formulation by enumerating all the paths: it also leads to a single-level MIP model but unfortunately this formulation is not practical given the massive number of paths even in midsize instances. In a second step, the presentation turns to the key innovations in P-PATH. First, P-PATH only enumerates two specific sets of paths: paths that are adopted by riders given their choice models and paths that are profitable for the transit agency but rejected by the riders. Second, P-PATH reasons about paths implicitly in the MIP formulation by using arc variables. The two path sets enumerated by P-PATH are small in size, which makes the path enumeration and the MIP model tractable even for large-scale instances.

The remaining part of this paper proceeds as follows: Section 2 presents the relevant literature. Section 3 introduces the ODMTS-DA formulation. Section 4 introduces the proposed path-based formulations: C-PATH and P-PATH. Section 5 presents preprocessing techniques to reduce the size of the path sets and the MIP formulation. Sections 6 and 7 consider real case studies on medium-sized and large-scale instances, respectively, to demonstrate the performance of P-PATH. Section 8 concludes the paper.

2. Literature Review

Transit network design is an important problem in urban planning and transportation that produces a connected network for cities (Farahani et al. 2013). Examples of transit network design on traditional transit modes, such as buses and rails, can be found in various works such as Borndörfer et al. (2007), Fan et al. (2018), and Almasi et al.

(2021). The emergence of Mobility-as-a-Service (MaaS) over the past decade has led to a substantial transformation of the transit landscape (Shaheen and Chan 2016). Consequently, the transit network design problem has garnered increased attention with the integration of MaaS services (Stiglic et al. 2018, Pinto et al. 2020, Liu and Ouyang 2021, Najmi et al. 2023).

ODMTSs are transit systems that integrate on-demand services and conventional transit modes. A key distinction between ODMTSs and the majority of MaaS systems lies in the fact that ODMTSs are entirely operated by transit agencies, as opposed to on-demand services provided by Transit Network Companies (TNCs) (Van Hentenryck et al. 2023). In this regards, the core concepts underlying the emerging ODMTSs can be traced back to the multimodal hub location and hub network design problems (Alumur et al. 2012). Those were extended by ODMTSs to employ the on-demand shuttles for addressing the first-and-last-mile problem, and establishing hub-to-hub bus networks where shuttles are used as feeders (Mahéo et al. 2019). Continuing from this line of work, Dalmeijer and Van Hentenryck (2020) incorporated bus arc frequencies, passenger transfer limit, and backbone transit lines such as existing rail services. These ODMTS studies consider a fixed transit demand and aim to discover a network design that minimizes a weighted combination of the passenger convenience and operating cost from the transit agency perspective. However, neither of these studies take into account the latent demand when solving the design problem.

Like the first two ODMTS studies, many methods for transit network design problems are based on the assumption that the passenger demand is known and fixed (Schöbel 2012). Network design problem with fixed demand can also be generalized to other fields such as supply chain, energy distribution, and telecommunications. For instance, Gudapati et al. (2022) present a path-based approach to solve large-scale network design problem with service requirements. On the other hand, another group of studies aims at integrating the transit network design problem together with latent demand, that is, potential customers who might switch to the proposed transit networks once they are built. For example, instead of solving an optimization problem with a fixed demand, Klier and Haase (2008) modeled the impact of travel time on transit demand using a linear demand function of the expected travel time. A series of studies presented by Canca et al. comprehensively discussed network designs and latent demand. Their first study presented a model that concurrently determines the network design, line planning, capacity, fleet investment, and passengers mode and route choices (Canca et al. 2016). In the two follow-up studies, the authors proposed an adaptive large neighborhood search meta-heuristic for the previous model and a bilevel meta-heuristic which is designed to solved a revised model with large-scale data, respectively (Canca et al. 2017, 2019).

Among the studies that consider latent ridership, three systematic studies have investigated ODMTS with rider adoption awareness (Basciftci and Van Hentenryck 2020, 2023; Guan et al. 2022). The first two studies indicate that bilevel optimization frameworks need to be utilized to model the problem to ensure a fair design for the transit system (Basciftci and Van Hentenryck 2020, 2023). In general, a bilevel formulation is a mathematical program in which several variables are constrained by the solution of another optimization problem (Kleinert et al. 2021). Hence, a bilevel formulation is useful when modeling hierarchical decision process and has been widely utilized in the field of transportation. To model adoptions, in the first study, a personalized choice model which associates adoption choices with the time and cost of trips in the ODMTS is incorporated into the bilevel optimization framework (Basciftci and Van Hentenryck 2020). In the second study, the choice model was then redesigned such that the rider adoption choices solely depend on the trip duration under a fixed pricing strategy of the transit agency (Basciftci and Van Hentenryck 2023). These two studies propose formulations that are called the ODMTS Design with Adoptions (ODMTS-DA) problem. The authors of these two studies also provide significant insights into the design of exact algorithms that decompose and solve the problem under different choice model assumptions. However, the complexity and combinatorial nature of the exact algorithms led to computational difficulties: the exact algorithms can only provide high-quality solutions to medium-sized ODMTS-DA models. In order to address this difficulty, another study further investigated the properties of the optimal solution and proposed five heuristics to rapidly approximate the optimal solutions of large-scale instances (Guan et al. 2022). In all three papers (Basciftci and Van Hentenryck 2020, 2023; Guan et al. 2022), detailed case studies with realistic data are conducted to demonstrate the advantages and practicability of different methods under varied circumstances; however, the possibilities of discovering the optimal solution for large-scale ODMTS-DA problem still remains elusive. Therefore, this paper aims to explore the availability of modeling and computational methods that can both rapidly solve normal-sized instances and obtain the optimal solution for large-scale instances in a reasonable amount of time. It should be noted that the model presented in the work by Basciftci and Van Hentenryck (2023) (the second ODMTS-DA study) is referenced in the subsequent sections of this paper as the *bilevel model*, serving as the established framework for defining the ODMTS-DA problem.

Generally speaking, the newly proposed path-based methods in this paper intends to reformulate the bilevel formulation of the ODMTS-DA problem into a single-level MIP. The reformulation not only provides an alternative view of the original problem but can also ease the computational difficulties. Multiple previous studies have applied reformulations on bilevel optimization in the field of transportation science and demonstrated their advantages, especially on enhancing computational efficiency on large-scale instances (Meng et al. 2001, Marcotte et al. 2009, Calvete et al. 2014, Alizadeh and Nishi 2019). These reformulations mainly utilize Karush-Kuhn-Tucker (KKT) conditions of the lower level problem or strong-duality theorem. For instance, Goerigk and Schmidt (2017) presented two different reformulation approaches to solve the line planning with route choice problem. The first reformulation utilizes dualization of the inner-level routing problem, and the second one is based on additional shortest-path constraints. In another example, Kara and Verter (2004) first formulated the hazardous materials transportation problem as a bilevel model; they then utilized slackness conditions to reformulate the problem as a single-level MIP.

In summary, the proposed methods introduced in this paper are novel from the following perspectives to address the ODMTS-DA problem: (i) the paper proposes a path-based reformulation to develop single-level optimization models satisfying the optimality of the lower level problem, (ii) the reformulation only requires a small number of critical paths, (iii) the approach generalizes previous studies by considering a more flexible bilevel formulation and enabling choice function for riders as a black-box, and (iv) the paper proposes preprocessing techniques that significantly reduce the formulation sizes, resulting in single-level MIP reformulations that can be directly solved by commercial solvers.

3. The ODMTS Design with doptions (ODMTS-D) Problem

This section introduces the main problem of this study—ODMTS Design with Adoptions (ODMTS-DA). The goal for proposing this problem arises from a desire to construct an ODMTS design $\mathbf{z}$ that concurrently serves the existing transit users and captures latent demand with the choices of riders. In general, the ODMTS-DA determines the locations (modeled by using arc variables $\mathbf{z}$) for the deployment of new fixed-route buses and provides multimodal pathways comprising fixed routes and on-demand shuttles to riders. The key assumption in ODMTS-DA is: it is framed around two distinct rider groups: (i) core riders, committed to utilizing the ODMTS, and (ii) latent riders, who have the option to choose between the ODMTS and other transportation modes. These mode choices are modeled using a predefined choice model that can be considered as a black-box. The primary aim of the ODMTS-DA problem is to minimize an aggregate weighted cost incurred by the system, stemming from three components: (i) operation of the newly-designed fixed routes, (ii) serving the core ridership, and (iii) serving the adopted latent demand.

The comprehensive details and rationale for the ODMTS-DA are elaborated in the subsequent subsections. Table 1 outlines all nomenclature used in this paper, and they are specifically discussed in Sections 3, 4, and 5. In this section, Section 3.1 first introduces the problem setting, and Section 3.2 summarizes a bilevel optimization framework that formally defines the ODMTS-DA problem.

3.1. ODMTS Preliminaries

The ODMTS considers a set of nodes N, that represents the stops for on-demand shuttles to pick-up or drop-off riders. Given a pair of stops $i, j \in N$, the direct travel time and travel distance between these stops are represented by t_{ij} and d_{ij}, respectively, which correspond to the travel time and distance for on-demand shuttles serving these stops. A subset $H \subseteq N$, which corresponds to the hubs, can be employed to establish high-frequency buses or rails. The decision on hub locations must be made prior to any computational processes, usually involving the selection of a few key locations to ensure comprehensive coverage throughout the service area. There are two types of hub-to-hub connections considered in the ODMTS. The first type is defined as *newly-designed bus arcs* or simply *new bus arcs*, where their connections are decided by the optimization models. Note that the ODMTS normally does not design any new rail arcs due to practical considerations. The second type is referred as *backbone arcs*, they represent preselected existing transit lines such as rails and Bus Rapid Transit (BRT), and can be integrated into the ODMTS as fixed arcs, that is, not decided by the optimization models. Similarly, for buses and rails operating between hubs $h, l \in H$, t'_{hl} and d'_{hl} are defined to represent time and distance values. The inclusion of t'_{hl} and d'_{hl} for hub-to-hub connections is necessitated by the potential variation in travel time and distance between the same origin-destination pair for different modes.

The set T includes all trips that are considered by the ODMTS, including both the existing and latent demand. Three values are associated with each trip $r \in T$: (i) an origin stop $or^r \in N$, (ii) a destination stop $de^r \in N$, and (iii) the number of riders taking that trip $p^r \in \mathbb{Z}_+$. In order to serve a particular trip r, an ODMTS path π that connects or^r and de^r is proposed by the transit agency, and it is usually multimodal. For example, a typical ODMTS path consists of three trip-legs: a shuttle leg that serves the first-mile, a hub leg in the middle (taking rail or bus), and another

Table 1. Nomenclature Used in This Study

Symbol	Definition
Sets:	
N	The set of ODMTS stops
H	The set of ODMTS hubs, $H \subseteq N$
T	The set of all trips where each trip r is defined by an unique O-D pair: $or^r, de^r \in N$
T_{latent}	The set of latent trips, i.e., the latent demand
T_{core}	The set of core trips, i.e., existing transit demand
Z	The set of hub arcs that are considered in the ODMTS-DA problem
Z_{fixed}	The set of hub arcs that are fixed in the ODMTS design $Z_{fixed} \subseteq Z$
V^r, E^r	The sets of nodes or edges for a particular trip r
L^r	The set of allowed on-demand shuttle arcs for trip r
^{r}	The set of all valid ODMTS paths that connect or^r and de^r ($r \in T_{latent}$)
A^r	The set of ODMTS paths that a latent trip r would **adopt** ($A^r \subseteq {}^r\ \forall r \in T_{latent}$)
AP^r	The set of **profitable** ODMTS paths that a latent trip r would **adopt**
ANP^r	The set of **nonprofitable** ODMTS paths that a latent trip r would **adopt**
RP^r	The set of **profitable** ODMTS paths that a latent trip r would **reject** ($RP^r \subseteq {}^r\ \forall r \in T_{latent}$)
RNP^r	The set of **nonprofitable** ODMTS paths that a latent trip r would **reject**
$T_{latent,red}, L^r_{red}, Z^r_{red}$	The **reduced** versions of the previously introduced trip and arc sets after preprocessing
${}^r_{red}, A^r_{red}, RP^r_{red}$	The **reduced** versions of the previously introduced path sets after preprocessing
$x(\pi)$	The set of hub legs used by path π for trip r
$y(\pi)$	The set of shuttle legs used by path π for trip r
$\mathcal{O}^r_f(\mathbf{z})$	The set of optimal paths for trip r given a design $\mathbf{z}$
Parameters:	
θ	Convex combination factor for weighted cost, $\theta \in [0,1]$
d_{ij}	Car travel distance between stops $i,j \in N$
t_{ij}	Car travel time between stops $i,j \in N$
d'_{hl}	Bus travel distance between hubs $h,l \in H$, not applicable to rail
t'_{hl}	Bus or rail travel time between hubs $h,l \in H$
p^r	Number of passengers that take trip r
n_{hl}	Number of buses or trains operating between hubs $h,l \in H$ during the planning horizon
t^{wait}_{hl}	A rider's expected waiting time for a bus or a train that operates between hubs $h,l \in H$
b_{dist}, b_{time}	Operating cost for newly-designed bus arcs, with different measurements
β_{hl}	Weighted cost of investing a bus or train leg between hubs $h,l \in H$
τ^r_{hl}	A rider's weighted cost when traveling between hubs $h,l \in H$ with buses or rail for trip r
γ^r_{ij}	A rider's weighted cost when traveling between stops $i,j \in N$ with shuttles for trip r
$\omega_{dist}, \omega_{time}$	Shuttle operating cost, with different measurements
ϕ	Fixed ticket price charged by the transit agencies
φ	Transit agency's revenue for each rider in terms of weighted cost
$\overline{g}^r$	Upper bound for the weighted cost for a trip r
$\underline{g}^r$	Lower bound for the weighted cost for a trip r
t^r_π	A rider's travel time for trip r under the ODMTS path π
l^r_π	Number of transfers for trip r under the ODMTS path π
l^r_{ub}	Transfer tolerance for a latent trip $r \in T_{latent}$ in deciding ODMTS adoption, used in choice model (4)
^{r}	Adoption factor of a latent trip $r \in T_{latent}$, used in choice models (3) and (4)
t^r_{cur}	Travel time of a latent trip $r \in T_{latent}$ with its current travel mode, used in choice models (3) and (4)
Decision variables:	
z_{hl}	A binary variable indicating if a hub arc between hubs $h,l \in H$ is open
x^r_{hl}	A binary variable indicating if a hub arc (h, l) is used by trip r
y^r_{ij}	A binary variable indicating if a shuttle arc (i, j) is used by trip r
t^r	Travel time of trip r for a rider, dynamically based on ODMTS paths
g^r	Weighted cost of trip r for a rider, dynamically based on ODMTS paths
m^r	Minimal weighted cost of trip r for a rider, dynamically based on ODMTS design $\mathbf{z}$
δ^r	A binary variable indicating if latent trip $r \in T_{latent}$ adopts
f^r_π	A binary variable indicating if path π is **feasible** for latent trip r, dynamically based on $\mathbf{z}$
λ^r_π	A binary variable indicating if path π is assigned to latent trip r

Note. For simplicity, π^r is simplified as π in variables and parameters.

shuttle leg that serves the last-mile. A path is allowed to include multiple hub legs, while shuttle legs can only be employed to serve as the first or last legs. In this study, the term *direct shuttle path*, indicates an ODMTS path that is covered by a single-leg on-demand shuttle that connects or^r and de^r. In summary, on-demand shuttles are only allowed for three types of trips: (i) the direct trip, (ii) a first leg from trip origin to a hub, and (iii) a last leg from a hub to trip destination. In addition, note that paths solely served with two on-demand shuttle legs are not considered. For example, a path in the form *origin* → *hub* → *destination* with two shuttles are not allowed.

To have a weighted objective considering both convenience and cost aspects of the ODMTS design, a parameter $\theta \in [0,1]$ is employed such that the trip duration associated with convenience are multiplied by θ and expenses are multiplied by $1 \quad \theta$. When a new bus arc connects hubs $h,l \in H$, n_{hl} denotes the number of buses operating along this route within the planning horizon, where n_{hl} is an input parameter defined by the transit agency for setting the hourly frequency. For instance, setting $n_{hl} = 12$ in a four-hour time horizon implies a bus departing from hub h to hub l every 20 minutes ($4 \times 60/12$). Alternatively, the agency could also specify an hourly frequency of 3 for the bus service between h and l, $n_{hl} = 3 \times 4$. It follows that a weighted cost β_{hl} is considered by the transit agency using equation $\beta_{hl} = (1 \quad \theta) n_{hl}\, d'_{hl}\, b_{dist}$, where b_{dist} is the cost of operating a bus per kilometer and d'_{hl} is the bus travel distance from h to l. It is worth pointing out that β_{hl} can be alternatively modeled by $\beta_{hl} = (1 \quad \theta) n_{hl}\, t'_{hl}\, b_{time}$, where b_{time} and t'_{hl} represent the cost of operating a bus per hour and bus travel time from hub h to l, respectively. Furthermore, for backbone hub connections between h and l, for example, a rail arc that is a part of the existing rail system, $\beta_{h,l} = 0$ is applied instead of the abovementioned equations.

In this paper, maintaining n_{hl} as a constant for all hub arcs is required. However, as noted by Dalmeijer and Van Hentenryck (2020), to give more flexibility to the network designers, the ODMTS can be extended to accommodate multiple predefined hourly frequency choices for each hub arc. Consequently, the optimization model could select the most suitable option from these choices. Since this study considers a complex transit network design problem by integrating rider choice decisions within a bilevel optimization framework, the decisions involving the selection of hourly frequency for hub arcs, shuttle ride-sharing, shuttle flow reservations, and shuttle fleet sizing decisions of on-demand shuttles are omitted while designing the network. These decisions can be either addressed by modifying constraints or considering them in the subsequent levels of decision making, which involve finer level operational planning under a given network design.

In addition to the weighted cost of opening legs between the hubs, serving each trip has its weighted cost to the transit agency. This cost consists of two parts depending on the path utilized by the trip, which are hub legs and shuttle legs. For the former one, as operating cost of the open legs between hubs are considered within the investment, only inconvenience cost of each trip $r \in T$, denoted as τ^r_{hl}, is computed by $\tau^r_{hl} = \theta(t'_{hl} + t^{wait}_{hl})$, where t^{wait}_{hl} is the expected waiting time of a rider for the bus or train between hubs h and l. On the other hand, for the parts of the trips traveled by on-demand shuttles, the transit agency incurs γ^r_{ij} for each trip $r \in T$ using equation $\gamma^r_{ij} = (1 \quad \theta)\omega_{dist}\, d_{ij} + \theta t_{ij}$, where ω_{dist} is the cost of operating a shuttle per kilometer. Moreover, γ^r_{ij} can be modeled alternatively by using $\gamma^r_{ij} = (1 \quad \theta)\omega_{time}\, t_{ij} + \theta t_{ij}$, where ω_{time} is the shuttle operating cost per hour.

3.2. Problem Description and Bilevel Formulation

This section formally defines the ODMTS-DA problem as a bilevel optimization problem. The formulation generalizes the proposal by Basciftci and Van Hentenryck (2023) to allow for more flexibility in the mode choice models. Equations (5)–(6) presents the bilevel optimization framework for ODMTS-DA problem.

3.2.1. Decision Variables. Binary variable z_{hl} represents whether a hub arc between $h,l \in H$ is open. Backbone arcs form a set $Z_{fixed} \subseteq \{(h,l) : \forall h,l \in H\}$, and their corresponding z_{hl} are always open. For each trip $r \in T$, binary variables x^r_{hl} and y^r_{ij} indicate whether trip r takes the hub leg between $h,l \in H$, and the shuttle leg between stops $i,j \in N$, respectively. An ODMTS path π in the network for trip r is defined as a sequence of distinct opened hub arcs and shuttle arcs that connect a sequence of vertices starting from origin or^r and ending at destination de^r. Under a given solution, a path of trip r can be constructed from the open arcs specified by the x^r_{hl} and y^r_{ij} variables.

3.2.2. doption Choices. To integrate the adoption choices of riders into the formulation, the trip set T is divided into two subsets with respect to the riders' adoption characteristics: (i) a *core trip set* $T_{core} = T \setminus T_{latent}$ representing the existing demand and (ii) a *latent trip set* T_{latent} representing the latent demand. Core trips correspond to the set of riders of the existing transit system: they are assumed to continue adopting the ODMTS. Latent trips correspond to the set of riders who are currently traveling with other modes and might switch to transit due to the deployment of ODMTS. The transit agency charges each rider a ticket fare ϕ to use the ODMTS, irrespective of the assigned paths. Hence, a fixed value φ, computed by $\varphi = (1 \quad \theta)\phi$, becomes an additional weighted revenue to the transit agency for riders switching to the ODMTS.

In this study, the adoption behavior of a latent trip $r \in T_{latent}$ is assumed to depend on the features of an assigned path π, such as trip duration t^r_π and the number of transfers l^r_π, and these features can be computed by the following equations:

$$t^r_\pi = \sum_{h,l \in H} (t'_{hl} + t^{wait}_{hl}) x^r_{hl} + \sum_{i,j \in N} t_{ij} y^r_{ij} \tag{1}$$

$$l^r_\pi = \sum_{h,l \in H} x^r_{hl} + \sum_{i,j \in N} y^r_{ij} \quad 1 \tag{2}$$

For each latent trip r, a choice model $\mathcal{C}^r$ returns a binary adoption decision δ^r, which is equal to 1 when its riders decide to adopt the ODMTS given the proposed path π. The structure of $\mathcal{C}^r$ can vary from a threshold model to a more complicated machine learning model. The choice model can be arbitrarily complicated in this paper since it is abstracted in the path enumeration. Basciftci and Van Hentenryck (2023) proposed the choice model

$$\mathcal{C}^r(\mathbf{x}^r, \mathbf{y}^r) = \mathcal{C}^r(\pi) \equiv \mathbb{1}(t^r_\pi \leq \quad ^r t^r_{cur}) \quad \forall r \in T_{latent} \tag{3}$$

that directly compares the ODMTS trip duration t^r_π with the travel duration t^r_{cur} under the current mode multiplied by a parameter $\quad ^r$. In the computational experiments, this paper also explores the following choice model

$$\mathcal{C}^r(\mathbf{x}^r, \mathbf{y}^r) = \mathcal{C}^r(\pi) \equiv \mathbb{1}(t^r_\pi \leq \quad ^r t^r_{cur}) \quad \mathbb{1}(l^r_\pi \leq l^r_{ub}) \quad \forall r \in T_{latent} \tag{4}$$

which takes transfers into account, that is, riders reject the ODMTS when they are assigned to a path with too many transfers. For each $r \in T_{latent}$, the transfer tolerance is denoted as l^r_{ub} and l^r_π represents the number of transfers in a path π. For example, consider a path π with 4 transfers ($l^r_\pi = 4$) offered to riders in trip r: *shuttle* $\to$ *bus* $\to$ *rail* $\to$ *bus* $\to$ *shuttle*. If the riders in r have a transfer tolerance of 2 ($l^r_{ub} = 2$), then according to choice model (4), the riders in r will reject the proposed path π.

3.2.3. The Bilevel Optimization Model. The Bilevel Optimization Model for the ODMTS Design with Adoption (ODMTS-DA) Problem is presented below:

$$\min_{z_{hl}} \quad \sum_{h,l \in H} \beta_{hl} z_{hl} + \sum_{r \in T_{core}} p^r g^r + \sum_{r \in T_{latent}} p^r \delta^r (g^r \quad \varphi) \tag{5a}$$

$$\text{s.t.} \quad \sum_{l \in H} z_{hl} = \sum_{l \in H} z_{lh} \quad \forall h \in H \tag{5b}$$

$$z_{hl} = 1 \quad \forall (h,l) \in Z_{fixed} \tag{5c}$$

$$\delta^r = \mathcal{C}^r(\mathbf{x^r}, \mathbf{y^r}) \quad \forall r \in T_{latent} \tag{5d}$$

$$z_{hl} \in \{0,1\} \quad \forall h,l \in H \tag{5e}$$

$$\delta^r \in \{0,1\} \quad \forall r \in T_{latent} \tag{5f}$$

where $(\mathbf{x^r}, \mathbf{y^r}, g^r)$ are a solution to the optimization problem

$$(\mathbf{x^r}, \mathbf{y^r}, g^r) \in \arg\min_{x^r_{hl}, y^r_{ij}} \quad g^r = \sum_{h,l \in H} \tau^r_{hl} x^r_{hl} + \sum_{i,j \in N} \gamma^r_{ij} y^r_{ij} \tag{6a}$$

$$\text{s.t.} \quad \sum_{\substack{h \in H \\ \text{if } i \in H}} (x^r_{ih} \quad x^r_{hi}) + \sum_{i,j \in N} (y^r_{ij} \quad y^r_{ji}) = \begin{cases} 1 & \text{if } i = or^r \\ \quad 1 & \text{if } i = de^r \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in N \tag{6b}$$

$$x^r_{hl} \leq z_{hl} \quad \forall h,l \in H \tag{6c}$$

$$x^r_{hl}, y^r_{ij} \in \{0,1\} \quad \forall i,j \in N \quad \forall h,l \in H \quad \forall r \in T \tag{6d}$$

The leader problem (Equations (5a)–(5f)) designs the network between the hubs for the ODMTS, whereas the follower problem (Equations (6a)–(6d)) obtains paths for each trip $r \in T$ by considering the legs of the network design and the shuttles to serve the first and last miles. The objective of the leader problem minimizes the sum of (i) the weighted cost of opening new bus legs, (ii) the weighted cost of the core trips, and (iii) the weighted net cost of the riders with choice who are switching to the ODMTS. Note that the third component of the objective function includes a nonlinear term as the binary variable δ^r is utilized to represent the riders that adopt ODMTS. Constraint (5b) ensures weak connectivity between the hubs such that each hub has the same number of incoming and outgoing open arcs. Constraint (5c) enforces the fixed arcs to be opened. Constraint (5d) represents the adoption decision of the riders in T_{latent} depending on the ODMTS path constructed by $\mathbf{x}^r$ and $\mathbf{y}^r$.

For each trip r, the objective of the follower problem (6) minimizes the cost and inconvenience (weighted cost), with the solution of the leader problem ($\mathbf{z}$) serving as inputs. Constraint (6b) guarantees flow conservation for the bus, rail, and shuttle legs used in trip r for origin, destination, and each of the intermediate points. Constraint (6c) ensures that a path only uses open legs between the hubs. Given that the follower problem has a totally unimodular constraint matrix, it can be solved as a linear program.

The bilevel optimization model can be understood intuitively as follows. The transit agency obtains an ODMTS design by minimizing the weighted costs of the design and the trips considered. The paths derived from this design are proposed to riders, and riders with choice determine whether they adopt the ODMTS based on their suggested paths or stay with their previous modes. It is worth noticing that the paths are proposed by the agency instead of the riders, this is the key difference between ODMTS and regular transit systems. The subproblem guarantees that the proposed paths are minimizing the cost and inconvenience of the riders, preventing the formulation to propose suboptimal paths that riders would systematically reject. The optimal design can thus be viewed as an equilibrium point for such a game between the transit agency and the latent demand.

3.2.4. The Original Bilevel Optimization Model. The follower subproblem in (6) may have multiple optimal solutions that may lead to routes with different travel durations. Basciftci and Van Hentenryck (2023) eliminates this issue by proposing the subproblem depicted in Formulation (7).

$$\underset{x^r_{hl}, y^r_{ij}, g^r, t^r}{\text{lex-min}} \quad \langle g^r, t^r \rangle \tag{7a}$$

$$\text{s.t.} \quad g^r = \sum_{h,l \in H} \tau^r_{hl} x^r_{hl} + \sum_{i,j \in N} \gamma^r_{ij} y^r_{ij} \tag{7b}$$

$$t^r = \sum_{h,l \in H} (t'_{hl} + t^{wait}_{hl}) x^r_{hl} + \sum_{i,j \in N} t_{ij} y^r_{ij} \tag{7c}$$

$$\text{(6b)–(6d)}$$

For each trip r, the objective of the follower problem (7) minimizes the lexicographic objective function $\langle g^r, t^r \rangle$, where g^r corresponds to the cost and inconvenience of trip r and t^r breaks potential ties between the solutions with the same value of g^r by returning the most convenient path for the rider of trip r. Basciftci and Van Hentenryck (2023) broke ties this way because it aligned with the choice model they used (i.e. (3)), that is solely based on travel duration. They also showed that the lexicographic minimizer of problem (7) exists and results in a unique (g^r, t^r) solution. This unicity property simplifies the design of algorithms and is exploited by Basciftci and Van Hentenryck (2023).

To solve this resulting bilevel optimization problem, Basciftci and Van Hentenryck (2023) proposed an exact decomposition algorithm that combines Benders decomposition algorithm with a combinatorial cut generation procedure to integrate rider adoption constraints. Despite the addition of valid inequalities and the application of several preprocessing techniques, the computational studies revealed some of the limitations of this approach for large-scale instances (e.g., metropolitan cities): those instances have large optimality gaps and run times.

The formulations and algorithms proposed in this study do not require this unicity property for the subproblem. This provides a more flexible framework where the subproblems do not need to have unique optimal solutions. The experimental results will consider both the general formulation and the more specialized model when comparing execution times.

4. Path-Based Formulations

This section presents path-based formulations for addressing the ODMTS-DA problem. Section 4.1 discusses the nature of the paths and the set of paths useful for the path-based formulations along with their properties. Sections 4.2 and 4.3 present the two path-based formulations.

4.1. Paths and Their Properties

4.1.1. Paths. Given a design $\mathbf{z}$, the follower subproblem (6) returns a set of paths for each rider. For trip r, these paths are specified by the vectors $\mathbf{x^r}$ and $\mathbf{y^r}$, that is, they specify which hub legs and which shuttle legs comprise the path. In fact, given the restrictions on shuttle legs (which can only be the first and last legs of a path or a direct path), a path is uniquely specified by its set of hub legs. This property is used by the path formulations.

The path formulations reason directly in terms of paths, in addition to the decision variables $\mathbf{x^r}$ and $\mathbf{y^r}$. For a rider r, a path π specifies which hub legs and shuttle legs trip r uses if the path is adopted. The notations

$$x(\pi) = \{(h,l) \in \pi : h,l \in H\} \tag{8a}$$

$$y(\pi) = \{(i,j) \in \pi : i,j \in N\} \tag{8b}$$

denote the set of hub legs and shuttle trips used by path π respectively. The paper also uses $\mathcal{C}^r(\pi)$ to denote whether riders in trip r adopt path π.

4.1.2. Set of Paths. The path formulations use some specific sets of paths to solve the ODMTS-DA problem. These sets are only constructed for riders in T_{latent}. The set of paths for trip $r \in T_{latent}$ are constructed from a graph $G^r = (V^r, E^r)$ that contains all the hub legs and the allowed shuttle trips, where the set of nodes is defined as $V^r = \{or^r, de^r\} \cup H$, the set of arcs as $E^r = Z \cup L^r$, with L^r is the set of allowed shuttle trips, that is,

$$L^r = \{(or^r, de^r)\} \cup \{(or^r, h) : \forall h \in H\} \cup \{(h, de^r) : \forall h \in H\}$$

The path-based formulations rely on three sets:

$$\ ^r = \text{the set of all paths in } G^r \text{ connecting } or^r \text{ to } de^r. \tag{9a}$$

$$A^r = \{\pi \in \ ^r : \mathcal{C}^r(\pi) = 1\}. \tag{9b}$$

$$RP^r = \{\pi \in \ ^r : \mathcal{C}^r(\pi) = 0 \quad g(\pi) < \varphi\}. \tag{9c}$$

where

$$g(\pi) = \sum_{(h,l) \in x(\pi)} \tau_{hl} + \sum_{(i,j) \in y(\pi)} \gamma_{ij}.$$

The set A^r represents the set of all paths that trip r would adopt. The set RP^r denotes the set of all paths that trip r would reject but which profitable from the perspective of the transit agency in terms of weighted cost. Algorithm PE (Path Enumeration) in Online Supplement A can be used to obtain these sets. In practice, these path sets are generated using various preprocessing techniques for computational efficiency, as discussed in Section 5. Why these sets are useful will become clear subsequently. The first path-based formulation uses sets $\ ^r$ and A^r for each latent trip $r \in T_{latent}$, whereas the second formulation is based on RP^r and A^r. In general, RP^r is smaller than $\ ^r$ because of the rare appearances of profitable paths. For proving the results, it is also useful to define the set

$$RNP^r = \{\pi \in \ ^r : \mathcal{C}^r(\pi) = 0 \quad g(\pi) \geq \varphi\}. \tag{10a}$$

Note that A^r, RP^r, and RNP^r form a partition of $\ ^r$. It is also useful to further partition A^r into AP^r and ANP^r with

$$AP^r = \{\pi \in A^r : g(\pi) < \varphi\}, \tag{11a}$$

$$ANP^r = \{\pi \in A^r : g(\pi) \geq \varphi\}. \tag{11b}$$

4.1.3. Path Formulations. It is important to recall that the bilevel nature of the ODMTS-DA problem is only due to the riders with choice: the bilevel formulation could include Equations (6b) and (6c) for riders in T_{core} in the leader problem. Indeed, since riders in T_{core} must use the ODMTS, the optimization model will necessarily minimize their objective terms. This is not the case of riders of choice, since their choice function decides whether they adopt the ODMTS. Without a bilevel model, the single-level optimization, which directly incorporates the lower level constraints to the upper level problem for riders in T_{latent}, can intentionally propose paths to riders with choices that they will reject, if no profitable paths exist for them. This happens as the optimality of the lower level is not ensured in this single-level problem for these riders, and the suggested routes and choices of the riders are evaluated only from the perspective of the transit agency. This is discussed at length by Basciftci and Van Hentenryck (2023), where the authors studied this single-level variant of the problem for comparison. Their computational study presents that this single-level formulation only evaluates the suggested routes and choices of the riders from the perspective of the transit agency, who consequently can suggest longer routes to the riders with choice if serving them is not profitable. Thus, their inconvenience is explicitly omitted in the system, which is undesirable for ensuring the access to the transit system. Consequently, the authors observe significantly less riders adopting ODMTS under this formulation, in comparison with the bilevel model. On the other hand, the bilevel formulation eliminates this pathological and unfair behavior. This is aligned with the objectives of many transit agencies which aims at using ODMTS to improve mobility for underserved communities.

As a result, in any single-level reformulation and solution algorithm, the bilevel nature of the problem and the combinatorial choice functions of the riders need to be carefully addressed. To this end, in this paper, both path formulations only reason about paths for riders with choices: they continue to use decision variables $\mathbf{x^r}$ and $\mathbf{y^r}$ for riders in T_{core}.

4.1.4. Path Properties. It is also useful to characterize the behavior of the leader problem (5) and the subproblem (6). Given a design $\overline{\mathbf{z}}$ and a trip $r \in T_{latent}$, the subproblem returns a set of optimal paths $\mathcal{O}^r_f(\overline{\mathbf{z}})$. The following two properties of the bilevel problem (5) are important in case $\mathcal{O}^r_f(\overline{\mathbf{z}})$ is not a singleton, that is, multiple paths have the same $g(\pi)$ value.

Property 1. *Given a design $\overline{\mathbf{z}}$, if there exists $\pi \in AP^r \cap \mathcal{O}^r_f(\overline{\mathbf{z}})$, then the optimization selects π since it decreases the objective* (5a) *(the third term in the objective function is negative).*

Property 2. *Given a design $\overline{\mathbf{z}}$, if $AP^r \cap \mathcal{O}^r_f(\overline{\mathbf{z}}) = \emptyset$ and there exists $\pi \in RNP^r \cap \mathcal{O}^r_f(\overline{\mathbf{z}})$, then the optimization selects π, since it does not increase the objective value.*

Property 1 favors the selection of a path in AP^r over a path in RP^r, while Property 2 prefers a path in RNP^r over a path in ANP^r, since the latter would induce a positive weighted cost in the third term of the objective function.

4.2. Formulation C-P TH

This section introduces the first path-based formulation, C-Path, that reasons over the sets ^{r} and A^r for each trip $r \in T_{latent}$. By reasoning about these paths, and not over the variables $\mathbf{x^r}$ and $\mathbf{y^r}$ for riders in T_{core}, C-Path can be expressed as a single-level formulation. The key to C-Path is to make sure that only optimal paths (i.e., those returned by the follower subproblem in the bilevel formulation) are selected for riders with choice. A high-level presentation of C-Path is presented in Formulation (12). The formulation uses two abbreviations

$$Feasible(\pi, \mathbf{z}) \equiv \forall (h,l) \in x(\pi) : z_{hl} = 1$$

and

$$Optimal(\pi, \mathbf{z}) \equiv \mathbb{1}(g(\pi) = \min\{g(\pi') : \pi' \in {}^r \quad Feasible(\pi', \mathbf{z})\})$$

Feasible$(\pi, \mathbf{z})$ holds if path π is feasible under design $\mathbf{z}$. Note that there is no need to consider the shuttle arcs since they are always available. *Optimal*$(\pi, \mathbf{z})$ holds if path π has an optimal objective value among all the feasible paths of design $\mathbf{z}$.

$$\min_{z_{hl}} \quad \sum_{h,l \in H} \beta_{hl} z_{hl} + \sum_{r \in T_{core}} p^r g^r + \sum_{r \in T_{latent}} \sum_{\pi \in A^r} p^r \lambda^r_\pi (g(\pi) \quad \varphi) \tag{12a}$$

$$\text{s.t.} \quad \sum_{l \in H} z_{hl} = \sum_{l \in H} z_{lh} \quad \forall h \in H \tag{12b}$$

$$z_{hl} = 1 \quad \forall (h,l) \in Z_{fixed} \tag{12c}$$

$$g^r = \sum_{h,l \in H} \tau^r_{hl} x^r_{hl} + \sum_{i,j \in N} \gamma^r_{ij} y^r_{ij} \quad \forall r \in T_{core} \tag{12d}$$

$$\sum_{\substack{h \in H \\ \text{if } i \in H}} (x^r_{ih} \quad x^r_{hi}) + \sum_{i,j \in N} (y^r_{ij} \quad y^r_{ji}) = \begin{cases} 1 & \text{if } i = or^r \\ \quad 1 & \text{if } i = de^r \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in N, r \in T_{core} \tag{12e}$$

$$x^r_{hl} \le z_{hl} \quad \forall h,l \in H, r \in T_{core} \tag{12f}$$

$$\lambda^r_\pi \le Feasible(\pi,\mathbf{z}) \quad Optimal(\pi,\mathbf{z}) \quad \forall r \in T_{latent}, \pi \in \;^r \tag{12g}$$

$$\sum_{\pi \in \;^r} \lambda^r_\pi = 1 \quad \forall r \in T_{latent} \tag{12h}$$

$$z_{hl}, x^r_{hl}, y^r_{ij} \in \{0,1\} \quad \forall h,l \in H, i,j \in N, r \in T_{core} \tag{12i}$$

$$\lambda^r_\pi \in \{0,1\} \quad \forall r \in T_{latent}, \pi^r \in \;^r \tag{12j}$$

In formulation (12), variables z_{hl} and variables x^r_{hl}, y^r_{ij}, and g^r for $r \in T_{core}$ are directly adopted from the bilevel formulation (5). The constraints for riders with choice are expressed in terms of paths. Variable λ^r_π indicates whether path π is selected among the optimal solutions of the follower problem (6) for trip $r \in T_{latent}$. Constraints (12h) ensure that only one such path is selected. Constraints (12g) guarantee that the selected paths are indeed feasible and optimal for the follower subproblem. The first two terms of the objective function are the same as in the bilevel formulation. The third term is more interesting. The selected path π for a trip $r \in T_{latent}$ only contributes to the objective if it belongs to A^r, that is, it is adopted by trip r. This achieves the same effect as the choice function in the bilevel formulation. It is interesting to note that formulation (12) is single-level: the use of paths has eliminated the need for follower problems.

It remains to linearize the two abbreviations. The linearization of $Feasible(\pi,\mathbf{z})$ introduces a variable f^r_π for each trip $r \in T_{latent}$ and each path $\pi \in \;^r$ and the following constraints:

$$f^r_\pi \le z_{hl} \quad \forall r \in T_{latent}, \pi \in \;^r, (h,l) \in x(\pi) \tag{13a}$$

$$valid(\pi,\mathbf{z}) \le f^r_\pi \quad \forall r \in T_{latent}, \pi \in \;^r \tag{13b}$$

where

$$valid(\pi,\mathbf{z}) \equiv \sum_{(h,l)\in x(\pi)} z_{hl} \quad |x(\pi)| + 1$$

This last constraint expresses that the path π is allowed by design $\mathbf{z}$. In other words, a path π cannot be chosen if it is infeasible ($f^r_\pi = 0$) under design $\mathbf{z}$. The linearization of $Optimal(\pi,\mathbf{z})$ introduces a variable m^r for all $r \in T_{latent}$ and the constraints

$$f^r_\pi \to m^r \le g(\pi) \quad \forall r \in T_{latent}, \pi \in \;^r$$

where m^r represents the optimal objective value of the subproblem. To ensure that the selected paths are feasible, the linearization of (12g) adds the constraints

$$\lambda^r_\pi \le f^r_\pi \quad \forall r \in T_{latent}, \pi \in \;^r.$$

To guarantee that the selected paths are optimal, the linearization adds the constraints

$$\lambda^r_\pi \to m^r \ge g(\pi) \quad \forall r \in T_{latent}, \pi \in \;^r.$$

Indeed, if $m^r < g(\pi)$, then it must be the case that $\lambda^r_\pi = 0$ and path π cannot be selected.

Theorem 1. *Formulation* (12) *returns a design that is optimal for Formulation* (5).

Detailed proof of C-Path's optimality and its full formulation can be found in Online Supplement B. The presentation of C-Path, the abbreviations used in C-Path, and the constraints linearization together serve as a transitional step, linking the bilevel formulation to the P-Path formulation, the main contribution of this paper. The similarity between C-Path and the bilevel formulation is that $(\mathbf{x^r}, \mathbf{y^r})$ are used for core trips in T_{core}, while P-Path deviates from this approach.

4.3. Formulation P-Path

$$\min_{z_{hl}} \quad \sum_{hl \in H} \beta_{hl} z_{hl} + \sum_{r \in T_{core}} p^r g^r + \sum_{r \in T_{latent}} \sum_{\pi \in A^r} p^r \lambda^r_\pi (g(\pi) \quad \varphi) \tag{14a}$$

$$\text{s.t.} \quad \sum_{l \in H} z_{hl} = \sum_{l \in H} z_{lh} \quad \forall h \in H \tag{14b}$$

$$z_{hl} = 1 \quad \forall (h,l) \in Z_{fixed} \tag{14c}$$

$$g^r = \sum_{hl \in H} \tau^r_{hl} x^r_{hl} + \sum_{i,j \in N} \gamma^r_{ij} y^r_{ij} \quad \forall r \in T \tag{14d}$$

$$\sum_{\substack{h \in H \\ \text{if } i \in H}} (x^r_{ih} \quad x^r_{hi}) + \sum_{i,j \in N} (y^r_{ij} \quad y^r_{ji}) = \begin{cases} 1 & \text{if } i = or^r \\ \ \ 1 & \text{if } i = de^r \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in N, \ \forall r \in T \tag{14e}$$

$$x^r_{hl} \leq z_{hl} \quad \forall h,l \in H, \ \forall r \in T \tag{14f}$$

$$g^r \leq \min\{g(\pi) : \pi \in A^r \cup RP^r \quad Feasible(\pi, \mathbf{z})\} \quad \forall r \in T_{latent} \tag{14g}$$

$$\lambda^r_\pi = Selected(\pi, \mathbf{x^r}) \quad \forall r \in T_{latent}, \pi \in A^r \tag{14h}$$

$$z_{hl}, x^r_{hl}, y^r_{ij} \in \{0,1\} \quad \forall h,l \in H, \ \forall i,j \in N, \ \forall r \in T \tag{14i}$$

$$\lambda^r_\pi \in \{0,1\} \quad \forall r \in T_{latent}, \pi \in A^r \tag{14j}$$

This section proposes the second path-based formulation, P-Path, that utilizes the sets RP^r and A^r for each trip $r \in T_{latent}$. A high-level presentation of P-Path is presented in Formulation (14). The critical novelty in P-Path is that variables $(\mathbf{x^r}, \mathbf{y^r})$ are used for every trip $r \in T$. In other words, P-Path computes the paths for every rider through constraints (14b), (14c), (14d), (14e), and (14f). The role of the sets A^r and RP^r is to ensure that the selected paths for riders with choice are optimal solutions to the follower subproblem. This is achieved in two steps. First, the constraints (14g) ensures that the selected path for trip r is no worse than all the feasible paths in $A^r \cup RP^r$. Second, constraints (14h) link decision variables $(\mathbf{x^r}, \mathbf{y^r})$ with the paths and their selection variables λ^r_π used in the objective function. The formulation uses the following abbreviation

$$Selected(\pi, \mathbf{x^r}) \equiv \forall (h,l) \in x(\pi) : x^r_{hl} = 1 \quad \forall (h,l) \notin x(\pi) : x^r_{hl} = 0$$

which indicates that path π has been selected for trip r.

Theorem 2. *Formulation* (14) *returns a design that is optimal for Formulation* (5).

Proof. The proof shows that every design has the same objective value in both formulations. Consider a design $\bar{\mathbf{z}}$ and a trip $r \in T_{latent}$. The proof makes a case analysis for each trip $r \in T_{latent}$. Riders in T_{core} do not raise any issue as discussed earlier.

Assume first that trip r has one or more profitable paths in $\mathcal{O}^r_f(\bar{\mathbf{z}})$. Such a path necessarily belongs to $AP^r \cup RP^r$. By property 1, if there exists such a path in A^r, the bilevel formulation would select it. P-Path also selects such a path π. Indeed, constraints (14g) ensures that π has the best objective value among the feasible paths in $AP^r \cup RP^r$. Moreover, the objective function drives λ^r_π to 1 to decrease the objective function (the third term becomes negative). Furthermore, and importantly, constraints (14h) ensure that a single path from AP^r is selected in order to avoid decreasing the objective with multiple profitable paths. If only RP^r contains optimal paths, the follower problem selects a path in RP^r that is profitable but rejected by riders of trip r. Here, constraint (14h) also ensures that all the feasible paths in AP^r are not selected. The objective terms in both the bilevel model and the P-Path formulation are thus the same.

If there is no profitable path for trip r in $\mathcal{O}^r_f(\bar{\mathbf{z}})$, the optimal path returned by the follower problem must belong to $ANP^r \cup RNP^r$. If such an optimal path exists in RNP^r, by Property 2, the bilevel formulation will choose one of them. P-Path ensures this automatically since the optimization has no incentive to assign the λ variables to 1. If no such path exists, the P-Path model is constrained to select the optimal path in A^r, which is exactly the choice that is made by the bilevel formulation. Again constraints (14h) makes sure that the proper λ^r_π variable is set to 1, all others being zeros. □

$$\min_{z_{hl}} \sum_{h,l \in H} \beta_{hl} z_{hl} + \sum_{r \in T_{core}} p^r g^r + \sum_{r \in T_{latent}} \sum_{\pi \in A^r} p^r \lambda^r_\pi (g(\pi) - \varphi) \tag{15a}$$

s.t. (14b)–(14f)

$$f^r_\pi \le z_{hl} \quad \forall r \in T_{latent}, \pi \in A^r \cup RP^r \tag{15b}$$

$$\sum_{(h,l) \in x(\pi)} z_{hl} - |x(\pi)| + 1 \le f^r_\pi \quad \forall r \in T_{latent}, \pi \in A^r \cup RP^r \tag{15c}$$

$$f^r_\pi \rightarrow g^r \le g(\pi) \quad \forall r \in T_{latent}, \pi \in A^r \cup RP^r \tag{15d}$$

$$\lambda^r_\pi \le x_{hl} \quad \forall r \in T_{latent}, \pi \in A^r, (h,l) \in x(\pi) \tag{15e}$$

$$\lambda^r_\pi \le (1 - x_{hl}) \quad \forall r \in T_{latent}, \pi \in A^r, (h,l) \notin x(\pi) \tag{15f}$$

$$\lambda^r_\pi \ge \sum_{(h,l) \in x(\pi)} x^r_{hl} - |x(\pi)| - \sum_{(h,l) \notin x(\pi)} x^r_{hl} + 1 \quad \forall r \in T_{latent}, \pi \in A^r \tag{15g}$$

(14i)–(14j)

$$f^r_\pi \in \{0,1\} \quad \forall r \in T_{latent}, \ \forall \pi \in A^r \cup RP^r \tag{15h}$$

Constraints (14g) can be linearized easily by using the same variables f^r_π as in model C-Path. Constraints (14h) can be linearized with the set of constraints (15e)–(15g).

The final formulation for P-Path is given in Formulation (15), where the implication constraints can be replaced with their linear big-M transformations. The number of variables and constraints for P-Path are given by:

$$\#\text{Variables}: |H|^2 + (|H|^2 + |N|^2 + 1) \cdot |T| + \sum_{r \in T_{latent}} (2|A^r| + |RP^r|)$$

$$\#\text{Binary Variables}: |H|^2 + (|H|^2 + |N|^2) \cdot |T| + \sum_{r \in T_{latent}} (2|A^r| + |RP^r|)$$

$$\#\text{Constraints}: |H| + |Z_{fixed}| + (1 + |N| + |H|^2) \cdot |T| + \sum_{r \in T_{latent}} (|A^r|(|H|^2 + 4) + 3|RP^r|)$$

It is worth pointing out that constraint (15e) can be replaced by the following constraints to reduce model size:

$$\lambda^r_\pi \le f^r_\pi \quad \forall r \in T_{latent}, \pi \in A^r \tag{16}$$

This is valid because a path cannot be selected if it is not feasible ($f^r_\pi = 0$).

Remark 1. To demonstrate the equivalence of the P-Path formulation with the original bilevel formulation proposed by Basciftci and Van Hentenryck (2023) with the follower problem (7) (as presented in Section 3.2), the lexicographic objective can be integrated into the calculation of g^r and $g(\pi)$ values. This can be achieved by replacing τ^r_{hl} and γ^r_{ij} values with $\hat{\tau}^r_{hl} := M\tau^r_{hl} + t^r_{hl} + t^{wait}_{hl}$ and $\hat{\gamma}^r_{ij} := M\gamma^r_{ij} + t_{ij}$, respectively, for a sufficiently large big-M value. To adjust the objective function value in (15a), β_{hl} and φ values can be further replaced with $\hat{\beta}_{hl} := M\beta_{hl}$ and $\hat{\varphi} := M\varphi$, respectively. These modifications apply for the C-Path formulation as well to demonstrate its equivalence with the original bilevel formulation.

To solve P-Path, commercial solvers can be directly applied. Utilizing a warm-start approach by giving an initial solution as $z_{hl} = 0$ for $h, l \in H$ such that $(h,l) \notin Z_{fixed}$ can be beneficial. Additionally, employing priority branching in the branch and bound tree by prioritizing network design decisions $\mathbf{z}$ over the remaining binary variables improve the computational performance.

Remark 2. For addressing the P-Path problem, one might think of applying a solution algorithm, which can incrementally add the necessary paths and constraints for more efficiency. To this end, a lazy constraint-generation algorithm can be designed by first considering a subset of paths from the set $A^r \cup RP^r$ for every trip $r \in T_{latent}$, namely $A^r_{temp} \cup RP^r_{temp}$, and then iteratively incorporating paths to this initial problem. In each iteration, for every trip $r \in T_{latent}$, the optimality of the selected path under the given network design can be evaluated by solving the follower's problem under the given design. If the path is not optimal, then the necessary paths are added to $A^r_{temp} \cup RP^r_{temp}$ to be considered in the subsequent iterations. This step corresponds to evaluating Constraint (14g) of the P-Path formulation and incrementally adding the relevant paths. The algorithm stops when the paths selected for every trip $r \in T_{latent}$ correspond to the optimal paths under the transit network design identified by the model. However, experimental results have shown that such approaches do not offer computational advantages in solving the P-Path problem. This situation is due to the bilevel nature of the original problem: feasibility and optimality of the selected paths under the resulting network design need to be ensured. A detailed discussion on the algorithm and its computational results can be found in Online Supplement F.

5. Preprocessing Techniques

This section introduces preprocessing techniques that reduce problem sizes significantly.

5.1. Elimination of Shuttle Trips and Paths

The ODMTS-DA problem always admits feasible paths for all riders: they include direct shuttle trips and trips that are only using fixed hub legs in Z_{fixed}. Let $\overline{\Pi}^r$ be the set of these paths and $\overline{g}^r$ be the minimal cost among them, that is,

$$\overline{g}^r = \min_{\pi \in \overline{\Pi}^r} g(\pi).$$

Here, $\overline{g}^r$ provides an upper bound on the weighted cost of the optimal path of trip r under any network design, as addition of hub legs to fixed hub legs in Z_{fixed} can only improve the desired value. To this end, the set L^r can be reduced by removing all shuttle legs whose weighted costs are greater than $\overline{g}^r$, that is,

$$\{(i,j) \in L^r \mid \gamma_{ij} > \overline{g}^r\}.$$

Furthermore, this upper bound can be used to reduce the set Π^r for every trip $r \in T_{latent}$ by removing the paths

$$\{\pi \in \Pi^r \mid g(\pi) > \overline{g}^r\}.$$

By applying an analogous relationship, the sets A^r and RP^r can be significantly reduced as well. Note that the proposed approach in preprocessing the sets Π^r, A^r, RP^r is valid under any choice function (viewed as a black-box). These sets can be further reduced by leveraging the structure of the choice function, for example, exploiting the trip duration and transfer limits in (4). These reduced sets are referred as Π^r_{red}, A^r_{red}, RP^r_{red} in the case studies in Sections 6 and 7 and Online Supplement C to showcase the impact of these preprocessing techniques over different instances.

5.2. Path Assignments

Some paths are guaranteed to be selected by Models C-Path and P-Path. Let

$$\underline{g}^r = \min_{\pi \in \Pi^r} g(\pi).$$

Here, $\underline{g}^r$ provides a lower bound on the weighted cost of the optimal path of trip r under any network design. If $\overline{g}^r = \underline{g}^r$ and there exists a path $\overline{\pi} \in \overline{\Pi}^r$ such that

$$\arg\min_{\pi \in \overline{\Pi}^r} g(\pi) = \{\overline{\pi}\},$$

then $\overline{\pi}$ is an optimal assignment for trip r. If $\overline{\pi} \in A^r$, then this assignment is an adopting path and riders in r are guaranteed to adopt. This condition can be generalized when there are multiple optimal paths by reasoning about the path profitability. More specifically, if there exists such a path $\overline{\pi} \in \arg\min_{\pi \in \overline{\Pi}^r} g(\pi) \cap AP^r$, then that rider will adopt the ODMTS under any network design. Otherwise, that rider will reject the ODMTS.

5.3. Rider Removal

A trip r can be removed from Models C-Path and P-Path when they are guaranteed (i) to adopt a specific path or (ii) not to adopt any path. The first case realizes when there exists an optimal path assignment from adopting paths as discussed above. This can be precomputed and added to the resulting objective of the optimization models. The latter case occurs when the optimal path assignment belongs to set $RP^r \cup RNP^r$. The guaranteed rejection can further happen when A^r is initially empty or A^r_{red} becomes empty after preprocessing. After removing these trips from the set T_{latent}, the reduced set is referred to as $T_{latent,red}$.

5.4. Removal of Hub Legs

It is also possible to remove hub legs that are too far from the origin and destination of a rider. Let $\underline{g}^r_{or,h}$ be the weighted cost from origin *or* to hub h and $\underline{g}^r_{l,de}$ be the weighted cost from hub l to destination *de*, where these costs are computed by identifying the shortest paths between these locations when all hub legs are available. If

$$\underline{g}^r_{or,h} + \tau_{hl} > \overline{g}^r \vee \underline{g}^r_{l,de} + \tau_{hl} > \overline{g}^r,$$

then variable x^r_{hl} can be removed from Models C-Path and P-Path, as its inclusion will not contribute to the objective function.

Combining these preprocessing techniques, the sizes of the formulations C-Path and P-Path can be significantly reduced, providing computational efficiency. Under the generalized setting (subproblem (6)), when solving P-Path with linearized constraints, the big-M values introduced to linearize constraint (15d) for each trip r can simply be chosen as $\overline{g}^r$, given that all paths with greater weighted cost than $\overline{g}^r$ are already eliminated.

When enumerating useful paths for P-Path, applying the previously mentioned preprocessing techniques are beneficial. Additionally, leveraging the underlying structure of the choice model $\mathcal{C}^r$ can also be advantageous. In the case studies in Sections 6 and 7, the Algorithm PE-DCM (Path Enumeration Dedicated to Choice Model (4)) is applied because the case studies focus on Choice Model (4). This tailored algorithm can efficiently generate A^r and $RP^r \ \forall r \in T$ by employing both preprocessing techniques and leveraging the structure of choice model (4). In general, the preprocessing techniques provide a significantly smaller search space for each trip when enumerating its paths for P-Path. Moreover, due to the significance of the transfer tolerance parameter (l^r_{ub}) in the choice model, the algorithm has the capability to eliminate all paths containing more than $l^r_{ub} + 1$ arcs. The technical details of this algorithm is outlined in Online Supplement A.

6. The Case Study in Ypsilanti

This section presents a case study using a realistic data set from AAATA,[1] the transit agency serving Ypsilanti area and the broader of Ann Arbor of Michigan, USA. More specifically, Section 6.1 first describes the experimental settings used in this case study. Section 6.2 then demonstrates the workability and computational advantages of the path-based formulations by comparing their computational results with the exact decomposition algorithm by Basciftci and Van Hentenryck (2023) and the path-based methods presented in Sections 4 and 5.

6.1. Experimental Settings

The experimental settings for this case study are mainly summarized in Tables 2 and 3. This regular-sized case study is based on the AAATA transit system that operates over 1,267 ODMTS stops. In order to design an ODMTS, 10 stops at high density corridors are selected as ODMTS hubs, and the other stops are only accessible to shuttles. No specific restriction is applied on new bus arcs; therefore, the 10 hubs lead to 90 z_{hl} variables. Moreover, all existing bus lines are assumed to be eliminated, that is, there is no backbone lines preserved in this case study. For new bus arcs, d'_{hl} and t'_{hl} are all assumed to be equal to d_{hl} and t_{hl}. The data set entails trips between 6 p.m. and 10 p.m., primarily consisting of commuting trips from work locations to home.

For ridership, the data set includes 1,503 trips (distinct O-D pairs) for a total of 5,792 riders. The mode preference of a rider depends on her income level, that is to say, a rider from a lower income level has a higher tolerance to travel time. There are 476 low-income, 819 middle-income, and 208 high-income trips with 1,754, 3,316, and 722 riders respectively. The classification of income level are introduced in Online Supplement C. An ^{r} value associated with each choice model (i.e., Equation (3)) is assigned to each class. Note that all trips in the low-income class are treated as members of the core trips set T_{core}; hence, no ^{r} value is required for them. For the middle-income and the high-income classes, 2.0 and 1.5 are employed as the ^{r} values, respectively.

To evaluate the performance of the proposed methodology under different configurations, four instances are generated as described in Table 3. In particular, when the ridership is doubled, the number of riders of each trip is twice as large. The core trips percentage parameter for each income level is utilized to divide the data set into core trips and latent trips by varied partitions.

The on-demand shuttle price is set as \$1 per kilometer and each shuttle is assumed to only serve one passenger. For buses, the operating fee is \$3.87 per kilometer and four buses are assumed to operate between open legs, resulting in an average of 7.5 minutes waiting time (t^{wait}_{hl}). A fixed \$2.5 ticket price that is in line with the current AAATA

Table 2. The Sets and Parameter Values Applied to the Ypsilanti Case Study

Sets and parameter	Set size or parameter value	Additional notes
N	1,267	Visualized in Online Supplement C
H	10	Visualized in in Online Supplement C, 90 z_{hl} variables need to be determined
T	1,503	—
T_{core}	1,194 (Instances 1 and 3), 937 (Instances 2 and 4)	More details can be found in Table 3
T_{latent}	309 (Instances 1 and 3), 566 (Instances 2 and 4)	More details can be found in Table 3
Z_{fixed}	$\emptyset$	No backbone lines are considered in Ypsilanti
n_{hl}	16	New buses' headway is assumed to be 900 s (15 min) ODMTS operation lasts for 4 hours
t_{hl}^{wait}	450 s	The expected waiting time computed from buses' headway
b_{dist}	\$3.87 per kilometer	—
w_{dist}	\$1 per kilometer	—
r	1.5 and 2.0 for middle and high income households	Households' income level are presented in Online Supplement C
ϕ	\$2.5 per rider	Consistent with AAATA's current ticket price

system is selected, regardless of the travel length and multimodality of the trip. Inconvenience is measured in seconds, and the inconvenience and cost parameter θ is set as 0.001.

The proposed approaches are applied to each of the four instances presented in Table 3. First, a benchmark run on the original bilevel formulation (see subproblem (7)) is carried out with the exact algorithm proposed in Basciftci and Van Hentenryck (2023). The exact algorithm is set to terminate once the gap reaches 0.1% or the running time exceeds 6 hours, and the upper bound of its result is reported as the objective value. Secondly, runs with P-Path Model are applied to those four instances. To compare these reformulations with the benchmark approach, P-Path are adjusted to compute the lexicographic optimal appeared in original subproblem (7), as discussed in Remark 1. On the other hand, an additional run with P-Path is conducted when the generalized bilevel formulation (see subproblem (6)) is considered. All models and algorithms were firstly programmed with Python 3.7 then updated to 3.10, and Gurobi 9.5 is selected as the solver. The online repository Guan et al. (2024) contains all the code, along with a sample test case for reference.

6.2. Computational Results

This section presents the computational study over the four instances in Table 3. Note that instances 1 and 3 were used in the previous study (Basciftci and Van Hentenryck 2023) under the original bilevel framework, where the authors articulate the benefits of deploying an ODMTS. This paper only focuses on computational aspects.

6.2.1. Path Enumeration Efficiency. Table 4 presents the model size and reports the computation times for enumerating the paths for P-Path model. For black-box path enumeration, Algorithm PE (see Online Supplement A) can be utilized with all preprocessing techniques applied. The enumeration algorithms may use the choice models as

Table 3. The Experimental Setups for the Four Instances

		Low income core trips			Medium income core trips			High income core trips		
Instance	Ridership	% trips	# trips	# riders	% trips	# trips	# riders	% trips	# trips	# riders
Instance 1	Regular	100%	476	1,754	75%	614	2,842	50%	104	434
Instance 2	Regular	100%	476	1,754	50%	409	2,262	25%	52	258
Instance 3	Doubled	100%	476	3,508	75%	614	5,684	50%	104	868
Instance 4	Doubled	100%	476	3,508	50%	409	4,524	25%	52	516

Notes. For doubled ridership, the number of riders for each O-D pair is multiplied by 2. The core trips percentages for each income level are [100%, 75%, 50%] and [100%, 50%, 25%] for low, medium, and high income trips, respectively.

Table 4. The Number of Paths, Variables, and Constraints in the P-Path Model After Preprocessing

						Path Enum. Time (min)	
P-Path	$\sum_{r\in T_{latent,red}} \|RP^r_{red}\|$	$\sum_{r\in T_{latent,red}} \|A^r_{red}\|$	# Vars.	# Binary Vars.	# Constrs.	Black-box	Dedicated
Instance 1	200	675	136,427	135,041	203,243	27.72	0.02
Instance 2	284	1,190	126,110	124,825	237,707	46.90	0.04
Instance 3	200	675	136,427	135,041	203,243	27.61	0.02
Instance 4	284	1,190	126,110	124,825	237,707	49.02	0.04

black-boxes or they can exploit their underlying structures to prune the search space earlier. This is especially helpful when using the P-Path model since it only takes A^r_{red} and RP^r_{red} as inputs, and these sets impose strong conditions on the choice models. In particular, for the choice model (3), applying k-shortest path algorithms on graphs in terms of t^r and g^r makes it possible to enumerate all paths in A^r_{red} and RP^r_{red} almost instantly (see Algorithm PE-DCM in Online Supplement A for details). Table 4 describes the significant improvements in running times when the structure of the choice models is used during the enumeration processes. The k-shortest path algorithms used in this study are by Yen (1971). Corresponding results for the C-Path model is presented in Online Supplement C.

6.2.2. Computational Efficiency. Table 5 summarizes the computation results, where the benchmark is the approach proposed by Basciftci and Van Hentenryck (2023). Under the original bilevel formulation, P-Path improves the running time by at least two orders of magnitude. In fact, P-Path finds the optimal solution in a few minutes, while the benchmark has not reached the 0.1% gap within the time limit of 6 hours. These results demonstrate the workability and computational efficiency of P-Path. Under the generalized bilevel formulation, P-Path finds optimal solutions even faster. Interestingly, the two bilevel formulations return the same optimal design on this case study. Note also that, under generalized bilevel formulation, the discrete variables x^r_{hl} and y^r_{ij} can be relaxed to be continuous variables since the subproblem (6) becomes totally unimodular when the design decisions are fixed. This further improves the computational efficiency of the P-Path model.

6.2.3. The Importance of Preprocessing Techniques. Table 6 demonstrates the substantial benefits of the preprocessing techniques: it shows the number of paths that are eliminated for the various sets. The first observation is that more than 35% of the latent trips can be eliminated through path assignments and rider removal. Moreover, the reductions on A^r and RP^r are also helpful, where the size reduction comes from two sources: (i) the smaller number of latent trips in $T_{latent,red}$, and (ii) the reduced number of allowed shuttle arcs in L^r_{red}. Finally, the number of eliminated x_{hl} variables is nonnegligible. In practice, computing $\overline{g}^r$, reducing L^r, and eliminating x^r_{hl} are performed separately, and the total computational time is less than 30 seconds. All other set reductions are carried out within

Table 5. Computational Efficiency of the P-Path Model

Instance	Bilevel subproblem type	Run	Optimality gap (%)	Run time (min)	Objective
Instance 1	Original	Benchmark	0.65	364.10	19,012.91
		P-Path	0.00	2.19	19,012.91
	Generalized	P-Path	0.00	2.70	19,012.91
Instance 2	Original	Benchmark	2.48	367.52	16,635.73
		P-Path	0.00	7.80	16,635.73
	Generalized	P-Path	0.00	5.04	16,635.73
Instance 3	Original	Benchmark	0.99	363.22	34,732.09
		P-Path	0.00	2.62	34,732.09
	Generalized	P-Path	0.00	1.86	34,732.09
Instance 4	Original	Benchmark	1.85	360.29	29,962.79
		P-Path	0.00	3.34	29,962.79
	Generalized	P-Path	0.00	2.32	29,962.79

Notes. Original stands for the original bilevel framework with subproblem (7); thus, the lexicographic optimal are considered. Generalized stands for the bilevel framework with subproblem (6) that is compatible with a black-box choice function.

Table 6. The Impact of Preprocessing on the P-PATH Model: Reductions of the Set Sizes

# Reduction on	$\lvert T_{latent}\rvert$	$\sum_{r\in T}\lvert L^r\rvert$	$\sum_{r\in T_{latent}}\lvert RP^r\rvert$	$\sum_{r\in T_{latent}}\lvert A^r\rvert$	$\sum_{r\in(T_{core})\cup T_{latent,red}}\sum_{h,l\in H} x^r_{hl}$
Instances 1 & 3	117	20,002	27	289	1,839
Instances 2 & 4	218	20,002	55	498	3,761

the path enumeration. In summary, these results highlight the effectiveness of preprocessing and their importance when solving ODMTS-DA problems.

7. Large-scale Case Study in tlanta

This section presents a large-scale case study conducted with travel data corresponding to a regular workday in Atlanta, Georgia, USA. The experiment is designed to evaluate the performance of P-PATH on large-scale cases. Section 7.1 explains the experimental setting, while Section 7.2 presents the computational results and the benefits of preprocessing techniques. Online Supplement E reports results on the designed ODMTS for those readers interested in the outcomes of the optimizations. This complements the results on Ypsilanti detailed in Basciftci and Van Hentenryck (2023).

7.1. Experimental Setting

This section presents the experimental settings for the case study, which are mainly summarized in Tables 7 and 8. The Metropolitan Atlanta Rapid Transit Authority (MARTA) is the major agency that provides transit services for the Atlanta Metropolis. The test cases all include the MARTA rail system as the backbone lines, given their importance to the city. The operating price for on-demand shuttle and buses are fixed at \$0.621 per kilometer and \$72.15 per hour, respectively. The inconvenience and cost parameter θ is fixed at $7.25/(60+7.25)$. Contrary to Section 6, this case study uses minute as the unit of time instead of second. The θ value and the costs just presented are adopted from an Atlanta-based ODMTS study conducted by Auad et al. (2021). There is a unique bus frequency of six buses per hour, giving an average waiting time (t^{wait}_{hl}) of five minutes. The ODMTS thus operates 24 buses between h and l during the 4 hours' time horizon, when the bus arc (h, l) is selected. The bus arcs must all involve a connection from or to the rail and are best viewed as rapid bus transit lines expanding the rail system. The bus travel distance (d'_{hl}) and travel time (t'_{hl}) are assumed to be equal to d_{hl} and t_{hl}, and these values are obtained using Graph-hopper.[2] Given that the rail system in the ODMTS is directly adopted from the existing transit system, its operating costs is omitted during the design of the ODMTS, since it is a constant. Thus, $\beta_{hl}=0$ if a rail leg connects hubs h and l, and its travel time t'_{hl} is derived from public rail schedules or the General Transit Feed Specification (GTFS) files. Lastly, shuttles are allowed to connect hubs in this case study, and each shuttle is assumed to only serve 1 passenger, that is, ride-sharing is not available. A \$2.5 fee is charged for each ODMTS rider, which is the ticket price of MARTA. Additional information related to the data set can be found in Online Supplement D.

To test the scalability of P-PATH, six instances (see Table 8) were built by controlling two additional parameters: (i) the number of nearby rail hubs a bus-only-hub can connect with and (ii) the transfer tolerance (l^r_{ub}) of the

Table 7. The Sets and Values Applied to the East Atlanta Case Study

Sets and parameter	Set size or parameter value	Additional notes
N	2,426	Visualized in Figure 3b in Online Supplement D
H	58	Visualized in Figure 3c in Online Supplement D
T	55,871	—
T_{core}	15,478	—
T_{latent}	36,283	—
Z_{fixed}	692	Derived from four backbone rail lines, see Online Supplement D
n_{hl}	24	New buses' headway is assumed to be 10 min ODMTS operation lasts for 4 hours
t^{wait}_{hl}	5 min	The expected waiting time computed from buses' headway
b_{time}	\$72.15 per hour	—
w_{dist}	\$0.621 per kilometer	Equivalent to \$1 per mile
^{r}	1.5	See choice model (4)
l^r_{ub}	2 or 3	See choice model (4) and Table 8
ϕ	\$2.5 per rider	Consistent with MARTA's current ticket price

Table 8. The Six Experimental Instances Considered in the Atlanta Case Study

Instance	# Nearby rail hubs	# z_{hl}	# Undecided z_{hl}	Transfer tolerance l^r_{ub} in choice model (4)
Instance 1	1	732	40	2
Instance 2	1	732	40	3
Instance 3	2	774	82	2
Instance 4	2	774	82	3
Instance 5	3	828	136	2
Instance 6	3	828	136	3

passengers in the choice model (4). Values 1, 2, and 3 for the first parameter correspond to 732, 774, and 828 possible hub arcs. The two values used for l^r_{ub} are 2 and 3. Note that larger transfer tolerance values are generally impractical for transit riders, especially for riders who reside in a metropolitan region. All passengers are assumed to share the same transfer tolerance. Overall, these two parameters lead to six instances with different problem sizes and complexities. In addition, for the adoption factor in the choice model, a constant value $^r = 1.5$ is applied to all passengers. More details related to these two parameters are presented in Online Supplement D. All models and algorithms are again implemented and solved using Python 3.7 (later upgraded to 3.10) and Gurobi 9.5 (Guan et al. 2024).

7.2. Computational Results

This section presents the computational results of the six instances described in Table 8. They used all the preprocessing techniques.

7.2.1. Model Size. The size of the P-Path model over the six instances are reported in Table 9. They show that profitable rejecting paths (paths in RP^r_{red}) are rarely observed in large-scale instances and their numbers remain relatively small as the problem size grows. This highlights again the benefits of the P-Path formulation, since it only relies on a small number of paths. In addition, greater value for the transfer tolerance parameter (instances 2, 4, and 6) leads to considerably more complex problems since the numbers of adopting paths increase significantly. The path enumeration leverages the structure of the choice models. For choice model (4), bounding the path length by $l^r_{ub} + 1$ makes it possible to construct A^r_{red} rapidly, while k-shortest path algorithms can be employed for RP^r_{red}. Table 9 shows that the preprocessing (i.e., reducing path sets and trip sets during path enumeration) takes nonnegligible times, yet it is critical to reduce the sizes of the MIP models. Observe that the numbers of MIP variables are in the range of 14 to 18 millions and the numbers of constraints in the range of 18 to 38 millions.

7.2.2. MIP Solving Time. Table 10 reports the computational efficiency of the P-Path model. The key takeaway is that P-Path can solve all instances to optimality, demonstrating its scalability. The computational times increase significantly as the problem size grows. Instances 5 and 6 which have more bus arcs are particularly challenging, but can still be solved optimally. Table 10 also reports the computational time needed to reach a 1% optimality gap. The results highlight that the two largest instances spend the majority of the computation times on the last 1% and the optimality proof. *These results show that it is now in the realm of optimization technology to solve the ODMTS-DA problem at the scale of large metropolitan areas.* Moreover, Table 10 presents the time for the solver to find the optimal solution as the upper bound during the branch-and-bound procedures. It is evident that the model often reaches the 1% gaps first. P-Path also provides a benchmark to evaluate the quality of fast heuristics that were proposed in Guan

Table 9. The Number of Paths, Variables, and Constraints in the P-Path Model After Preprocessing

P-Path	$\sum_{r\in T_{latent,red}} \lvert RP^r_{red}\rvert$	$\sum_{r\in T_{latent,red}} \lvert A^r_{red}\rvert$	# Vars.	# Binary Vars.	# Constrs.	Path Enum. Time (min)
						Dedicated
Instance 1	33	7,827	14,206,406	16,420	18,648,588	8.85
Instance 2	33	11,115	14,311,802	22,996	21,121,733	30.11
Instance 3	51	11,663	15,126,181	22,152	21,620,809	9.39
Instance 4	51	17,098	15,270,518	35,022	26,652,670	32.16
Instance 5	78	16,042	17,228,315	32,991	27,797,908	12.03
Instance 6	78	27,538	17,585,987	55,979	37,357,885	44.47

Table 10. Computational Efficiency of the P-PATH Model

Instance	Bilevel subproblem type	Run	Computational time (hours) For 0% gap	For 1% gap	For Opt. Sol.	Objective
Instance 1	Generalized	P-PATH	1.94	1.93	1.93	217,944.89
Instance 2	Generalized	P-PATH	2.35	2.34	2.35	218,811.60
Instance 3	Generalized	P-PATH	3.59	2.84	3.04	217,196.06
Instance 4	Generalized	P-PATH	6.09	4.25	5.98	218,387.18
Instance 5	Generalized	P-PATH	46.75	9.90	35.01	212,917.59
Instance 6	Generalized	P-PATH	116.48	31.52	29.98	214,463.71

Note. Generalized stands for the bilevel framework with subproblem (6) that is compatible with a black-box choice function.

et al. (2022). Additional details regarding the use of P-PATH for assessing the quality of heuristic solutions can be found in Online Supplement G.

7.2.3. Preprocessing Results. The impact of preprocessing is shown in Table 11. After applying path assignments and rider removal, a vast number of trips are recognized as being ODMTS adoptions or rejections and are removed from T_{latent}. Most of the adoptions are related to local O-Ds in East Atlanta for which direct shuttle paths can be assigned to the riders. The rejections typically correspond to long paths that are cost-effective for the agency but inconvenient for the riders. Table 11 also shows that preprocessing eliminates about 50% of the variables (about 15 to 18 millions of variables). This is not surprising: many hub arcs are irrelevant for a given rider, since they are far away from the origin and destination of the trip. Both of these elimination methods utilize paths that exclusively involve backbone arcs and shuttles. This preprocessing technique underscores an advantage of ODMTSs. As emphasized by Van Hentenryck et al. (2023) in a realistic ODMTS pilot, connecting shuttle legs to rails is prominent, and servicing local trips is also a common scenario. These two types of paths can significantly reduce the problem size before solving P-PATH. Applying all preprocessing techniques usually take about 1.5–3 hours for the six instances. The most time consuming part is variable elimination (see Section 5, removal of hub legs) because every single hub leg needs to be analyzed with the trip set $T_{core} \cup T_{latent,red}$. However, this is still highly beneficial considering the reduction in the number of variables and the computational requirements of the MIP model, which is large-scale.

8. Conclusion

This paper considered the ODMTS Design with Adoption problem (ODMTS-DA) proposed by Basciftci and Van Hentenryck (2023) to capture the latent demand in on-demand multimodal transit systems. The ODMTS-DA is a bilevel optimization problem and Basciftci and Van Hentenryck (2023) proposed an exact combinatorial Benders decomposition to solve it. Unfortunately, their proposed algorithm only finds high-quality solutions for medium-sized cities and is not practical for large metropolitan cities. The main difficulty is in the tension between the design by the transit agency, which minimizes a combination of cost and inconvenience, and the choice model of the riders that expresses their tolerance to, for instance, transit time and the number of transfers.

This paper revisited the ODMTS-DA problem and presented a novel path-based optimization model, called P-PATH, aimed at solving large-scale instances. The key idea behind P-PATH is to replace the follower subproblems by enumerating certain types of paths that capture the essence: adopting paths and profitable paths rejected by the choice model. The paper showed that, with the help of these two sets, the bilevel formulation can be replaced by a single-level optimization which can be formulated as a MIP model. In addition, the paper presented a number of

Table 11. The Impact of Preprocessing on P-PATH Model

# Reduction on	$\lvert T_{latent} \rvert$	$\sum_{r \in T} \lvert L^r \rvert$	$\sum_{r \in T_{latent}} \lvert RP^r \rvert$	$\sum_{r \in T_{latent}} \lvert A^r \rvert$	$\sum_{r \in T_{core} \cup T_{latent,red}} \sum_{h,l \in H} x^r_{hl}$
Instance 1	30,642	4,805,467	10,175	1,117,353	14,990,648
Instance 2	30,501	4,805,467	10,175	2,587,646	15,110,231
Instance 3	30,261	4,805,467	10,188	1,137,236	15,931,908
Instance 4	30,082	4,805,467	10,188	2,621,658	16,103,498
Instance 5	29,235	4,805,467	10,229	1,171,093	18,106,680
Instance 6	27,746	4,805,467	10,229	2,693,458	18,521,145

preprocessing techniques that help in reducing the sizes of these path sets and hence the number of variables and constraints in the MIP model. P-PATH was evaluated on two comprehensive case studies: the midsize transit system of the Ann Arbor – Ypsilanti region (which was studied in prior work) and the city of Atlanta. For the Ann Arbor case study, the experimental results show that P-PATH solves the ODMTS-DA instances in a few minutes: this contrast with the existing approach that cannot solve the problem optimally after 6 hours. For Atlanta, the results show that P-PATH can solve the ODMTA-DA instances optimally: these instances are obviously out of scope for the existing method. The resulting MIPs are large-scale (about 17 millions of variables and 37 millions of constraints), but can be solved in a few hours, except for the large instances which may take longer. These results show the computational benefits of P-PATH, which provides a scalable approach to the design of on-demand multimodal transit systems with latent demand.

As pointed out in Section 3.1, the existing ODMTS-DA exhibits several limitations regarding modeling aspects, including the construction of fixed-routes with multiple bus frequencies, balancing shuttle flows, determining the optimal shuttle fleet size, and enabling ride-sharing for shuttles. While these limitations can be addressed in subsequent problems after resolving the ODMTS-DA (i.e., multiple bus frequencies as in Dalmeijer and Van Hentenryck (2020), design with shuttle ride-sharing and fleet-sizing as in Auad-Perez and Van Hentenryck (2022), and real-time shuttle dispatching with ride-sharing as in Riley et al. (2019)), it remains interesting for future research to explore integrating these elements into the ODMTS-DA framework. This exploration, particularly extending the P-PATH approach on shuttle ride-sharing and bus frequencies, could impact on the results of network design.

Endnotes

[1] See https://www.theride.org (Last Visited Date: January 16, 2024).

[2] See https://www.graphhopper.com/ (Last Visited Date: January 16, 2024).

References

Alizadeh R, Nishi T (2019) Dynamic p + q maximal hub location problem for freight transportation planning with rational markets. *Adv. Mech. Engrg.* 11(2):1687814018822934.

Almasi MH, Oh Y, Sadollah A, Byon YJ, Kang S (2021) Urban transit network optimization under variable demand with single and multi-objective approaches using metaheuristics: The case of Daejeon, Korea. *Internat. J. Sustain. Transportation* 15(5):386–406.

Alumur SA, Kara BY, Karasan OE (2012) Multimodal hub location and hub network design. *Omega* 40(6):927–939.

Auad R, Dalmeijer K, Riley C, Santanam T, Trasatti A, Van Hentenryck P, Zhang H (2021) Resiliency of on-demand multimodal transit systems during a pandemic. *Transportation Res. Part C Emerg. Tech.* 133:103418.

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

Basciftci B, Van Hentenryck P (2020) Bilevel optimization for on-demand multimodal transit systems. Hebrard E, Musliu N, eds. *Integration of Constraint Programming, Artificial Intelligence, and Operations Research* (Springer-Verlag, Berlin, Heidelberg), 52–68.

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

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

Calvete HI, Galé C, Iranzo JA (2014) Planning of a decentralized distribution network using bilevel optimization. *Omega* 49:30–41.

Canca D, De-Los-Santos A, Laporte G, Mesa JA (2016) A general rapid network design, line planning and fleet investment integrated model. *Ann. Oper. Res.* 246(1):127–144.

Canca D, De-Los-Santos A, Laporte G, Mesa JA (2017) An adaptive neighborhood search metaheuristic for the integrated railway rapid transit network design and line planning problem. *Comput. Oper. Res.* 78:1–14.

Canca D, De-Los-Santos A, Laporte G, Mesa JA (2019) Integrated railway rapid transit network design and line planning problem with maximum profit. *Transportation Res. Part E Logist. Transportation Rev.* 127:1–30.

Dalmeijer K, Van Hentenryck P (2020) Transfer-expanded graphs for on-demand multimodal transit systems. *Internat. Conf. Integration of Constraint Programming Artificial Intelligence Oper. Res.* (Springer, Cham, Switzerland), 167–175.

Fan W, Mei Y, Gu W (2018) Optimal design of intersecting bimodal transit networks in a grid city. *Transportation Res. Part B Methodological* 111:203–226.

Farahani RZ, Miandoabchi E, Szeto W, Rashidi H (2013) A review of urban transportation network design problems. *Eur. J. Oper. Res.* 229(2):281–302.

Goerigk M, Schmidt M (2017) Line planning with user-optimal route choice. *Eur. J. Oper. Res.* 259(2):424–436.

Guan H, Basciftci B, Van Hentenryck P (2022) Heuristic algorithms for integrating latent demand into the design of large-scale on-demand multimodal transit systems. Preprint, submitted December 7, https://arxiv.org/abs/2212.03460.

Guan H, Basciftci B, Van Hentenryck P (2024) Path-based formulations for the design of on-demand multimodal transit systems with adoption awareness. https://dx.doi.org/10.1287/ijoc.2023.0014.cd, https://github.com/INFORMSJoC/2023.0014.

Gudapati NV, Malaguti E, Monaci M (2022) Network design with service requirements: Scaling-up the size of solvable problems. *INFORMS J. Comput.* 34(5):2571–2582.

Kara BY, Verter V (2004) Designing a road network for hazardous materials transportation. *Transportation Sci.* 38(2):188–196.
Kleinert T, Labbé M, Ljubić I, Schmidt M (2021) A survey on mixed-integer programming techniques in bilevel optimization. *EURO J. Comput. Optim.* 9:100007.
Klier MJ, Haase K (2008) Line optimization in public transport systems. *Oper. Res. Proc. 2007* (Springer, Berlin, Heidelberg), 473–478.
Liu Y, Ouyang Y (2021) Mobility service design via joint optimization of transit networks and demand-responsive services. *Transportation Res. Part B Methodological* 151:22–41.
Mahéo A, Kilby P, Van Hentenryck P (2019) Benders decomposition for the design of a hub and shuttle public transit system. *Transportation Sci.* 53(1):77–88.
Marcotte P, Mercier A, Savard G, Verter V (2009) Toll policies for mitigating hazardous materials transport risk. *Transportation Sci.* 43(2):228–243.
Meng Q, Yang H, Bell MG (2001) An equivalent continuously differentiable model and a locally convergent algorithm for the continuous network design problem. *Transportation Res. Part B Methodological* 35(1):83–105.
Najmi A, Rashidi TH, Waller T (2023) A multimodal multi-provider market equilibrium model: A game-theoretic approach. *Transportation Res. Part C Emerg. Tech.* 146:103959.
Pinto HK, Hyland MF, Mahmassani HS, Verbas IÖ (2020) Joint design of multimodal transit networks and shared autonomous mobility fleets. *Transportation Res. Part C Emerg. Tech.* 113:2–20.
Riley C, Legrain A, Van Hentenryck P (2019) Column generation for real-time ride-sharing operations. *Integration Constraint Programming Artificial Intelligence Oper. Res. 16th Internat. Conf. CPAIOR 2019*, Thessaloniki, Greece, June 4–7, 2019, 16 (Springer, Cham, Switzerland), 472–487.
Schöbel A (2012) Line planning in public transportation: Models and methods. *OR Spectrum* 34(3):491–510.
Shaheen S, Chan N (2016) Mobility and the sharing economy: Potential to facilitate the first-and last-mile public transit connections. *Built Environ.* 42(4):573–588.
Stiglic M, Agatz N, Savelsbergh M, Gradisar M (2018) Enhancing urban mobility: Integrating ride-sharing and public transit. *Comput. Oper. Res.* 90:12–21.
Van Hentenryck P, Riley C, Trasatti A, Guan H, Santanam T, Huertas JA, Dalmeijer K, Watkins K, Drake J, Baskin S (2023) Marta reach: Piloting an on-demand multimodal transit system in Atlanta. Preprint, submitted August 4, https://arxiv.org/abs/2308.02681.
Yen JY (1971) Finding the k shortest loopless paths in a network. *Management Sci.* 17(11):712–716.