An operation-agnostic stochastic user equilibrium model for Mobility-on-Demand networks with congestible capacities

Bingqing Liu[1], David Watling[2], Joseph Y. J. Chow[1*]
[1]C2SMARTER University Transportation Center, New York University, New York, NY 10012, US
[2]Institute for Transport Studies, University of Leeds, Woodhouse, Leeds LS2 9JT, UK
[*]Corresponding author email: joseph.chow@nyu.edu

Abstract

Evaluating the impact of privately-owned Mobility-on-Demand (MoD) services is important from a regulatory perspective. There is a need to model multimodal equilibria with MoD to support policymaking. While there exists a large body of literature on MoD services focusing on service design under equilibrium modeling, these studies commonly adopt assumptions of MoD operational policies. However, such policies might not be shared with regulatory agencies due to commercial privacy concerns of private operators. We model multimodal equilibrium with MoD systems in an operation-agnostic manner based on empirical observations of flow and capacity. This is done with a Flow-Capacity Interaction (FC) matrix that captures systematic effect of *congestible capacities*, a phenomenon in MoD systems where capacities are affected by flows. The FC matrix encapsulates the operation and demand patterns by capturing the empirical equilibrium relationship between flows and capacities. An operation-agnostic logit-based stochastic user equilibrium (SUE) formulation is proposed and proof of equivalence of the SUE formulation is derived. The proof shows that, unlike static capacities, path delays are not just the sum of the Lagrange multipliers of the links on the paths, but dependent on the whole network. We name this phenomenon as the "*non-separable link delays*". A solution algorithm that finds SUE with a bounded path set is proposed, with a custom Frank-Wolfe algorithm to solve the non-linear SUE formulation. Since the FC matrix cannot be directly observed, an inverse optimization problem is introduced to estimate it with observed flow and capacity data. Two numerical examples are provided with sensitivity tests. An empirical example with yellow taxi data of downtown Manhattan, NY is provided to demonstrate effectiveness of estimating the FC matrix from real data, and for determining the equilibrium that captures the underlying flow-capacity dynamics.

Keywords: Mobility-on-Demand**,** Stochastic User equilibrium, Congestible Capacities, multimodal traffic assignment, Inverse optimization

1 Background

Mobility-on-demand (MoD) refers to transportation systems that provide customers with on-demand, point-to-point mobility services, including bikeshare, micro-mobility, carshare, ride-hailing, ridesharing, ridepool/microtransit, carpools, among others, typically using digital technologies for booking, dispatching, and tracking. Different from traditional transit service, these systems are not based on fixed lines and schedules. Operational planning decisions include service region coverage and fleet distribution catering to individual or shared trips. Supported by Information and Communication Technologies (ICT), the operation of MoD are more flexible and complicated with real-time information, leading to a large body of literature on operational models of MoD, looking into aspects like rebalancing idle vehicles (Chow and Sayarshad, 2014; Sayarshad and Chow, 2017), updating prices and vehicle routes (Sayarshad and Chow, 2015), and updating vehicle schedules (Allahviranloo and Chow, 2019).

Equilibrium modeling of privately-owned MoD services is vital from a regulatory perspective, but it is difficult without the operation policy knowledge shared by the MoD operators. Equilibrium modeling of MoD in the literature typically require modeling the MoD operational policies. However, operational policies and relevant data are often considered commercially confidential by the MoD operators, making such models nonapplicable to the regulators. For example, when LADOT created guidelines for private mobility companies to share such data with them, Uber sued them in response (Teale, 2020). We focus on equilibrium modeling of multimodal transportation systems that involve MoD operators from a regulatory perspective. It is assumed that the regulator does not have full knowledge of all operators' policies but can observe multimodal flows and steady state capacities.

We define MoD as mobility services whose supply distribution responds to demand and is affected by demand. For example, e-hailing services deploy vehicles to pick-up customers, so customer trip demand impacts the spatial distribution of the vehicle supply. Another example is bike-sharing/car-sharing. A company deploys bikes/vehicles to depots considering expected demand, and customer trips also impact the distribution of bikes/vehicles. Our research subject includes various types of services that exhibit such microscopic supply and demand interaction, including ride-hailing, bike-sharing, car-sharing, microtransit, etc.

Network equilibrium modeling is typically based on the modeling of congestion effect at microscopic (link, node, zone) level. The congestion effect under steady state models is often reflected as static link capacities or link costs as functions of flows. For MoD, with flexible movements of supply (e.g. vehicles cruising, shared bikes rebalancing), congestion effects are more complicated. The earliest MoD equilibrium studies looked at street-hailing taxi services (Yang and Wong, 1998; Wong et al., 2001; Yang et al., 2002; Wong et al., 2008; Yang et al., 2010; Yang and Yang, 2011). In street-hailing, congestion effect are typically reflected by search behavior and frictions, which are modeled by the probability of drivers choosing to meet customers at a location (Yang and Wong, 1998; Wong et al., 2001; Wong et al., 2008), meeting rates of drivers and customers (Yang et al., 2010; Yang and Yang, 2011), and customers' waiting functions (Wong et al., 2001; Yang et al., 2002; Wong et al., 2008).

With the emergence of ICT, the research focus of MoD switched to e-hailing. In e-hailing, congestion effect is even more dependent on the deployment and matching policies of operators. Such policies include centralized dispatching (Ban et al., 2019; Di and Ban, 2019) and matching based on different behavioral and operational assumptions (Xu et al., 2019; Zhang and Khani, 2021; Liu et al., 2021). Customers' wait times are modeled with different forms, typically functions of number of hailing passengers and number of idle vehicles. He and Shen (2015) studied the equilibrium with both street-hailing taxis and e-hailing taxis, modeling intra-zonal e-hailing matching. They assumed that there is little search friction for e-hailing, and modeled that through a large constant number within the Cobb–Douglas type meeting function similar to Yang et al. (2010). Ban et al. (2019) modeled traffic assignment with ride-sourcing with capacity constraints of allocated fleet sizes, assuming a centralized ride-sourcing platform that deploys vehicles to maximizes profit. Di and Ban (2019) proposed an equilibrium framework considering driving a personal vehicle, riding a personal vehicle, and e-hailing, also assuming centralized fleet deployment from the e-hailing platform maximizing profit. Xu et al. (2019) proposed a network equilibrium model that considers the cruising and deadheading trips of ride-sourcing vehicles, including intranode and internode matching. Ke et al. (2020) modeled the demand-supply equilibrium of ride-sourcing platforms with and without car-pooling in an aggregate context without considering network structures. Monopoly (platform maximize revenue) and social optimum solutions determine trip fares and fleet sizes. Liu et al. (2021) proposed an equilibrium of traditional taxis, app-based taxis, and ride-sourcing. They employ the relationship between passengers' and drivers' wait time derived from Yang et al. (2010).

Zhang and Khani (2021) studied a stochastic user equilibrium (SUE) in which ride-sourcing services complement the transit service as an access mode. The equilibrium includes the stochastic mode choice of riders and logit-based zone choice of drivers. Wait time is assumed to be a reciprocal function of average number of available TNC vehicles in the zone.

In addition to ride-sourcing, there are studies looking at the ride-sharing user equilibrium (RUE), which is the equilibrium of solo drivers, ridesharing drivers, and ridesharing passengers. Congestion and inconvenience cost of ridesharing are modeled as functions of link flows (Xu et al., 2015) or shared travel time and distances (Ma et al., 2020; Li et al., 2020). Congestion delay cost can also be modeled with ridesharing capacity constraints (Sun and Szeto, 2021) or meeting rate functions (Chen and Di, 2021; Noruzoliaee and Zou, 2022). Bike-sharing and other micro-mobility studies mostly look at optimal policies from the operator's perspective, including rebalancing, fleet sizing, dock locating (Lin and Yang, 2011; Lin et al., 2013; Chow and Sayarshad, 2015; Frade and Ribeiro, 2015; Park and Sohn, 2017), pricing (Pfrommer et al., 2014, Singla et al., 2015, Haider et al., 2018), and competition between different operators (Jiang and Ouyang, 2022; Zhang et al., 2023), rather than equilibrium modeling from a regulator's perspective.

Assumptions are the basis of model structures and outcomes. Existing equilibrium models of MoD adopt centralized deployment assumption or different matching/behavioral assumptions for operators. Such a modeling paradigm is not agnostic to the operation of more complicated mobility operation forms and more players in the mobility market, especially from a regulator's perspective.

Firstly, in real cases, MoD often shows complicated combinations of centralized deployment and drivers' free cruising rather than dominated by one. The equilibrium outcome is highly dependent on the specific operation policy of the MoD operator, which the regulators cannot know due to data-sharing concerns. This concern is further exacerbated in a multimodal environment.

With the emergence of technological breakthroughs like innovative electric vehicles (EV) charging infrastructure and connected and automated vehicles (CAV), multimodal equilibrium becomes increasingly complicated from an analytical perspective. Assumptions regarding different modes and operation types need to coexist in one model. Such an overlay of assumptions leads to cumbersome models that are less applicable and scalable. For example, with a mixture of charging stations and enroute charging devices, the routing of EV-based mobility services becomes more complicated. For CAV-based mobility services, deployment and driving behavior would be different from operator to operator, leading to further assumptions that are hard to be validated without data sharing.

With the emergence of deep learning, learning-based approaches are applied to user equilibrium modeling, which would potentially resolve the overlay of assumptions problem for complicated mobility systems. For example, Liu et al. (2023) proposed a learning framework with neural networks to capture travelers' path choices and compute user equilibrium flows. Learning-based approaches are more generic and flexible when it comes to multimodal equilibrium with different operation types. However, large amount of data is needed to train the model as a compensation for lack of interpretable model structures (1536 sets of equilibrium flows are used for training in the numerical tests of Liu et al. (2023)). With big data generated by private mobility companies, such methods would be helpful for the MoD companies. However, for a regulator, such methods would not be applicable without such strong data collection capabilities.

With the above challenges from a regulator's perspective, we aim to model multimodal equilibrium with MoD under the following circumstance: 1) without operation information shared by private mobility operators (agnostic to operational assumptions of multiple operators); 2) with small data needs for calibration; 3) with a unified and simple model structure that can generalize to emerging complex operational settings.

We introduce the Flow-Capacity Interaction matrix (FC matrix) to capture the microscopic interaction between demand and supply in MoD systems without explicitly modeling the operation policies that are unknown to a regulator (i.e. "operation-agnostic"). The construction of FC matrix is based on the concept of "*congestible capacity*" from Xu and Chow (2021). Congestible capacity describes the phenomenon where capacity distribution in MoD systems is not static, but is affected by customer flows and rebalancing/matching policies. Xu and Chow (2021) studied a real-time version of congestible capacity. They proposed an offline-online estimation method to capture the real-time impacts that flows have on capacities in a subsequent time interval. This approach is not capturing the equilibrium state but modeling the real-time states. In this study, we model the static network equilibrium state with "congestible capacities" through the proposed forward model. In short, Xu and Chow (2021) only proposed a cost function whereas in this study we propose an entirely new network equilibrium model that makes use of that cost function. We consider the equilibrium state and model congestible capacities in MoD as link capacity constraints, in which the capacities are dependent on flows in the network. The dependency of capacities on flows is described by the FC matrix in a linear form. We introduce a non-linear stochastic user equilibrium (SUE) model with congestible capacity constraints (labeled as the "forward model"). An exact solution algorithm is proposed. A new inverse optimization model (labeled as the "inverse model") is formulated to calibrate the FC matrix with multimodal trip and capacity data.

The FC matrix encapsulates complicated interaction between demand patterns and supply patterns including matching, rebalancing, and drivers' behavior. We acknowledge that the relationship between capacity and network-wide flows is indeed complex, and a linear approximation may not fully capture the dynamics of flow propagation through the network. We have adopted a linear assumption for simplicity, to ensure the uniqueness of solution and an efficient solution algorithm. In some cases linear functions are enough to capture the interaction of components in very complicated systems. The idea of FC matrix is similar to how the technical efficiency matrix in aggregate economic input-output models describe the structure of the technology contributions from empirical observation (Leontief, 1936). In Leontief's IO model, the interaction between different economic industries can be quite complex and nonlinear, but for modeling localized, incremental impacts a linear model is deemed sufficient. This is also the justification for use in linear regression models. In our model, we also keep our analysis to localized, incremental changes. We cannot use the model to predict disruptive system changes, but we can use it to analyze incremental increase/decrease on capacity effects. Moreover, Hazelton and Watling (2004) investigated the stochastic nature of traffic flows using Markov models and derived equilibrium distributions, which provide approximations of network flow covariances based on linear filters. Their approach highlights how approximating the equilibrium distribution of flows through linear approximations can still yield reasonable predictions. An FC matrix is also similar to the neural network used by Liu et al. (2023) in a way that they both encapsulate complicated behavioral patterns with limited interpretability. However, the FC matrix requires much less data to calibrate due the simpler model structure. With inverse optimization approach, only one set of equilibrium flow observation is needed along with a prior set of parameters. Capacity observations are optional. Such characteristic allows the regulators to apply the model with limited trip data.

When congestible capacities are modeled with access/egress links to represent entering and exiting of a MoD service, subnetworks of different modes can be connected to apply the proposed SUE model to multimodal equilibrium modeling. Calibrating the FC matrix to multimodal trip data captures operation strategies of different modes and their impact on each other without knowing their individual policies, much like how Leontief's input-output model captures inter-industry interactions at a macroscopic level without needing to know firms' behaviors.

The rest of the paper is organized as follows. Section 2 introduces the concept of congestible capacities, the FC matrix, SUE model with congestible capacities (the forward model), the solution algorithm, and the estimation method of the FC matrix (the inverse model). Section 3 shows two illustrative numerical examples. Section 4 presents a case study in which the FC matrix is estimated with yellow taxi data from downtown Manhattan, NY to demonstrate the effectiveness and scalability of the forward and inverse models with real data. Section 5 concludes.

2 METHODOLOGY

2.1 Preliminaries: Congestible capacity

In existing MoD equilibrium models, congestion effects are modeled in three ways: meeting functions (Yang et al., 2010; Xue and Zeng, 2019; Liu et al., 2021), wait time functions (Di and Ban, 2019; Xu et al., 2019; Ke et al., 2020), and capacity constraints (Ban et al., 2019). Meeting functions and wait time functions assume continuous congestion effect, while capacity constraints assume fixed search/wait cost before the capacity binds. Binding capacity means demand exceeds supply at a location or in a zone. Lagrange multiplier of capacity constraint represent the extra matching cost when the capacity binds.

We choose capacity constraints to model MoD congestion effects. MoD services can be categorized into 2 types: with fixed stations and without fixed stations. The approach of capacity constraints is more generalizable to diverse multi-operator settings: for MoD with fixed stations such as docked bike-sharing and car-sharing, capacity constraints fit better, since the time of picking up a bike/car does not vary much before all the bikes/cars are taken from a station. For MoD without fixed stations such as ride-hailing and dockless bike-sharing, assuming evenly distributed vehicles, search/wait time remains steady before all vehicles/bikes in the zone are occupied, which aligns with the characteristics of capacity constraints. Furthermore, capacity constraints can also model fixed route transit services which we can also include in our multimodal model.

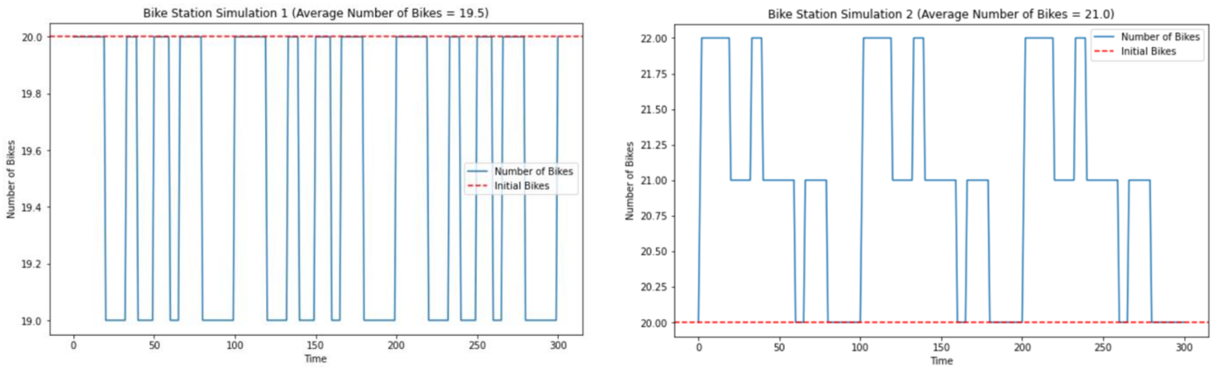
As defined by Xu and Chow (2021), congestible capacity refers to the phenomenon where capacity distribution in a network does not remain static but is affected by flows in the network. MoD generally exhibits such a characteristic due to superposition effects of demand patterns and operation policies, leading to complicated interdependency between flows and capacities. Not all extra capacity dropped by finished trips and rebalancing can be transformed into available capacity. The transformation is dependent on the sequence of customer arrival, departure, and rebalancing (for those familiar with traffic signal capacity modeling, this is similar to how steady state delays are dependent on a host of dynamic factors like arrival platoon behavior or distance between signals).

We use the 5 bike sharing station simulations in **Figure 1** as illustrations. We simulate the number of bikes at a bike sharing station. Initially, there are 20 bikes at the bike sharing station.

There are 3 sequences that impact the number of bike available: user departure that takes away bikes, user arrival that drops off bikes, and rebalancing efforts made by the operator that drop off bikes. We set fixed frequencies for the 3 sequences for all the 5 cases, which are 0.05, 0.03, and 0.02 bikes per unit time, respectively. The distributions of the 3 sequences are different, described as follows:

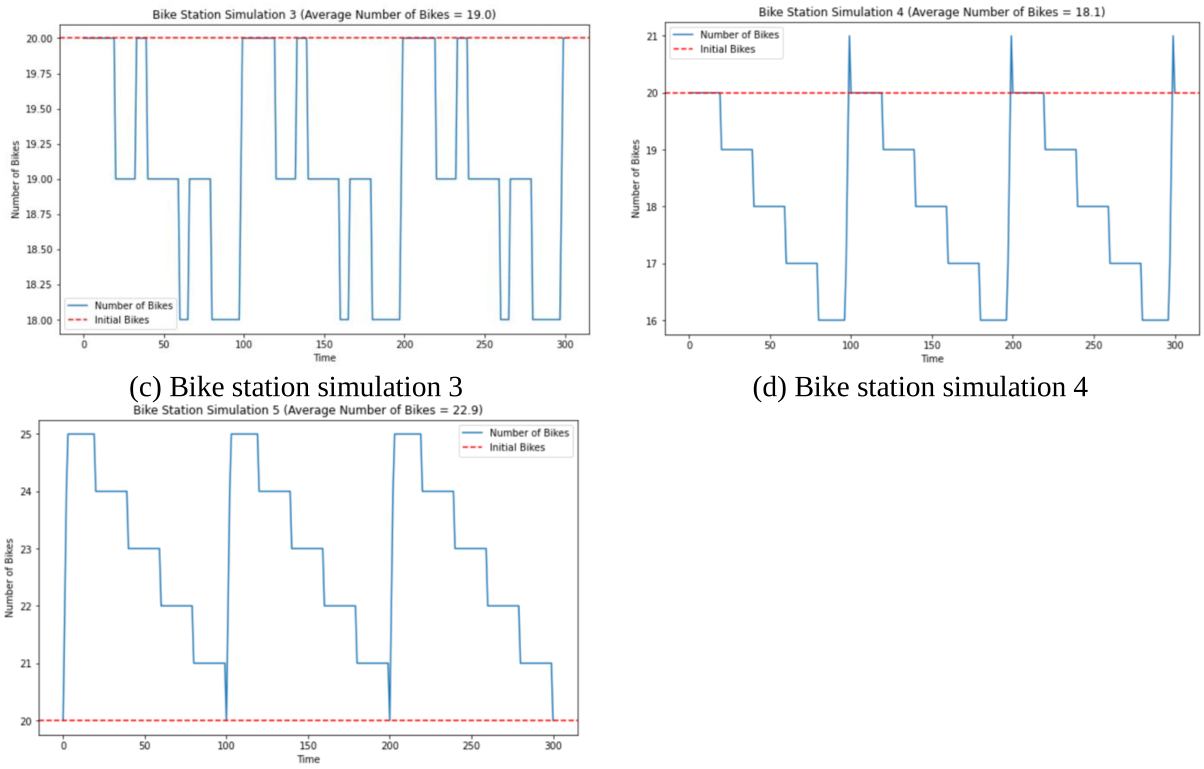
- Bike station simulation 1: evenly distributed user take-aways; evenly distributed user drop-offs; evenly distributed rebalancing drop-offs.
- Bike station simulation 2: evenly distributed user take-aways; evenly distributed user drop-offs; rebalancing drop-offs happen in the first 2 units per 100 units of time (1 bike/unit time).
- Bike station simulation 3: evenly distributed user take-aways; evenly distributed user drop-offs; rebalancing drop-offs happen in the last 2 units per 100 units of time (1 bike/unit time).
- Bike station simulation 4: evenly distributed user take-aways; user drop-offs happen in the last 3 units per 100 units of time (1 bike/time unit); rebalancing drop-offs happen in the last 2 units per 100 units of time (1 bike/unit time).
- Bike station simulation 5: evenly distributed user take-aways; drop-offs happen in the first 3 units per 100 units of time (1 bike/unit time); rebalancing drop-offs happen in the first 2 units per 100 units of time (1 bike/unit time).

Figure 1 shows how the number of bikes available at the bike station changes with time for the 5 cases. Since the total drop-off frequency equals the take-away frequency, the overall number of bikes available at the station should remain 20. However, the average number of bikes available across the time units are: (1)19.5; (2)21.0; (3)19.0; (4)18.1; (5)22.9. The cases show that, even with the same collective frequencies of take-aways and drop-offs, the equilibrium capacities that are experienced by the users can be different. All 5 cases are with evenly distributed user take-aways. The case with the highest average number of bikes is case (5), which has all the drop-offs happening earlier than most of the take-aways, so the dropped-off bikes are more available to the coming users. For case (4), the drop-offs happen largely after the take-aways, making the dropped-off bikes unavailable to the users, leading to lower average number of available bikes. Hence, equilibrium capacity is determined by the interaction between the three sequences. Generally, when more arrivals and input rebalancing happen before the departures, a higher portion are transformed as available capacity.



(a) Bike station simulation 1

(b) Bike station simulation 2



(c) Bike station simulation 3

(d) Bike station simulation 4

(e) Bike station simulation 5

Figure 1. Capacity transformation illustration: Bike station simulation

Except for bike-sharing, a broad range of mobility services which we define as MoD also exhibit similar capacity effects. The capacity dynamics of MoD services are all affected by the following 3 processes, similar to the bike sharing example in **Figure 1**:

- Process 1: capacity decrease due to users taking away vehicles/bikes/scooters, etc.;
- Process 2: capacity increase due to users dropping off vehicles/bikes/scooters, etc.;
- Process 3: capacity change due to operation policies such as rebalancing/interzonal matching.

The different characteristics of these different services would only lead to different distributions of the three processes and different interactions between them. Some other examples are as follows.

- Ride-sourcing: Capacity of a zone is the number of drivers available in the zone.
  - Process 1: When a driver is occupied by a customer, one unit of capacity is taken away from the zone.
  - Process 2: When a driver drops off a customer at her destination, this driver adds to the capacity of the destination zone.
  - Process 3: When a driver at zone A is matched with a customer at zone B, this driver becomes one unit of capacity of zone B.
- Dockless bike/scooter sharing: Capacity of a zone is the number of bikes/scooters available in the zone.
  - Process 1: When a bike/scooter is taken by a customer, one unit of capacity is taken away from the zone.
  - Process 2: When a customer drops off a bike/scooter in a zone, this adds to the capacity of the zone.

    - Process 3: When the operator takes bikes/scooters from zone A to zone B, the capacity of zone A is decreased, and capacity of zone B is increased.
- CAV-based ride-sourcing: Capacity of a zone is the number of CAVs available in the zone.
    - Process 1: When a CAV is matched with a customer in a zone, one unit of capacity is taken away from the zone.
    - Process 2: When a CAV drops off a customer in a zone, this adds to the capacity of the zone.
    - Process 3: When a CAV at zone A is matched with a customer at zone B, this CAV becomes one unit of capacity of zone B.

In this study, we aim to model the steady states of such dynamic capacities. As a result of the interaction between departures, arrivals, and rebalance/interzonal matching, congestible capacity at equilibrium is similar to the average capacity that we calculate for the cases in **Figure 1**: the capacity experienced by the users at a static equilibrium state. Two elements should be captured while modeling congestible capacities: demand patterns and operation policies. Demand patterns includes customer departure and arrival flows in the MoD network. The impact of demand patterns on equilibrium capacities can be captured by modeling flow-capacity relationships. Operation policies that cause capacity movements are designed to respond to customer flows, which are dependent on customer flows. Hence, operation policies can also be implicitly captured by modeling flow-capacity relationships. Note that without observable operation policy knowledge, we model only customer flows which could be observed through trip data.

As a result, we model congestible capacities at equilibrium as a relationship between customer flows and capacities. We define the equilibrium capacity of a MoD zone as a linear combination of customer link flows in the network, with coefficients representing the observed, aggregated impact of all link flows in the network on the capacity, leading to the construction of the FC matrix. Details are discussed in Section 2.2.

2.2 Problem description

Assume that the service region of a MoD provider can be divided into a number of zones. Each zone is represented by a centroid node. Neighboring nodes are connected by links with link cost as the MoD service cost between the zones represented by the nodes (note that the links are not street segments but OD-level paths between zone centroids). The cost of entering and exiting a MoD service is represented by access/egress links connecting the walking modal network with the MoD network. In a multimodal setting, each mode or service is represented by a subgraph and are all connected with the walking network with access/egress links at the access/egress locations (e.g. MoD pick-up/drop-off points, MoD zone centroids if no specific pick-up/drop-off points, bike-sharing stations).

We define the above multimodal network as a directed graph $G(V,E)$. Each link i has a capacity s_i and a constant undersaturated travel cost t_i when capacity is not binding. The link travel cost t_i is the generalized sum of all the costs for a customer to traverse the link transformed into the same units (e.g. $). For links within a mode/service, link cost includes travel time cost, fare cost (if applicable), comfort cost etc. For access links, link cost includes the uncongested time cost and comfort cost of accessing a mode (e.g. time of unlocking a shared bike, average ride-hailing wait time in a zone when demand is smaller than supply). For egress links, link cost includes the uncongested time cost and comfort cost of egressing a mode (e.g. time of locking up a bike, time of returning a car). All elements of link costs are converted to a common unit (e.g. $).

We assume s_i as infinity (uncapacitated) for walking links and links connecting MoD nodes. We do not consider crowding congestion in the in-vehicle links of MoD nor the contribution of MoD vehicles to the background traffic congestion. We consider s_i as finite for MoD access links and some MoD egress links. Congestible capacity effect generally exists on the MoD access/egress links. The capacity of MoD access links represents the amount of available supply at a location, which is affected by customer flows and operation. The capacities of MoD egress links are typically infinity (e.g. there is no restriction in getting off a vehicle), but in some cases they also exhibit congestible capacity effects (e.g. limited docks might cause extra waiting while dropping off a bike, and the number of available docks at a station is affected by customer flows and operation).

Travel demand is deterministic and is classified by a set of origin-destination (OD) pairs W starting and ending from the walking network. We denote the travel demand of OD pair $w \in W$ as r_w. Dummy links can be setup between OD pairs to capture elastic demand that does not participate in a trip. Trips between OD pairs are made along paths $j \in K$ formed by sequences of links $i \in E_j$, where E_j denotes the set of all the links on path j. The path set connecting OD pair $w \in W$ is denoted as K_w. Link flow of links $i \in E_j$ is denoted as v_i. Path flow of path $j \in K$ is denoted as h_j. When link flows v_i reach the link capacity, non-zero Lagrange multipliers m_i may manifest in these links, resulting in path delays d_j. The uncongested path cost of path $j \in K$ is denoted as T_j, and congested path cost is denoted as $c_j = T_j + d_j$.

We model congestible capacities as functions of link flows, $s_i = f(v)$, where the function characterizes an equilibrium state structure. Xu and Chow (2021) used a linear combination to update congestible capacities in real-time cases, which is a linear function of inbound and outbound flows of the zone and the capacity of the last time interval. More generally, we assume that the flow of all links in the network may contribute to the capacity of each link within the network to varying degrees. For example, in ride-sourcing systems, the number of available vehicles of a zone may not just be affected by the inbound and outbound flows of the zone, they may impact neighboring zones via deadhead cruising.

We represent the capacity of a link $i \in E$ as the sum of an exogenous capacity s_{i0} (e.g. initial allocation of vehicles at the start of a period) and the contribution from all the links within the network. The flow contributions to the capacity of a link is modeled with linear functions. The congestible capacity s_i of a link $i \in E$ at equilibrium is modeled as **Eq. (1)**, where p_{ik} is a "contribution efficiency" parameter (referred to as efficiency in this study), representing how much the flow on link $k \in E$ contributes to the capacity of link $i \in E$, and n is the number of links in the network. Not every link has contribution to the capacities of every other link, which means there would be zero efficiencies. Specification of the efficiencies is dependent on the network structure and intermodal interactions. Examples are given in Section 3.

$$s_i = p_{i1}v_1 + p_{i2}v_2 + \cdots + p_{in}v_n + s_{i,0} \tag{1}$$

If we write **Eq. (1)** for all links $i \in E$ as a system of equations in matrix form, we have **Eq. (2)**. Vector $\boldsymbol{S}$ represents the link capacities. Vector $\boldsymbol{v}$ represents the link flows. Matrix $\boldsymbol{F}$ and vector $\boldsymbol{s_0}$ are defined in **Eq. (3)** and **Eq. (4)**. Matrix $\boldsymbol{F}$ is the Flow-Capacity Interaction matrix (FC matrix). FC matrix $\boldsymbol{F}$ is composed of efficiencies $p_{ik}, i \in E, k \in E$, which describes the systematic pattern of how equilibrium capacities depend on equilibrium flows. The vector $\boldsymbol{s_0}$ shown in **Eq. (3)** is the vector of exogenous capacities s_{i0}, $i \in E$. For an uncapacitated link i, $s_i = s_{i,0} = \infty$.

$$\boldsymbol{S} = \boldsymbol{F}\boldsymbol{v} + \boldsymbol{s_0} \tag{2}$$

$$\boldsymbol{F} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1n} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2n} \\ p_{31} & p_{32} & p_{33} & \cdots & p_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & p_{n3} & \cdots & p_{nn} \end{bmatrix} \tag{3}$$

$$\boldsymbol{s_0} = \begin{bmatrix} s_{1,0} \\ s_{2,0} \\ \vdots \\ s_{n,0} \end{bmatrix} \tag{4}$$

The FC matrix $\boldsymbol{F}$ can be estimated with trip and capacity data to capture the combined impact of observed demand patterns and unobserved operation policies on capacities. For example, a MoD system with low support for real-time information and poor matching algorithms might see higher absolute values of efficiencies than one with better information and matching, since less rebalancing generally leads to more imbalanced capacity distribution. When all the efficiencies are 0, it indicates that the capacity distribution remains constant at equilibrium, meaning that there is no congestible capacity effect in the system, for which one example is fixed-route transit services (which can thus be included in our multimodal models). If flow on link k takes away the capacity of link i, $p_{ik} = -1$ represents perfect efficiency in transferring that flow on link k to the capacity loss of link i. If link k contributes to the capacity of link i, $p_{ik} = 1$ represents perfect efficiency in transferring the flow on link k to the capacity gain of link i. As shown by **Figure 1**, The imperfect efficiencies are caused by the sequences of pick-ups, drop-offs, and rebalancing and inter-zone matching. Generally, more rebalancing or inter-zone matching leads to p_{ik} closer to 0, otherwise closer to -1 and 1 depending on the relationship of taking and dropping.

As mentioned in section 1, we aim to incorporate a broad range of mobility services which we define as MoD. All these different types of on-demand mobility systems that exhibit the effect of congestible capacities have distinct operational characteristics and policies. However, all these systems that similar dynamics of link/node capacities. Different operation characteristics leads to different distributions of the 3 sequences and different interactions between them. For example, for an EV-based car-sharing service, nodes and links closer to a charging hub might receive more efficient rebalancing. Comparing ride-hailing and bikeshare, capacity of a bikeshare station is only affected by the drop-offs and rebalance at the station, while the capacity of a node in a ride-hailing network would be impacted by drop-offs at the node and drop-offs at other nodes in a buffer to different extents. All these different effects would be captured as different coefficients in the FC matrix, to reflect different response strategies of demand.

Section 2.3 presents the forward model of a SUE model given a pre-estimated FC matrix. Section 2.5 presents the inverse model in which the FC matrix is estimated from multimodal trip data and capacity data.

2.3 Forward Model: Formulation of Stochastic User Equilibrium with Congestible Capacities

We define a path-based SUE similar to Bell (1995): a SUE is reached when the assignment of demand among alternative paths conforms to the logit model shown in **Eq. (5)**. There also exist other logit-based equilibrim models that account for path overlap, such as nested logit (Beckhor and Prashker, 2001), dogit (Chu, 2012), and c-logit (Zhou et al. 2021). For simplicity, we adopted the multinomial logit form to focus on the change from the standard SUE model, and subsequent extensions can be adopted to capture more realistic route choice behaviors.

$$ln\left(\frac{h_j}{h_{j'}}\right) = -\alpha\left(c_j - c_{j'}\right) = -\alpha\left(T_j + d_j - T_{j\prime} - d_{j\prime}\right) \qquad (5)$$

where $j \in K_w$ and $j' \in K_w$ are alternative paths connecting the same OD pair $w \in W$ and $\alpha > 0$ is a given parameter.

We have the path-based SUE with congestible capacity formulation (P_1) in **Eqs. (6) – (9)**, in which $\boldsymbol{A}$ is the link-path incidence matrix and $\boldsymbol{B}$ is the OD-path incidence matrix.

$$P_1 : \text{minimize } \Phi = \sum\nolimits_{j \in K} h_j\left(\ln h_j - 1\right) + \alpha \boldsymbol{t}^{\boldsymbol{T}} \boldsymbol{A} \boldsymbol{h} \qquad (6)$$

$$s.t.$$

$$\boldsymbol{r} = \boldsymbol{B}\boldsymbol{h} \qquad (7)$$

$$\boldsymbol{F}\boldsymbol{A}\boldsymbol{h} + \boldsymbol{s}_{\boldsymbol{0}} \geq \boldsymbol{A}\boldsymbol{h} \qquad (8)$$

$$\boldsymbol{h} \geq \boldsymbol{0} \qquad (9)$$

The objective **Eq. (6)** is composed of the flow-spreading term and the system total uncongested travel cost. **Eq. (7)** is the demand constraint. **Eq. (8**) is the congestible capacity constraint. **Eq. (9)** is the path flow non-negativity constraint. The proposed model is a generalization in which SUE with static capacities is a special case as noted in **Lemma 1** (proof is trivial).

Lemma 1. As $\boldsymbol{F} \rightarrow \boldsymbol{0}$, P_1 becomes the SUE formulation with static capacities, where $\boldsymbol{s}_{\boldsymbol{0}}$ becomes the vector of static capacities.

We define the Lagrange multiplier of the congestible capacity constraint (**Eq. (8**)) of link i as m_i. The complementary slackness condition should hold by definition; at SUE, if $m_i > 0$, $v_i = s_i$; if $v_i < s_i$ then $m_i = 0$. The Lagrangian is shown in **Eq. (10)**, where $\boldsymbol{v} = \boldsymbol{A}\boldsymbol{h}$.

$$\boldsymbol{L} = \sum\nolimits_{j \in K} h_j\left(\ln h_j - 1\right) + \alpha \boldsymbol{t}^{\boldsymbol{T}} \boldsymbol{v} + \boldsymbol{l}^{\boldsymbol{T}}(\boldsymbol{r} - \boldsymbol{B}\boldsymbol{h}) + \boldsymbol{m}^{\boldsymbol{T}}\left((\boldsymbol{F} - \boldsymbol{I})\boldsymbol{A}\boldsymbol{h} + \boldsymbol{s}_{\boldsymbol{0}}\right) \qquad (10)$$

If we denote $\boldsymbol{G} = (\boldsymbol{F} - \boldsymbol{I})\boldsymbol{A}$, and $\boldsymbol{g}_{\boldsymbol{j}}$ is the *j*th column of $\boldsymbol{G}$, the KKT condition is shown in **Eq. (11)**.

$$\frac{\partial \boldsymbol{L}}{\partial h_j} = \ln h_j + \alpha \boldsymbol{t}^{\boldsymbol{T}} \boldsymbol{a}_{\boldsymbol{j}} - \boldsymbol{l}^{\boldsymbol{T}} \boldsymbol{b}_{\boldsymbol{j}} + \boldsymbol{m}^{\boldsymbol{T}} \boldsymbol{g}_{\boldsymbol{j}} = \boldsymbol{0} \qquad (11)$$

At SUE, $h_j > 0$ should be satisfied for all paths $j \in K$. **Eq. (11)** ensures the result with the log term. **Eq. (11)** can be reduced to **Eq. (12)**.

$$ln\, h_j = -\alpha T_j + \boldsymbol{m}^{\boldsymbol{T}} \boldsymbol{g}_{\boldsymbol{j}} + l_w \qquad (12)$$

where $w \in W$ is the OD pair connected by path $j \in K_w$, l_w is the corresponding Lagrange multiplier, $\boldsymbol{m}$ is the vector of Lagrange multipliers of the congestible capacity constraints. **Eq. (12)** leads to the logit path flows shown in **Eq. (13)**.

$$h_j \;=\; r_k \frac{\exp(-\alpha T_j + \boldsymbol{m^T g_j})}{\sum_{j' \in K_k} \exp(-\alpha T_{j'} + \boldsymbol{m^T g_{j'}})} \tag{13}$$

If $\boldsymbol{m^T g_j} = -\alpha d_j$ holds, P_1 yields SUE. **Proposition 1** proves that, for any path $j \in K$, $\boldsymbol{m^T g_j} = -\alpha d_j$ is necessary and sufficient for P_1 to yield SUE.

Proposition 1 (SUE with congestible capacities): *P_1 yields a SUE assignment with congestible capacities if and only if the Lagrange multipliers* $\mathbf{m}$ *associated with the following constraints $\boldsymbol{FAh} + \boldsymbol{s_0} \geq \boldsymbol{Ah}$ satisfies $\boldsymbol{m^T g_j} = -\alpha d_j, \forall w \in W$, for all paths $j \in K_w$, where d_j is the SUE path delay.*

Proof: We follow a similar logic to Bell (1995). At SUE, we rank all the paths $j \in K_w, w \in W$, according to the value of the function shown in **Eq. (14)**.

$$f(j) \;= ln\, h_j + \;\alpha T_j, \qquad j \in K_w, w \in W \tag{14}$$

(i)

Suppose $f(j)$ is not the greatest among the paths $j \in K_w$. We assume the bottleneck of path $j \in K$ is link $i \in E_j$. In this case, a small increase in the capacity of link i would allow the objective function to be reduced by shifting some trips from another path $j' \in K_w$ to path j. To minimize the objective function, the shift would be made from the path offering the largest reduction in the objective. Assume that a small shift of Δh trips is made from path j' to path j. Ignoring terms in Δh^2 and higher powers of Δh, $\ln(h + \Delta h) \approx \ln h + \frac{\Delta h}{h}$, so the change in objective can be given by **Eq. (15)**. Obviously, $f(j')$ is the greatest among the paths $j \in K_w$ (Bell, 1995).

$$dL \;= \Delta h \left(ln\, h_j - ln\, h_{j'}\right) + \Delta h\, \alpha(T_j - T_{j'}) \tag{15}$$

This is where the proof differs from Bell (1995). Different from networks with static capacities, the switch causes changes in link capacities across the whole network.

Consider that Δh trips are added to path j, we denote the capacity of link $q \in E$ before adding Δh as s_q, and the capacity after as ${s_q}'$. Hence, the capacity change is $\Delta s_q = {s_q}' - s_q$.

For any link $q \in E \backslash E_j$, Δs_q can be given by **Eq. (16)**. Since v_q does not change, the change in capacity resource caused by adding Δh is Δs_q.

$$\Delta s_q \;= \sum_{i \in E_j} p_{qi}\, h_j - \sum_{i \in E_j} p_{qi}\,(h_j + \Delta h) = \sum_{i \in E_j} p_{qi}\, \Delta h \tag{16}$$

For any link $q \in E_j$, the capacity constraint before adding the trips can be written as **Eq. (17)**.

$$v_q \leq s_q \tag{17}$$

The capacity constraint after adding the trips becomes **Eq. (18)**. To compute the resource change of the constraint before and after adding the trips, the left-hand side of the constraint should be the same. The conversion is shown in **Eq. (18)**.

$$v_q + \Delta h \leq s_q{}' \rightarrow v_q \leq s_q{}' - \Delta h \tag{18}$$

Hence, change in the capacity resource of link q on path j is $(s_q{}' - \Delta h) - s_q$, which equals to $(\sum_{i \in E_j} p_{qi} - 1)\Delta h$, according to **Eq. (16)**.

Therefore, the change in objective can be written as **Eq. (19)**.

$$dL' = \sum_{q \notin E_j} \sum_{i \in E_j} p_{qi} \, \Delta h \, m_q + \sum_{q \in E_j} \left(\sum_{i \in E_j} p_{qi} - 1 \right) \Delta h \, m_q \tag{19}$$

Let us denote the element in qth row and jth column of $\boldsymbol{G} = (\boldsymbol{F} - \boldsymbol{I})\boldsymbol{A}$ as g_{qj}**.** For link $q \in E_j$, $g_{qj} = \sum_{i \in E_j} p_{qi} - 1$. For link $q \notin E_j$, $g_{qj} = \sum_{i \in E_j} p_{qi}$. Hence, **Eq. (19)** can be written as **Eq. (20)**.

$$dL' = \boldsymbol{m^T g_j} \, \Delta h \tag{20}$$

Change in the objective caused by a switch of Δh from path *j'* to path *j* is then written as **Eq. (21)**.

$$dL = (\boldsymbol{m^T g_j} - \boldsymbol{m^T g_{j'}})\Delta h \tag{21}$$

Combining **Eq. (15)** and **Eq. (21)**, we have **Eq. (22)**.

$$\boldsymbol{m^T g_j} - \boldsymbol{m^T g_{j'}} = \ ln\left(\frac{h_j}{h_{j'}}\right) + \alpha\left(T_j - T_{j'}\right) \tag{22}$$

According to SUE, trips are allocated to paths according to the logit model shown in **Eq. (5)**. Hence, we have **Eq. (23)** for SUE with congestible capacities, where $j' \in K_w$ is the path that maximizes $f(j) = ln(h_j) + \alpha T_j$ among all $j \in K_w$.

$$ln\left(\frac{h_j}{h_{j'}}\right) = -\alpha(T_j + d_j - T_{j'} - d_{j'}) \tag{23}$$

Combining **Eq. (22)** and **Eq. (23)**, we end up with **Eq. (24)**.

$$\boldsymbol{m^T g_j} - \boldsymbol{m^T g_{j'}} = -\alpha(d_j - d_{j'}) \tag{24}$$

Eq. (24) can be re-written as **Eq. (25)**, where a_w is an OD-specific constant.

$$\boldsymbol{m}^{\boldsymbol{T}}\boldsymbol{g}_{\boldsymbol{j}} = -\alpha d_j + a_w,\ \ j \in K_w, w \in W \tag{25}$$

With a logit model as the path choice model of P_1, path flows h_j are dependent on relative costs between the paths in K_w that connect the same OD pair $w \in W$. This means that only the differences between path costs $(T_j + d_j)$ impact path flows. Hence, the OD-specific constant a_w can be omitted to arrive at **Eq. (26)**.

$$\boldsymbol{m}^{\boldsymbol{T}}\boldsymbol{g}_{\boldsymbol{j}} = -\alpha d_j,\ \ j \in K_w, w \in W \tag{26}$$

(ii)
Alternatively, suppose $f(j)$ the greatest among the paths $j \in K_w$. In that case, path j is path j*'* in (i), and **Eqs. (24)**, **(25)** and **(26)** obviously hold. ■

As shown in **Eq. (25)**, unlike static capacities, path delays are not just the sum of the Lagrange multipliers of the links on the paths, making it impossible for us to identify unique link delays. This point shows the core characteristic of the concept "congestible capacities": the distribution of capacities in the network is affected by the flows, leading to a single path delay dependent on the whole network. We call this phenomenon "non-separable link delays". This is a unique phenomenon that we observe from such systems. As for solution uniqueness, the proof from Bell (1995) remains applicable.

Lemma 2. *The Lagrange multipliers* $\boldsymbol{m}^*$, $\boldsymbol{l}^*$ *and path flows* $\boldsymbol{h}^*$ *at SUE are unique if and only if all the constraints that bind are linearly independent.*

Through the proof of Bell (1995) we only know that the Lagrange multipliers are unique if and only if all the constraints that bind are linearly independent, as opposed to having unique link delays. **Lemma 2** does not assure link delays are unique. However, with unique Lagrange multipliers we can get unique path delays. Unique path delays then lead to a unique SUE flow assignment.

2.4 Proposed solution algorithm with $\boldsymbol{\rho}$-bounded shortest path generation

At the SUE point solved from P_1, all the paths $j \in K$ are assigned non-zero flows. A reasonable path set needs to be determined. Like Bell (1995), we first considered iterative balancing with column generation. However, iterative balancing is not applicable with non-separable link delays. It leads to a nonconvex constraint space where intermediate iterative balancing solutions may get stuck. Column generation is not applicable either, since unique link delays cannot be found through solving P_1 due to **Lemma 2**. The non-separable link delays also make more conventional methods like the Method of Successive Averages (MSA) hard to be applied. As shown by Proposition 1, path delays are not just determined by the links along the paths, but all the links in the network. Such a property makes it very complicated to assign penalties, since path delays are not just determined by the links along the paths, but all the links in the network. It is hard to assign separate penalties to links to reflect the non-separable links delays across the iterations. Hence, we consider other choice set generation approaches.

Methods from the literature include choice set pre-generation, bounded rational models, and endogenous choice set restriction to determine the path set for SUE. Choice set pre-generation refers to the method of identifying a path set before running the SUE algorithm, and iteratively improve the path set by adjusting link costs according to the SUE solutions (Prato and Bekhor, 2006; Bovy, 2009). The disadvantage is that the travel cost inputs for choice set generation are inconsistent over iterations (Watling et al., 2018). The bounded rational models give a space of flow solutions, assuming that travelers are indifferent to path cost differences within an indifference band (Mahmassani and Chang, 1987; Lou et al., 2010; Di et al., 2013; Di and Liu, 2016). However, this method does not give a point solution or a probability distribution of the solution space. The endogenous choice set restriction method determines a path set by adding constraints to the SUE formulations (Pel and Chaniotakis, 2017; Rasmussen et al., 2015), but the existence of SUE solution cannot be guaranteed. This shortcoming is solved by Watling et al. (2018) through a Bounded Choice Model, which integrates a path utility bound to determine the equilibrium path set, as well as conditions to guarantee SUE existence.

With P_1, we use the choice set pre-generation method for path set determination. We use the k-shortest path generation algorithm from Yen (1971) for convenience which is modified to avoid cycles by checking repeated nodes, combined with a bounding ratio ρ to limit the number of paths generated. The bounding ratio is defined as the upper bound of the cost ratio of the paths with the largest and smallest path costs within a path set of an OD pair. With computational consideration, we want to limit the number of paths for large-scale real network and eliminate the longer paths that are less realistic. The bounding ratio ρ implicitly limits the number of paths and removes long paths. In contrast to k-shortest path algorithm, it leads to desirable properties that the number of paths will be OD-specific (e.g. distant OD pairs have more paths than close OD pairs) and context-specific (e.g. rise of cost leads to rise of path cost upper bound and number of paths). The ρ parameter could be set according to the specific case and modeling requirement. The ρ-bounded shortest path generation also ignores the impact of path overlap, which could be addressed by embedding other more complicated path generation in future research. For example, incorporating Bounded Path Size (BPS) route choice models like the bounded path size SUE model proposed by Duncan et al. (2024) is one possibility. The Bounded Choice Model with Local Detour Threshold (BCM-LDT) proposed by Rasmussen et al. (2024) could also be considered to incorporate local detouredness. Whether the routes are realistic is dependent on the ρ value. There might be some unrealistic routes assigned non-zero flows when ρ is set too large. However, the actual parameter that controls the level of stochasticity is the parameter α in the objective function. When α is calibrated properly, unrealistic path in terms of length will be assigned positive flows that are very close to zero. In real applications, ρ could be applied to ignore such paths. Other types of unrealistic paths, which are not in terms of length, could be dealt with through having link costs that represent an overall cost to include other aspects such as comfort level. In a multimodal or MaaS setting, path finding algorithms with time-dependent, label-constraining, or disjointness features could also be considered (Sherali et al, 1998; Sherali et al, 2003).

In the solution algorithm we propose, for all OD pairs, a path set that satisfies the bounding ratio is generated using the k-shortest path algorithm to form an $\boldsymbol{A}$ matrix. Paths connecting OD pair $w \in W$ are generated starting from the shortest path until the path cost of the latest generated path exceeds ρ times the shortest path cost. In the next step, we solve P_1 with the bounded path set to obtain a SUE with the $\boldsymbol{A}$ matrix. The proposed algorithm is shown in **Algorithm 1**. **Proposition 2** proves that when $\rho \rightarrow \infty$, Step 2 of **Algorithm 1** gives a full path set which leads to a SUE with congestible capacities.

Algorithm 1. Solution algorithm with ρ-bounded shortest path generation

```
Step 1 (Initialization): Initialize empty path sets for all OD pairs.
Step 2 (Path set generation)
 Repeat the following for each OD pair $w \in W$:
 -  Generate the next shortest path using undersaturated link costs with the k-shortest path
    generation method. If no more paths could be generated, move on to the next OD pair $w + 1$.
 -  If the generated path is the first path of OD pair $w$ whose undersaturated cost is $T_{w0}$, add the path
    to the path set of OD pair $w$ and move on to the next OD pair $w + 1$. If the generated path is not
    the first path of OD pair $w$, check if its cost is greater than $\rho T_{w0}$. If so, move on to the next OD
    pair $w + 1$; if not, add the path to the path set of OD pair $w$.
Step 3 (SUE finding)
 -  Form $\boldsymbol{A}$ and $\boldsymbol{B}$ matrices based on the path sets generated from Step 2.
 -  Solve $P_1$ with $\boldsymbol{A}$ and $\boldsymbol{B}$ using Algorithm 2 to obtain $\boldsymbol{h}, \boldsymbol{m}$.
Step 4 (Output SUE flows and capacities)
 -  Equilibrium path flows: $\boldsymbol{h}$
 -  Equilibrium link flows: $\boldsymbol{v} = \boldsymbol{Ah}$
 -  Equilibrium capacities: $\boldsymbol{s} = \boldsymbol{FAh} + \boldsymbol{s_0}$
```

Proposition 2 (Algorithm convergence): ***Algorithm 1*** *converges to a SUE assignment with congestible capacities with a complete path set, i.e. a solution to P1, when ρ → ∞.*

Proof: When ρ → ∞, the upper bound of path cost of OD pair w ∈ W is $\rho T_{w0} \to \infty$. All the paths connect OD pair w ∈ W will be enumerated. With a full path set, solving P_1 gives SUE assignment with congestible capacities. ■

As discussed before, P_1 cannot be solved by Bell's iterative balancing (Bell, 1995) or MSA. In Step 3 of **Algorithm 1**, P_1 is solved through a Frank-Wolfe algorithm (Frank and Wolfe, 1956) (not implemented in the same traffic assignment-based approach from LeBlanc et al., 1975) as shown in **Algorithm 2**. Since the objective of P_1 is convex and constraints are linear, Frank-Wolfe algorithm is applicable. As shown in **Algorithm 2**, while solving for the step size X, the equation $\frac{\partial \Phi(\boldsymbol{h^{n-1}} + X(\boldsymbol{y}^* - \boldsymbol{h^{n-1}}))}{\partial X} = 0$ has a log term. We use the bisection method to solve for X, which stops when the sections are smaller than a threshold ε.

Algorithm 2: Frank-Wolfe algorithm applied to solve $\boldsymbol{P_1}$

```
Start with an initial feasible guess $\boldsymbol{h^0}$, $n = 1$, $X = \infty$.
$count = 0$.
While $X \le \epsilon$ and $count < N$ do
    -  Solve $\min_{\boldsymbol{y}} \sum_{j \in K} \frac{\partial \Phi}{\partial h_j}(h_j^n) y_j$, subject to $\boldsymbol{FAy} + \boldsymbol{s_0} \ge \boldsymbol{Ay}$, $\boldsymbol{r} = \boldsymbol{By}, \boldsymbol{y} \ge \boldsymbol{0}$, to obtain $\boldsymbol{y}^*$.
    -  Solve $\min_{X} \Phi(\boldsymbol{h^{n-1}} + X(\boldsymbol{y}^* - \boldsymbol{h^{n-1}}))$ by solving $\frac{\partial \Phi(\boldsymbol{h^{n-1}} + X(\boldsymbol{y}^* - \boldsymbol{h^{n-1}}))}{\partial X} = 0$ , which can be
       simplified to $\sum_{j \in K}(\log(h_j^{n-1} + X(y_j^* - h_j^{n-1})) + \alpha \boldsymbol{t^T a_j})(y_j^* - h_j^{n-1}) = 0$ with bisection
       method (stops when sections smaller than $\varepsilon$).
    -  Update $h$: $\boldsymbol{h^n} = \boldsymbol{h^{n-1}} + X(\boldsymbol{y}^* - \boldsymbol{h^{n-1}})$.
    -  If $X \le \epsilon$, $count = count + 1$, else, $count = 0$.
Return $\boldsymbol{h^n}$.
```

2.5 Inverse model: estimation of the Flow-Capacity Interaction Matrix

In application, FC matrices need to be estimated for equilibrium modeling. We propose an estimation method of the FC matrix given observed equilibrium path flows, equilibrium capacities, and initial capacities by formulating the inverse optimization problem of the forward model (P_1), solving for $\boldsymbol{F}$ with the parameter α preset. An inverse optimization model takes an optimization problem with observed decision variables and a prior set of parameters and solves for the parameters such that the observed variables are optimal. Inverse optimization has a rich literature starting from Burton and Toint (1992) and Ahuja and Orlin (2001) with inverse shortest path problems. Reviews are provided in Xu et al. (2018) and Chan et al. (2023). No such inverse problem has been formulated for SUE models, much less for the more generalized proposed model.

We denote the observed flow of path $j \in K$ as $h_{O,j}$, vector of observed path flows as $\boldsymbol{h_O}$, observed equilibrium capacity of link $i \in E$ as $s_{O,i}$, initial capacity of link $i \in E$ as $s_{0,i}$. The FC matrix is estimated with an initial prior denoted as $\boldsymbol{F}^0$, in which the element at the ith row and kth column is denoted as p_{ik}^0. The choice of prior $\boldsymbol{F}^0$ may affect the estimation results, and additional data can help improve estimation with more robust and repeated observations of the prior (see traffic monitoring application in Xu et al., 2018, 2021). A default choice of $\boldsymbol{F}^0$ can be a $n \times n$ all-zero matrix.

Decision variables of the inverse optimization problem include the perturbations of selected FC matrix elements, and the Lagrange multipliers of the constraints in P_1: $\boldsymbol{m}, \boldsymbol{l}$. Perturbations include positive and negative perturbations of selected elements of the FC matrix. The set of selected elements are defined as $(i,k) \in Z$. Selection of the elements depends on the situation, considering network structure, rebalancing policies of the service, and other assumptions, which will be illustrated in the examples of section 3. Positive perturbation of the element at the ith row and kth column of the FC matrix is denoted as p_{ik}^+, while negative perturbation of the element is denoted as p_{ik}^-. Both p_{ik}^+ and p_{ik}^- are non-negative. We denote the perturbed FC matrix as $\boldsymbol{F}^\pm$, in which the element at the ith row and kth column is $(p_{ik}^0 + p_{ik}^+ - p_{ik}^-)$. The jth column of $(\boldsymbol{F}^\pm - \boldsymbol{I})\boldsymbol{A}$ is denoted as $\boldsymbol{g_j^\pm}$.

The basic inverse optimization formulation for the FC matrix has an objective of minimizing the total perturbation, and KKT conditions of P_1 as constraints. The KKT conditions of P_1 are shown in **Eqs. (26) – (31).**

$$\ln h_j + \alpha \boldsymbol{t^T a_j} - \boldsymbol{l^T b_j} - \boldsymbol{m^T g_j} = 0 \tag{26}$$

$$\boldsymbol{h} > 0 \tag{27}$$

$$\boldsymbol{m} \geq 0 \tag{28}$$

$$\boldsymbol{m}\big((\boldsymbol{F} - \boldsymbol{I})\boldsymbol{Ah} + \boldsymbol{s_0}\big) = \boldsymbol{0} \tag{29}$$

$$(\boldsymbol{F} - \boldsymbol{I})\boldsymbol{Ah} + \boldsymbol{s_0} \geq 0 \tag{30}$$

$$\boldsymbol{r} = \boldsymbol{Bh} \tag{31}$$

In practice, **Eq. (26)** can cause infeasibility of the inverse optimization problem. For path j_1 and j_2 connecting the same OD pair, $\boldsymbol{l^T b_{j_1}} = \boldsymbol{l^T b_{j_2}}$. From **Eq. (26)**, we have $\ln h_{j_1} + \alpha \boldsymbol{t^T a_{j_1}} - \boldsymbol{m^T g_{j_1}} = \ln h_{j_2} + \alpha \boldsymbol{t^T a_{j_2}} - \boldsymbol{m^T g_{j_2}}$. When $\boldsymbol{m^T g_{j_1}} = \boldsymbol{m^T g_{j_2}} = 0$, $\ln(\frac{h_{j_1}}{h_{j_2}}) = -\alpha \boldsymbol{t^T}(\boldsymbol{a_{j_1}} - \boldsymbol{a_{j_2}})$, indicating that the observed path flows need to yield logit flows to make **Eq. (26)** hold. However, in real data, observed path flows hardly yield logit flows exactly. To be compatible with non-logit path flows, we add the term $\beta \sum_{j \in K} (\ln h_{O,j} + \alpha \boldsymbol{t^T a_j} - \boldsymbol{l^T b_j} - \boldsymbol{m^T g_j^\pm})^2$ to the objective function to fit the observed path flows to logit flows and eliminate **Eq. (26)** from the constraints. Note that the network flow observations do not need to be a full set that covers all the paths. Partial observations would also work, but with some possible accuracy loss. If only partial observations are obtained with only paths $j \in K' \subset K$, this term can be the sum over $j \in K'$. In practice, critical paths with significant demand could be observed to ensure the overall accuracy. The parameter β controls how close the resulting logit flows are from the observed flows, which can be set according to specific estimation needs.

Moreover, solving the basic inverse optimization problem ensures that the estimated FC matrix leads to the observed flows (decision variables in the forward model), but not the observed capacities. We add another term $\gamma \sum_{k\in E}\left(\sum_{j\in K}\sum_{k\in E}(p_{ik}^{0}+p_{ik}^{+}-p_{ik}^{-})a_{kj}h_{O,j}+s_{0,i}-s_{O,i}\right)^2$ to the objective function to minimize the sum of square of the differences between equilibrium capacities computed from the estimated FC matrix and observed equilibrium capacities (the capacity-fitting term). The parameter γ is the weight of capacity fitting. If capacity data is available, this term would help fit a FC matrix that leads to both observed flows and capacities. If capacity data is not available, the term could be neglected ($\gamma = 0$). The fitted FC matrix would still lead to the observed path flows.

The resulting inverse optimization problem is formulated as P_2. **Eq. (31)** is eliminated since demand is observed through path flows in this case.

$$\begin{aligned} P_2\text{: min} \sum\nolimits_{(i,k)\in Z}(p_{ik}^{+}+p_{ik}^{-}) + \beta \sum\nolimits_{j\in K}\left(\ln h_{O,j} + \alpha \boldsymbol{t^T a_j} - \boldsymbol{l^T b_j} - \boldsymbol{m^T g_j^{\pm}}\right)^2 \\ + \gamma \sum_{k\in E}\left(\sum\nolimits_{j\in K}\sum\nolimits_{k\in E}(p_{ik}^{0}+p_{ik}^{+}-p_{ik}^{-})a_{kj}h_{O,j}+s_{0,i}-s_{O,i}\right)^2 \end{aligned} \quad (32)$$

Subject to

$$(\boldsymbol{F}^{\pm}-\boldsymbol{I})\boldsymbol{A}\boldsymbol{h_O}+\boldsymbol{s_0} \geq 0 \quad (33)$$

$$\boldsymbol{m}\left((\boldsymbol{F}^{\pm}-\boldsymbol{I})\boldsymbol{A}\boldsymbol{h_O}+\boldsymbol{s_0}\right) = 0 \quad (34)$$

$$p_{ik}^{+},\ p_{ik}^{-} \geq 0, \qquad \forall (i,k)\in Z \quad (35)$$

$$\boldsymbol{m} \geq 0 \quad (36)$$

P_2 is a quartic program with quadratic constraints. The term $(\boldsymbol{m^T g_j^{\pm}})^2$ in the objective function **Eq. (32)** is quartic. Constraint of **Eq. (34)** is quadratic due to the term $\boldsymbol{mF^{\pm}}$. To solve P_2, we use a commercial solver LINGO which tackles non-linear optimization problems with non-linear constraints. In practice, there are other methods that can be considered to solve P_2. For example, McCormick envelopes (McCormick, 1976) can transform bilinear terms ($\boldsymbol{mF^{\pm}}$) and squares of bilinear terms ($(\boldsymbol{m^T g_j^{\pm}})^2$) into a set of linear upper and lower bound constraints. For the computational experiments in our case study, we found a commercial solver was adequate, and leave more efficient algorithmic development to future research.

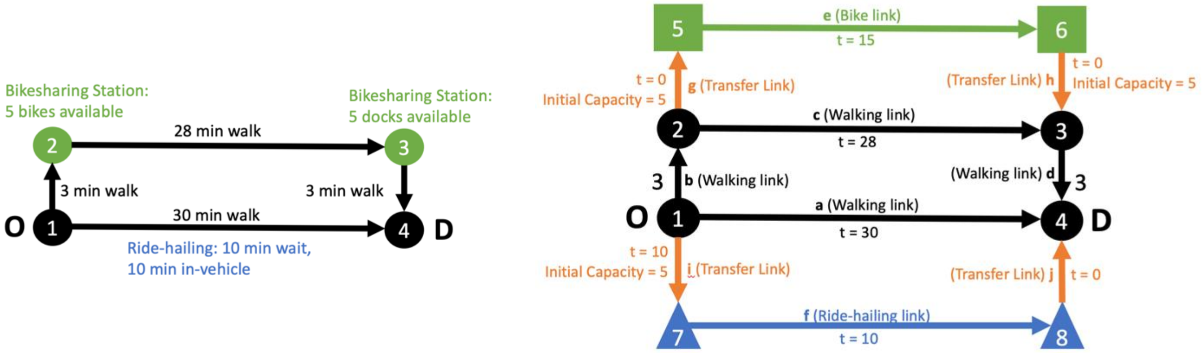
Compared with learning-based approaches (Liu et al., 2023), the proposed inversed optimization approach requires much smaller amount of data. Only one set of equilibrium flows along with prior estimates are needed to estimate the FC matrix. Capacity observations are optional. The solution of P_2 is non-unique with respect to the capacities, since the capacity of a link with congestible capacities may be affected by a set of links, but all the solutions equally capture the flow-capacity interdependency caused by demand patterns and operation policies. Due to such non-uniqueness, the accuracy of the equilibrium capacities computed from the estimated FC matrix is dependent on the parameter γ, since the structure of P_2 is derived from optimality conditions of path flows, not capacities. The non-unique FC matrix estimated from the same observed path flows lead to non-unique equilibrium capacities from estimated FC matrix. This is a known challenge with inverse optimization where further regularization and additional data or repeated observations (see discussion in Xu et al., 2018) can help to address. We illustrate this non-uniqueness in Section 3.2. Such non-uniqueness would not bias the reproduced equilibrium flows significantly. Since the structure of P_2 imbeds the KKT conditions of the forward SUE assignment model. This ensures that the estimated FC matrix is the optimal one in terms of producing the observed path flows, given the prior of the FC matrix and the same constraints in additional to the KKT conditions.

Due to such non-uniqueness, the estimated FC matrix by P_2 has limited interpretability. The major purpose of the estimated FC matrix is for flow prediction. However, although the matrix does not give explicit information regarding operational policies, it can imply some important information. Adding the constraints while estimating the matrix makes the estimated matrix more interpretable. This makes operational policy inferences possible. For example, as shown in the downtown Manhattan FC matrix calibration in section 4.3, sign constraints could also be added to define if it is a contribution or reduction. Another example would be to constrain the relative relationship between the coefficients that reflects the contribution to the same capacity from different links.

3 Numerical Examples

3.1 Illustrative small example

Before application to larger networks, we use a toy network to demonstrate how the models and solution algorithm works. The network is shown in **Figure 2**, in which **Figure 2(a)** is a demonstration of the initial network, and **Figure 2(b)** is the expanded network connecting subgraphs of different modes with transfer links. In the toy network, there are three modes: walk, bike-share, and ride-hail. There are 4 nodes in the original network: 1,2,3, and 4. There are 2 bike-sharing stations where the travelers can pick up and drop off bikes: 2 and 3. There are 10 units of demand going from 1 to 4 (1 to 8 in (b)). Customers can choose to walk, to go to the bike-sharing station at 2 and ride a shared bike, or to ride-hail directly from 1.



(a) Original Network (b) Expanded Network

Figure 2. Toy network.

In **Figure 2(b)**, the green part is the bike-sharing subgraph, the black part is the walking subgraph, the blue part is the ride-hailing subgraph, and the orange links are the transfer links connecting the subgraphs. The cost of transfer link accessing/egressing bike-sharing is the time of picking up/dropping off a bike (we assume both 0 here). The cost of a transfer link accessing ride-hailing is the wait time before being picked up by a ride-hailing vehicle (we assume 10 here), and the cost of the transfer link egressing ride-hailing is 0 (no restriction on getting off). Links in bike-sharing, walking, and ride-hailing subnetworks are uncapacitated (no restriction on traveling in these modes through the real network (a)). Access link g is capacitated due to limited number of bikes available at 5 (initial capacity assumed as 5). Egress link h is capacitated due to limited number of vacant docks at 6 (initial capacity assumed as 5). Access link i accessing ride-hailing is restricted by the number of vehicles available near 1 (initial capacity assumed as 5), while egress link j is incapacitated. The undersaturated travel times and initial capacities of the links are labeled in **Figure 2(b)**.

3.1.1 Forward model illustration

The capacities of the access links g, i, and egress link h are dependent on the flows. The bikes at the bike-sharing station at node 2 are taken away by the customers traversing link h. The ride-hailing vehicles available near node 1 are taken by the customers traversing link i. The vacant docks at the bike-sharing station at node 2 are taken away by the customers traversing link h. With rebalancing of bike-sharing and matching of ride-sourcing providers, we assume a 0.1 absolute efficiency for link g, i, and h. The above capacity dependency pattern leads to the FC matrix $\boldsymbol{F}$ shown in **Table 1**. With such an efficiency, the capacity change of link g, i, and h would be 10% of the flows on themselves, which complies with the condition of localized, incremental changes.

Table 1. FC matrix ($\boldsymbol{F}$) assumed for the toy network.

	a	b	c	d	e	f	g	h	i	j
a	0	0	0	0	0	0	0	0	0	0
b	0	0	0	0	0	0	0	0	0	0
c	0	0	0	0	0	0	0	0	0	0
d	0	0	0	0	0	0	0	0	0	0
e	0	0	0	0	0	0	0	0	0	0
f	0	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	-0.1	0	0	0
h	0	0	0	0	0	0	0	-0.1	0	0
i	0	0	0	0	0	0	0	0	-0.1	0
j	0	0	0	0	0	0	0	0	0	0

In this case, we do not apply **Algorithm 1** to determine a bounded path set. Instead, we solve the model with different path sets to illustrate how different path sets affect the equilibrium. There are 4 paths going from 1 to 4: i-f-j, b-g-e-h-d, a, and b-c-d. We add the paths one by one from the shortest to the longest and use **Algorithm 2** to solve the SUE, assuming $\alpha = 1$. The results are shown in **Table 2**.

Table 2. SUE with different path set

Number of paths	**Paths generated**	**Path flows**	**Link flows (*a,b,c,d,e,f,g,h,i,j*)**	**Equilibrium capacities (*a,b,c,d,e,f,g,h,i,j*)**	$\boldsymbol{m}^{\boldsymbol{T}}\boldsymbol{g}_{\boldsymbol{j}} + \alpha\boldsymbol{T}_{\boldsymbol{j}}$	**Objective value**
1	*ifj*	-	-	-	-	Infeasible
2	*ifj*, *bgehd*	-	-	-	-	Infeasible
3	*ifj*, *bgehd*, *a*	4.545, 4.545, 0.909	0.909, 4.545, 0, 4.545, 4.545, 4.545, 4.545, 4.545, 4.545, 4.545	∞, ∞, ∞, ∞, ∞, ∞, 4.545, 4.545, 4.545, ∞	28.39, 28.39, 30	217.315
4	*ifj*, *bgehd*, *a*, *bcd*	4.545, 4.545, 0.893, 0.017	0.893, 4.562, 0.017, 4.562, 4.545, 4.545, 4.545, 4.545, 4.545, 4.545	∞, ∞, ∞, ∞, ∞, ∞, 4.545, 4.545, 4.545, ∞	28.372, 28.372, 30, 34	217.298

Static capacity	*ifj*, *bgehd*, *a*, *bcd*	4.999, 5, 0.001, 0	0, 4.999, 0, 4.999, 4.999, 5, 4.999, 4.999, 5, 5	∞, ∞, ∞, ∞, ∞, ∞, 5, 5, 5, ∞	21.332, 21.332, 30, 34	211.094

As stated in Section 2.2, the value of $\frac{\boldsymbol{m}^T \boldsymbol{g}_j}{\alpha}$ represents a path delay of path j. The value $\boldsymbol{m}^T \boldsymbol{g}_j + \alpha \boldsymbol{T}_j$ should give us the trip utility as shown in **Table 2**. It can be verified that the assignment and trip utilities conform to the logit model shown in **Eq. (5)**. The flows assigned with all 4 paths are shown in **Figure 3**.

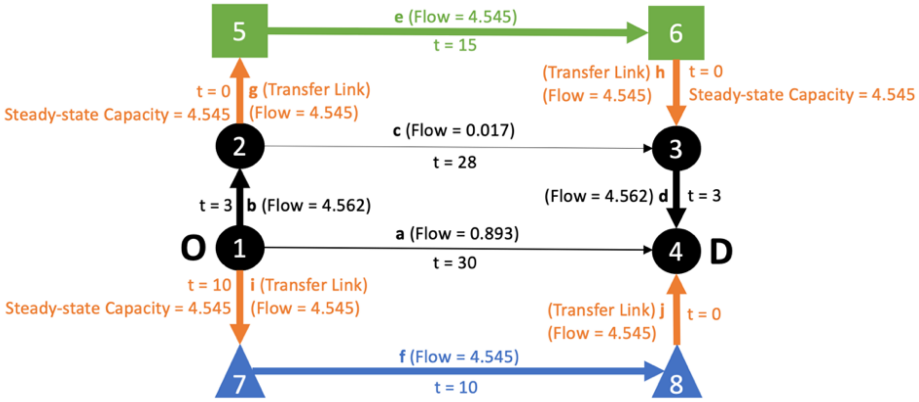


Figure 3. Equilibrium flows and capacities of the toy network.

With one path or two paths, the assignment is not feasible. The reason for those 2 scenarios to be infeasible is due to insufficient capacities of bikeshare and ride-hailing. When applied to an actual multimodal network with car paths, such an issue would only happen when some of the capacities are exceeded, which could be the capacity of any part of the multimodal network: road link capacity, fleet capacity, etc. The infeasibility shows that either the path set is not realistic (more paths are needed), or the network capacities are actually insufficient. With three paths, the assignment becomes feasible with an optimal objective value of 217.315. With all four paths, adding the longest path *bcd* leads to a reduction of only 0.0017 to the objective value. The flow on path *bcd* is 0.017, which is only 0.17% of the total demand. In real cases, insignificant paths like path *bcd* can be left out by running **Algorithm 1** with a reasonable bounding ratio. In this case, the corresponding bounding ratio would be around 1.6.

Comparing the SUE with congestible capacities with the SUE with static capacities (both with all four paths), for OD 1, the capacities of link g,i,h are reduced by the congestible capacity effect, which leads to more flow on path *ifj* and *bgehd*.

3.1.2 Inverse model illustration

We use the equilibrium flows as with 4 paths solved from the forward model as observed path flows to estimate the FC matrix under two circumstances: without capacity and with capacity fitting. Without capacity fitting, $\beta = 1$ and $\gamma = 0$. With capacity fitting, $\beta = \gamma = 1$.

Table 3 shows the estimated FC matrix without capacity fitting. Compared with the FC matrix that generates the equilibrium flows (original FC matrix), there is one different element: p_{gg}, which is -0.1 in the original FC matrix but estimated as 0. Equilibrium flows computed from the estimated FC matrix is the same as the observed path flows. The reason is that, on path *bgehd*, there is another link with congestible capacity which have the same estimated efficiencies as the original efficiencies: link *h*. Congestible capacity constraint of link *h* restricts the flow on path *bgehd* to the same as the observed flow.

With capacity fitting, the estimated FC matrix is identical to the original FC matrix shown in **Table 1**.

Table 3. Estimated FC matrix (***F***) without capacity fitting.

	a	b	c	d	e	f	g	h	i	J
a	0	0	0	0	0	0	0	0	0	0
b	0	0	0	0	0	0	0	0	0	0
c	0	0	0	0	0	0	0	0	0	0
d	0	0	0	0	0	0	0	0	0	0
e	0	0	0	0	0	0	0	0	0	0
f	0	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	**0**	0	0	0
h	0	0	0	0	0	0	0	-0.1	0	0
i	0	0	0	0	0	0	0	0	-0.1	0
j	0	0	0	0	0	0	0	0	0	0

3.2 Nguyen-Dupuis network

To comply with the condition of localized, incremental changes, we test 0.1 and 0.2 absolute efficiencies in this example.

3.2.1 Base case: 0.1 absolute efficiency with different bounding ratios

We further test the model and solution algorithm on the Nguyen-Dupuis network (Nguyen and Dupuis, 1984) shown in **Figure 4** with a multimodal setting. The network has 13 nodes and 19 links. Undersaturated travel costs of the links are labeled in red by the side of the links, which includes time costs converted to dollars and fare cost if applicable. The numbers on the links are link IDs. There are four OD pairs in the network (OD1:*1-2*, OD2: *1-3*, OD3: *4-2*, OD4: *4-3*). The demands of the OD pairs are 400, 800, 600, 200, respectively.

We assume that the Nguyen-Dupuis network in **Figure 4** is the walking network, from which the travelers can transfer to three other modes: bike-sharing, ride-hailing, and microtransit. Travelers can transfer to bike-sharing by picking up a bike at node 5 and can transfer back to the walking network by dropping off the bikes at node 11. There are three paths they can take while riding a shared-bike: 5-6-7-11 (undersaturated travel cost is $8.5), 5-6-10-11 (undersaturated travel cost is $11), and 5-9-10-11 (undersaturated travel cost is $12.5). The ride-hailing service picks up and drops off passengers at nodes 1, 2, 3, 4, 5, 11, and 12. Travelers can transfer to ride-hailing and transfer back to walking from these nodes. Between each pair of pick-up and drop-off nodes of ride-hailing, the drivers take the shortest path. Hence, the subnetwork of ride-hailing is composed of links connecting all pairs of pick-up and drop-off nodes, the travel costs of which are shown in **Table 4**. Microtransit runs from node 12 to 11 through a fixed route (12-6-7-11) with an undersaturated travel cost of $5. Travelers can transfer to micro-transit at node 12 and transfer back to walking at node 11.

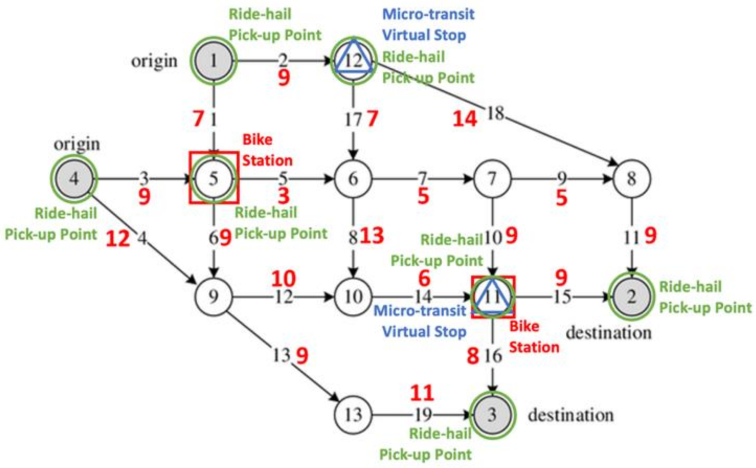


Figure 4. Nguyen-Dupuis network with multiple modes.

The network in **Figure 4** is expanded into the complete network with 24 nodes and 54 links as shown in **Figure 5**, which is composed of four subnetworks (walking subnetwork in black, microtransit subnetwork in blue, ride-hailing subnetwork in green, and bike-sharing subnetwork in red) and the transfer links between modes (in orange). All the links within the subnetworks are uncapacitated, while the capacities of the transfer links are dependent on the availability of the three modes.

Initially, there are 400 shared bikes available at node 5, and 300 vacant docks available at node 11. We assume it takes 1 min to pick up/drop off a shared bike. Hence, the initial capacities of link (5,16) and (17,11) are 400 and 300, respectively, and the undersaturated travel costs are $1 for both.

For microtransit, the initial number of seats available at node 12 is 100, and there is no restriction on getting off microtransit at node 11. The average access cost is $5. The initial capacities of link (12,14) and (15,11) are 100 and infinity (uncapacitated), respectively. The travel cost of (12,14) is $5, and (15,11) has $0 cost.

For ride-hailing, the initial number of vehicles available at nodes 1, 2, 3, 4, 5, 11, and 12 are all 200. A typical cost of waiting is $3. The initial capacities of the transfer links accessing ride-hailing (links (4,21), (1,18), (5,22), (12,24), (11,23)) are 200, and the travel costs are $3. Since nodes 3 and 4 are destinations, we ignore the access links to ride-hailing there. The transfer links egressing ride-hailing are uncapacitated and have $0 travel cost due to no restrictions getting off.

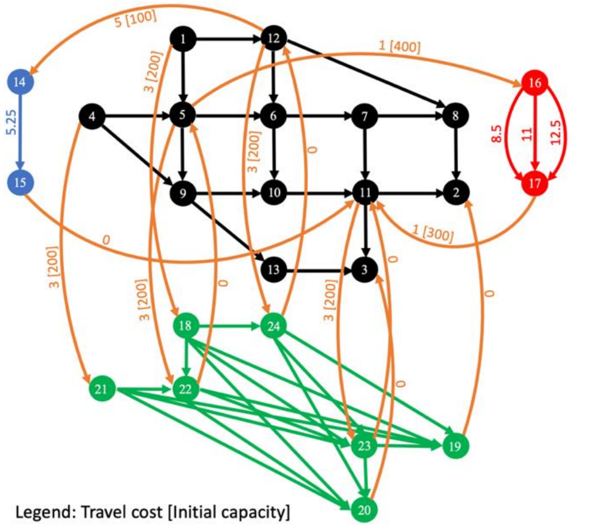
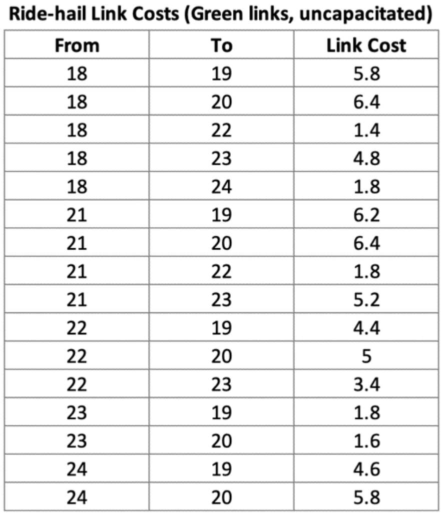


Figure 5. Complete network of the Nguyen-Dupuis network example.

Table 4. Costs of ride-hailing links in the complete network



Ride-hail Link Costs (Green links, uncapacitated)

From	To	Link Cost
18	19	5.8
18	20	6.4
18	22	1.4
18	23	4.8
18	24	1.8
21	19	6.2
21	20	6.4
21	22	1.8
21	23	5.2
22	19	4.4
22	20	5
22	23	3.4
23	19	1.8
23	20	1.6
24	19	4.6
24	20	5.8

Assumptions regarding the FC matrix are similar to the toy network in section 3.1. Capacities of the access links of micro-transit, bike-share, and ride-hailing are taken away by a proportion of the flows traversing them. Capacity of the bike-share egress link is taken away by a proportion of the flow traversing it. All the absolute values of the efficiencies above are assumed to be 0.1, which means that the capacity taken away/added is 10% of the flow that affects the capacity.

Assuming $\alpha = 1$, we solve the assignment using **Algorithm 1** with 17 different bounding ratios ρ: from 3.4 to 5.0 with an interval of 0.1. The objective value change and system total travel cost change are shown in **Figures 6-7**. We pick four scenarios to show the detailed results: Scenario(a): $\rho = 3.5$, Scenario(b): $\rho = 4$, Scenario(c): $\rho = 4.5$, Scenario(d): $\rho = 5$. Parameter settings are: $\epsilon = 0.001$, $N = 5$, $\varepsilon = 0.001$. The algorithm was run on a laptop with 2.3 GHz Quad-Core Intel Core i7 and 32GB installed RAM, with Python 3.7. The results are shown in **Figure 8**. Note that **Algorithm 1** is not able to obtain the Lagrange multipliers ($\boldsymbol{m}$).

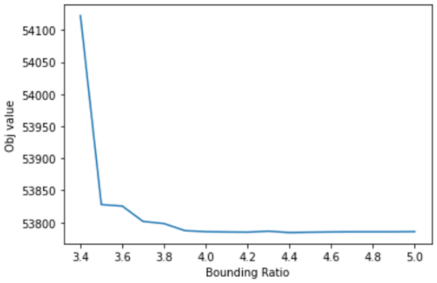


Figure 6. Objective value change with bounding ratio.

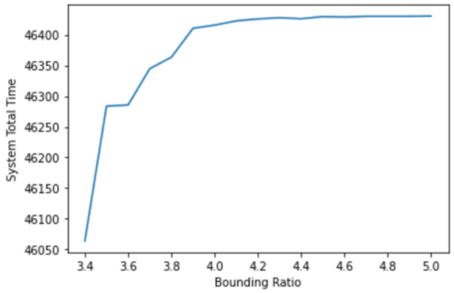


Figure 7. System total cost change with bounding ratio.

From **Figures 6-7**, we observe that the objective value is significantly reduced when the bounding ratio is increased to 3.5, indicating that 3.5 can be considered a sufficient value of bounding ratio to capture enough path choice diversity. After the bounding ratio is increased to 4, both the objective value and system total cost hardly change. The method is effective in eliminating unreasonable paths.

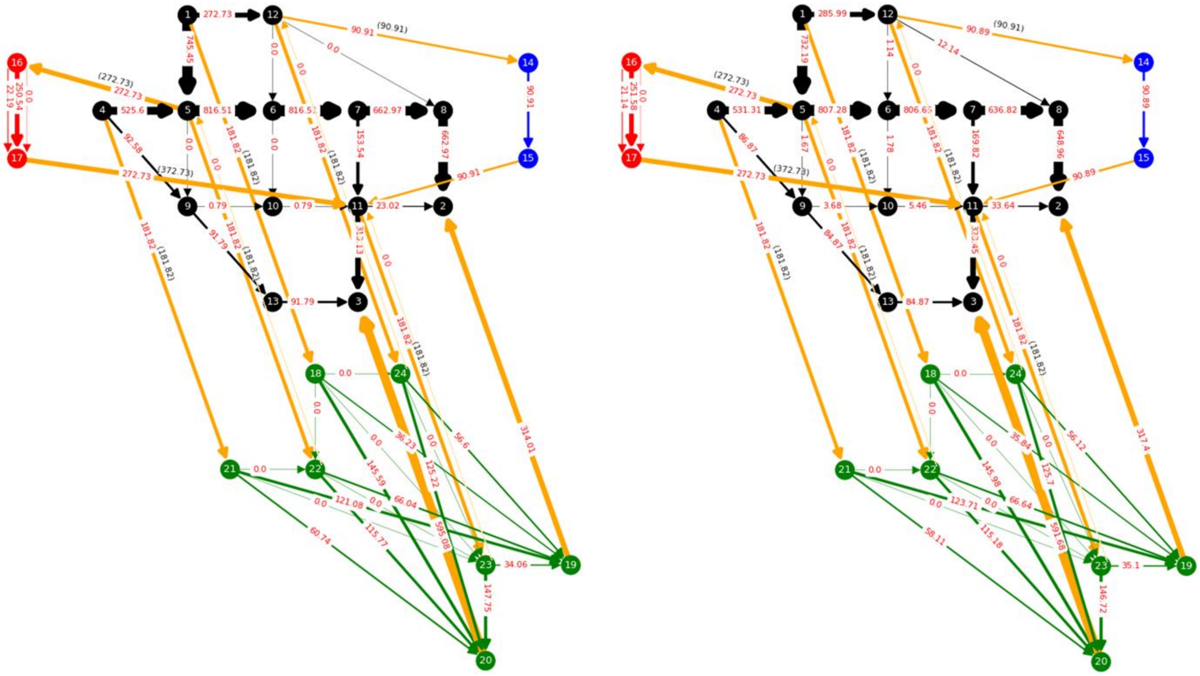
We also observe that the feasibility of assignment depends on the relationship between demand and base capacities with fixed flow-capacity patterns. With current demand and initial capacities, the assignment is infeasible when bounding ratio is smaller than 3.4. This is because the capacity-flow dependency could lead to very small capacities at certain links, which is a situation that the MoD operators would like to avoid.

Remark 1. *Feasible assignments obtained by changing base capacities indicate that proper initial fleet deployment could avoid inadequate capacities, which shows a potential application of the SUE model with congestible capacities.*

As shown in **Figure 8**, equilibrium link flows and capacities change very little across Scenario (b), (c) and (d), indicating that the assignment with the 16 paths in Scenario (b) basically incorporate all effective paths. In Scenario (d), the paths in Scenario (a) are assigned 97.79% of the total demand. Looking at **Figure 8**, with higher bounding ratio, flows on links with major flows switch to links with minor flows. For example, link (4,5),(5,6),(6,7),(7,8) form a major corridor, since they are more likely to be a part of shorter paths. Flows on these major links generally decrease when more paths are generated, switching to longer paths.

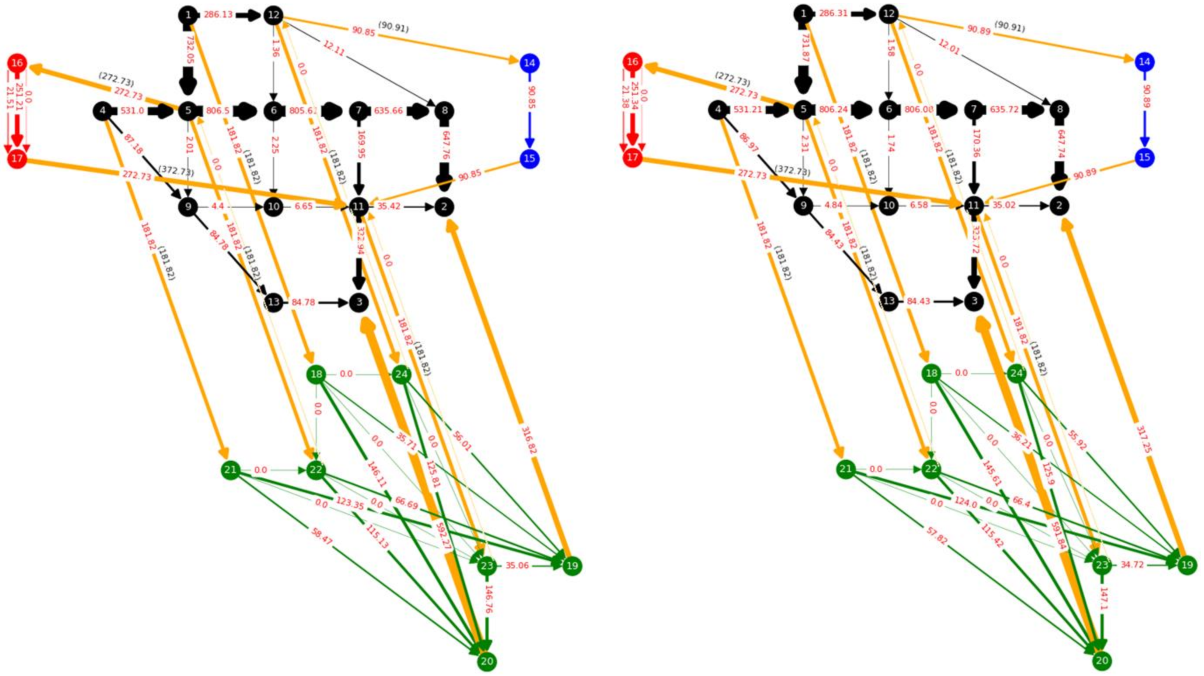
3.2.2 Sensitivity test: 0.2 absolute efficiency

To illustrate how the FC matrix affects the assignment results, we change all the non-zero efficiency from 0.1 to 0.2. The assignment is shown in **Figure 9**. The bounding ratio ρ is set to 5 to be compared with Scenario (d). **Algorithm 1** is applied with the same configuration, taking 35.59 sec. Equilibrium link flows are shown in **Figure 9**. The optimal objective value is 55187.13.



(a) Bound ratio = 3.5

(b) Bound ratio = 4



(c) Bound ratio = 4.5 (d) Bound ratio = 5

Figure 8. Link flow and link capacities at SUE (0.1 absolute efficiency).

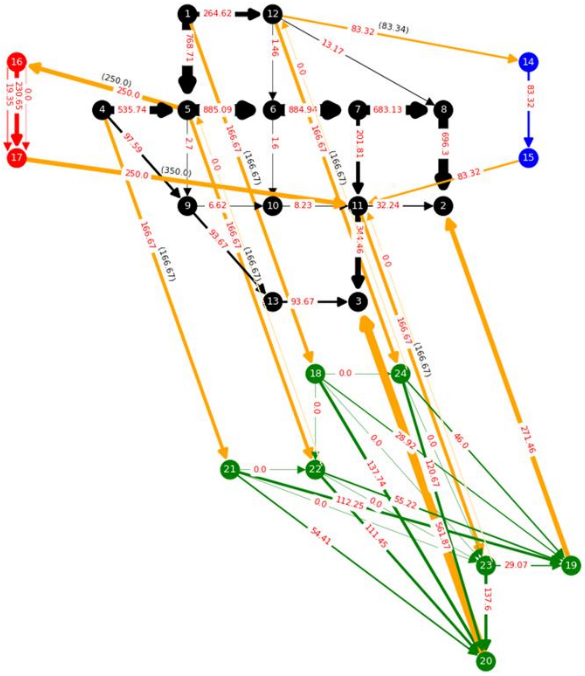


Figure 9. Link flow (in red) and link capacities (in black) at SUE with congestible capacities (0.2 absolute efficiency).

Larger absolute efficiency results in more decrease in access capacities, which reflects less rebalancing effort in real cases. Compared with Scenario (d), ride-hailing, bike-share, and micro-transit have less flows due to reduced equilibrium capacities of access links. Equilibrium capacities of ride-hailing access links (4,21),(5,22),(1,18),(12,24),(11,23) are reduced from 181.82 to 166.67.

Equilibrium capacities of bike-share access link (5,16) is reduced from 272.73 to 250.0. Equilibrium capacities of bike-share access link (12,14) is reduced from 90.91 to 83.34. Flows within ride-hailing, bike-share, and micro-transit subnetworks significantly reduced. More flows are assigned to the walking subnetwork.

3.2.2 Sensitivity test of α

To show how α affects the assignment results, we set $\alpha = 0.9$ to obtain SUE with bounding ratio and FC matrix the same as Scenario (d). Equilibrium link flows are shown in **Figure 10.** The optimal objective value is 49138.18.

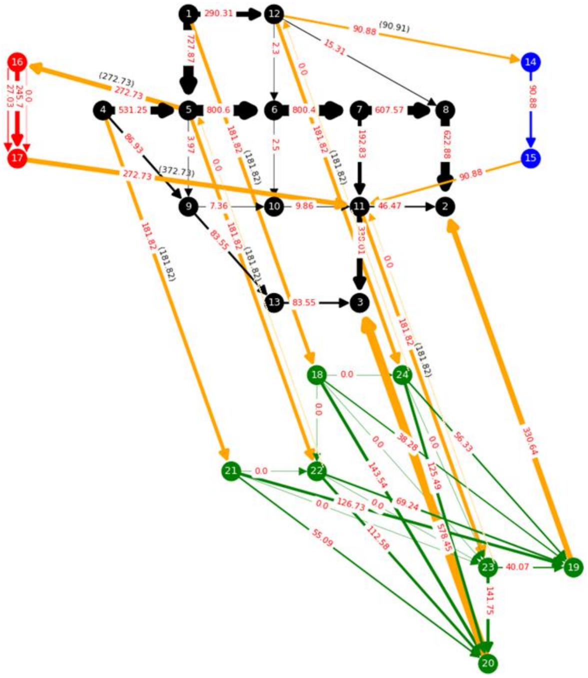


Figure 10. Link flow (in red) and link capacities (in black) at SUE with congestible capacities ($\alpha = 0.9$).

Compared with Scenario (d) ($\alpha = 1.0$), the 10% change in α does not lead to change in capacities, but significant change in flows. Less randomness in route choice leads to more concentrated flow distribution. With larger α, generally, flows on links with major flows become larger, while flows on links with minor flows becomes smaller. Such effect can be observed in **Figure 8**. For example, the major corridor formed by link (4,5),(5,6),(6,7),(7,8) is assigned less flow compared with Scenario (d). While links with small amounts of flows such as (9,10),(10,11),(11,2) are assigned more flows.

Parameter α represents customers' perception of travel costs, which need to be pre-estimated from local data. The parameter can be adopted from mode choice/path choice models from local planning frameworks. When there are more factors than travel time and fare such as comfort level, weighted travel cost αt_i can be extended to a linear combination.

4 Case study with yellow taxi data from Downtown Manhattan, NY

To demonstrate application of the model to a real network, we selected an area composed of 19 taxi zones in downtown Manhattan, NY and extracted OD demand from yellow taxi data of July

1st, 2021 (TLC, 2021). The study area is shown in **Figure 11**, in which the zone IDs are labeled in red. OD demands aggregated to origin and destination are shown in **Figure 12**.

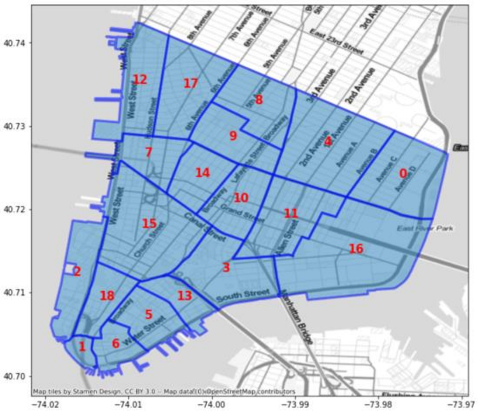
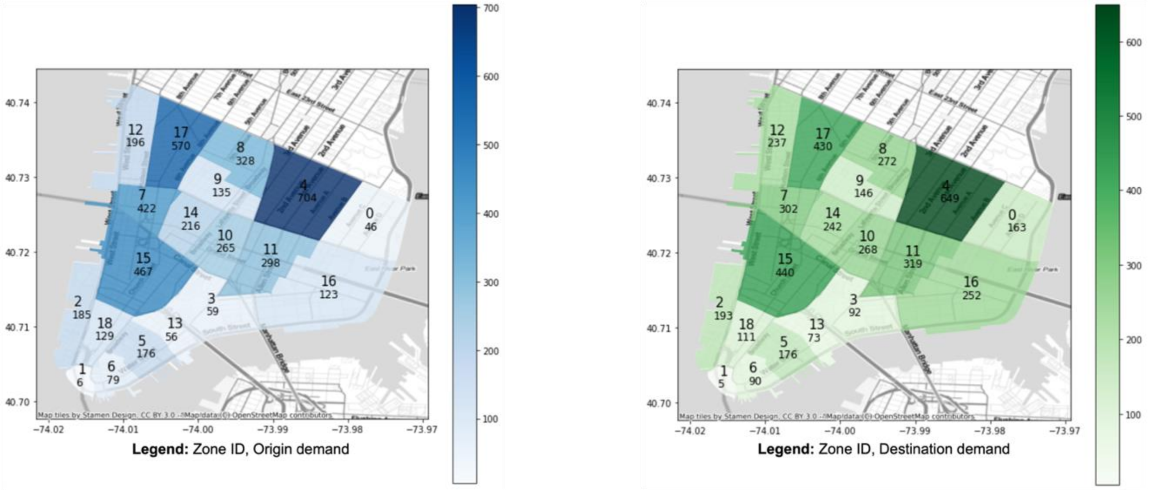


Figure 11. Study area composed of 19 taxi zones in downtown Manhattan.



(a) OD demand aggregated to origins (b) OD demand aggregated to destinations

Figure 12. Origin and destination demand aggregated from yellow taxi data on July 1st, 2021.

We treat each taxi zone as a node and form a network by connecting all pairs of neighboring zones with bidirectional links as shown in **Figure 13**. We assume that the origins/destinations of the OD demand from yellow taxi data are origins/destinations of the trips, which may involve first-mile/last-mile walking. This assumption is inconsistent with the data but is acceptable since this case study is for demonstration of model application instead of practical evaluation. For real cases, trip OD demand should be applied.

To model the trips composed of first-mile walking, taxi ride, and last-mile walking, we constructed the network as shown in **Figure 14**. The network has 57 nodes and 287 links, including five parts: the access network for first-mile walking, taxi network for taxi rides, egress network for last-mile walking, transfer-in links connecting the access network with the taxi network, and transfer-out network connecting the taxi network with the egress network. Links within the access network, egress network, and taxi network are bidirectional, while the transfer-in and transfer-out links are single-directional. The 19 nodes in the access network serve as origins while the 19 nodes in the egress network are destinations, so the trips would have a structure of first-mile walking→taxi ride→last-mile walking. Link costs in access network and egress network are set as the walking times of the shortest path connecting the centroids of the zones, assuming an average walking speed of 3 mph. Link costs in the taxi network are set as the driving times of the shortest paths connecting the centroids of the zones, assuming an average driving speed of 12.7 mph. Taxi fare is ignored in this case.

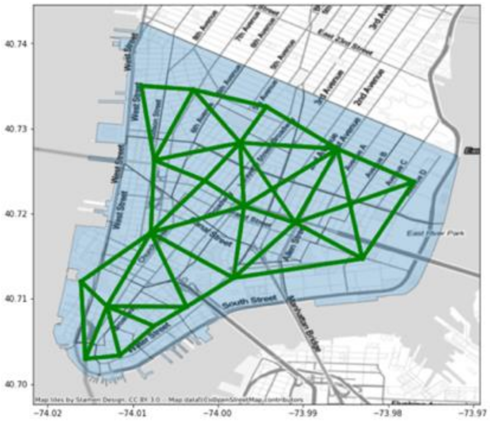


Figure 13. Downtown Manhattan network.

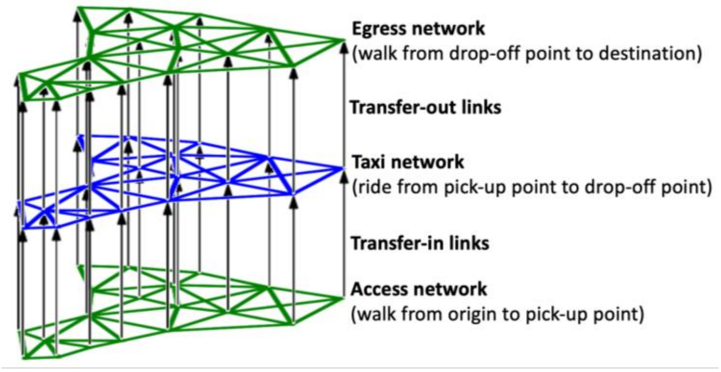

Egress network
(walk from drop-off point to destination)
Transfer-out links
Taxi network
(ride from pick-up point to drop-off point)
Transfer-in links
Access network
(walk from origin to pick-up point)

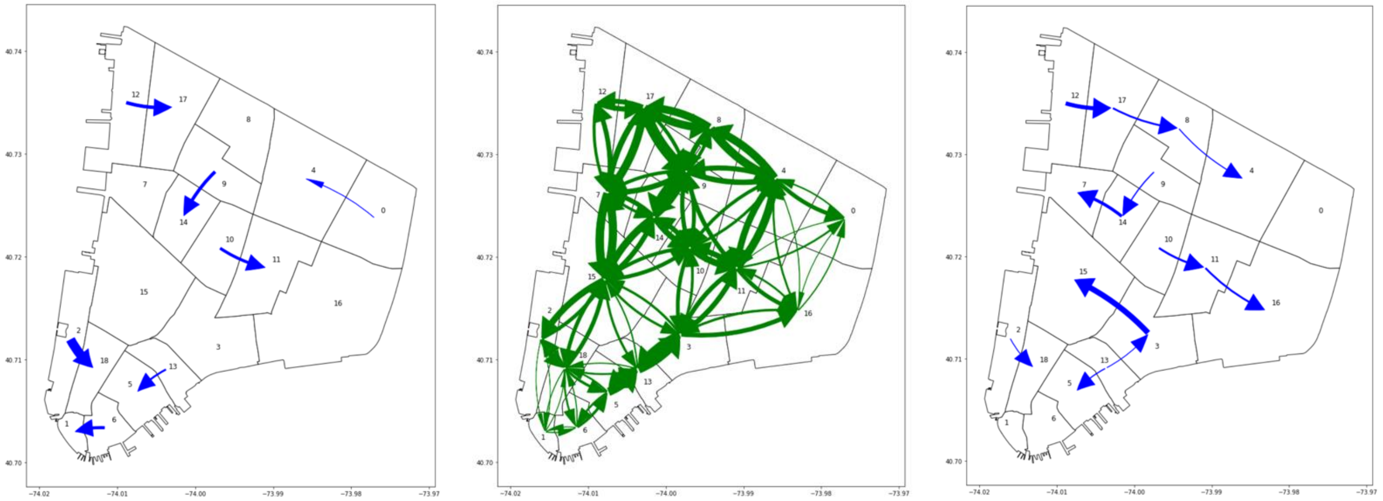
Figure 14. Complete network for trip modeling.

Shortest paths are computed from OpenStreetMap using python packages OSMnx and NetworkX. Links within the access, egress, and taxi networks are uncapacitated, since we assume that the modeled flows is only a small proportion of the background traffic. The transfer-out links are also incapacitated since there are no restrictions on getting off the taxis. The transfer-in links are the ones with congestible capacities, whose initial capacities are 750. We assume the undersaturated link costs of transfer-in links are \$1, representing a waiting time without congestion.

4.1 Base case: 0.75 absolute efficiency

We assume a 0.75 absolute efficiency for all zones in the research area. Efficiency from all transfer-in links to itself is -0.75, and efficiency from all transfer-out link to the transfer-in link of the same node is 0.75. We run **Algorithm 1** with the same configuration as the Nguyen-Dupuis example. Bounding ratio is set to be 1.1, generating 660 paths. Parameter settings are: $\epsilon = 0.01$, $N = 5$, $\varepsilon = 0.001$. The algorithm converged after 65 iterations. Run time is 16min 48s (6min 30s on path generation, 10min 18s on Frank-Wolfe algorithm). **Figure 15** and **Table 5** show the equilibrium flows in the access, egress, and taxi networks. **Figure 16** shows the saturation and flows of transfer-in flows in each zone. **Figure 17** shows the capacity change caused by the flows ($s_i - s_{0,i}$). The maximum absolute capacity change is 105.19, which is 14.03% of the initial capacity. The changes also falls into a range of localized, incremental change.

Taxi flows are larger on the northwestern part of the study area, which is where the high OD demand is as shown in **Figure 12**. Access and egress walking flows are very small, indicating that first-mile/last-mile walking only occurs on the longer paths. This happens when taxi is significantly faster than walking. None of the transfer-in links have binding capacities, indicating that there is no extra wait caused by limited capacities ($\boldsymbol{m^T g_j} = 0$). Undersaturated path costs determine the logit flows. The highest saturation rate appears at zone 4 (99.3%), which is the zone with the highest origin and destination demand. Zones with higher demand tend to have higher saturation rates due to more capacities taken away by the incoming flow, which can be observed from **Figures 11** and **12**. Capacities move from northwest to southeast given the demand distribution. The northwestern part has high origin demand as well as high destination demand, while in the southeastern part, destination demand is significantly higher than origin demand, which leads to more taxis being left in the southeast.



(a) Access walking flows (b) Taxi flows (c) Egress walking flows

Figure 15. Link flows (0.75 absolute efficiency)

Table 5. Link flows (0.75 absolute efficiency)

Access walking links			**Taxi links**														
Start	**End**	**Flow**	**Start**	**End**	**Flow**	**Start**	**End**	**Flow**	**Start**	**End**	**Flow**	**Start**	**End**	**Flow**			
0	4	0.00	0	4	28.00	5	13	220.48	10	9	94.50	15	14	125.25			
2	18	0.18	0	11	8.84	5	18	35.04	10	11	167.15	15	10	94.89			
6	1	0.05	0	16	6.15	6	1	10.23	10	3	91.51	15	3	52.57			
9	14	0.07	1	2	4.58	6	5	60.39	10	15	88.33	15	13	52.51			
10	11	0.06	1	6	16.68	6	18	14.80	10	14	119.41	15	18	92.85			
12	17	0.07	1	18	7.65	7	12	116.80	11	3	114.12	15	2	162.63			
13	5	0.02	2	1	5.61	7	17	213.44	11	4	208.61	16	3	55.18			
Egress walking links			2	15	89.29	7	9	136.56	11	10	140.69	16	11	40.53			
			2	18	104.73	7	14	62.47	11	16	100.33	16	0	20.34			
Start	**End**	**Flow**	3	10	59.85	7	15	222.92	11	0	39.18	16	4	13.65			
2	18	0.02	3	11	127.60	8	17	206.62	12	7	67.84	17	12	120.11			
3	15	0.44	3	13	194.69	8	9	126.17	12	17	127.69	17	7	149.12			
8	4	0.04	3	15	53.47	8	4	223.26	13	3	266.95	17	9	294.23			
9	14	0.05	3	16	111.30	9	17	84.95	13	15	26.09	17	8	209.50			
10	11	0.13	4	0	100.48	9	8	77.54	13	18	16.18	18	2	40.62			
11	16	0.08	4	8	212.93	9	4	137.89	13	5	179.34	18	1	12.02			
12	17	0.31	4	9	125.83	9	11	109.81	14	7	115.49	18	6	36.49			
13	3	0.01	4	10	80.08	9	10	104.83	14	9	147.58	18	5	59.02			
13	5	0.05	4	11	169.88	9	14	316.38	14	10	147.34	18	13	37.96			
14	7	0.21	4	16	40.85	9	7	82.34	14	15	187.01	18	15	103.34			
17	8	0.11	5	6	43.30	10	4	63.59	15	7	217.18						

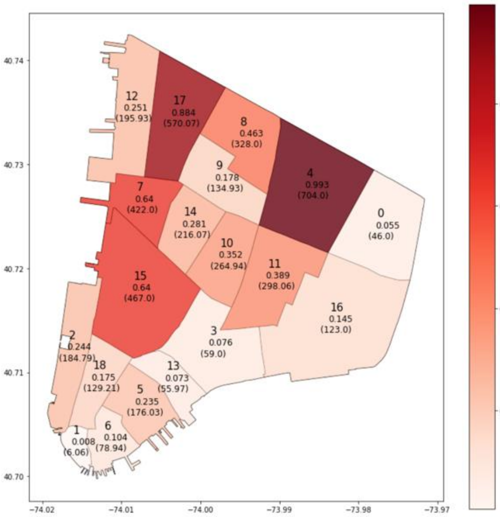


Figure 16. Saturation rates of transfer-in links (0.75 absolute efficiency).

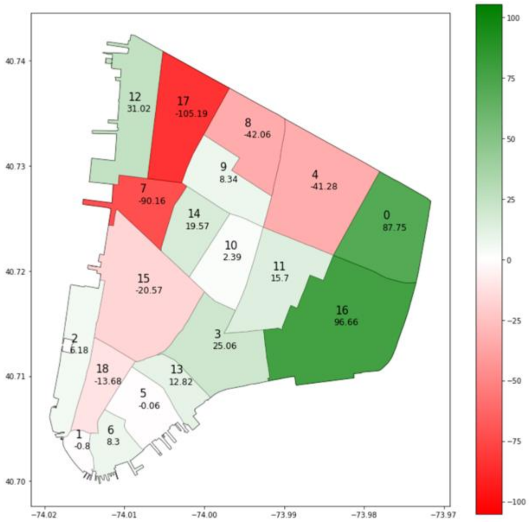


Figure 17. Capacity change ($s_i - s_{0,i}$) of transfer-in links (0.75 absolute efficiency).

4.2 Comparison case: 0.5 absolute efficiency

To test how the efficiency parameters affect the assignment, we change the absolute efficiencies to 0.5 to compare with the base case. Efficiency from all transfer-in links to itself is -0.5, and efficiency from all transfer-out link to the transfer-in link of the same node is 0.5. Path set is the same as base case. Computation configuration and parameter settings are the same as the base case. Run time of **Algorithm 1** is 16min 41sec (6min 30s on path generation, 10min 11s on Frank-Wolfe algorithm). **Algorithm 2** converged after 65 iterations. **Figure 18** compares the saturation rates and flows of transfer-in links with the base case. **Figure 19** compares capacity changes caused by the flows ($s_i - s_{0,i}$) with the base case. The maximum absolute capacity change is 70.13, which is 9.35% of the initial capacity. The changes also falls into a range of localized, incremental change.

With the same path set as the base case in section 4.1, and no binding link (as shown in **Figure 14**), path flows are the same as the base case, determined by undersaturated path costs. The difference is the equilibrium capacities. Comparing the saturation rates of 2 cases (**Figures 16** and **18**), with more rebalancing efforts (0.5 absolute efficiency), zones that lose capacities have lower saturation rates, and zones that gains capacities have higher saturation rates. It can be observed from **Figure 19** that the capacity shift from the northeast to the southwest is less significant compared with the base case due to more rebalancing effort. Capacity distribution becomes less sensitive to the flows, leading to a more even distribution of the capacities.

Remark 2. *Decreases in absolute efficiencies can be observed by the proposed model to lead to more even distribution of capacities, implying more rebalancing or inter-zone matching.*

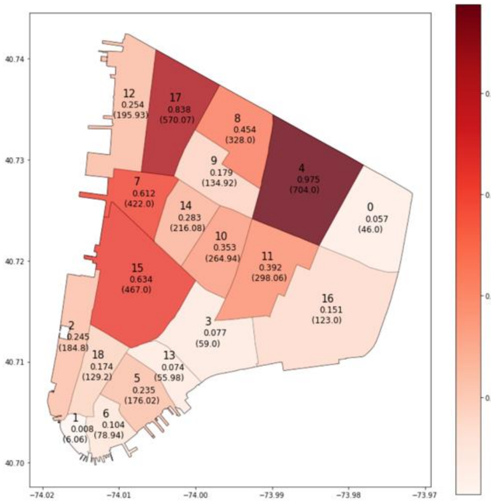


Figure 18. Saturation rate and flows of transfer-in links (0.5 absolute efficiency).

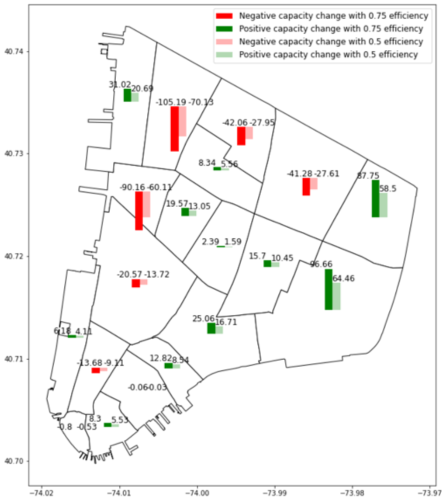


Figure 19. Capacity change ($s_i - s_{0,i}$) of transfer-in links (0.5 absolute efficiency).

4.4 System evaluation with demand change: 95% demand

To demonstrate the application of estimated FC matrices, we changed the travel demand to 95% of the base case (all OD pairs) to evaluate system changes with the same FC matrix from the base case. All the other parameters are kept the same as the base case. Run time of **Algorithm 1** is 17min 41sec (6min 9s on path generation, 11min 32s on Frank-Wolfe algorithm). **Algorithm 2** converged after 75 iterations. **Figure 20** compares the saturation rates and flows of transfer-in links with the base case. **Figure 21** compares capacity changes caused by the flows ($s_i - s_{0,i}$) with the base case.

As shown in **Figure 20**, compared with the base case, all the saturation rates and flows are lower with lower demand. As shown in **Figure 21**, absolute capacity changes are all smaller except zone 5, which is also very close. Smaller demand leads to smaller flows, which has smaller impact on the capacity distributions. This case shows how the model with estimated FC matrices could be applied to evaluate the impact of system demand changes, with unchanged travel and operation patterns.

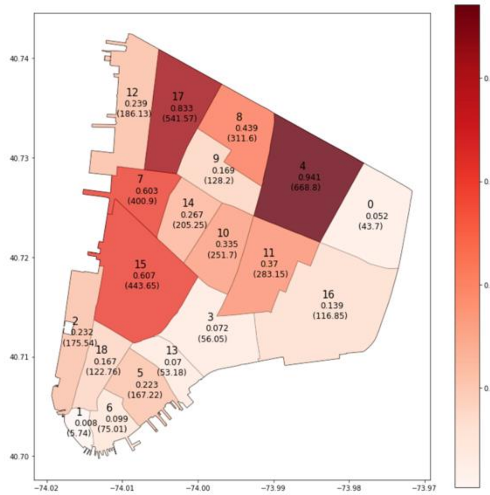


Figure 20. Saturation rate and flows of transfer-in links (0.5 absolute efficiency).

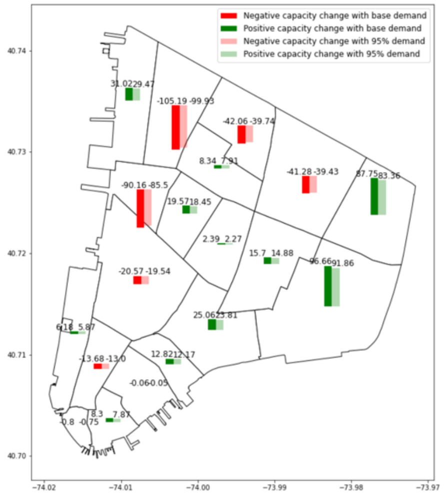


Figure 21. Capacity change ($s_i - s_{0,i}$) of transfer-in links (0.5 absolute efficiency).

4.3 FC matrix estimation

To demonstrate the use of P_2 for estimation of the FC matrices with trip and capacity data, we adopt the equilibrium path flows and capacities solved from the base case in section 4.1 as the observed path flows and capacities to estimate an FC matrix. Two different sets of estimations are applied to demonstrate their impact on the resulting equilibrium capacities: 1) $\beta = 1, \gamma = 1$; 2) $\beta = 1, \gamma = 100$.

The initial guess of the FC matrix is an all-zero matrix for both cases, representing static capacities. In addition to the constraints in P_2, we add the following set of constraints (**Eqs. (37 – 38)**). For all pairs of (i, i) in which i is a transfer-in link, we define a set Z_{in}. For all pairs of (i, j) in which i is a transfer-in link and j is the transfer-out link at the same MoD node, we define a set Z_{out}. These constraints ensure the interpretability of the estimation results by making sure that there is no over-rebalancing (over-compensating the reduced capacities/over-alleviate the increased capacities). All other elements of the FC matrix are assumed to be 0 (no perturbation).

$$-1 \leq p_{ii}^0 + p_{ii}^+ - p_{ii}^- \leq 0, \qquad (i, i) \in Z_{in} \tag{37}$$

$$0 \leq p_{ij}^0 + p_{ij}^+ - p_{ij}^- \leq 1, \qquad (i, j) \in Z_{out} \tag{38}$$

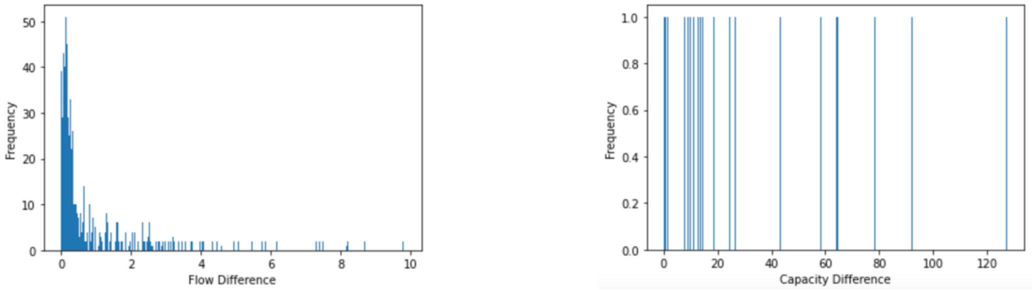
We use the commercial solver LINGO to solve P_2 for these instances. Results are shown in **Table 6**.

Table 6. Estimated efficiencies of the base FC matrix.

Node ID (q)	Efficiencies of transfer-in link (i) flow to the capacity at node q (p_{ii})		Efficiencies of transfer-out link (j) flow to the capacity at node q (p_{ij})	
	Case (1): $\beta = 1, \gamma = 1$	Case (2): $\beta = 1, \gamma = 100$	Case (1): $\beta = 1, \gamma = 1$	Case (2): $\beta = 1, \gamma = 100$
0	0	0	0.179	0.538
1	-0.100	-0.131	0	0
2	-0.007	0	0	0.032
3	0	0	0.068	0.271
4	0	-0.059	0.133	0
5	0	-0.000	0.004	0
6	-0.007	0	0	0.092
7	-0.028	-0.214	0	0
8	-0.047	-0.128	0	0
9	-0.011	0	0	0.057
10	0	0	0.016	0.009
11	0	0	0.009	0.049
12	-0.063	0	0	0.131
13	0	0	0.027	0.176
14	-0.023	0	0	0.081
15	-0.015	-0.044	0	0
16	0	0	0.128	0.384
17	-0.023	-0.185	0	0
18	0	-0.106	0.006	0

The 2 estimated FC matrices lead to the same equilibrium flows, whose absolute errors compared with observations are shown in **Figure 22(a)**. Both estimated FC matrices are very different from the FC matrix defined in the base case which generated the observed flows and capacities, while the equilibrium path flows generated have small differences as shown in **Figures**

22(a). When other system settings are the same, different FC matrices could lead to the same equilibrium flows and capacities as discussed in section 2.5. In this case, since we chose the all-zero matrix as the initial guess, minimizing perturbation leads to less non-zero parameters than expected. Between the transfer-in link and transfer-out link of a MoD node, only one link is estimated to have impact on the capacity for all MoD nodes in both cases. Comparing the two cases, the resulting equilibrium path flows are the same, while accuracy of equilibrium capacities are different. Capacity accuracy comparison of Case 1 and 2 are shown in **Figure 22(b)** and **Figure 23**. Case (2) has significantly more accurate capacities due to a larger weight γ of the capacity-fitting term.



(a) Histogram of path flow differences (observed vs. solved with estimated FC matrix)

(b) Histogram of equilibrium capacity differences (observed vs. solved with estimated FC matrix)

Figure 22. Flow and capacity accuracy of case (1): $\beta = 1, \gamma = 1$.

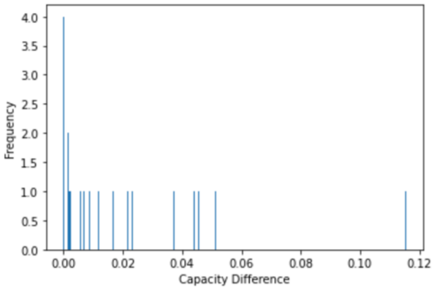


Figure 23. Histogram of equilibrium capacity differences of case (2): $\beta = 1, \gamma = 100$. (observed vs. solved with estimated FC matrix)

Remark 3. *Inverse optimal parameters from even the same initial prior can have the same equilibrium path flows but different capacity accuracies, with higher γ weights for improved capacity fitting.*

In practice, selection of the initial prior of the FC matrix affects the estimation result significantly. Interpretable initial priors representing basic rebalancing/matching policies and customer/driver arrival patterns would help get interpretable results. For example, negative efficiencies for nearby transfer-in links and positive efficiencies for nearby transfer-out links may make sense. Additional constraints can be added to P_2 to ensure interpretability. For example, zones further away can be constrained to have smaller absolute efficiency than zones closer. Efficiency p_{ij} can be assumed proportional to the distance between i and j to help avoid aggregation of congestible capacity effect to one link. In some cases where the capacity distribution data are not available, the capacity-fitting term from the objective of P_2 can be removed.

For real MoD and multimodal networks with path flow and capacity data, the FC matrix can be estimated as a digital profile capturing the combined effect of travel behavior and operation strategy. Different FC matrices with different combinations of travel behavior and operation policies can be estimated to build a library of digital profiles. The library could facilitate planning and design of new/extended service areas by extracting digital profiles that align with the intended service area, which should have similar factors that determine travel behavior (e.g. demographics, land use information, travel surveys). Such digital profiles could be used to model equilibrium flows to compare different operating policies or for use as priors.

5 Conclusion

This study proposed a nonlinear optimization formulation that yields a logit-based SUE for MoD systems with congestible capacities, which is a phenomenon that capacity distribution depends on the flows. The capacitated SUE model from Bell (1995) is generalized to have capacities as functions of flows through an input-output-based FC matrix that captures a combined effect of travel behavior and operation policies, making it possible to model MoD equilibrium without operation information form the MoD providers. Proof of equivalence of the formulation and SUE is given, which shows how the path delays can be obtained through a combination of Lagrange multipliers of the capacity constraints. The uniqueness of the solution is discussed. The model has a simple structure which is easy to scale and apply.

Because of the unique structure of the non-separable capacity effects, balancing algorithms like Bell (1995) are shown to be ineffective. Instead, a solution algorithm that generates a ρ-bounded path set is proposed. A Frank-Wolfe algorithm is applied to solve the nonlinear SUE formulation. An inverse optimization formulation is derived to estimate the FC matrices with observed equilibrium path flows and capacities along with a prior.

Three examples are given, including a small toy network to illustrate the formulation and solution algorithm, one larger numerical example to demonstrate the application to multimodal systems, and one example of downtown Manhattan, NY to show the application to real networks. Sensitivity tests are conducted to demonstrate the influence of the travel time coefficient α and the efficiency parameters which defines how much link flows affect capacities. The proposed estimation method of FC matrix is tested in the downtown Manhattan case.

Compared to the literature, there are 2 major contributions. First, the model provides a generalized equilibrium framework to empirically model equilibrium state interactions of dynamic capacities and flows across modes. Secondly, without knowing the operation policies of the MoD operators, the estimation of the FC matrix requires limited observations without having to make assumptions on operators' policies, which allows modeling for regulators without data shared by operators. Such an approach addresses the problem of MoD equilibrium modeling needed by public policy-makers under the data privacy limitations from private operators. Ultimately, we hypothesized that we could model these complex dynamics in a steady state model using linear functions, much like how complex inter-industry interactions are modeled with simple linear Input-Output models by Leontief (1936) or linear regression models for that matter (the subjects of those models typically are nonlinear to some extent). The point of the design is that it can capture the steady state effects of these multimodal, MoD systems. Through the numerical experimentation with the use of the proposed inverse models, we empirically prove that we can get good fits to the real data that can result in interpretable analysis.

There are some shortcomings that could be addressed in further studies. The proposed logit-based formulation retains the independence of irrelevant alternative (IIA) property, which leads to issues addressing path correlation. Other choice models such as the c-logit logit model could be applied to the SUE formulation. Other more effective path set identifying methods can also be considered. The equilibrium assignment and capacity distribution are dependent on the path set, indicating that path recommendation under a MaaS setting could influence the equilibrium, which could be another future direction. Considering the dependency of feasibility on the relationship between the demand and base capacities, further research could look at determining the feasibility criterion of the assignment, which could be helpful for the initial fleet deployment. In addition, to achieve more accurate modeling, another future direction would be to explore other forms of flow-capacity relationships other than the simple linear form. With a more accurate non-linear model, we would not have to restrict attention to just local and incremental changes, but at the price that we can no longer prove uniqueness or guarantee such an efficient algorithm.

Interpretability of the FC matrices can be poor due to the complex factors that affect them. However, with data from different regions with different operations, it is possible to model the relationship between the FC matrices and the factors affecting them (e.g. demographics, land-use information, travel survey data, and operation policies). A further exploration can be conducted to identify additional counterfactual scenarios that can be analyzed with this model framework.

While this study provides a steady-state model of MoD networks, it does not consider the impact of operation cost and pricing. It can determine the resulting route flows for a given operator policy, not considering its stability with respect to cost allocation options available to the operator. The work of Liu and Chow (2023) does consider cost allocation but maintains simple route flow decisions without congestible capacity effects. Future research may tie these two efforts together with an assignment game model that exhibits customer behavior under congestible capacities.

Acknowledgments

B. Liu and J. Chow were partially supported by the C2SMARTER Center (USDOT #69A3551747124).

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: B. Liu, J.Y.J. Chow, D. Watling; analysis and interpretation of results: B. Liu, J.Y.J. Chow; draft manuscript preparation: B. Liu, J.Y.J. Chow, D. Watling. All authors reviewed the results and approved the final version of the manuscript.

References

Ahuja, R.K., Orlin, J.B. (2001). Inverse optimization. *Operations Research* 49(5): 771–783.

Allahviranloo, M., Chow, J.Y.J. (2019). A fractionally owned autonomous vehicle fleet sizing problem with time slot demand substitution effects. *Transportation Research Part C* 98, 37-53.

Ban, X.J., Dessouky, M., Pang, J.S., & Fan, R. (2019). A general equilibrium model for transportation systems with e-hailing services and flow congestion. *Transportation Research* Part B 129, 273-304.

Bekhor, S., & Prashker, J. N. (2001). Stochastic user equilibrium formulation for generalized nested logit model. Transportation Research Record, 1752(1), 84-90.

Bell, M.G. (1995). Stochastic user equilibrium assignment in networks with queues. Trans. Res. Part B 29(2), 125-37.

Bovy, P.H. (2009). On modelling route choice sets in transportation networks: a synthesis. Transport Reviews 29(1), 43-68.

Burton, D., Toint, P.L. (1992). On an instance of the inverse shortest paths problem. Mathematical Programming 53(1–3): 45–61.

Chan, T. C., Mahmood, R., & Zhu, I. Y. (2023). Inverse optimization: Theory and applications. Operations Research.

Chen, X., Di, X. (2021). Ridesharing user equilibrium with nodal matching cost and its implications for congestion tolling and platform pricing. Transportation Research Part C 129, 103233.

Chow, J.Y.J., Sayarshad, H.R. (2014). Symbiotic network design strategies in the presence of coexisting transportation networks. Transportation Research Part B 62, 13-34.

Chu, Y. L. (2012). Network equilibrium model with dogit and nested logit structures. Transportation research record, 2302(1), 84-91.

Duncan, L. C., Watling, D. P., Connors, R. D., Rasmussen, T. K., & Nielsen, O. A. (2024). Formulation and solution method of bounded path size stochastic user equilibrium models–consistently addressing route overlap and unrealistic routes. *Transportmetrica A: transport science*, *20*(2), 2178240.

Di, X., Ban, X.J. (2019). A unified equilibrium framework of new shared mobility systems. Transportation Research Part B 129, 50-78.

Di, X., Liu, H.X., Pang, J.S., Ban, X.J. (2013). Boundedly rational user equilibria (BRUE): mathematical formulation and solution sets. Transportation Research Part B 57, 300-13.

Di, X., Liu, H.X. (2016). Boundedly rational route choice behavior: A review of models and methodologies. Transportation Research Part B 85, 142-79.

Frade, I., Ribeiro, A. (2015). Bike-sharing stations: A maximal covering location approach. Transportation Research Part A 82, 216-227.

Frank, M., Wolfe, P. (1956). An algorithm for quadratic programming. Naval research logistics quarterly, 3(1-2), 95-110.

Haider, Z., Nikolaev, A., Kang, J.E., Kwon, C. (2018). Inventory rebalancing through pricing in public bike sharing systems. European Journal of Operational Research, 270(1), 103-117.

Hazelton, M. L., & Watling, D. P. (2004). Computation of equilibrium distributions of Markov traffic-assignment models. *Transportation Science*, *38*(3), 331-342.

He, F., Shen, Z.J.M. (2015). Modeling taxi services with smartphone-based e-hailing applications. Transportation Research Part C 58, 93-106.

Jiang, Z., Ouyang, Y. (2022). Pricing and resource allocation under competition in a docked bike-sharing market. Transportation Research Part C 143, 103833.

Ke, J., Yang, H., Li, X., Wang, H., Ye, J. (2020). Pricing and equilibrium in on-demand ride-pooling markets. Transportation Research Part B 139, 411-431.

LeBlanc, L.J., Morlok, E.K., Pierskalla, W.P. (1975). An efficient approach to solving the road network equilibrium traffic assignment problem. Transportation research, 9(5), 309-318.

Leontief, W.W. (1936). Quantitative input and output relations in the economic systems of the United States. The Review of Economic Statistics 18(3), 105-125.

Li, Y., Liu, Y., Xie, J. (2020). A path-based equilibrium model for ridesharing matching. Transportation Research Part B 138, 373-405.

Lin, J.R., Yang, T.H. (2011). Strategic design of public bicycle sharing systems with service level constraints. Transportation research part E 47(2), 284-294.

Lin, J.R., Yang, T.H., Chang, Y.C. (2013). A hub location inventory model for bicycle sharing system design: Formulation and solution. Computers & Industrial Engineering, 65(1), 77-86.

Liu, B., Chow, J.Y.J. (2023). On-demand Mobility-as-a-Service platform assignment games with guaranteed stable outcomes. arXiv preprint arXiv:2305.00818.

Liu, X., Yang, H., Xiao, F. (2022). Equilibrium in taxi and ride-sourcing service considering the use of e-hailing application. Transportmetrica A 18(3), 659-675.

Liu, Z., Yin, Y., Bai, F., & Grimm, D. K. (2023). End-to-end learning of user equilibrium with implicit neural networks. Transportation Research Part C: Emerging Technologies, 150, 104085.

Lou, Y., Yin, Y., Lawphongpanich, S. (2010). Robust congestion pricing under boundedly rational user equilibrium. Transportation Research Part B 44(1), 15-28.

Ma, J., Xu, M., Meng, Q., Cheng, L. (2020). Ridesharing user equilibrium problem under OD-based surge pricing strategy. Transportation Research Part B 134, 1-24.

Mahmassani, H.S., Chang, G.L. (1987). On boundedly rational user equilibrium in transportation systems. Transportation Science 21(2), 89-99.

McCormick, G.P. (1976). Computability of global solutions to factorable nonconvex programs: Part I—Convex underestimating problems. Mathematical programming, 10(1), 147-175.

Nguyen, S., Dupuis, C. (1984). An efficient method for computing traffic equilibria in networks with asymmetric transportation costs. Transportation Science, 18(2), 185-202.

Noruzoliaee, M., Zou, B. (2022). One-to-many matching and section-based formulation of autonomous ridesharing equilibrium. Transportation Research Part B 155, 72-100.

Park, C., Sohn, S.Y. (2017). An optimization approach for the placement of bicycle-sharing stations to reduce short car trips: An application to the city of Seoul. Trans. Research Part A 105, 154-166.

Pel, A.J., Chaniotakis E. (2017). Stochastic user equilibrium traffic assignment with equilibrated parking search routes. Transportation Research Part B 101, 123-39.

Pfrommer, J., Warrington, J., Schildbach, G., Morari, M. (2014). Dynamic vehicle redistribution and online price incentives in shared mobility systems. IEEE T-ITS 15(4), 1567-1578.

Prato, C.G., Bekhor, S. (2006). Applying branch-and-bound technique to route choice set generation. Transportation Research Record 1985(1), 19-28.

Rasmussen, T.K., Watling, D.P., Prato, C.G., Nielsen, O.A. (2015). Stochastic user equilibrium with equilibrated choice sets: Part II–Solving the restricted SUE for the logit family. Trans. Res. Part B 77, 146-65.

Rasmussen, T. K., Duncan, L. C., Watling, D. P., & Nielsen, O. A. (2024). Local detouredness: A new phenomenon for modelling route choice and traffic assignment. *Transportation Research Part B: Methodological*, *190*, 103052.

Sayarshad, H.R., Chow, J.Y.J. (2015). A scalable non-myopic dynamic dial-a-ride and pricing problem. Transportation Research Part B 81, 539-54.

Sayarshad, H.R., Chow, J.Y.J. (2017). Non-myopic relocation of idle mobility-on-demand vehicles as a dynamic location-allocation-queueing problem. Transportation Research Part E 106, 60-77.

Sherali, H.D., Hobeika, A.G., & Kangwalklai, S. (2003). Time-dependent, label-constrained shortest path problems with applications. Transportation Science, 37(3), 278-293.

Sherali, H.D., Ozbay, K., Subramanian, S. (1998). The time-dependent shortest pair of disjoint paths problem: Complexity, models, and algorithms. Networks: An International Journal, 31(4), 259-272.

Singla, A., Santoni, M., Bartók, G., Mukerji, P., Meenen, M., Krause, A. (2015). Incentivizing users for balancing bike sharing systems. In Proceedings of the AAAI Conference on Artificial Intelligence 29(1).

Sun, S., Szeto, W.Y. (2021). Multi-class stochastic user equilibrium assignment model with ridesharing: Formulation and policy implications. Transportation Research Part A, 145, 203-227.

Teale, C. (2020, March 26). Uber sues Ladot over data-sharing requirements. Smart Cities Dive. https://www.smartcitiesdive.com/news/uber-jump-sues-los-angeles-mobility-data-sharing-requirement/574893/

TLC (2021). TLC Trip Record Data. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Watling, D.P., Rasmussen, T.K., Prato, C.G., Nielsen, O.A. (2018). Stochastic user equilibrium with a bounded choice model. Transportation Research Part B 114, 254-80.

Wong, K.I., Wong, S.C., Yang, H. (2001). Modeling urban taxi services in congested road networks with elastic demand. Transportation Research Part B 35(9), 819-842.

Wong, K. I., Wong, S.C., Yang, H., Wu, J.H. (2008). Modeling urban taxi services with multiple user classes and vehicle modes. Transportation Research Part B 42(10), 985-1007.

Xu, H., Pang, J.S., Ordóñez, F., Dessouky, M. (2015). Complementarity models for traffic equilibrium with ridesharing. Transportation Research Part B 81, 161-182.

Xu, S.J., Nourinejad, M., Lai, X., Chow, J.Y.J. (2018). Network learning via multiagent inverse transportation problems. Transportation Science, 52(6), 1347-1364.

Xu, S.J., Chow, J.Y.J. (2021). Online route choice modeling for Mobility-as-a-Service networks with non-separable, congestible link capacity effects. IEEE T-ITS 23(8), 11518-11527.

Xu, S.J., Xie, Q., Chow, J.Y.J., Liu, X. (2021). Empirical validation of network learning with taxi GPS data from Wuhan, China. IEEE ITS Magazine 13(1), 42-58.

Xu, Z., Chen, Z., Yin, Y., Ye, J. (2021). Equilibrium analysis of urban traffic networks with ride-sourcing services. Transportation science, 55(6), 1260-1279.

Xue, Z., Zeng, S. (2019). Equilibrium of the ride-sourcing market considering labor supply. In 2019 16th ICSSSM (pp. 1-6). IEEE.

Yang, H., Wong, S.C. (1998). A network model of urban taxi services. Trans.Res.Part B 32(4), 235-246.

Yang, H., Wong, S.C., Wong, K.I. (2002). Demand–supply equilibrium of taxi services in a network under competition and regulation. Transportation Research Part B 36(9), 799-819.

Yang, H., Leung, C.W., Wong, S.C., Bell, M.G. (2010). Equilibria of bilateral taxi–customer searching and meeting on networks. Transportation Research Part B 44(8-9), 1067-1083.

Yang, H., Yang, T. (2011). Equilibrium properties of taxi markets with search frictions. Trans. Res. Part B 45(4), 696-713.

Yen, J.Y. (1971). Finding the k shortest loopless paths in a network. Management Science 17(11), 712-716.

Zhang, Y., Khani, A. (2021). Integrating transit systems with ride-sourcing services: A study on the system users' stochastic equilibrium problem. Transportation Research Part A 150, 95-123.

Zhang, J., Meng, M., Wang, D.Z., Zhou, L., Han, L. (2023). Optimal bike allocations in a competitive bike sharing market. Journal of Cleaner Production, 384, 135602.

Zhou, Z., Chen, A., & Bekhor, S. (2012). C-logit stochastic user equilibrium model: formulations and solution algorithm. Transportmetrica, 8(1), 17-41.