Design of Multimodal Network for Mobility-as-A-Service: First/Last Mile Free Floating Bikes and On-Demand Transit

Article *in* SSRN Electronic Journal · January 2017

DOI: 10.2139/ssrn.3075645

Qi Luo
University of Michigan

Design of Multimodal Network for Mobility-as-a-Service: First/Last Mile Free Floating Bikes and on-Demand Transit

Shukai Li[a], Qi Luo[b], Robert Hampshire[b,c,*]

[a] *Peking University*
[b] *Industrial & Operations Engineering, University of Michigan*
[c] *Transportation Research Institute, University of Michigan*

Abstract

Mobility-as-a-Service (MaaS) is a solution that integrates multiple modes of transport into seamless trip chains. In a trip chain, the smooth transfer from one mode to another is critical. In particular, we study a multimodal system in which passengers use free-floating bikes for first/last-mile, and transfer to one-demand transport (shuttles or taxis) at certain hubs. We formulate the fleet management as a bi-level optimization on a closed queuing network. The objective is to minimize the capital expenditures as well as the operating expenses. On the pricing level, we use a corrected approximation method to calculate the stationary distributions of vehicles in a fixed network, and find the minimum expense and its corresponding prices. On the hub selection level, a greedy algorithm finds the near-optimal transfer hub locations. Under mild assumptions, we are able to find the near-optimal prices and network structures in a case study of New York City.

Keywords: Multimodal transportation, network pricing, free-floating bike-sharing, on-demand transport, closed network.

1. Introduction

Mobility-as-a-Service (MaaS) allows a shift from personally-owned vehicles towards easy mobility services by combining transportation services from

*Corresponding Author. Tel: (734) 763-7746
Email addresses: `luoqi@umich.edu` (Qi Luo), `hamp@umich.edu` (Robert Hampshire)

public and private providers through a unified way. MaaS is promoted by a number of emerging low-cost mobility modes including free-floating bikes, carpool, and e-hailing services. Such services aim to provide more affordable mobility solutions by expanding supply from idle vehicles or capacity. In a MaaS system, users usually pay for the transport service with a single account. Expenses are often measured by time or distance and can be lowered through customers' payment share. In addition, MaaS enables the integration of multiple modes of transport into seamless trip chains.

Among all modes of mobility services, the bike-sharing system is one of the most popular, which has been installed in many cities of the world. Free-floating bikes are known for its advantage of convenient parking and storage as well as easy maintenance (Pal and Zhang (2017), Reiss et al. (2015)). One major limitation for bike-sharing is its short-distance-oriented nature. Carpool, another popular mobility mode, is the sharing of one car by more than one travelers. It makes better use of the originally underused taxi capacity due to travelers who travel alone and thereby is more cost-saving compared with conventional taxi service. Carpool has at least two basic forms: one is car owners giving a ride to riders (termed as "ride-hailing"), and the other is more than one customer taking on-demand transport together (e.g., dial-a-ride shuttles, taxis) (termed as "ride-pooling"). Distinguishing from ride-hailing or public transport, in the ride-pooling scenario, drivers play no role in the route selection. One major challenge for carpool is the matching of sparsely distributed customers.

These shared mobility services are viable in populated cities thanks to the introduction of information technology. However, under-use of the asset (especially free-floating bikes) and inadequate allocation of service providers potentially impede moving towards an ecosystem for MaaS. On the top of that, there are arguments on the impact of shared mobility on traffic congestion (Li et al. (2016)), as well as social welfare (Hampshire et al. (2017)). Therefore, MaaS operators appeal to integrate different sources as an unified system to enhance the utilization of all modes. For example, Beijing announced a strict regulation on the qualification of private car drivers for DiDi Chuxing, the largest transportation network company in China, to reduce the traffic congestion in the city in 2017 (Shirouzu et al. (2016)). On the other hand, free-floating bike-sharing and on-demand shuttle are rising rapidly. Users of the previous service may change their traveling habits because of these regulations, and the company consequently provides an interface that combines bike-sharing with existing ride-hailing or public transport to adapt to this transition. Compared to connections of bikesharing to public transport literature, there is no systematic way to optimize this aggregation of two or more modes of shared mobility.

We aim to build a multimodal transport system that combines ride-pooling with free-floating bikes. Travelers use free-floating bikes for first/last-mile (FM/LM), then "pool" together to on-demand transport at hubs as the "line haul", as shown in Figure 1. Such a multimodal system is advantaged in speed compared with biking or public transport, and more cost-saving than conventional taxi service. In our study, we use taxi service as an example of the on-demand transport in our multimodal transport system. However, in other scenarios, the on-demand transport can be other services like dial-a-ride shuttles.

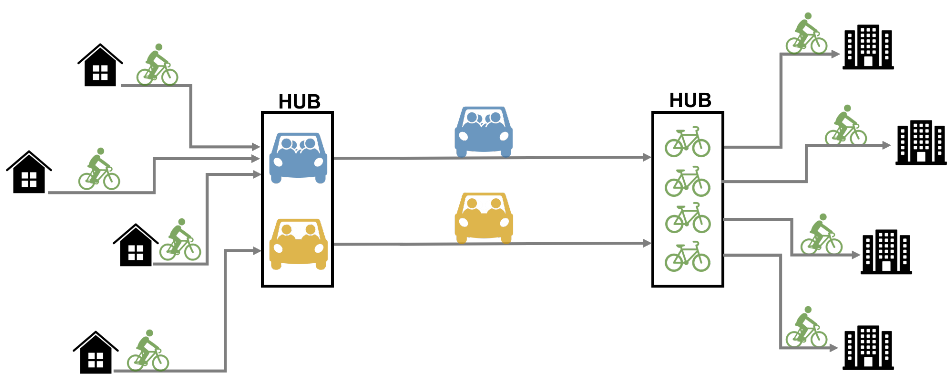

Figure 1: Customers ride a bike in both the first and last miles, and ride-pool between these two hubs.

In this paper, Section 2 summarizes the most relevant contributions provided by the literature on MaaS and multimodal transport service. Section 3 summarizes our proposed multimodal system's behavioral impacts on customers. Section 4 develops the the model. Section 5 provides the approximation method and correction. In Section 6, we use NYC as a case study. Conclusions are reported in Section 7.

2. Literature Review

Many studies have investigated the operations of the bike-sharing and on-demand transit systems. The major concerns with the station-based bike-sharing system design are the hub location, inventory, and re-balancing of bikes over time such that the appropriate number of bikes and docks are available to users. Banerjee et al. (2016) formulate a general approximation framework to design pricing policies in vehicle-sharing systems. Lin et al. (2013) develop a hub location inventory model. The key design decisions considered include the number and locations of bike stations, the creation of bicycle lanes, and users' route selection. Schuijbroek et al. (2017) combine two aspects that have previously been handled separately by the literature: determining service level requirements at bike stations, and designing (near-)optimal bike routes to re-balance the bikes. Pal and Zhang (2017) tackle the static re-balancing with one vehicle. They present a hybrid nested large neighborhood search with variable neighborhood descent algorithm. Freund et al. (2017) deals with a more strategic question of how to (re-)allocate dock-capacity in bike-sharing systems. The impacts of free-floating bike-sharing are also studied as it grows to be popular in recent years. Faghih-Imani et al. (2017) study the factors that affect the usage of the bikes. They employ a mixed linear model to estimate the impact of bicycle infrastructure, sociodemographic status and land-use characteristics on customers' demand for bikes.

As for carpooling in a ride-sharing system, Stiglic et al. (2015) introduce meeting points into the transportation network. In their proposed system, riders can be picked up and dropped off by drivers either at their origin, destination or at a meeting point which is within a certain distance from their origin or destination. They find meeting points significantly increase efficiency through additional matches of drivers with riders as well as less stops drivers needs to make. Santi et al. (2014) quantify the benefit of vehicle pooling by analyzing the New York City taxi trip data. The simulation results reveal the vast potential of a new taxi system where trips are routinely combined with mildly prolonging average travel time.

A handful of studies have investigated the integration of different modes of transportation service. Stiglic et al. (2018) examine the integration of ride-sharing with public transit in suburban areas. They find the proposed system can significantly improve mobility, increase the usage of public transport, and reduce the negative externalities related to car travel. Miramontes et al.

(2017) examine customers' acceptance of the Mobility Station, a multimodal mobility hub connecting public transport with new shared mobility services in the City of Munich, as well as customers' short and long term effects on mobility behavior. The results provide some insights for the awareness of the multimodal traveling system among users. For example, most users are young and highly educated. Campbell and Brakewood (2017) evaluate the impact of bike-sharing systems on bus ridership. They employ a natural experiment of the phased implementation of a bike-sharing system to different areas of NYC and find either bike-sharing members are substituting bike-sharing for bus trips or that bike-sharing may affect the travel choice of non-members like private bicyclists.

According to our knowledge, although many studies have examined bike-sharing systems connected with public transport, no literature has investigated integrating bike-sharing with on-demand mobility modes. This paper contributes by proposing a new multimodal transport system where passengers use free-floating bikes for first/last-mile, then pool together at on-demand transport at hubs, where vehicles are waiting there. We aim to achieve maximum operations profit from the multimodal system by pricing as well as optimizing the system's network. In this paper, we use taxi service as an example of the on-demand transport in our multimodal transport system. However, in other scenarios, the on-demand transport can be other services like dial-a-ride shuttles.

3. Behavioral impact on travelers

In general, whether an individual chooses the taxi service is decided by two factors: (a) U_s the utility obtained from the service; (b) C_s the payment for the service. For example, a conventional assumption is the individual will choose the service only if $U_s - C_s > 0$, where U_s depicts the customer's willingness to pay. From this perspective, customers in taxi trip data can be regarded as having a high utility or willingness to pay for taxi service, while those who rarely take a taxi or have to take the public transportation have a relatively lower willingness to pay. Our multimodal system also aims to make use of such underused demand due to customers whose willingness to pay is below the cost of taking taxi alone by providing customers with the affordable alternative of ride-pool service. Customers with higher willingness to pay can still choose to hail a taxi alone for less waiting time, while those with lower willingness to pay can choose the ride-pool service for less expense. Note our multimodal service system also benefits non-ride-pool customers by the advantage of zero waiting time and the reliability of waiting taxis at hubs as well as the potentially lower payment compared with conventional taxi services. This also indicates that travelers' choice between our multimdal system and public transport is complementary as public transport has fixed routes with large fleet and our system is more flexible in connecting hubs that have no direct public transport in between.

From service providers' side, participating in this hub-based ride is also valuable. By waiting in the line at hubs, they have shorter expected waiting time for the next ride, as well as saving energy and gasoline in cruising for passengers. Consequently, the traffic congestion caused by these activities is also reduced, which has otherwise contribute significantly to local congestion (Wong et al. (2014)).

We also divide ride-pool customers into "followers" and "starters", and offer starters the priority to decide the terminal hub. Followers can only join starters whose terminal hub is a good fit, but compared with starters, followers wait for less time. Since starters and followers have different priorities, we impose starters and followers different weight for payment share, and allow customers themselves to decide whether to become a starter or follower based on their preference in terminal priority, waiting time and payment share.

4. Model

4.1. Setting of Multimodal Service System

4.1.1. Networks of bike stations and taxi hubs

As shown in Figure 2, we model the whole travel system as a multimodal system with two connected parts: bike-sharing system and taxi-hailing system. In our model, the whole city is divided by a grid into different regions: $\mathcal{I} = \{1, 2, \ldots, I\}$. The center of each region is a bike station, and bikes only travel between stations. Among all bike stations, a subset of them, denoted as $\boldsymbol{h}$, are selected to function as taxi hubs at the same time. Hubs are indexed by r. Taxis only move between hubs, and empty taxis keep waiting at hubs for customers.

4.1.2. Customers' route choices

In the system, a customer who travels from his local region i to his final destination region j will first ride a bike from the bike station at region i to a starting hub around, then take a taxi to a terminal hub near region

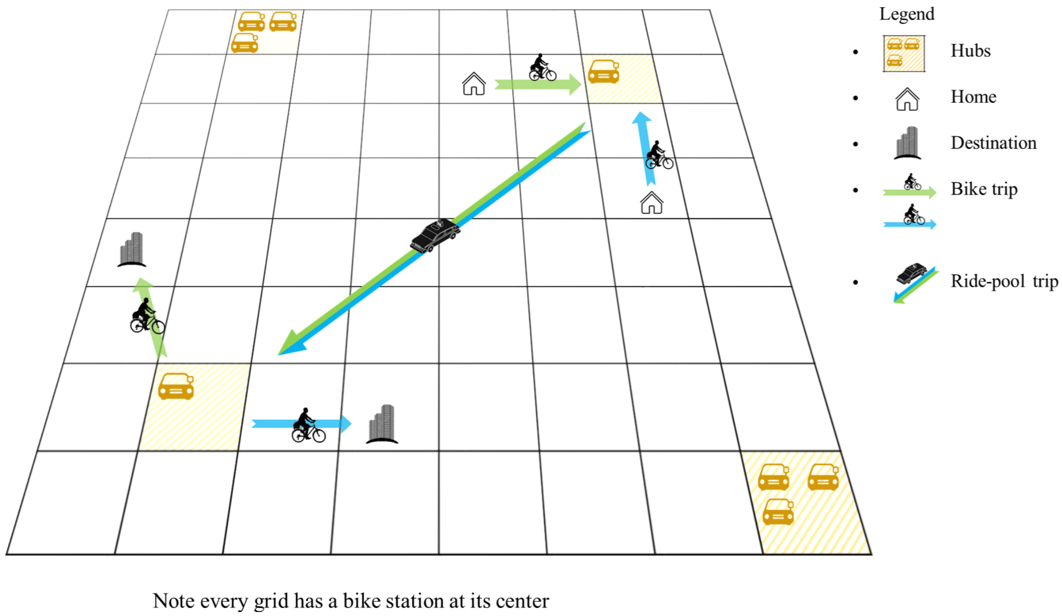

Figure 2: The multimodal service system consists of a bike-sharing system and a taxi system, and provides ride-pool service for customers

j, and then ride another bike to region j. When he arrives at the starting taxi hub, he can choose either to take a taxi alone for less waiting time, or to ride-pool with others for less expense. Note the customer does not necessarily ride a bike if his local region i or his final destination region j exactly has a taxi hub. For each customer, the starting hub is randomly selected among hubs within the distance of a threshold biking time t_1 from the customer's local region i. Closer hubs have a higher chance to be selected as the starting hub, and t_1 represents the maximum biking time customers accept. Specially, when region i exactly has a taxi hub, that hub will be selected as the starting hub. The terminal hub will be the nearest hub within the distance of a threshold biking time t_2 from the final destination region j. Specially, when the customer finds the terminal hub within the distance of t_1 from region i, he will directly ride a bike to region j instead of taking a taxi between. Customers who fail to find a feasible path are missed, which indirectly imposes a loss to the system. This route choice mechanism are presented in Figure 3.

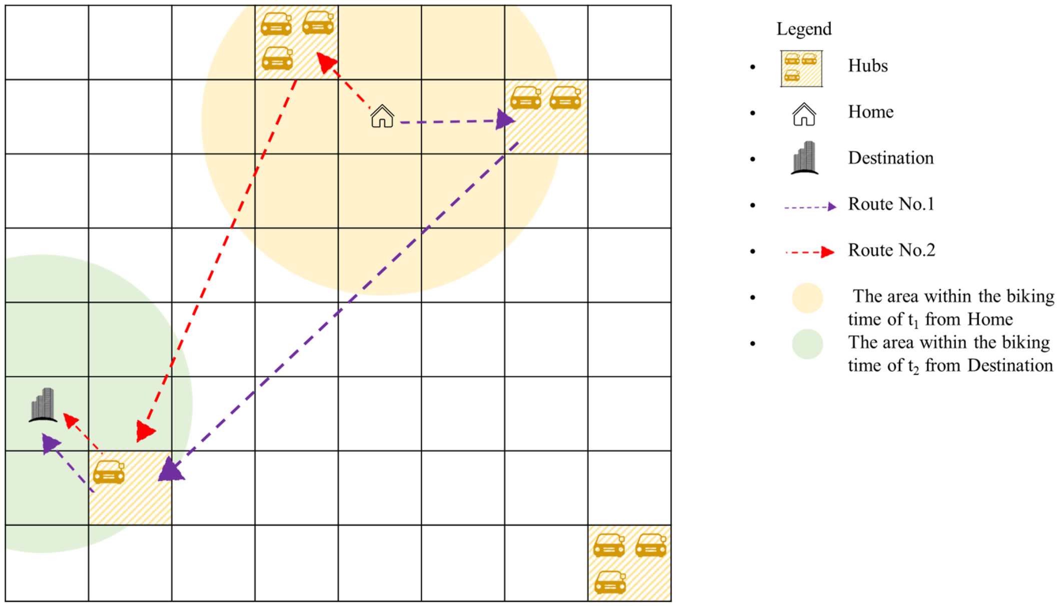

Figure 3: Customers select one hub within the distance of t_1 from Home as the starting hub, and select the nearest hub within the distance of t_2 from Destination as the terminal hub.

4.1.3. Customers' taxi service choices

As shown in Figure 4, customers at taxi hubs are provided with two service choices: 1) he can choose the ride-pool service, and then go through a matching system for ride-pool mates before hailing a taxi, or 2) he can hail a taxi directly without ride-pool. Ride-pool customers can be "followers" or "starters". When a new follower joins the matching system, he will search for starters whose terminal hub r_1 matches with his terminal hub r_2 under a certain matching rule $\Phi(r_1, r_2)$, where $\Phi(r_1, r_2) \in \{0, 1\}$. One specific case is $\Phi(r_1, r_2) = 1$ when and only when biking from hub r_1 to r_2 takes time less than a threshold time t_M. Starters have the priority to decide the terminal hub for the trip, and when a follower successfully matches with a starter, his terminal hub will be adjusted to be the same with the starter's. Each starter has a maximum number C_M of followers who can ride-pool with him due to reasons including taxi capacity. If there is no fitting starter who can take on extra followers, new followers will wait in a queue for fitting starters.

Starters do not join others and are required waiting in the matching system for certain time for potential ride-pool mates, after which the starter goes to hail a taxi with his followers. Such a group of customers with one starter and some followers are referred to as a "ride-pool group". A "taxi order" is defined as one request for taxi from a non-ride-pool customer who has just finished the FM stage, or one request from a ride-pool group who have just completed the matching stage. All trips are made by individuals acting independently of each other. Starters and followers share the payment for their taxi trip by weight w_A for starters and unit for followers.

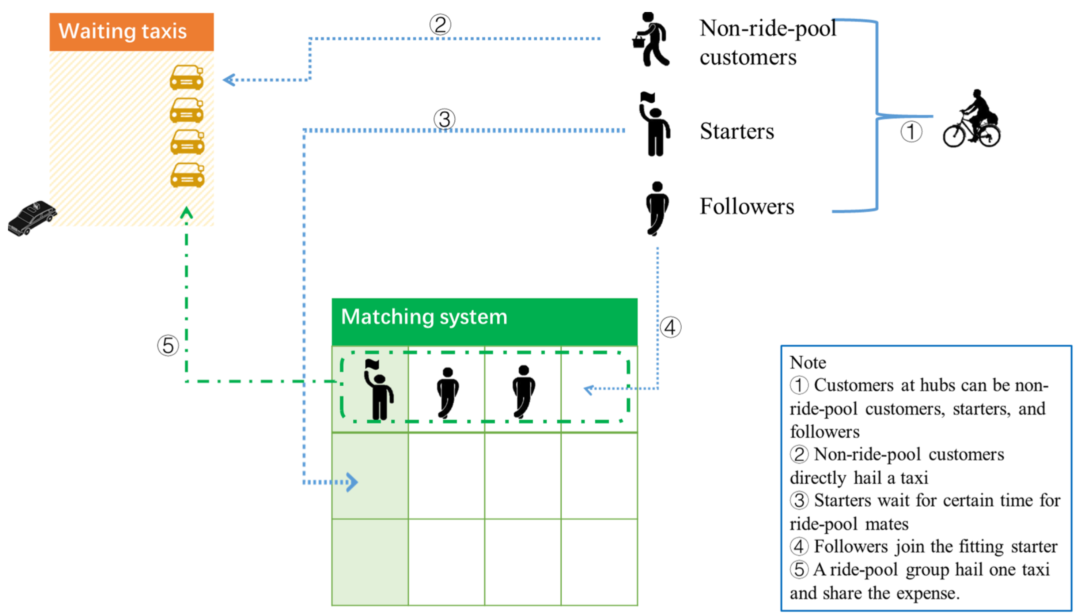

Figure 4: Customers at taxi hubs can hail the taxi for less waiting time ("non-ride-pool customers"), or wait in the matching system for ride-pool mates for less expense ("followers" and "starters").

4.2. Bilevel Optimization for Network Design

We assume the operator's profit only comes from taxi operation: each taxi in operation will charge its customers for price p per unit time. Note ride-pool customers can share such payment and thereby reduce personal traveling cost. Taxis waiting at hubs do not bring profit. Costs have two sources: 1) penalty ℓ each time a non-ride-pool customer or a ride-pool group find there is not available taxi at their hub when they want to hail a taxi; 2) cost for establishing and maintaining hubs. Both ride-pool and non-ride-pool customer's demand rates are impacted by the hub network desgin and price p. In addition, given the hub network and price p, the demand rates of starters and followers are also impacted by matching rule Φ, their expected waiting time, and the payment share policy w_A. In our model, we assume π is fixed due to exogenous factors. The mechanism of how these factors influence starters' and followers' demand rates will be discussed below.

From the operator's perspective, the objective of network design is to maximize the net profit or to minimize the net costs in establishing and operating the multimodal system. The operating expenses, however, are embedded in the structure of the hub network. To decouple the two types of costs, we use a bilevel optimization: In the inner level, we minimize the net costs for any given set of hub locations; In the upper level, we select hub locations to minimize the total net costs.

4.2.1. Inner Optimization

In this section, we formulate the pricing in a closed queuing network as a non-convex optimization to minimize the operating costs. Fixed hubs connected by routes of shared vehicles formulate a network. Passengers arrive at each hub, with appropriate relaxation below, following Poisson processes. The operator's object in inner optimization is to minimize the total cost under the stationary distributions.

Relaxing FM/LM bike-sharing inventory constraint. Assuming that free-floating bike-sharing is a more voluminous resource comparing to ride-sharing, we model the clustering (biking first-mile to nearest hubs) and declustering (biking last-mile from nearest hubs) processes as follows. Each region $i \in \mathcal{I}$, the intensity of travel demand is $\tilde{\Lambda}_i$. For each hub $r = 1, 2, \ldots H$, the regions where bikeshares are assigned to this hub is denoted as $\mathcal{I}_r \subset \mathcal{I}$. The resulting departing intensity of taxis from hub r to others is Λ_r. By Poisson process assumption, the probability that a biking trip from region i to hub r is $\tilde{\Lambda}_i/\Lambda_r$ and $\Lambda_r = \sum_{i\in\mathcal{I}_r} \tilde{\Lambda}_i$. Reversely, the arrivals rate of taxis at the hub r is μ_r and $\mu_r = \sum_{i\in\mathcal{I}_r} \tilde{\mu}_i$. We denote the number of bikes available at each region at time t as $\boldsymbol{X}_r(t) = \{X_i(t); i \in \mathcal{I}_r\}$, and each element X_i is subject to a birth-death process with birth rate $\tilde{\mu}_i$ and death rate $\tilde{\Lambda}_i$. Supposing that the system operator can manipulate either the first-mile or last-mile biking path by control $\boldsymbol{\alpha}$ to all $i \in \mathcal{I}_r$. For simplicity of notations, we assume that $\tilde{\Lambda}_i$ is fixed and $\tilde{\mu}_i(\alpha_i)$ for all i[1]. We briefly prove that there exist a deterministic control policy with finite control cost to rebalance bikes without losing any customer at stationary states[2].

A stationary policy to control this system is a sequence of $\{\boldsymbol{\alpha}_t; t \geq 0\}$ of transition probabilities associated with initial states $\boldsymbol{X}_r(0)$. The cost function associated with controls $\boldsymbol{\alpha}$ is $c_t(\boldsymbol{X}, \boldsymbol{\alpha})$, and the β-discounted cost function is:

$$V_{\boldsymbol{\alpha}}(\boldsymbol{x}_0) = \mathbf{E}\left[\sum_{t=0}^{\infty} \beta^t c_t(\boldsymbol{X}(t), \boldsymbol{\alpha}(t)) | \boldsymbol{X}_0 = \boldsymbol{x}_0\right]. \tag{1}$$

It is a well-known result from control of Markov Chain that there exists a deterministic stationary optimal β-discounted average cost per-unit-time policy $\{\boldsymbol{\alpha}(t); t \geq 0\}$ in infinite horizon (i.e. $\boldsymbol{\alpha} = \arg\inf_{\boldsymbol{\alpha}} V_{\boldsymbol{\alpha}}$). We can also find a deterministic policy satisfying that the stationary distribution $P(\boldsymbol{X}(\infty)) = 0$[3]

We have the optimality equation:

$$\begin{aligned} V(x_0) &= \lim_{t\to\infty} V_t(x_0) \qquad (2) \\ &= \lim_{t\to\infty} \min_{\boldsymbol{\alpha}} \{c(\boldsymbol{x}(t), \boldsymbol{\alpha}) + \beta \sum_{\boldsymbol{x}(t+1)} p(\boldsymbol{x}(t)|\boldsymbol{\alpha}, \boldsymbol{x}(t-1)) V_{t-1}(\boldsymbol{x}(t-1), \boldsymbol{\alpha})\}. \end{aligned}$$

The exact policies are not of interest in this paper and we refer previous literature with comprehensive discussion on the bikeshare rebalancing policies (Schuijbroek et al. (2017), Spieser et al. (2016)). The resulting arrivals of customers who have just finished the FM stage at each hub r follow Poisson processes with intensity Λ_r. Assume for any trip from region i to region j, the arrivals of non-ride-pool customers, starters, and followers independent during their FM stage. Then at any hub r_1, the arrivals of non-ride-pool

[1] In practice, it can be interpreted as the system operator offers an incentive mechanism so that passengers will park the bikes to rebalance the bike shares locally.

[2] More strictly, there exists controls $\boldsymbol{\alpha}$ for all regions associated with hub r such that $P(X_i = 0) = 0$ for all $i \in \mathcal{I}_r$. Here we presume that the initial number of bikes at each region i satisfies $X_i(0) \neq 0$.

[3] Under the boundedness condition (Serfozo (1981)) $\lim_{t\to\infty} \inf_{\boldsymbol{\alpha}} \mathbf{E}_{\boldsymbol{\alpha}} \left[\sum_{k=t}^{\infty} \beta^t \max\{0, c_t(\boldsymbol{X}(k), 0)\} | \boldsymbol{x}_0\right] = 0$.

customers, starters, and followers who have just finished their FM stage and wants to go to hub r_2 follows Poisson process.

Poisson arrivals of taxi orders. Let $d_{r,s}^{(k)}(t)$ be the departure rate of the k th starter who wants to go to hub s but is now waiting in the matching system at hub r at time t, and the total number of such starters are $n_{r,s}(t)$. To simplify the system, we assume $\sum_{k=1}^{n_{r,s}(t)} d_{r,s}^{(k)}(t)$ equals a constant $D_{r,s}$ when $n_{r,s}(t) > 0$. In practice, the system may assign $D_{r,s}$ to all starters weighted by the number of followers each starter has matched. With these assumptions, the number of starters who wants to go to hub s but is now waiting at hub r follows the $M/M/1/\infty$. Thus, the departures of ride-pool groups from the matching system follow Poisson process. Recall the arrivals of non-ride-pool customers who has just finished the FM stage also follows Poisson process. Then "taxi orders" for trips from hub r to hub s also comes in a Poisson process. As will be discussed below, this pleasant property of ride-pool groups' Poisson departures will largely simplify future calculation. By contrast, one alternative idea is to assume starters can "follow" starters, and to allow the ride-pool group to leave once its number of customers reaches C_M. In this case, we may have finite customers waiting in the system. However, the Poisson departure no longer holds.

Theorem 1 *Consider an irreducible queuing network $\boldsymbol{Q}$ which has a set of departure flows $\mathcal{L}$ and a set of states $\mathcal{S}$. Suppose the departure flow $l \in \mathcal{L}$ happens only when the state changes. For l, if there exists a state $s \in \mathcal{S}$ such that any other state's changing into s cannot result in a departure of l, then l cannot form a Poisson process.*

Corollary 1 *Consider an irreducible queuing network $\boldsymbol{Q}$ which has a set of departure flows $\mathcal{L}$ and a set of states $\mathcal{S} = \{\boldsymbol{s}_i\}$, where $\boldsymbol{s}_i$ is defined by $(q_1, q_2, ..., q_m) \in A \subseteqq \mathbb{N}^m$. Suppose the departure flow $l \in \mathcal{L}$ happens only when at least one of $\{q_k\}$ stricly decreases. Then if $|\mathcal{S}| < \infty$, l cannot form a Poisson process.*

Let (s, f) be the vector of the number of starters and followers in the matching system who want to go from hub r to hub s. Since the departure of a ride-pool group will decrease either s or f, and $|\{(s, f)\}| \leqslant (C_M + 1)^2$, the departure flow of ride-pool groups cannot form a Poisson process.

Impact factors for customer flows. As discussed above, both ride-pool and non-ride-pool customer's demand rates are impacted by the hub network design. In addition, given the hub network, $\mu^{(A)}_{r,s}$ the demand rates of starters from hub r to s, and $\mu^{(B)}_{r,s}$ the demand rates of followers from hub r to s are also impacted by their expected waiting time $T^{(A)}_{r,s}$ for starters and $T^{(B)}_{r,s}$ for followers, p price, w_A the payment share policy , $l_{r,s}$ the expectation of the number of followers with each starter. $T^{(A)}_{r,s}$, $T^{(B)}_{r,s}$, and $l_{r,s}$ are decided by the matching rule $\Phi(\cdot)$, $D_{r,s}$ the requested waiting time, $\{\mu^{(A)}_{r,s}\}$, $\{\mu^{(B)}_{r,s}\}$ and C_M the maximum number of followers allowed. Let square matrix $\boldsymbol{\phi} = (\phi_{r,s})$, where $\phi_{r,s} = \Phi(r,s)$. Let matrix $\boldsymbol{M}^{(A)} = (\mu^{(A)}_{r,s})$, $\boldsymbol{M}^{(B)} = (\mu^{(B)}_{r,s})$, and $\boldsymbol{D} = (D_{r,s})$. Then, we can write these relationships as:

$$\begin{cases} \boldsymbol{M}^{(A)} = f(\boldsymbol{M}^{(A)}, \boldsymbol{M}^{(B)}, \boldsymbol{\phi}, \boldsymbol{D}, C_M, w_A, p) \\ \boldsymbol{M}^{(B)} = g(\boldsymbol{M}^{(A)}, \boldsymbol{M}^{(B)}, \boldsymbol{\phi}, \boldsymbol{D}, C_M, w_A, p) \end{cases} \tag{3}$$

$\boldsymbol{M}^{(A)}$ and $\boldsymbol{M}^{(B)}$ are the solutions of the above function as well as the demand rates when both starters' and followers' flow reach an equilibrium. $(\boldsymbol{\phi}, \boldsymbol{D}, C_M, w_A)$ can be summarized as a general ride-pool policy variable $\pi \in \Pi$. In our model, we assume π is fixed due to exogenous factors, and the above two equations has one unique solution (or the starters' and followers' flows have one unique equilibrium) for any given π and p. Then, we can simplify the above relationships as

$$\begin{cases} \boldsymbol{M}^{(A)} = F(p) \\ \boldsymbol{M}^{(B)} = G(p) \end{cases} \tag{4}$$

In our model, we assume functions $F(\cdot)$ and $G(\cdot)$ are already obtained based on data from field surveys.

Stationary Rideshare Distribution in Closed Networks. With the above assumptions, we model the system as a time varying closed queuing network. The key input parameters of the closed queueing network model are: $\boldsymbol{h}$ a configuration of hubs chosen for bike-taxi transfer, each indexed by r, $|\boldsymbol{h}| = H$; λ^0_r, λ^1_r, and λ^2_r respectively the arrival rate of non-ride-pool customers, starters, and followers at hub r; P^0_{rs}, P^1_{rs}, and P^2_{rs} respectively the conditional probability that a non-ride-pool customer, a starter, and a follower want to go to hub s, given they arrive at hub r. Note all above arrivals refer to those of customers who have just finished the FM stage or the matching stage. λ_r^0, λ_r^1, λ_r^2, P_{rs}^0, P_{rs}^1, and P_{rs}^2 are all functions of price p. Let μ_{rs} be the average arriving rate during a taxi trip from hub r to s; N the total number of taxis in the network; $\boldsymbol{n} \in \mathbb{R}^H$ a sample of possible outcome of waiting taxis at each hub. Each element n_r stands for the number of taxis waiting at hub r, $n_r \in \mathbb{N}$. $\|\boldsymbol{n}\|_1 \leq N$; $\aleph$ is sample space for $\boldsymbol{n}$. Measure of each configuration is $P(\boldsymbol{n})$, and v_r is an intermediate variable interpreted as the relative mean rate of departure from hub r.

For simplicity of notation, we denote

$$\lambda_r = \lambda_r^0 + \lambda_r^1 \tag{5}$$

$$P_{rs} = \frac{\lambda_r^0 \cdot P_{rs}^0 + \lambda_r^1 \cdot P_{rs}^1}{\lambda_r}, \tag{6}$$

where λ_r can be interpreted as the arrival rates of taxi orders at station r, and P_{rs} can be interpreted as the conditional probability that a taxi order is to go to hub j, given it is from hub i. Note $\{\lambda_r\}$ and $\{P_{rs}\}$ are all functions of price p. Denote matrix (P_{rs}) as $\bar{\boldsymbol{P}}$. Posner and Bernholtz (1968) show that the measure of possible outcome of waiting taxis at hubs in the steady state is:

$$P(\boldsymbol{n}) = C_0 \prod_{r=1}^{H} \left(\frac{v_r}{\lambda_r}\right)^{n_r} \frac{\hat{W}^{N-\sum_{i=1}^{H} n_i}}{\left(N - \sum_{i=1}^{H} n_i\right)!}, \tag{7}$$
$$\text{where } v_s = \sum_{r=1}^{H} P_{rs} v_r \text{ and } \hat{W} = \sum_{r=1}^{H}\sum_{s=1}^{H} \frac{v_r P_{rs}}{\mu_{rs}}.$$

Pricing on a fixed network. Unit price is set to be uniform for all routes. For a fixed network h whose cardinality is H, the inner optimization's objective function is:

$$\min_p \sum_{\boldsymbol{n}\in\aleph} P(\boldsymbol{n}) \sum_{r=1}^{H} \mathbb{1}\{n_r = 0\} \cdot \lambda_r \cdot \ell - \sum_{\boldsymbol{n}\in\aleph} P(\boldsymbol{n})(N - \sum_{r=1}^{H} n_r) \cdot p. \tag{8}$$

The first term is the penalty for missed customers who find no taxi waiting at the hub when they request for one. The second term is the negative value of expected profit that collected from taxis on road.

As the penalty for losing passengers is fixed, p can also be interpreted as the implicit trip fare multiplier. The elasticity curve of λ_r in terms of price p follows the Constant Elasticity of Substitution curve: λ_r approaches zero when price goes to infinity.

4.2.2. Upper-Level Optimization

Define that a $\boldsymbol{h} \in \mathcal{I}$ is irreducible if every hub in $\boldsymbol{h}$ can be reached by a taxi from every other hub after some trips with its customers. Define Γ as $\{\boldsymbol{h} \in \mathcal{I} | \boldsymbol{h} \text{ is irreducible}\}$. Since we aim to establish a sufficiently connected hub network, we only view all $\boldsymbol{h} \in \Gamma$ to be potential hub locations. Then, the upper-level network design problem is

$$\min_{\boldsymbol{h}\in\Gamma}\Bigg\{\boldsymbol{c}_h^{\intercal} \cdot \mathbb{1} + \min_{p}\Big\{\sum_{\boldsymbol{n}\in\aleph} P(\boldsymbol{n}) \sum_{r=1}^{|\boldsymbol{h}|} \mathbb{1}\{n_r = 0\} \cdot \lambda_r \cdot \ell - \sum_{\boldsymbol{n}\in\aleph} P(\boldsymbol{n})(N - \sum_{r=1}^{|\boldsymbol{h}|} n_r) \cdot p\Big\}\Bigg\}, \tag{9}$$

where $\boldsymbol{c}_h$ is the cost vector of establishing and maintaining hubs $\boldsymbol{h}$. Apparently the combinations of $\boldsymbol{h} \in \mathcal{I}$ grow exponentially with cardinality of $\mathcal{I}$. Therefore, we propose an approximation method in Section 5 to make the upper-level optimization tractable.

To determine the irreducibility of $\boldsymbol{h}$, we outline a technique to solve this problem next. Define the matrix $\bar{\boldsymbol{P}}$ is sufficiently connected if every principal minor $\boldsymbol{A}$ of $(\bar{\boldsymbol{P}} - \boldsymbol{I})$ has $\boldsymbol{A} \cdot \mathbb{1} \neq \mathbb{0}$, except $(\bar{\boldsymbol{P}} - \boldsymbol{I})$ itself. Here $\mathbb{1}$ and $\mathbb{0}$ are column vectors with the same number of rows as $\boldsymbol{A}$.

Theorem 2 *The following four conditions are equivalent:*

1. $\boldsymbol{h}$ *is irreducible;*
2. $\bar{\boldsymbol{P}}$ *is sufficiently connected;*
3. $rank(\bar{\boldsymbol{P}} - \boldsymbol{I}) = |\boldsymbol{h}| - 1$*;*
4. *every principal minor of* $(\bar{\boldsymbol{P}} - \boldsymbol{I})$ *is full rank except* $(\bar{\boldsymbol{P}} - \boldsymbol{I})$ *itself.*

The Theorem 2 provides four equivalent features of the hub network. Condition 1 depicts an explicit structural feature of the hub network; condition 2 reflects such a feature in the matrix $\bar{\boldsymbol{P}}$; condition 3 relates this feature with the solution space for $\{v_r\}$ and $\hat{W}$ (recall equation 9); condition 4 sheds more light on how to determine the irreducibility of $\boldsymbol{h}$ as well as how to solve $\{v_r\}$ and $\hat{W}$ in calculation practice. When determining the irreducibility of $\boldsymbol{h}$ or equivalently the rank of $(\hat{\boldsymbol{P}}-\boldsymbol{I})$, we only need to get the determinant of matrix $\boldsymbol{Q}$, where $\boldsymbol{Q}$ is obtained by replacing the first column of $(\hat{\boldsymbol{P}}-\boldsymbol{I})$ with a H-dimension column vector $\boldsymbol{q}^T = (1, 0, ..., 0)$. It is easy to prove $det(\boldsymbol{Q}) \neq 0$ if and only if condition 4 hold. In addition, we can get one solution for $\{v_r\}$ when $det(\boldsymbol{Q}) \neq 0$ by solving $\boldsymbol{v}$ in

$$\boldsymbol{Q}^T \cdot \boldsymbol{v} = \boldsymbol{q} \tag{10}$$

5. Approximation method and correction

5.1. Approximation for Stationary Rideshare Distribution

Approximation method for closed finite queuing networks with time lags. In general, the form of $P(\boldsymbol{n})$ is very complicated, and as a result, the marginal distributions $P(\boldsymbol{n})$ must be obtained by performing an N-fold multiple summation on $P(\boldsymbol{n})$. If N and H are fairly large, this procedure become prohibitively tedious[4]. Posner and Bernholtz outlined a technique for obtaining an approximate solution for the absolute marginal probabilities. By their approximation method, we can rewrite the joint probability function in Equation (7) in the form:

$$P(\boldsymbol{n}) = C \prod_{r=1}^{H+1} B_r(n_r)$$
$$\text{where } n_r \in \mathbb{N} \text{ and } \sum_{r=1}^{H+1} n_r = N \tag{11}$$

[4]For example, a scenario with 20 hubs and 1000 taxis would require 3.39×10^{41} summation on $P(\boldsymbol{n})$

here $B_r(n_r) = \left(\dfrac{v_r}{\lambda_r}\right)^{n_r}, r = 1, \ldots, H$; while $B_{H+1}(n_{H+1}) = \dfrac{\hat{W}^{n_{H+1}}}{n_{H+1}!}$. After normalizing B_r to Ω_r, the joint probability function in Equation (11) can be written in the form

$$P(\boldsymbol{n}) = C' \prod_{r=1}^{H+1} \Omega_r(n_r) \tag{12}$$

Let U_r, $r = 1, 2, ..., H+1$ be a family of independent discrete random variables whose probability distribution functions are Ω_r correspondingly. Let the mean and the variance of U_r be β_r and σ_r^2 respectively. In addition, we denote $\mathbf{E}(|U_r|^3) = \varrho_r$. Since $P(U_r > N) = 0$, $\beta_r < \infty$ and $\varrho_r < \infty$. For H is sufficiently large, by the means of Central Limit Theorem,

$$P_r(n_r) \approx C''\Omega_r(n_r)\exp\{-\frac{(N - n_r - \nu_r)^2}{2\delta_r^2}\} \tag{13}$$

$$\text{where} \begin{cases} \nu_r = \sum_{k=1}^{H+1} \beta_k - \beta_r \\ \\ \delta_r^2 = \sum_{k=1}^{H+1} \sigma_k^2 - \sigma_r^2 \end{cases}$$

For simplicity, we denote

$$E_r = \sum_{n_r=1}^{H+1} P_r(n_r) \cdot n_r \tag{14}$$

$$\widetilde{P}_r(n_r) = C''\Omega_r(n_r)\exp\{-\frac{(N - n_r - \nu_r)^2}{2\delta_r^2}\} \tag{15}$$

$$\widetilde{E}_r = \sum_{n_r=1}^{H+1} \widetilde{P}_r(n_r) \cdot n_r \tag{16}$$

Here $P_r(n_r)$ represents the probability that n_r taxis are waiting at hub r, and E_r represents the expectation of the number of taxis waiting at hub r, while $\widetilde{P}_r(n_r)$ represents the approximated probability that n_r taxis are waiting at hub r, and $\widetilde{E}_r$ represents the approximated expectation of the number of taxis waiting at hub r.

However, this method does not always produce satisfying approximation. Recall in our model, $rank(P_{rs}) = H - 1$. Thus, $\{v_r\}$ and $\hat{W}$ have one degree of freedom, represented by a set of initial value $\{v_r^0\}$ and $\hat{W}^0$ with an unfixed multiplier $\alpha > 0$. The quality of the approximation largely depends on α. For example, given H, when α is sufficiently small, e.g., when v_r is designed to sum to unit, in Function (15), ν_r and δ_r^2 are both close to 0. Thus, $\widetilde{E}_r \approx N$, $r = 1, 2, ..., H + 1$. When α is sufficiently large, $\widetilde{E}_r \approx 0, r = 1, 2, ..., H + 1$.

• Correction for the approximation method

To pick the α that generates the best quality of approximation, we impose a constraint on α. We aim to minimize the error between the original distribution and the approximate distribution:

$$\min_{\alpha>0} \sum_{r=1}^{H+1} \sum_{n_r=0}^{N} ||\widetilde{P}_r(n_r) - P_r(n_r)||_d \tag{C-0}$$

where function $||\cdot||_d$ is a measure of the error between the true distribution of n_r and the approximate distribution. One specific example is

$$||\widetilde{P}_r(n_r) - P_r(n_r)||_d = n_r \cdot ||\widetilde{P}_r(n_r) - P_r(n_r)|| \tag{17}$$

However, since $P_r(n_r)$ has a complicated form, it is hard to directly use (C-0) to optimize α. In stead, we relax it by putting the two summation operations inside $|| \cdot ||$. Then we can write (C-0) as

$$\min_{\alpha>0} || \sum_{r=1}^{H+1} (\widetilde{E}_r - E_r)|| \tag{18}$$

or $\sum_{r=1}^{H+1} \widetilde{E}_r = \sum_{r=1}^{H+1} E_r$. Since $\sum_{r=1}^{H+1} E_r = N$, we have the constraint:

$$\sum_{r=1}^{H+1} \widetilde{E}_r = N \tag{C-1}$$

Due to the complicated form of distribution $\widetilde{P}_r$, it is hard to directly calculate α based on (C-1). Recall in Equation (15), the peak point for

$\exp\{-(N-k-\nu_r)^2/\delta^2\}$ is $N-\sum_{r=1}^{H+1}\mathbf{E}(U_r)+\mathbf{E}(U_r)$, and the expectation of Ω_r is $\mathbf{E}(U_r)$. Thus, when $\sum_{r=1}^{H+1}\mathbf{E}(U_r)=N$, $E_r\approx\mathbf{E}(U_r)$, $r=1,2,...H+1$. Let constraint (C-2) be

$$\sum_{r=1}^{H+1}\mathbf{E}(U_r)=N \tag{C-2}$$

Since 1) $\mathbf{E}(U_r)$ is monotone with α, 2) $\lim\limits_{\alpha\to 0}\mathbf{E}(U_r)=0$, and 3) $\lim\limits_{\alpha\to\infty}\mathbf{E}(U_r)=N$, constraint (C-2) has one unique solution. Then constraint (C-2) can capture one approximate solution for (C-1). The solutions of (C-0) and (C-1) are both limited within (C-1), although we lack the information of the former. When H is sufficiently large, most $\alpha>0$ should generate "high-quality" approximation. Otherwise, the set of α constrained by (C-1) should has small size. Thus, (C-0) and (C-2) should provide approximate α as both are constrained by (C-1).

Recall the functions B_r and Ω_r, when N is sufficiently large, Ω_{H+1} approximates a Poisson distribution with mean equal to $\hat{W}$, and Ω_r approximates a geometry distribution with mean equal to $\frac{1}{1-v_r/\lambda_r}$ if v_r/λ_r is smaller than 1 for all $r=1,2,...,H$. Thus, constraint (C-2) can be simplified as

$$\begin{cases}\sum_{r=1}^{H+1}\dfrac{1}{1-\alpha(v_r^0/\mu_r^0)}+\alpha\hat{W}_0=N\\[2ex] 1-\alpha\left(\dfrac{v_r^0}{\mu_r^0}\right)>0, r=1,2,...,H\end{cases} \tag{C-3}$$

It is easy to prove (C-3) still has one unique solution. In this case, we have a quick approximate solution for E_r and $P_r(n_r)$:

$$\begin{cases}E_r\approx\alpha\times(v_r^0/\mu_r^0),\\ P_r(n_r=0)\approx 1-\alpha\times(v_r^0/\mu_r^0),\\ r=1,2,...,H.\end{cases} \tag{19}$$

and

$$\begin{cases} E_{H+1} \approx \alpha \times \hat{W}_0, \\ P_{H+1}(n_{H+1} = 0) \approx \exp\{-\alpha \times \hat{W}_0\}. \end{cases} \tag{20}$$

In calculation practice, constraint (C-2) also brings other calculation convenience. Recall $\widetilde{P}_r(n_r)$ has a product form of a exponential term with a geometric or Poisson term. When H and N is sufficiently large, both of these two terms can easily reach an enormous level. One possible solution is to find the maximum value and standardize it into 1. This is easy when $r < H + 1$ as the geometric term is essentially an exponential term, and we only need to find out the maximum exponent, which is a quadratic term for n_r. However, $\widetilde{P}_{H+1}$ includes a Poisson term, and after employing the Stirling's approximation, the exponent has the complicated form:

$$-\frac{n_{H+1}^2}{2\delta_{H+1}^2} + (\ln H - 1 + \frac{N - \nu_{H+1}}{\delta_{H+1}^2})n_{H+1} + (n_{H+1} + \frac{1}{2})\ln(n_{H+1}) \tag{21}$$

To minimize function (20) can be tedious. By contrast, under (C-2), the maximum value of the Poisson term $\frac{H^{n_{H+1}}}{n_{H+1}!}e^{-H}$ and the maximum value of exponential term $\exp\left(-\frac{(N-n_{H+1}-\nu_{H+1})^2}{2\delta_{H+1}^2}\right)$ are both reached when $n_{H+1} = \beta_{H+1}$ (note $\beta_{H+1} \approx H$, and $\beta_{H+1} = N - \nu_{H+1}$). Constraint (C-3) with the quick solution further simplifies this process by avoiding the calculation of $\widetilde{P}_r(n_r)$.

Recall the designed property of Poisson departure in the last section. This property has largely simplified the product-form solution into the form depicted in function (9) by avoiding incorporating the state of the matching systems at each hub and reducing the state space into $\aleph$. In addition, our correction of the approximation method has been simplified by the assumption of the irreducibility of $\boldsymbol{h}$ to the most simple case: only one degree of freedom for $\{v_r\}$ and $\hat{W}$.

• Numerical Experiment for Stationary Rideshare Distribution

Suppose a city has two commercial districts and twenty residential districts. Each commercial or residential district has a taxi hub. A total of 100 taxis are in the system. Customers only commute between a commercial district and a residential district so that there are in total 80 possible directed routes. For similarity, we assume: 1) the arrival rates are unit and identical on all directed routes; 2) the demand rates of customers are unit and identical on all directed routes.

We aim to calculate the expectation and variance for n_A the number of taxis waiting at each commercial district, n_B the number of taxis waiting at each residential district, and n_W the number of taxis in transit with passengers. We respectively employ the discrete event simulation, the unconstrained approximation method with some casually designed α, the approximation method with constraints we proposed in the last section (C-1, C-2 and C-3), and the quick solution of equation (18) and (19). In the discrete event simulation, we run 1000 iterations till t=10, 20, 25, 30 respectively. As for the unconstrained approximation method, we try three casually designed constraint for α: 1)$\sum_{r=1}^{H} v_r = 1$; 2)$\min_r\{\frac{v_r}{\lambda_r}\} = 1$; 3) $\min_r\{\frac{v_r}{\lambda_r}\} = 2$.

The results are presented in Table 1. The first panel "Discreet Event Simulation" shows under the equilibrium distribution, $E(n_A) \approx 2.034$, $E(n_B) \approx 2.063$, and $E(n_W) \approx 54.666$. The second panel "Unconstrained Approximation" shows when we have small α by setting $\sum_{r=1}^{H} v_r = 1$, the approximate expectations of n_A, n_B and n_W are unreasonably large. When we have large α by setting $\min_r\{\frac{v_r}{\lambda_r}\} = 1$ or $\min_r\{\frac{v_r}{\lambda_r}\} = 2$, the approximate expectations of n_A, n_B and n_W are unreasonably small. The third panel "Constrained Approximation" generates more pleasant results with the three proposed constraints. For example, $E(n_A) \approx 2.073$ under (C-1), $E(n_A) \approx 2.011$ under (C-2), and $E(n_A) \approx 2.189$ under (C-3). The fourth panel "Quick Solution" provides a more time-saving solution although the error is enlarged at the same time. For example, $E(n_A) \approx 2.415$.

5.2. Greedy Algorithm for Hub Selection

Selecting the optimal hub locations in the upper-level optimization is NP-hard (the capacitated single allocation hub location problem itself is a classic NP-hard problem). Therefore, we need an approximation algorithm to hasten the enumeration of $\boldsymbol{h}$ over Γ. A greedy scheme is the most intuitive in the case that the cost function does not have pleasant convexity properties in general. We start from two initial hubs $h_0, h_1 \in \Gamma$ in iteration $k = 1$. Then each iteration $k + 1$ we add new hub h_{k+1} such that:

Table 1: Results of Numerical Experiment for Stationary Rideshare Distribution (C.T. is computational time, measured by minutes)

Method		$\mathbf{E}(n_A)$	$\mathbf{E}(n_B)$	$\mathbf{E}(n_W)$	$\mathrm{Var}(n_A)$	$\mathrm{Var}(n_B)$	$\mathrm{Var}(n_W)$	C. T.
Discrete Event Simulation	t=10	2.141	2.044	54.849	6.084	5.348	36.658	105.213
	t=20	2.014	2.062	54.738	5.295	5.729	41.539	203.827
	t=30	2.034	2.063	54.666	5.349	5.958	36.436	299.826
Unconstrained Approximation	$\sum_{r=1}^{H} v_r = 1$	93.173	93.173	98.42	1.269	1.269	0.303	0.010
	$\min_r\{\frac{v_r}{\lambda_r}\} = 1$	16.167	16.167	75.445	255.494	255.494	73.144	0.009
	$\min_r\{\frac{v_r}{\lambda_r}\} = 2$	0.339	0.339	0	0	0	0	0.008
Constrained Approximation	C-1	2.073	2.073	54.401	6.076	6.076	42.73	0.072
	C-2	2.011	2.011	53.985	5.72	5.72	39.115	0.059
	C-3	2.189	2.189	55.663	6.331	6.331	32.011	0.081
Quick Solution	-	2.415	2.415	46.872	3.417	3.417	46.872	0.066

$$
\begin{aligned}
h_{k+1} = \mathop{\arg\min}_{h_{k+1}\in\Gamma/\{h_0,\dots,h_k\}} & \Bigg\{ \boldsymbol{c}_{\boldsymbol{h}'}^{\intercal} \cdot \mathbb{1} + \\
& \min_{p} \Bigg\{ \sum_{\boldsymbol{n}\in\aleph} \widetilde{P}(\boldsymbol{n}) \sum_{r=1}^{|\boldsymbol{h}'|} \mathbb{1}\{n_r = 0\} \cdot \lambda_r \cdot \ell - \sum_{\boldsymbol{n}\in\aleph} \widetilde{P}(\boldsymbol{n})(N - \sum_{r=1}^{|\boldsymbol{h}'|} n_r) \cdot p \Bigg\} \Bigg\}
\end{aligned}
\tag{22}
$$

where $\boldsymbol{h}' = \{h_{k+1}\} \cup \{h_0, \dots, h_k\}$. Note the greedy algorithm may not be robust if the first two initial hubs are randomly selected. Consider a city has an two parts: region A and region B. Most of the taxi trips are within region A while region B only accounts for a small proportion. Trips of commuting between A and B are rare and take a relatively long time compared with those within a certain region. If the two initial hubs are set in B, the following hubs will all be selected from B if the commuting time are sufficiently long or the commuting trips are sufficiently rare. To avoid such unreasonable situation, we set the first two initial hubs as the O-D pair of hubs that have the largest customer flow.

We briefly prove that the loose bound for the approximation ratio in bi-level optimization is $O\left(\log|\Gamma| \cdot \max\{\varrho\} \cdot \bar{H}^{-2} \cdot (\max\sigma)^{-3}\right)$, which is a pseudo-polynomial algorithm depending on the constant $\bar{H}$ (the maximum required hubs in design). Suppose an optimal solution $\hat{\boldsymbol{h}}$ contained κ hubs. Our greedy algorithm finds a subset with at most $\kappa \log|\Gamma|$ sets. For each cost evaluation in the inner optimization, we use the approximation for cost function based on central limit theorem of $(H+1)$ non-identically distributed summands. According to Berry-Esseen theorem, the error of approximation on distribution is bounded by $0.4097 \cdot (\sum_{\boldsymbol{h}} \sigma_r^2)^{-3/2} \cdot \sum_{\boldsymbol{h}} \varrho_r$ in case of non-identical distributions. A loose bound takes the maximum of second and third momentum of $U_r, r \in \Gamma$. Since in practice we set $\bar{H} \ll |\Gamma|$, and $\log |\Gamma|$ is a very tight bound, the implementation performance is acceptable.

The complexity of this algorithm is mainly constrained by the searching space in each iteration of upper-level optimization, as well as the convergence rate of tuning α^* as a parameter in the inner-level stationary distribution. The complexity is expressed as $O(|\Gamma|^2 \cdot \log_2(\frac{N^2}{\hat{W}_0^2 \cdot \epsilon}))$. The first term $|\Gamma|^2$ is the complexity of greedy method , which is similar to a knapsack problem; in each iteration we solve α^* binds by the complexity of bisection search. ϵ is the sufficient small error in the halting conditions. Since it takes logarithm of the inverse of ϵ, computing the optimal prices within each iteration of the greedy algorithm is rather efficient even for the case study of New York city.

6. A Case Study: Bike-Ride-Share in New York City

6.1. Description of the Data

We use Manhattan as a case study. We first divide the whole Manhattan land area into 225 regions with a grid. Each region has a bike station at its center. Suppose Manhattan has 1,000 taxis in total and wants to build 20 hubs. We aim to find the optimum hub network and price to maximize net profit.

We use the average hourly customer flow from 8 am to 17 pm during weekdays in August 2015 (20 days included in total) as the initial demand rate of non-ride-pool customers who want to go from region i to region j. The data is obtained from Taxi and Limousine Commission (TLC), New York City. The TLC created in 1971, is the agency licensing and regulating NYC's taxicabs, for-hire vehicles, commuter vans, and paratransit vehicles. It licenses and regulates over 50,000 vehicles and approximately 100,000 drivers. We assume p, price for per hour's taxi operation, has a linear effect on every non-ride-pool customer demand: when $p = 0$, the actual demand are twice the initial demand rate; when $p = 10$, the actual demand rate becomes zero. Negative demand rates when $p > 10$ are truncated into 0. Given the fixed general policy π and price p, we suppose for every customer flow from region

i to region j, 20% of customers are non-ride-pool customers, 30% are starters and 50% are followers. These starters and followers may be those who cannot afford taking a taxi and take public transportation instead when there is no ride-pool service.

The penalty $\ell = 0.25$. As for threshold time, $t_1 = 1/12$ hour, and $t_2 = 1/12$ hour. The bike traveling time and taxi traveling time between different bike stations or hubs are obtained from Google Maps Api. When a customer who is not from a hub region selects his starting hub, the probability of selecting hub k is in proportion with $exp(-t_k)$, where t_k is the biking time to hub k. The cost of maintaing hubs is 50 per hub per hour.

We used the approximation methods presented in the last section in our case study. The constraint we used for α is (C-2).

6.2. Results

We find the optimum price $p_0^* = 9.193$ in this scenario. The optimum hub locations are presented in Figure 5. Note no selected hubs are located in regions where the hourly demand density (starting regions) is below 30% in quantile. This would bring a hourly net profit of W_0^* =8,295.

To test the robustness of our model over the time span, we randomly select 10 days' data to calculate the optimum hub selection, price p^* and net profit W^*. Then we use the hub locations and the price to calculate the net profit W' based on the rest 10 days' data. This process is repeated 10 times, and each time is indexed by $k = 1, 2, ..., 10$. We compare the p_k^*, W_k^* and $|W_k^* - W_k'|$ for each time. We find $|p_k^* - p_0^*|/p_0^*$ and $|W_k^* - W_0^*|/W_0^*$ remain in a tiny scale (<0.5%) as presented in Figure 6. In addition, $|W_k^* - W_k'|/W_k'$ also remains in a tiny scale ($\leqslant$0.31%). These results validate the robustness of our model over time span.

6.3. Sensitivity Analysis

We also do a sensetivity analysis by respectively changing total taxi number N from 500 to 3000, H from 5 to 30, ℓ from 0.10 to 0.35, $t_1$from 1/21 hour to 1/6 hour, t_2 from 1/21 hour to 1/6 hour, θ from 1 to 6, where θ is a divider for every demand rate. The results are presented in Figure 7. We find the optimum price decreases as N and θ increase, or as H, t_1, t_2 and ℓ decrease. The net profit with the optimum price will increase as N, H, $t_1$and t_2 increase, or as ℓ and θ decrease. The pattern of the net profit is clear: larger N, H, $t_1$or t_2 will bring more demands, while larger θ has the opposite effects. The ℓ decreases the net profit by increasing the cost due to

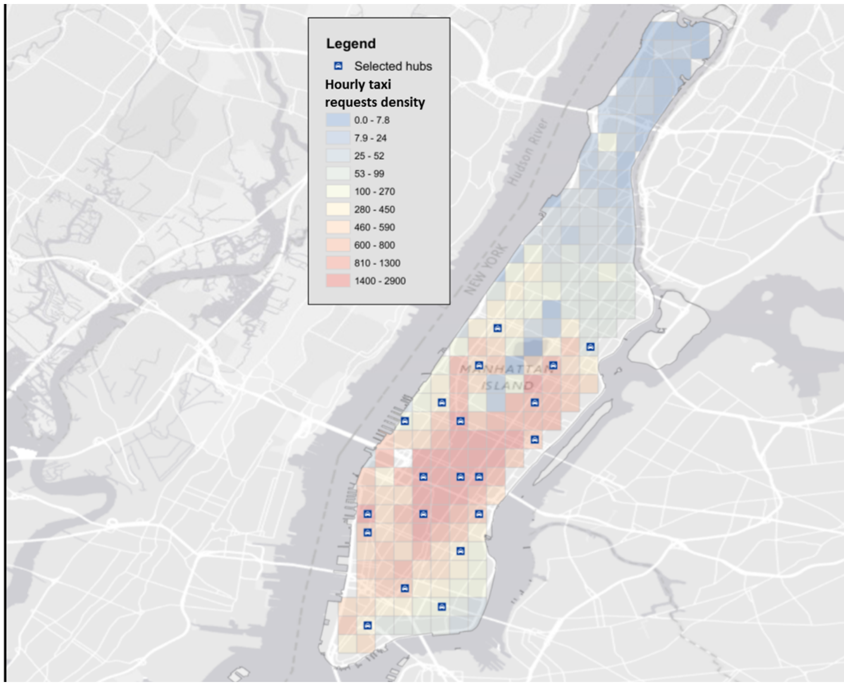

Figure 5: The optimum network has its hubs mainly located in areas with high demand density.

penalty and decreasing the number of taxis on road. As for optimum price, recall that $p_0^* = 9.193$ and the demand rate becomes 0 when $p = 10$. This relatively high price in the original case means the supply falls short of the potential demand. To reduce the price will bring more penalty due to missed customers than the extra profit due to more demands. A larger N or θ, and a smaller H, t_1 or t_2 can narrow such a supply-demand gap, and thus reduce the optimum price. A smaller l can reduce the penalty and thus drcreased the optimum price.

7. Conclusion

This paper contributes to the literature on transportation-as-a-service by presenting a new multimodal transportation system to integrate two popular mobility service: bike-sharing and vehicle-sharing. We document research undertaken toward answering two questions related to the efficiency of such a system: (a) pricing, or the optimum price for maximum net profit under

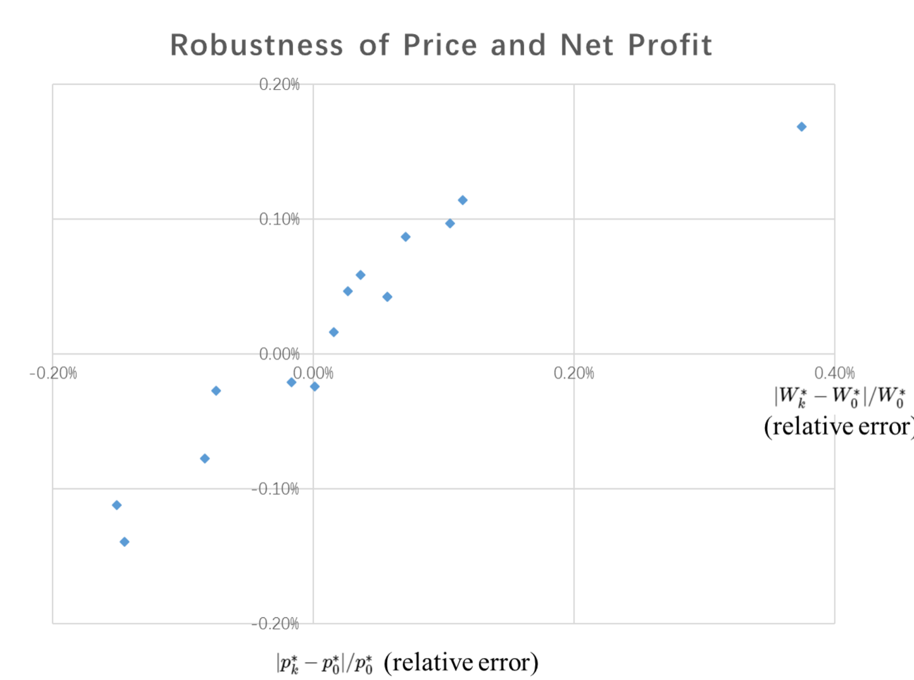

Figure 6: In the robustness test, $|p_k^* - p_0^*|/p_0^*$ and $|W_k^* - W_0^*|/W_0^*$ remain in a tiny scale (<0.5%).

a fixed taxi hub network, and (b) hub selection, or the locations of hubs in a optimum taxi hub network. The research questions also seek to find the near-optimal prices and network in a case study of New York City.

With some mild assumptions, we model the multimodal transportation system as a closed queuing network with time lags. One important feature of our model is the Poisson arrivals of taxi orders, which largely reduces the state space and simplifies the product-form solution. We limit the hub selections to those sufficiently connected hubs networks, which simplifies the conventional approximation method for our product form solution to the most simple case. We also find the instability of the conventional approximation method when the number of product factors are not sufficiently large, and propose a correction technique to improve the approximation.

To be sure, the study is not without its limitations. It is hard to estimate the segmentation of demand rates of non-ride-pool customers, starters and followers for a hypothetical multimodal transport system. For calculation

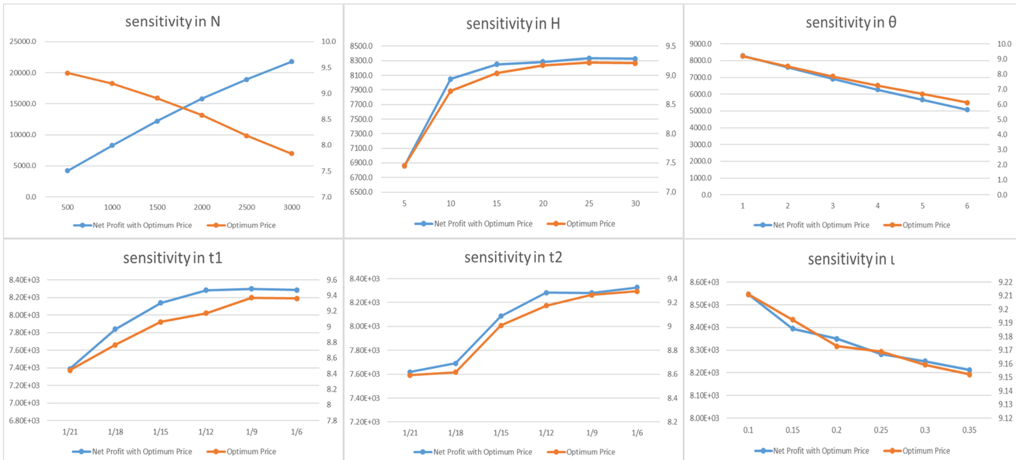

Figure 7: Sensitivity analysis for optimum price and net profit when N, H, θ, t_1, t_2, and ℓ change.

convenience, we use the average hourly customer flow in the taxi cab data as the initial demand rate of non-ride-pool customers, and assume for every customer flow, 20% of customers are non-ride-pool customers, 30% are starters and 50% are followers. In practice, those who take taxis in the taxi cab data may not necessarily become non-ride-pool customers in our multimodal system, especially for those whose willingness to pay is relatively low. In addition, the relationship of price with demand rates may not follow a simple monotone function like the linear functions we use. That is mainly because price can also indirectly influence the demand rates through the expected waiting time as discussed in the third subsection in section 4.2.1. Field survey and data are necessary for estimating the relationship of price with customers' demand. Such uncertainty of the demand and price elasticity disenables us to compare the efficiency and customers' expenses in our multimodal system with that for conventional taxi service.

Acknowledgement

This research was partially supported by DiDi Chuxing via Michigan Institute of Data Science, an initiative at University of Michigan. We thank our colleagues Xuan Sharon Di and Zhenyu Wang from Columbia University, and Aditi Misra from Transportation Research Institute, University of Michigan

who provided insight and expertise that greatly assisted the research.

Reference

A. Pal, Y. Zhang, Free-floating bike sharing: Solving real-life large-scale static rebalancing problems, Transportation Research Part C: Emerging Technologies 80 (2017) 92–116.

S. Reiss, F. Paul, K. Bogenberger, Empirical analysis of munichs free-floating bike sharing system: Gps-booking data and customer survey among bikesharing users, in: Transportation Research Board 94th Annual Meeting, 15-3741.

Z. Li, Y. Hong, Z. Zhang, An empirical analysis of on-demand ride sharing and traffic congestion (2016).

R. C. Hampshire, C. Simek, T. Fabusuyi, X. Di, X. Chen, Measuring the impact of an unanticipated suspension of ride-sourcing in austin, texas (2017).

N. Shirouzu, e. b. M. Heinrich, S. Powell, Beijing, shanghai propose curbs on who can drive for ride-hailing services, Reuters (2016).

S. Banerjee, D. Freund, T. Lykouris, Pricing and optimization in shared vehicle systems: An approximation framework, CoRR abs/1608.06819 (2016).

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

J. Schuijbroek, R. C. Hampshire, W.-J. Van Hoeve, Inventory rebalancing and vehicle routing in bike sharing systems, European Journal of Operational Research 257 (2017) 992–1004.

D. Freund, S. G. Henderson, D. B. Shmoys, Minimizing multimodular functions and allocating capacity in bike-sharing systems, in: International Conference on Integer Programming and Combinatorial Optimization, Springer, pp. 186–198.

A. Faghih-Imani, R. Hampshire, L. Marla, N. Eluru, An empirical analysis of bike sharing usage and rebalancing: Evidence from barcelona and seville, Transportation Research Part A: Policy and Practice 97 (2017) 177–191.

M. Stiglic, N. Agatz, M. Savelsbergh, M. Gradisar, The benefits of meeting points in ride-sharing systems, Transportation Research Part B: Methodological 82 (2015) 36–53.

P. Santi, G. Resta, M. Szell, S. Sobolevsky, S. H. Strogatz, C. Ratti, Quantifying the benefits of vehicle pooling with shareability networks, Proceedings of the National Academy of Sciences 111 (2014) 13290–13294.

M. Stiglic, N. Agatz, M. Savelsbergh, M. Gradisar, Enhancing urban mobility: Integrating ride-sharing and public transit, Computers & Operations Research 90 (2018) 12–21.

M. Miramontes, M. Pfertner, H. S. Rayaprolu, M. Schreiner, G. Wulfhorst, Impacts of a multimodal mobility service on travel behavior and preferences: user insights from munichs first mobility station, Transportation (2017) 1–18.

K. B. Campbell, C. Brakewood, Sharing riders: How bikesharing impacts bus ridership in new york city, Transportation Research Part A: Policy and Practice 100 (2017) 264–282.

R. C. P. Wong, W. Y. Szeto, S. Wong, H. Yang, Modelling multi-period customer-searching behaviour of taxi drivers, Transportmetrica B: Transport Dynamics 2 (2014) 40–59.

R. Serfozo, Optimal control of random walks, birth and death processes, and queues, Advances in Applied Probability 13 (1981) 61–83.

K. Spieser, S. Samaranayake, W. Gruel, E. Frazzoli, Shared-vehicle mobility-on-demand systems: a fleet operators guide to rebalancing empty vehicles, in: Transportation Research Board 95th Annual Meeting, 16-5987.

M. Posner, B. Bernholtz, Closed finite queuing networks with time lags, Operations Research 16 (1968) 962–976.