MODELLING AND DATA ANALYSIS FOR THE DESIGN OF MODERN URBAN MOBILITY

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Alyf Janmohamed

August 2024

© 2024 Alyf Janmohamed
ALL RIGHTS RESERVED

MODELLING AND DATA ANALYSIS FOR THE DESIGN OF MODERN URBAN MOBILITY

Alyf Janmohamed, Ph.D.

Cornell University 2024

This thesis has two parts. The first focuses on shared e-bike and e-scooter systems, which have become globally popular solutions for urban mobility. However, they introduce new operational challenges around charging and many charging mechanisms have been proposed (e.g. employees swapping batteries, charging stations, crowd-sourced, etc.). We develop a model that allows the direct comparison of different mechanisms and quantifies their impact on platform performance. In order to achieve this, novel ideas for analyzing the underlying stochastic process are introduced.

The second part considers an emerging idea in public transit, using on-demand shuttles for first- and last-mile. We develop a modelling framework to help cities assess the feasibility of integrating on-demand shuttles. It consists of two modular components. The first models individual regions serviced by shuttles and the second efficiently optimizes the deployment of shuttles throughout the day. A key feature of this framework is that it can be easily extended to key considerations of transit operators such as latent demand and equity across regions.

BIOGRAPHICAL SKETCH

Alyf Janmohamed was born and raised in Toronto, Canada, but also lived briefly in Nairobi, Kenya with his family. He completed his undergraduate degree at University of Toronto in Engineering Science with a major in Mathematics, Statistics, and Finance. During a summer research project in his first year, Alyf was exposed to mathematical modelling and optimization and immediately loved the ideas. Alyf also worked as a quantitative trading analyst at RBC Capital Markets for a little over a year. After completing his undergraduate degree in 2018, he moved to Ithaca, NY to enroll as a PhD student in the Operations Research and Information Engineering department at Cornell. After graduating, Alyf will move to New York City to work on the shared rides team at Uber.

This thesis is dedicated to my family and in loving memory of Kassamali Janmohamed.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my wonderful advisors Shane Henderson and David Shmoys for their continuous support, mentorship, and advice. David and Shane helped me grow as a problem solver by constantly encouraging me to think deeper and reflect on the real world systems my models represent. You have helped me grow as a researcher, a teacher, and a person.

I want to thank Siddhartha Banerjee for serving on my committee, sharing his insights into closed queueing, and giving me invaluable feedback. Thank you Samitha Samaranayake for teaching me about transit planning and microtransit, and for sharing your valuable insights. Thank you to the several faculty that I was fortunate enough to learn from and who supported me throughout my PhD.

During my PhD, I had the priviledge to work on the Cornell Covid modelling team, lead by Peter Frazier. Thank you to David and Shane who played a pivotal role as faculty advisors. I will miss working through these challenging problems with the wonderful team of students: Bonnie Akhavan, Massey Cashore, Ning Duan, Brian Liu, Henry Robbins, Jiayue Wan, and Yujia Zhang.

I found my passion for the world Operations as an undergraduate at the University of Toronto because of Timothy Chan and Justin Boutilier. Thank you for introducing me to and teaching me about the wonderful world of mathematical modelling and optimization. Additionally, the encouragement and support of Deirdre Timusk and Matthew Brennan during high school, which inspired me to explore math at a deeper level, laid the groundwork for this journey.

The last 6 years would not have been nearly as fun without the amazing ORIE department students. Thank you Sam Gutekunst and Andrew Daw for a wonderful visit weekend that convinced me to join ORIE and for their guid-

ance during my first couple years. Thank you Sean Sinclair for being my first semester problem set buddy. Thank you Venus Lo, Pamela Badian-Pessot, and Benjamin Grimmer for building me the rocketship chair that now becomes a department artifact. Thank you Hugo Mainguy for being a fantastic ORGA co-president and thank you to the Friday night board game crew, both old and new (Sander Aarts, Pamela Badian-Pessot, Benjamin Grimmer, Matthew Zalesak, Ishan Bansal, Amit Harlev, Priya Pulyassary, Tyler Sam, Ricky Shapley, and Varun Suriyanarayana). Lastly, thank you to the countless other students over the years for their advice and friendship and some friends outside Ithaca, Alisha, Sadiq, Alina, and Zafrin.

Thank you to Renee Mirka, Soumya Basu, and Abhishek Shetty for the countless memories, restaurant recommendations, and small sided soccer games followed by Ted Lasso. To my Mystery Machine teammates, thank you for the wonderful memories and many laughs. And thank you to all the soccer players in the ORIE department for the department games.

Lastly, I am particularly grateful to my family for their constant support and love. Azraa and Malik, you have been there every step of the way with laughter, advice, and support. To my parents, Amir and Zahida, thank you for your endless sacrifices and constant support. I could not have done it without you.

This research was supported in part by an NSERC PGS-D Fellowship, NSF grants DMS-2230023 and CMMI-2035086, and US Depart of Energy, Vehicle Technologies Office grant DE-EE0008464.

CONTENTS

Biographical Sketch . . . iii
Dedication . . . iv
Acknowledgements . . . v
Contents . . . vii
List of Figures . . . viii

1 Introduction **1**

2 Comparing (Power) Charging Mechanisms for Fleets of E-scooters **5**
2.1 Introduction . . . 5
2.2 Related Work . . . 8
2.3 Model Description and Preliminaries . . . 9
2.3.1 Stationary Distribution . . . 12
2.3.2 Pricing Problem . . . 14
2.4 Constant Service Rates . . . 27
2.4.1 Non-electric Vehicles . . . 27
2.4.2 Electric Vehicles . . . 36
2.4.3 Scaling Analysis . . . 41
2.5 State-dependent Service Rates . . . 46
2.5.1 Charging Mechanisms . . . 46
2.5.2 Performance Metrics . . . 49
2.5.3 Balance Point Approximation . . . 52
2.5.4 Scaling Limits . . . 52
2.6 Computational Results . . . 59
2.6.1 Exploratory Data Analysis . . . 59
2.6.2 Case Study . . . 75

3 Planning for On-Demand Multimodal Transit Systems **81**
3.1 Introduction . . . 81
3.2 Literature Review . . . 83
3.3 Individual Region Model . . . 85
3.3.1 Customer Arrivals . . . 86
3.3.2 Shuttle Service . . . 86
3.3.3 DTMC Formulation . . . 87
3.3.4 Stability and Performance Metrics . . . 88
3.3.5 Multiple Shuttles . . . 90
3.4 Optimizing Fleet Deployment . . . 91
3.4.1 Problem Input and Assumptions . . . 92
3.4.2 Integer Programming Formulation . . . 93
3.4.3 Min Cost Circulation Reduction . . . 95
3.4.4 Extending to time-varying shuttle availability . . . 99
3.5 Case Study . . . 104

LIST OF FIGURES

2.1 Graphs showing the approximation error as m grows while varying other parameters. 35
2.2 The distribution of micromobility trip start times in Austin, Chicago, Minneapolis and New York City in July. Minneapolis has no data between midnight and 4-am because the scooters are disabled at this time. 60
2.3 The distribution of trip duration (truncated at 1 hour) for Minneapolis e-scooter trips on weekdays in July 2021. 62
2.4 The distribution of trip distance (truncated at 5km) for Minneapolis e-scooter trips in July 2021. 63
2.5 The trips initiated and e-scooter allocation by neighborhood in Minneapolis in 2021. 65
2.6 The number of e-scooter trips by each origin (row) and destination (column) zone in Minneapolis at different times in July. . . . 70
2.7 The number of e-scooter trips by each origin (row) and destination (column) zone in Minneapolis at different times in October. . 74
2.8 Plots of the observed trip rate (between 4-8pm) vs the number of available vehicles at 4pm and 8 pm. Each daily observation corresponds to a blue dot (vehicles available at 4pm), an orange dot (vehicles available at 8pm), and a dashed horizontal red line connecting the two. The y-coordinate corresponds to observed trip rate. 77
2.9 The availability for each charging mechanism as the number of e-scooters increases. 80

3.1 The average daily entrances and exits for each station over the 5 time windows. 105
3.2 The distribution of trip start times for Chicago Bikeshare in July 2023. 106
3.3 The service regions for each subway station, drawn as a Voronoi diagram. Each blue point denotes a station and each orange point represents an intersection of the partitioning lines. 107
3.4 The service regions for each subway station, drawn as a Voronoi diagram. Each blue point denotes a station and each orange point represents an intersection of the partitioning lines. 108
3.5 The distribution of shuttles throughout the day according to solution for 1450 shuttles. 114

CHAPTER 1

INTRODUCTION

Smartphones and GPS technology are powering a revolution in the way people move, especially in urban settings. They have unlocked new types of mobility that have changed transportation globally. From ride-sharing (Uber, Lyft), to shared-vehicle systems (car, bike, and scooter), to mass transit, almost all forms of urban transportation are finding ways to leverage new technology to improve the quality of their services. Coupled with the rapid growth of electric vehicles (EVs), these systems raise new challenges around their design and operations. From optimizing charging for EVs, to routing for shared rides, to fleet sizing for on-demand mobility, Operations Research is playing a pivotal role in ensuring these systems reach their potential.

As new systems emerge and modern innovations are integrated into transportation, our models must adapt to reflect the new reality. We will need to find new model 'physics' to represent the world and rapidly changing ways that people interact with urban mobility. New algorithmic ideas will be required to solve and scale these solutions since simple changes in model 'physics' can break the algorithmic structures relied upon. The work in this thesis demonstrates how new models, analysis ideas, and algorithms can be used to guide the design of modern urban mobility.

This thesis is composed of two works. The first focuses on quantifying the performance trade-offs between different charging solutions for shared e-bike and e-scooter systems. The second introduces a simple, extensible model to help cities evaluate the potential of microtransit (using on-demand shuttles to replace buses for first and last mile trips). While considering two different trans-

portation systems and applications, these works are unified by building tools that help answer key design decisions for modern platforms.

The first project is inspired by the rapid rise of shared e-bike and e-scooter systems around the world. When using these systems, users typically locate the closest available vehicle via an app and walk to it to unlock it. After completing their trip, they park it on the sidewalk and do not have to return it to a dedicated depot or station. These systems are called dockless, referring to the lack of dedicated drop off or pick up locations. This new generation of urban mobility systems raises two key challenges beyond conventional bike sharing: charging bikes and operating a dockless system.

Looking around the world, there are three main models for charging and operations: 1) charging stations, 2) having an employee travel to vehicles to swap batteries, and 3) crowd-sourcing (paying someone to bring the vehicle home, charge it, and redeploy it). The decision of how to charge the system is critical because it determines the operations challenges and together these represent 60% of the operating cost of the system Schellong et al. [2023]. This project introduces the first model that allows a direct comparison of these different charging mechanisms, by representing them as queues with state-dependent service rates within a larger closed queueing network.

Closed queueing networks correspond to a class of stochastic models where a fixed number of jobs (vehicles) move between queues (representing locations/states) according to the random arrival of users. While closed queueing networks are frequently used for modelling vehicle-sharing applications, analyzing them typically requires constant service rates. This project develops several new insights into this class of stochastic systems. First, it shows that

the high-dimensional closed queueing network can be exactly represented by a one dimensional birth-death process. Second, a closed form expression for performance is developed assuming that the state-dependent service rates take a specific form. Only one other closed-form expression is known, requiring constant service rates which induce a uniform stationary distribution (some additional assumptions are required for both cases). Third, we extend results and algorithms from previous work to our model with charging.

However, this model requires a particularly strong assumption that all demand rates are constant. No real-world systems satisfies this assumption since there is a temporal aspect to demand for transportation. This assumption is problematic for some cities such as New York where bike-sharing demand is driven by many users commuting to and from work. This challenges the assumption because both the demand for trips and their origin-destination distribution vary throughout the day. However, for other cities (such as Minneapolis), these assumptions are more reasonable. Both the trip demand and the origin-destination distribution are more stable over time.

The second project is inspired by a successful microtransit pilot in Atlanta where on-demand shuttles were used to replace first- and last-mile bus trips. Typically, fixed bus lines are inefficient for these trips because of their low utilization and infrequent schedule. With the integration of GPS into public transit vehicles and the prevalence of smartphone apps, technology can facilitate a microtransit system that operates smoothly. As cities see the potential of this new urban mobility solution, this project develops a modelling framework to assess if this technology is right for them.

For a city evaluating the potential of microtransit, there are several consid-

erations such as: cost (fleet size, deployment), impact on performance metrics (waiting time, access to points of interest), resilience, and robustness. We provide a new two-stage modelling framework that can be extended to address the variety of considerations. First, we propose a new stochastic model to estimate the performance of on-demand shuttles for a given region. Second, we show how to efficiently optimize the deployment of shuttles throughout the day under time-varying demand.

This modelling framework has two key advantages. First, it is easily extensible to the broad range of questions a city might care about such as: equity between regions, latent demand, peak vs. off-peak performance, robustness to mis-specified parameters, and system or region-wide performance targets. Second, it is modular. You can replace the first part of the modelling framework (e.g. with a detailed simulation model) and still use the second part to optimize efficiently. However, in order to achieve this extensibility, detailed operational decisions (such as matching and routing) are abstracted away. Therefore, this model is only suited for understanding the impact and trade-off between key design decisions.

Through these projects, I have illustrated some ways that stochastic modelling and optimization can guide the design of modern urban mobility systems.

CHAPTER 2

COMPARING (POWER) CHARGING MECHANISMS FOR FLEETS OF E-SCOOTERS

2.1 Introduction

Shared mobility systems (car-sharing, bike-sharing, ride-sharing, etc.) are key elements of urban transportation infrastructure. The popularity of shared e-scooter programs has exploded since 2017 when several of the biggest companies in the space (Lime, Bird) were founded [Schellong et al., 2023]. In 2022, there were over 350 million e-scooter trips in the US and Europe alone [Heineke et al., 2023].

The rise of e-scooter platforms raises two new operational challenges compared to traditional bike sharing. First, the system is dockless: modern e-scooter systems typically do not have dedicated stations to pick up and return vehicles. Instead, vehicles can be left anywhere on the side of the road for the next user to pick up. Second, e-scooters must be charged. Four different mechanisms for charging fleets of e-scooters are used: crowd-sourcing, employees swapping batteries, charging stations, and vending machines. The choice of charging mechanism significantly impacts the operations of the system. Together, charging and operations represent 60% of the costs of running these systems [Schellong et al., 2023], which motivates the question of which charging mechanism(s) a platform should use?

In the electric vehicle charging literature, the model assumptions are tailored to one charging mechanism and therefore cannot be extended for a head-to-

head comparison. In this paper, we provide a model that allows for the direct comparison of different charging mechanisms on system performance. Quantifying and understanding these performance differences is central to selecting the best charging mechanism.

We assume a single provider and that the service area (city) is partitioned into n regions. The platform sets static prices for trips that can depend on origin and destination with the goal of maximizing steady-state throughput. There is a fixed number of vehicles, m, and their charge is binary (charged or uncharged). This binary state of charge partially reflects how many of these systems work in practice, where only e-scooters above a threshold battery level are shown to users. It is admittedly a strong assumption that, as we shall see, affords great tractability. Within each region, we use two queues to track the *number* of charged and uncharged vehicles respectively. Customers arrive to each region according to independent Poisson processes. An arriving customer will take a trip if they are willing to pay the price of the trip, and if they can find a sufficiently close charged vehicle. We do not track individual vehicle locations, so we model "sufficiently close" through state-dependent service rates in the queue of charged vehicles in each region. If these conditions are met, the customer takes the vehicle and travels to their destination. Otherwise, they abandon the system. After each trip is completed, there is a potentially origin-destination-dependent probability that the vehicle will not have enough charge for another trip. If the vehicle does not have sufficient charge for another trip, then it enters a queue for recharging in the destination region. Recharging occurs according to a birth-death process with state-dependent service rates, that reflect different charging mechanisms.

The key contributions of this paper are:

- Introducing a novel e-scooter sharing model with the flexibility to represent multiple charging mechanisms, allowing direct comparisons of charging mechanisms.
- Incorporating spatial demand for dockless systems and charging complexities and providing the first performance guarantees for this system. These dynamics require us to move beyond queues with constant service rates and the same relative throughput that, together, induce a uniform stationary distribution for the station queues. Most previous work requires this assumption.
- Demonstrating that this new method retains similar tractability to the traditional (non-EV, station-based) model of this system [George and Xia, 2011].
- Introducing a novel reduction of the $n^2 + 2n$ dimensional closed-queueing system to a 1-D capacitated open queue where performance metrics can be computed exactly in $O(mn)$ time.
- Providing the first closed-form expression for availability with non-constant service rates and the first scaling limits for a system with charging. These scaling limits are the same for all mechanisms, suggesting that any differences in the performance of charging mechanisms arise in the pre-limit regime.
- Providing computational results on city-scale experiments for data from Minneapolis.

2.2 Related Work

Although there is an extensive body of literature applying closed queueing models to shared-vehicle problems, the literature focuses on systems comprised of only single (station) and infinite server (transit) queues. The literature presented in this subsection assumes this structure for closed queueing networks. When considering state-dependent service rates, computational methods such as Denning and Buzen [1978] are relied upon.

There are several methods for analyzing closed queueing systems. First, heuristics that aim to satisfy a balanced demand property are used. George and Xia [2011] used a closed queueing model to describe a vehicle rental system and analyzed fleet sizing while only providing asymptotic guarantees. Zhang et al. [2018] considers rebalancing policies and system sizing.

Second, fluid limits are derived for the closed queueing system and optimal fluid policies are used. Özkan and Ward [2020] considers the problem of matching riders to drivers while Braverman et al. [2019] considers the optimal rebalancing of empty vehicles in ride-sharing. Both papers show the asymptotic optimality of the optimal fluid policies.

Last, open queue approximations are used. He et al. [2017] uses the open queueing network approximation and scales the arrival rates such that the expected vehicles in the open system is the number of vehicles in the closed system. This is a common approach for analyzing closed queueing systems.

Our work is most closely related to the following literature. Waserhole and Jost [2016] use pricing to maximize system throughput and obtain finite-unit ap-

proximation guarantees for the system. Specifically, they show that their policy is within a factor of $m/(m+n-1)$ of optimal. This paper does not consider travel time for vehicles. These results are extended by Banerjee et al. [2022] which shows the $m/(m+n-1)$ guarantee holds for a more general class of objective functions and extend the results to include travel time. They also characterize the scaling limits of the system as the number of vehicles and demand jointly go to infinity. Our results directly build upon this paper. Benjaafar et al. [2022] uses mean-value analysis [Reiser and Lavenberg, 1980] to give a recursive expression for availability and provides a closed form approximation for fleet sizing given an availability target. Benjaafar and Shen [2023] further optimizes the static pricing policy by parameterizing it with the number of vehicles in the system.

2.3 Model Description and Preliminaries

We model a dockless e-bike or e-scooter sharing company with $m \geq 1$ vehicles operating in a city partitioned into $n \geq 1$ non-overlapping regions as a closed queueing network where vehicles (e-bikes or e-scooters) are transported around the network by customers. There are 3 types of queues in this network: charged vehicles, uncharged vehicles, and transit queues. Each region has both a charged-vehicle queue and an uncharged-vehicle queue that are modeled as birth-death processes with state-dependent service rates. Transit queues exist for each pair of regions and are modeled as $\cdot/G/\infty$ queues. The state of this queueing network is $\mathbf{X} = [X_i^c, X_i^u, X_{ij}^t]$ where X_i^c (X_i^u) denotes the number of charged (uncharged) vehicles in region i and X_{ij}^t represents the number of vehicles in transit from i to j. The number of vehicles, m, is fixed, so $m = \sum_i \left(X_i^c + X_i^u\right) + \sum_i \sum_j X_{ij}^t$.

Customer Arrivals: Customers arrive wanting to travel from region i to j according to independent Poisson processes with rates ϕ_{ij} and their willingness to pay is drawn independently from a known distribution function F_{ij}. We assume that there is a 1-to-1 mapping from prices to the fraction, q_{ij}, of customers willing to pay the price for the trip. For state-independent pricing, the arrival rate of customers to region i willing to pay for their trip is $d_i = \sum_j \phi_{ij} q_{ij}$.

Searching for e-scooters: If the customer is willing to pay the price for a trip, then they will search for the closest charged vehicle (typically via an app). Customer demand is sensitive to walking, e.g., travelling 300m to a bike station decreases likelihood of use by 60% [Kabra et al., 2020]. We incorporate this spatial element of demand while suppressing geography in the model by letting the trip rate be a function of the number of charged vehicles in the region. In this setup, we view locations as random and assume that a customer who is willing to pay and who finds x charged vehicles in Region i upon arrival will take the ride with probability $\zeta_i(x)$, independently of all else. We impose two mild conditions on the function ζ_i for each i. First, we assume that it is monotone non-decreasing and concave (diminishing returns). Second, we assume that if there were infinitely many charged vehicles in the region, everyone would be able to find a sufficiently close vehicle, i.e., $\lim_{x\to\infty} \zeta_i(x) = 1$ for all regions i. We also assume that the state-dependent service rates converge linearly. A sequence x_ℓ converges to a finite limit x^* linearly if $x_\ell \to x^*$ as $\ell \to \infty$ and $x_\ell \geq x^*\ell/(\ell + c)$ for some constant c that can be queue dependent. The convergence rate is a technical assumption required for the proof of Lemma 3, though we suspect that it may not be necessary.

Trip time and completion: The trip travel time from Region i to j is drawn

independently from a known distribution with mean $\mu_{ij}^{-1} < \infty$. Upon completion of the trip, the vehicle does not have enough battery charge for another trip with probability p_{ij}, independently of all else. For convenience, we take $p_{ij} = p$ for all trips, but our results and analysis extend to the setting with heterogeneous p_{ij}. If the vehicle is uncharged, it cannot be used for a trip until it is recharged. Upon being recharged, the vehicle becomes available for use in the same region where its last trip was completed, so that we only consider recharging and not rebalancing.

Charging Dynamics: The recharging rate in Region i, as a function of the number, x, of uncharged vehicles in the region, is $\xi_i(x)$. Charging processes are independent in all regions. State-dependent charging rates are chosen to reflect real-world dynamics of e-scooter sharing systems. For example, if an employee travels between uncharged vehicles in a region to swap batteries, more uncharged vehicles means the average distance travelled between successive vehicles is smaller, and the overall recharging rate increases. We assume that the function ξ_i is monotone non-decreasing and that the recharging rate is eventually fast enough to keep up with demand in the region, i.e., $\xi_i(\infty) = \lim_{x\to\infty} \xi_i(x) \geq d_i$ for all regions i. If the maximum charging rate is finite, $\xi_i(\infty) < \infty$, then we require a similar technical condition of linear convergence to the limit. Since most e-scooter sharing systems have many vehicles, if charging is not fast enough in a region in the limit, then the region could be a bottleneck, in which case system operators would have an economic incentive to increase charging resources in the region.

Summary: For charging queues, there are two types of arrivals: trip completions with battery levels above threshold, and charged vehicles departing

from the uncharged queue in the same region. There is only one type of departure: trip initiations. For uncharged queues, an arrival occurs when a trip completes with battery below the threshold and departures occur when a vehicle is recharged. For transit queues, an arrival occurs when a trip is initiated and a departure occurs when a trip is completed.

2.3.1 Stationary Distribution

Under any static pricing policy, the network belongs to the class of BCMP networks that have product-form stationary distributions [Baskett et al., 1975]. We begin with preliminaries required for computing the stationary distribution. Consider a specific queue, in the network. Let the relative arrival rate be λ and the state-dependent service rates be $\mu(k)$. Considering only this birth-death process, and ignoring the network, we calculate a stationary distribution up to a normalization constant. We refer to this quantity as the unnormalized stationary distribution, $\tilde{\pi}$, whose mathematical form is

$$\tilde{\pi}(k) = \begin{cases} 1, & k = 0 \\ \dfrac{\lambda^k}{\prod_{l=1}^{k} \mu(l)}, & k \geq 1. \end{cases} \tag{2.1}$$

In what follows we calculate an unnormalized stationary distribution for each queue in our network for all values of $k \in \{0, 1, ...m\}$. It is important to note that this distribution, even when properly normalized, is not the same as the marginal stationary distribution of the queue. In the marginal distribution, the arrival rate to the queue is state-dependent. For example, if all vehicles are at the queue, then the arrival rate is 0.

First consider queues of uncharged vehicles. In region i, the arrival rate is $p\lambda_i$ and the recharging rate is $\xi_i(X_i^u)$. Therefore, the unnormalized stationary distribution is

$$\tilde{\pi}_i^u(k) = \begin{cases} 1, & k = 0 \\ \dfrac{p^k \lambda_i^k}{\prod_{l=1}^k \xi_i(l)}, & k = 1, 2, ..., m. \end{cases} \tag{2.2}$$

Second, for the queue of charged vehicles in region i, the arrival rate is λ_i and the rate at which trips begin is $d_i\zeta_i(X_i^c)$. (Recall that there are two sources of arrivals: vehicles completing trips with enough charge at rate $(1-p)\lambda_i$ and vehicles completing recharging at rate $p\lambda_i$.) Therefore, the unnormalized stationary distribution is

$$\tilde{\pi}_i^c(k) = \begin{cases} 1, & k = 0 \\ \dfrac{\lambda_i^k}{d_i^k \prod_{l=1}^k \zeta_i(l)}, & k = 1, 2, ..., m. \end{cases} \tag{2.3}$$

Third, for the $\cdot/G/\infty$ transit queues, the arrival rate is $\lambda_{ij} = \omega_{ij}\lambda_i$, where ω_{ij} is the probability, for rides originating in Region i, of selecting Region j as the destination and the mean service time is μ_{ij}^{-1}. The unnormalized stationary distribution has the same form as an infinite server queue and therefore is proportional to the truncated Poisson distribution. The exact form is

$$\tilde{\pi}_{ij}^t(k) = \frac{\left(\frac{\lambda_{ij}}{\mu_{ij}}\right)^k}{k!}, \quad k = 0, 1, ..., m. \tag{2.4}$$

Let π be the probability mass function of the product-form stationary distribution of the network and $S_{m,n} = \{\mathbf{x} \,|x \in \mathbb{Z}^n,\ x \geq 0,\ \sum_i \left(x_i^c + x_i^u\right) + \sum_i \sum_j x_{ij}^t = m\}$ be the set of all states where the m vehicles are distributed across the network. Then, for $x \in S_{m,n}$,

$$\pi(\mathbf{x}) = \frac{1}{G_{m,n}} \prod_{i=1}^n \tilde{\pi}_i^u(x_i^u) \prod_{i=1}^n \tilde{\pi}_i^c(x_i^c) \prod_{i=1}^n \prod_{j=1}^n \tilde{\pi}_{ij}^t(x_{ij}^t)$$

where $G_{m,n}$ is the normalization constant for the network with m vehicles and n regions and is given by

$$G_{m,n} = \sum_{\mathbf{x} \in S_{m,n}} \prod_{i=1}^{n} \tilde{\pi}_i^u(x_i^u) \prod_{i=1}^{n} \tilde{\pi}_i^c(x_i^c) \prod_{i=1}^{n} \prod_{j=1}^{n} \tilde{\pi}_{ij}^t(x_{ij}^t).$$

2.3.2 Pricing Problem

Our goal is to find a static pricing policy that has provably near-optimal performance relative to the optimal dynamic pricing policy.

We extend the flow approximation framework introduced in Banerjee et al. [2022] by incorporating charging dynamics and state-dependent service rates. Thus our development has a similar structure and notation to that of Banerjee et al. [2022]. For simplicity, we focus on the case of maximizing system throughput and will discuss later how this extends to other objective functions. Solving the optimal state-dependent pricing problem is challenging since, even in the simpler setting, this problem is high-dimensional and non-convex [Banerjee et al., 2022].

Let $OBJ_m(\mathbf{q})$ be the expected throughput rate of the system under pricing policy $\mathbf{q}$ with m total vehicles, i.e.,

$$OBJ_m(\mathbf{q}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \phi_{ij} q_{ij} \underbrace{\sum_{k=1}^{m} \left(\mathbb{P}_{\mathbf{q}}[k \text{ vehicles at } i] \zeta_i(k) \right)}_{A_{i,m}(\mathbf{q})}.$$

Similarly, $OBJ_\infty(\mathbf{q})$ represents the limiting availability achieved as the number of vehicles $m \to \infty$.

Optimal Dynamic Policy

Let the optimal state-dependent pricing policy be $\mathbf{q}^{m*}(\mathbf{x})$ with stationary distribution π^{m*}. Then, the optimal throughput of an m-unit system is

$$OPT_m(\mathbf{q}^*) = \sum_{\mathbf{x}\in S_{m,n}} \pi^{m*}(\mathbf{x}) \left(\sum_{i,j} \phi_{ij} q_{ij}^{m*}(\mathbf{x}) \zeta_i(x_i^c) \right).$$

Elevated Flow Relaxation

We use the same infinite-unit limit considered in Banerjee et al. [2022]. At least 1 queue in the network (a bottleneck) will accumulate an infinite number of vehicles. Since we assumed that the recharging rate is eventually faster than the incoming rate of uncharged vehicles ($\lim_{x\to\infty} \xi_i(x) > d_i$), an infinite number of vehicles will not accumulate in the uncharged queue. Similarly, an infinite number of vehicles will not accumulate in the transit (infinite) server queues. Therefore, any bottleneck has to be a charged queue. We show later that availability (the fraction of time that at least one vehicle is sufficiently close) across all charged queues is the same, even with different spatial demand profiles. Therefore, in this limit all charged queues accumulate an infinite number of vehicles and spatial demand complexities vanish due to our assumption that $\lim_{k\to\infty} \zeta_i(k) = 1$. We thus begin by finding the balanced demand policy that maximizes throughput, considering neither charging nor spatial demand complexities. This is found by solving the elevated flow relaxation,

$$\begin{aligned} \max_{q} \quad & \sum_{ij} \phi_{ij} q_{ij} \\ \text{s.t.} \quad & \sum_{k} \phi_{ki} q_{ki} = \sum_{j} \phi_{ij} q_{ij} \qquad \forall i \\ & q_{ij} \in [0,1] \qquad\qquad \forall i, j, \end{aligned} \tag{2.5}$$

which maximizes throughput subject to a circulation constraint. Let an optimal solution be denoted by $\hat{q}$, with objective value $\widehat{OBJ}(\hat{q})$. We begin by proving that the elevated flow relaxation is an upper bound on the finite-unit performance.

Lemma 1. *The optimal elevated flow yields an upper bound on optimal finite-m performance, i.e.,* $\widehat{OBJ}(\hat{q}) \geq OPT_m$.

The proof of this lemma can be found in [Banerjee et al., 2022] and is essentially as follows. Consider the state-independent policy, $\bar{q}$, where $\bar{q}_{ij}$ is defined by the long-run average fraction of customers arriving to take a ride from i to j who complete their trip under an optimal state-dependent policy (π^{m*}) We set $\hat{q}_{ij}(x_i) = 0$ if $x_i = 0$. Under the (finite m) stationary distribution, the average rate of trips beginning and ending in a region must be the same since there are a fixed number of vehicles. Therefore, $\bar{q}$ satisfies the circulation constraint and is a feasible solution to the elevated flow relaxation. By construction of the policy, the (finite-m) performance of the optimal state-dependent policy is equal to the elevated-flow objective function value of $\bar{q}$. Therefore, the optimal objective function value of the elevated flow relaxation is an upper bound on the objective function value of any optimal state-dependent policy for finite m.

Lemma 1 applies even to pricing policies that do not satisfy the balanced-flow assumption; such policies yield limiting flow rates $\bar{q}$ that *do* satisfy balanced flow, and thus the bound holds.

Demand-Balanced Policy

We focus on demand-balanced, static pricing policies $\mathbf{q}$ in which all locations are equally popular as origins and destinations, i.e. $\sum_j \phi_{ij} q_{ij} = \sum_k \phi_{ki} q_{ki}$ for all

regions i.

As introduced previously, $d_i = \sum_j \phi_{ij} q_{ij} > 0$ is the total rate of customers arriving to Region i requesting a trip and willing to pay the price for the trip. It also represents the rate at which trips would begin from Region i if there was an infinite number of charged vehicles there. Let Ω be the routing matrix with entries $\omega_{ij} = \phi_{ij} q_{ij} / d_i$ that correspond to the probability that a trip has destination j given that it began at i, so that $\sum_j \omega_{ij} = 1$ for all regions i. We assume that the routing matrix is irreducible. Relative arrival rates (λ_i) to regions are given by the invariant distribution corresponding to the stochastic matrix Ω. Then, we can define the relative throughput of each charged queue as $r_i^c = \lambda_i / d_i$ (which does not reflect the thinning of demand due to spatial demand complexities). As a consequence of balanced demand, λ_i is proportional to d_i and therefore all r_i^c are equal for all charged queues.

A central quantity for computing performance guarantees when the system operates with a finite number of vehicles, m, is *availability*. Availability refers to the probability that a user who is willing to pay for a trip can find a sufficiently close charged vehicle. Let $A_{i,m}(\mathbf{q})$ be the availability in Region i when the system operates with m vehicles and static pricing policy $\mathbf{q}$. Also, let $A_{i,\infty}(\mathbf{q}) = \lim_{m \to \infty} A_{i,m}(\mathbf{q})$, assuming the limit exists.

Lemma 2. *Balanced demand implies equal availabilities.*

Proof. Proof of Lemma 2 We begin by considering the marginal distribution of charged vehicles in region i. Let $S_{m,n}^{-i} = \{\mathbf{x} \in S_{m,n} \mid \sum_j \left(x_j^c + x_j^u\right) + \sum_j \sum_k x_{jk}^t = m,\ x_i^c = 0\}$ be the set of all states where there are no vehicles in the charged queue at location i and $\tilde{\pi}_{m,-i}(\mathbf{x}) = \tilde{\pi}_m(\mathbf{x}) / \tilde{\pi}_i^c(x_i^c)$ be the unnormalized stationary distribution

without the term associated with the charged queue at i. We begin by considering the marginal distribution of charged vehicles at i. Conditioned on having b vehicles in the charged queue at i, we know that the remaining $m - b$ vehicles must be distributed between the remaining queues, so

$$\mathbb{P}[X_i^c = b] = \sum_{\mathbf{x} \in S_{m,n}} \pi_m(x) \mathbb{1}[X_i^c = b] = \sum_{\mathbf{x} \in S_{m,n}} \frac{\tilde{\pi}_m(x) \mathbb{1}[X_i^c = b]}{G_{m,n}} = \sum_{\mathbf{y} \in S_{m-b,n}^{-i}} \frac{\tilde{\pi}_{m,-i}(\mathbf{y}) \tilde{\pi}_i^c(b)}{G_{m,n}}$$

With the marginal distribution of charged vehicles at i, we find that the availability in Region i is

$$\begin{aligned}
A_{i,m}(\mathbf{q}) &= \sum_{b=1}^{m} \mathbb{P}[X_i^c = b] \zeta_i(b) = \sum_{b=1}^{m} \left(\sum_{\mathbf{y} \in S_{m-b,n}^{-i}} \frac{\tilde{\pi}_{m,-i}(\mathbf{y}) \tilde{\pi}_i^c(b)}{G_{m,n}} \right) \zeta_i(b) \frac{d_i}{\lambda_i} \frac{\lambda_i}{d_i} \\
&= \sum_{b=1}^{m} \left(\sum_{\mathbf{y} \in S_{m-b,n}^{-i}} \frac{\tilde{\pi}_{m,-i}(\mathbf{y})}{G_{m,n}} \right) \left(\tilde{\pi}_i^c(b) \zeta_i(b) \frac{d_i}{\lambda_i} \right) \frac{\lambda_i}{d_i} \\
&= \sum_{b=1}^{m} \left(\sum_{\mathbf{y} \in S_{m-b,n}^{-i}} \frac{\tilde{\pi}_{m,-i}(\mathbf{y})}{G_{m,n}} \right) (\tilde{\pi}_i^c(b-1))\, r_i = \frac{r_i}{G_{m,n}} \sum_{b=1}^{m} \left(\sum_{y \in S_{m-b,n}^{-i}} \tilde{\pi}_{m,-i}(\mathbf{y}) \tilde{\pi}_i^c(b-1) \right) \\
&= r_i \frac{G_{m-1,n}}{G_{m,n}}.
\end{aligned}$$

To go from line two to three, we use properties of calculating stationary distributions of birth-death processes. On the third line, the sum over all values of b and over all remaining states is equivalent to summing over all states with total $m - 1$ vehicles (since the number of vehicles in the charged queue at i is $b - 1$). Recall, the r_i are equal for all regions i. Therefore, all regions have the same availability. □

We next want to show that as the number of vehicles $m \to \infty$, availability goes to 1 in all regions. This is intuitive, yet our proof is not straightforward. In the setting where the r_is are all distinct, Pollett [2000] provided a proof where queues are added, one at a time, inductively. We adapt that proof to our setting,

overcoming the challenge that we have tied r_i values, at least across all charged queues. We must also add the uncharged queues that have r_i values that are bounded above by the tied value for charged queues, and we must add the transit queues that have $r_i = 0$. Our proof relies on an assumption that the services rates in all non-transit queues converge at least polynomially quickly to their limiting service rates, which we conjecture could be relaxed.

Lemma 3. *Assume that for all queues with finite maximum service rates* ($\mu_i(\infty) < \infty$), *the service rates converge polynomially quickly to* $\mu_i(\infty)$. *Then* $\lim_{m\to\infty} A_{i,m}(\mathbf{q}) = 1$ *in all Regions, i.*

Proof. Proof of Lemma 3 For this proof only, we abuse the relative throughput notation that we had previously defined for each region based on its charged queue. Let $r_i = \lambda_i/\mu_i(\infty)$ for all queues i, and not just for charged queues. Divide the queues into 2 groups: $r_i > 0$ and $r_i = 0$ with n_1 and n_2 queues respectively, where $n_1 + n_2 = 2n + n^2$, corresponding to $\mu_i(\infty) < \infty$ for Group 1 and $\mu_i(\infty) = \infty$ for Group 2. Let the queues in Group 1 be labelled according to increasing relative throughput, $0 < r_1 \leq r_2 \leq r_3 ... \leq r_{n_1}$, breaking ties arbitrarily except that charged queues come last. Increasing relative throughput means the queues get progressively busier with the bottleneck being the queue with the largest relative throughput. In Lemma 4, we iteratively add queues in this order, showing that immediately after adding Queue i, $r_i \lim_{m\to\infty} G_{m-1,i}/G_{m,i} = 1$. Here, $G_{m,i}$ is the normalization constant for a system with m vehicles and i queues. We assumed that charging is fast enough in the limit, so the charged queues come last in Group 1. Since all charged queues have equal relative throughput, $r_{\max}$ say, it follows that $r_i \lim_{m\to\infty} G_{m-1,n_1}/G_{m,n_1} = 1$ for all charged queues i, i.e., all regions have limiting availability 1. After all queues from Group 1 have been added, we

add the queues from Group 2, and Lemma 5 shows that adding these queues does not change the limiting availability that we already calculated. □

Lemma 4. *Suppose queues with relative throughputs* $0 < r_1 \leq r_2 \leq ... \leq r_{n_1}$ *are added in this order. For each* $i = 1, 2 \ldots, n_1$*, immediately after adding the ith queue in Group 1,* $r_i \lim_{m\to\infty} G_{m-1,i}/G_{m,i} = 1$.

Proof. Proof of Lemma 4 We prove this claim by induction on the number of queues.

Base Case: 1 queue With one queue, there is only one way that m vehicles can be distributed. Therefore, the normalization constant $G_{m,1} = \tilde{\pi}_1(m)$. Recall, $r_1 = \lambda_1/\mu_1(\infty)$ and $\tilde{\pi}(m) = \tilde{\pi}(m-1)\lambda/\mu_1(m)$. Therefore, as $m \to \infty$,

$$\frac{r_1 G_{m-1,1}}{G_{m,1}} = \frac{r_1\tilde{\pi}_1(m-1)}{\tilde{\pi}_1(m)} = \frac{\lambda_1}{\mu_1(\infty)}\frac{\tilde{\pi}(m-1)}{\tilde{\pi}(m-1)\frac{\lambda_1}{\mu_1(m)}} = \frac{\mu_1(m)}{\mu_1(\infty)} \to 1.$$

Inductive Hypothesis: Assume the result is true for $1 \leq k \leq n_1 - 1$ queues. We want to show that $r_{k+1}G_{m-1,k+1}/G_{m,k+1} \to 1$. We partition into 2 cases, (1) $r_{k+1} > r_k$ and (2) $r_{k+1} = r_k$. The proof for the first case is adapted from Proposition 2 of Pollett [2000]. The proof of the second case is a new argument. For notational convenience, let $r = r_{k+1}$, $G_m = G_{m,k+1}$, $G_m^- = G_{m,k}$, and $\tilde{\pi}(\cdot) = \tilde{\pi}_{k+1}(\cdot)$.

Case 1: $r_{k+1} > r_k$. We use the recursive expression for the normalization constant

$$G_m = \sum_{l=0}^{m} G_l^- \tilde{\pi}(m-l)$$

to expand the expression for availability we derived previously. We also pull a factor of $\tilde{\pi}(m-1)$ and $\tilde{\pi}(m)$ out of the numerator and denominator respectively.

This yields

$$\frac{rG_{m-1}}{G_m} = \frac{r\sum_{l=0}^{m-1} G_l^- \tilde{\pi}(m-1-l)}{\sum_{l=0}^{m} G_l^- \tilde{\pi}(m-l)} = \frac{r\tilde{\pi}(m-1)}{\tilde{\pi}(m)} \frac{\sum_{l=0}^{m-1} G_l^- \frac{\tilde{\pi}(m-1-l)}{\tilde{\pi}(m-1)}}{\sum_{l=0}^{m} G_l^- \frac{\tilde{\pi}(m-l)}{\tilde{\pi}(m)}}.$$

Using the same argument as the base case, we know that $\lim_{m\to\infty} r\tilde{\pi}(m-1)/\tilde{\pi}(m) = 1$. Therefore, if we show that

$$\lim_{m\to\infty} \frac{\sum_{l=0}^{m-1} G_l^- \frac{\tilde{\pi}(m-1-l)}{\tilde{\pi}(m-1)}}{\sum_{l=0}^{m} G_l^- \frac{\tilde{\pi}(m-l)}{\tilde{\pi}(m)}} = 1,$$

then the inductive hypothesis holds. Let $d_m = \sum_{l=0}^{m} G_l^- \tilde{\pi}(m-l)/\tilde{\pi}(m)$, allowing us to re-write the expression as $\lim_{m\to\infty} d_{m-1}/d_m$. If $\lim_{m\to\infty} d_m$ exists and lies in $(0,\infty)$, then the result follows immediately. Define the power series $G^-(\gamma) = \sum_{l=0}^{\infty} G_l^- \gamma^l$ with radius of convergence γ_G, and $\Pi(\gamma) = \sum_{l=0}^{\infty} \tilde{\pi}(l)\gamma^l$ with radius of convergence $\gamma_\pi < \gamma_G$. From the inductive hypothesis, we know that $G_l^-/G_{l+1}^- = G_{l,k}/G_{l+1,k} \to r_k$ as $l \to \infty$, so by the ratio test, the power series G^- converges for all $\gamma < 1/r_k$. Similarly, since $\lim_{l\to\infty} \tilde{\pi}(l-1)/\tilde{\pi}(l) = r_{k+1}$, the power series Π converges for all $\gamma < 1/r_{k+1}$. We next show that $\lim_{m\to\infty} d_m = G^-(1/r_{k+1}) < \infty$, since $1/r_{k+1} < 1/r_k$. We begin by defining

$$f_l(m) = \begin{cases} \dfrac{\tilde{\pi}(m-l)}{\tilde{\pi}(m)}, & \text{if } l \le m \\ 0, & \text{if } l > m. \end{cases}$$

Then, for each l, $f(l) = \lim_{m\to\infty} f_l(m) = 1/r_{k+1}^l$. By Fatou's lemma,

$$\sum_{l=0}^{\infty} G_l^- f(l) = \sum_{l=0}^{\infty} G_l^- \frac{1}{r_{k+1}^l} \le \liminf_{m\to\infty} \sum_{l=0}^{\infty} G_l^- f_l(m) \implies G^-\left(\frac{1}{r_{k+1}}\right) \le \liminf_{m\to\infty} d_m.$$

Moreover,

$$\tilde{\pi}_{k+1}(m) = \tilde{\pi}_{k+1}(m-l)\frac{\lambda_{k+1}^l}{\prod_{z=m-l+1}^{m} \mu_{k+1}(z)} \implies \frac{\tilde{\pi}_{k+1}(m-l)}{\tilde{\pi}_{k+1}(m)} = \frac{\prod_{z=m-l+1}^{m} \mu_{k+1}(z)}{\lambda_{k+1}^l}.$$

Since $\mu_{k+1}(\cdot)$ is a non-decreasing function, the numerator is non-decreasing in m for all l. Thus, as $m \to \infty$, $\tilde{\pi}_{k+1}(m-l)/\tilde{\pi}_{k+1}(m) \uparrow 1/r_{k+1}^l$ and therefore,

$$\limsup_{m\to\infty} d_m \le \limsup_{m\to\infty} \sum_{l=0}^{m} \frac{G_l^-}{r_{k+1}^l} = G^-\left(\frac{1}{r_{k+1}}\right),$$

completing the proof of Case 1.

Case 2: $r_{k+1} = r_k$. We have that

$$\begin{aligned}
rG_{m-1} &= r\sum_{l=0}^{m-1} G_l^-\tilde{\pi}(m-l-1) = \sum_{l=0}^{m-1} G_l^-\tilde{\pi}(m-l)\left(\frac{\mu_{k+1}(m-l)}{\lambda_{k+1}}\right)r = \sum_{l=0}^{m-1} G_l^-\tilde{\pi}(m-l)\left(\frac{\mu_{k+1}(m-l)}{\mu_{k+1}(\infty)}\right)\\
&= \sum_{l=0}^{m-1} G_l^-\tilde{\pi}(m-l)\left[1 - \frac{\mu_{k+1}(\infty)-\mu_{k+1}(m-l)}{\mu_{k+1}(\infty)}\right]\\
&= \sum_{l=0}^{m-1} G_l^-\tilde{\pi}(m-l) - \sum_{l=0}^{m-1} G_l^-\tilde{\pi}(m-l)\frac{\mu_{k+1}(\infty)-\mu_{k+1}(m-l)}{\mu_{k+1}(\infty)}\\
&= \sum_{l=0}^{m} G_l^-\tilde{\pi}(m-l) - G_m^- - \sum_{l=0}^{m-1} G_l^-\tilde{\pi}(m-l)\frac{\mu_{k+1}(\infty)-\mu_{k+1}(m-l)}{\mu_{k+1}(\infty)}.
\end{aligned}$$

Thus,

$$\frac{rG_{m-1}}{G_m} = 1 - \frac{G_m^-}{G_m} - \frac{1}{G_m}\left(\sum_{l=0}^{m-1} G_l^-\tilde{\pi}(m-l)\frac{\mu_{k+1}(\infty)-\mu_{k+1}(m-l)}{\mu_{k+1}(\infty)}\right)$$

The inductive hypothesis follows if the last two terms go to zero in the limit. We begin with the first of these.

Claim 1: $\lim_{m\to\infty} G_m^-/G_m = 0$. We have

$$\begin{aligned}
G_m &= \sum_{l=0}^{m} G_l^-\tilde{\pi}(m-l) = G_m^-\tilde{\pi}(0) + G_{m-1}^-\tilde{\pi}(1) + \cdots + G_0^-\tilde{\pi}(m)\\
&= G_m^- + G_{m-1}^-\frac{\lambda_{k+1}}{\mu_{k+1}(1)}\frac{\mu_{k+1}(\infty)}{\mu_{k+1}(\infty)} + \cdots + G_0^-\frac{\lambda_{k+1}^m}{\prod_{l=1}^m \mu_{k+1}(l)}\frac{\mu_{k+1}(\infty)^m}{\mu_{k+1}(\infty)^m}\\
&= G_m^- + G_{m-1}^- r\frac{\mu_{k+1}(\infty)}{\mu_{k+1}(1)} + \cdots + G_0^- r^m\frac{\mu_{k+1}(\infty)^m}{\prod_{l=1}^m \mu_{k+1}(l)}
\end{aligned}$$

From the monotonicity in l of service rates $\mu_{k+1}(l)$, all ratios $\mu_{k+1}(\infty)/\mu_{k+1}(l) \ge 1$ for all l. Therefore,

$$\frac{G_m}{G_m^-} \ge 1 + \frac{rG_{m-1}^-}{G_m^-} + \frac{rG_{m-1}^-}{G_m^-}\frac{rG_{m-2}^-}{G_{m-1}^-} + \cdots + \prod_{l=1}^{m}\frac{rG_{m-l}^-}{G_{m-l+1}^-}$$

and we can write this expression recursively. Recall that $h(m) := rG^-_{m-1}/G^-_m \in [0,1]$ since it is the availability with m vehicles and the first k queues. Let $g(m) = G_m/G^-_m$, which gives us the recursive relationship $g(m) \geq 1 + h(m)g(m-1)$. In addition, $r_{k+1} = r_k$ which means $h(m) = rG^-_{m-1}/G^-_m \to 1$ as $m \to \infty$ from our inductive hypothesis. Since $h(m) \to 1$ as $m \to \infty$, $g(m)$ diverges, proving our claim.

Claim 2: Let

$$\Lambda_m = \frac{1}{G_m}\left(\sum_{l=0}^{m-1} G^-_l \tilde{\pi}(m-l)\frac{\mu_{k+1}(\infty) - \mu_{k+1}(m-l)}{\mu_{k+1}(\infty)}\right).$$

Then, $\lim_{m\to\infty} \Lambda_m = 0$.

We have

$$\begin{aligned}\Lambda_m &= \frac{1}{G_m}\left(\sum_{l=0}^{m-1} G^-_{m-1-l}\tilde{\pi}(l+1)\frac{\mu_{k+1}(\infty) - \mu_{k+1}(l+1)}{\mu_{k+1}(\infty)}\right)\\ &= \frac{G^-_m}{G_m}\left(\sum_{l=0}^{m-1} \frac{G^-_{m-1-l}}{G^-_m}\tilde{\pi}(l+1)\left(1 - \frac{\mu_{k+1}(l+1)}{\mu_{k+1}(\infty)}\right)\right)\end{aligned}$$

From Claim 1, we know the limit of $G^-_m/G_m = 0$ as $m \to \infty$ and from availability being upper bounded by 1, $G^-_{m-1-l}/G^-_m \leq (1/r_k)^{l+1}$. This yields

$$\Lambda_m \leq \frac{G^-_m}{G_m}\left(\sum_{l=0}^{m-1} \frac{1}{r^{l+1}}\tilde{\pi}(l+1)\left(1 - \frac{\mu_{k+1}(l+1)}{\mu_{k+1}(\infty)}\right)\right)$$

We use the variable transformation $\mu_{k+1}(j) = \mu_{k+1}(\infty)e^{a_j}$ where $a_j < 0$ and $a_j \uparrow 0$ as $j \to \infty$. Therefore, we can express

$$\tilde{\pi}(l) = \prod_{j=1}^{l}\frac{\lambda_{k+1}}{\mu_{k+1}(j)} = \frac{\lambda^l_{k+1}}{\mu_{k+1}(\infty)^l\left(e^{\sum_{j=1}^l a_j}\right)} = \frac{r^l}{e^{\sum_{j=1}^l a_j}},$$

and then,

$$\begin{aligned}\Lambda_m &\leq \frac{G^-_m}{G_m}\left(\sum_{l=0}^{m-1} e^{-\sum_{j=1}^{l+1} a_j}\left(1 - e^{a_{l+1}}\right)\right)\\ &\leq \frac{G^-_m}{G_m}e^{-\sum_{j=1}^{\infty} a_j}\sum_{l=1}^{m}\left(1 - e^{a_l}\right).\end{aligned}$$

We assumed that $\mu_{k+1}(l) \geq \mu_{k+1}(\infty)(1 - l^{-p})$, for some $p > 1$ and $l \geq \kappa$. Thus $1 - e^{a_l} \leq l^{-p}$ for $l \geq \kappa$, and then

$$\Lambda_m \leq \frac{G_m^-}{G_m} e^{-\sum_{j=1}^{\infty} a_j} \sum_{l=1}^{\infty} l^{-p}.$$

But $-\sum_{j=1}^{\infty} a_j < \infty$ because $-a_j \leq -\ln(1-j^{-p}) \sim j^{-p}$ as $j \to \infty$, so that $\sum_{j=1}^{\infty}(-a_j) < \infty$ if $\sum_{j=1}^{\infty} j^{-p} < \infty$; the latter holds because $p > 1$. Thus, Claim 2 follows from Claim 1.

□

Lemma 5. *Given the network of queues from Proposition 4, adding a new queue* $(k+1)$*, with* $\mu_{k+1}(\infty) = \infty$*, does not change the previously calculated availabilities.*

Proof. Proof of Proposition 5 Continuing with the same notation as in Proposition 4, we know that $r_i \lim_{m\to\infty} G_{m-1}^-/G_m^- = 1$ (where $r_i = r_{\max}$). We want to show that, after adding the new queue, $r_i \lim_{m\to\infty} G_{m-1}/G_m = 1$. We begin by expanding this expression and factoring out G_{m-1}^- and G_m^- from the numerator and denominator respectively, to get

$$r_i \frac{G_{m-1}}{G_m} = r_i \frac{\sum_{l=0}^{m-1} G_l^- \tilde{\pi}(m-l-1)}{\sum_{l=0}^{m} G_l^- \tilde{\pi}(m-l)} = r_i \frac{G_{m-1}^- \sum_{l=0}^{m-1} \frac{G_l^-}{G_{m-1}^-} \tilde{\pi}(m-l-1)}{G_m^- \sum_{l=0}^{m} \frac{G_l^-}{G_m^-} \tilde{\pi}(m-l)}$$
$$= r_i \frac{G_{m-1}^-}{G_m^-} \frac{\sum_{l=0}^{m-1} \frac{G_l^-}{G_{m-1}^-} \tilde{\pi}(m-l-1)}{\sum_{l=0}^{m} \frac{G_l^-}{G_m^-} \tilde{\pi}(m-l)}.$$

Since $r_i \lim_{m\to\infty} G_{m-1}^-/G_m^- = 1$, it suffices to show that

$$\lim_{m\to\infty} \frac{\sum_{l=0}^{m-1} \frac{G_l^-}{G_{m-1}^-} \tilde{\pi}(m-l-1)}{\sum_{l=0}^{m} \frac{G_l^-}{G_m^-} \tilde{\pi}(m-l)} = 1.$$

Let

$$e_m = \sum_{l=0}^{m} \frac{G_l^-}{G_m^-} \tilde{\pi}(m-l) = \sum_{l=0}^{m} \frac{G_{m-l}^-}{G_m^-} \tilde{\pi}(l)$$

which allows us to rewrite the condition as $\lim_{m\to\infty} e_{m-1}/e_m = 1$. This follows immediately if

$$0 < \lim_{m\to\infty} e_m = \sum_{l=0}^{\infty} \frac{\tilde{\pi}(l)}{r_i^l} < \infty.$$

The infinite sum in this expression is finite by the ratio test, since

$$\lim_{l\to\infty} \frac{\tilde{\pi}(l+1)}{r_i^{l+1}} \frac{r_i^l}{\tilde{\pi}(l)} = \frac{\lambda_{k+1}}{\mu_{k+1}(l+1)r_i} < 1,$$

which holds for all l above a threshold because $\lim_{l\to\infty} \mu_{k+1}(l) = \infty$. Now, we turn our attention to proving that $e_m \to \sum_{l=0}^{\infty} \tilde{\pi}(l)/r_i^l$ as $m \to \infty$. Let

$$g_l(m) = \begin{cases} \dfrac{G^-_{m-l}}{G^-_m}, & \text{if } l \le m \\ 0, & \text{if } l > m. \end{cases}$$

We have that

$$\frac{G^-_{m-l}}{G^-_m} = \left[r_i \frac{G^-_{m-l}}{G^-_{m-l+1}}\right]\left[r_i \frac{G^-_{m-l+1}}{G^-_{m-l+2}}\right] \cdots \left[r_i \frac{G^-_{m-1}}{G^-_m}\right] \frac{1}{r_i^l} \tag{2.6}$$

and for each term in the product, $\lim_{m\to\infty} r_i G^-_{m-l+s}/G^-_{m-l+s+1} = 1$ because $\lim_{m\to\infty} r_i G^-_{m-1}/G^-_m = 1$. Then, for each l, $g(l) = \lim_{m\to\infty} g_l(m) = 1/r_i^l$ since $r_i = r_{\max}$. Fatou's lemma then gives

$$\sum_{l=0}^{\infty} \tilde{\pi}(l) g(l) = \sum_{l=0}^{\infty} \frac{\tilde{\pi}(l)}{r_i^l} \le \lim_{m\to\infty} \sum_{l=0}^{\infty} \tilde{\pi}(l) g_l(m) \implies \sum_{l=0}^{\infty} \frac{\tilde{\pi}(l)}{r_i^l} \le \liminf_{m\to\infty} e_m.$$

In addition, since availability is bounded by 1, (2.6) ensures that $g_l(m) \le r_i^{-l}$, and then

$$\limsup_{m\to\infty} e_m \le \limsup_{m\to\infty} \sum_{l=0}^{m} \frac{\tilde{\pi}(l)}{r_i^l} < \infty.$$

□

Lemma 6. *For any state-independent pricing policy* $\mathbf{q}$ *satisfying balanced demand, the objective value of the infinite-unit system equals the objective value of the elevated flow relaxation,* $OBJ_\infty(q) = \widehat{OBJ}(\mathbf{q})$ *(Lemma 4, Banerjee et al., 2022).*

Proof. Proof of Lemma 6 Balanced demand and Lemma 3 imply that availability of all regions converges to 1 as $m \to \infty$. Therefore, $A_{i,\infty}(q) = 1$ and $OBJ_\infty(q) = \sum_{i,j} \phi_{ij} q_{ij} A_{i,\infty}(q) = \sum_{i,j} \phi_{ij} q_{ij} = \widehat{OBJ}(q)$. □

Lemma 7. *For any state-independent pricing policy* $\mathbf{q}$*, the approximation ratio is availability:*

$$\frac{OBJ_m(\mathbf{q})}{OBJ_\infty(\mathbf{q})} = A_m(\mathbf{q}).$$

Proof. Proof of Lemma 7 This follows exactly as in Banerjee et al. [2022], since

$$\frac{OBJ_m(\mathbf{q})}{OBJ_\infty(\mathbf{q})} = \frac{\sum_{i,j} \phi_{ij} q_{ij} A_{i,m}(\mathbf{q})}{\sum_{i,j} \phi_{ij} q_{ij} A_{i,\infty}(\mathbf{q})} = \frac{\sum_{ij} \phi_{ij} q_{ij} r_i \frac{G_{m-1}}{G_m}}{\sum_{i,j} \phi_{ij} q_{ij} (r_i / r_{\max})} = r_{\max} \frac{G_{m-1}}{G_m} = A_m(\mathbf{q}).$$

□

Since the infinite-unit limit has the same objective value as the elevated flow relaxation (which is an upper bound to the optimal state-dependent pricing policy), the approximation ratio of our state-independent pricing policy is at least the availability of the system, $A_m(\mathbf{q}) = r_i G_{m-1}/G_m$. This is the same expression as the simpler setting studied in previous work (without state-dependent service rates and charging). This implies that complexities around these additional features are contained within the normalizing constants.

In this section, the closed-queueing model was developed, and the stationary distribution and normalization constants were discussed. Then, the state-dependent pricing problem and the elevated flow relaxation (an upper bound corresponding to infinite vehicles) were introduced. We analyze a class of 'demand-balanced' static pricing policies which includes the elevated flow relaxation policy. We show that all regions have the same availability and that availability goes to 1 as the number of vehicles go to infinity. Lastly, we conclude by showing that the availability is the ratio of throughput of the finite and

infinite unit systems. Therefore, when analyzing the elevated flow's policy, the availability is an approximation ratio for the system.

2.4 Constant Service Rates

We begin by examining a simpler version of the system described previously, with no charging dynamics and constant (state-independent) service rates. This is the only setting that has been examined in literature to the best of our knowledge. Later in this section, we introduce charging dynamics while maintaining constant service rates.

2.4.1 Non-electric Vehicles

The setting without charging dynamics can be considered as trying to optimize a system with traditional, non-electric vehicles. The system has n queues for vehicles in regions and n^2 transit queues. Within this setup, we can compute the performance measures of interest from the stationary distribution of a *one-dimensional* open queue that tracks the number of vehicles in transit. This reduction represents a tremendous gain in computational tractability relative to computing the full stationary distribution of a closed queueing network with $n^2 + n$ queues. First, we show that all n^2 $\cdot/M/\infty$ transit queues can be collapsed into a single transit queue. (We use the term "collapse" to refer to computing the stationary distribution of a reduced description of the queueing system. For example, this first reduction corresponds to computing the stationary distribution of the number of vehicles in each region along with a single combined

number of vehicles in transit. This reduced set of quantities, when tracked over time as a stochastic process, does not yield a Markov process, but what is important is that the process still has a stationary probability distribution that we can compute.) Then, we show that the entire network can be collapsed, at least for the purpose of computing the performance measures we discuss, to a one-dimensional open queue for a system with instantaneous travel.

We collapse all n^2 $\cdot/M/\infty$ travel queues (each with an average travel time of $1/\mu_{ij}$) into a single $\cdot/M/\infty$ queue with relative arrival rate $\Lambda = \sum_{i,j} \lambda_{ij}$ and average service rate (μ_{avg}) given by the reciprocal of the average trip time,

$$\frac{1}{\mu_{\text{avg}}} = \frac{\sum_{i,j} \lambda_{ij} \frac{1}{\mu_{ij}}}{\sum_{i,j} \lambda_{ij}}.$$

When calculating the unnormalized stationary distribution, we only use the relative arrival rates and service rates of each queue. Therefore, this collapse of the transit queues does not affect the calculations for non-transit queues.

To be more precise, let $\tilde{\pi}_m$ denote the unnormalized stationary distribution of all $n+n^2$ region and transit queues and $\tilde{\pi}_{m,\text{comb}}$ be the unnormalized stationary distribution of the system with a single combined transit queue replacing the n^2 transit queues.

This new stationary distribution is defined over states $\mathbf{X}_{\text{comb}} = [\mathbf{X}^r, X^t]$ where $X^t = \sum_{i,j=1}^{n} \mathbf{X}^t_{ij} \in \{0, 1, 2, \ldots\}$ gives the total number of vehicles in transit, $\mathbf{X}^r$ is an n-vector giving the number of vehicles in the queues in each of the n regions, and $\mathbf{X}^t$ is a matrix giving the number of vehicles in transit between all pairs of regions. The product-form nature of $\tilde{\pi}_m$ allows us to express it as a product of 2 components - one representing the region dynamics and one representing the vehicles in transit. Therefore, $\tilde{\pi}_m(\mathbf{x}) = \tilde{\pi}^r_m(\mathbf{x}^r)\tilde{\pi}^t_m(\mathbf{x}^t)$ and

$\tilde{\pi}_{m,\text{comb}}(\mathbf{x}_{\text{comb}}) = \tilde{\pi}^r_m(\mathbf{x}^r)\tilde{\pi}^t_{m,\text{comb}}(x^t)$. Since we are only combining the transit queues, the regional component ($\tilde{\pi}^r_m$) is the same in both equations. The unnormalized stationary distribution of the combined transit queue is truncated Poisson, i.e., $\tilde{\pi}^t_{m,\text{comb}}(k) = (1/k!)(\Lambda^k/\mu_{\text{avg}})^k$, where $k = 0, 1, 2, \ldots, m$. This observation is formalized in the following result.

Lemma 8. *Fix a state* $\mathbf{x}$ *in which* $m - k$ *vehicles are distributed across the regions, and all remaining* k *vehicles are in transit. Let* $\mathbf{x}_{comb}$ *be the corresponding combined state. Then,* $\tilde{\pi}_{m,comb}(\mathbf{x}_{comb}) = \sum_{\mathbf{x}|\sum_{i,j} x^t_{ij}=k} \tilde{\pi}_m(\mathbf{x})$.

Proof. Proof of Lemma 8 Since both systems have the same configuration of units across regions, the $\tilde{\pi}^r_m(\mathbf{x}^r)$ components are the same. Therefore, we only need to show that $\sum_{\mathbf{x}|\sum_{i,j} x^t_{ij}=k} \tilde{\pi}^t_m(\mathbf{x}^t) = \tilde{\pi}^t_{m,\text{comb}}(k)$. Consider the transit queues corresponding to the number of users in transit from region i to region j, for $i, j = 1, 2, \ldots, n$. The unnormalized stationary distribution of the number of vehicles in these queues is exactly, up to a normalizing constant, that of independent Poisson random variables with mean λ_{ij}/μ_{ij}, $i, j = 1, 2, \ldots, n$. Moreover, this unnormalized stationary distribution takes values on the simplex $\{\mathbf{x} \in \mathbb{Z}^{n^2}, \mathbf{x} \geq 0, \sum_{i,j=1}^n x_{ij} = m\}$. Therefore, combining transit queues is mathematically identical to adding independent Poisson random variables, up to the maximal value m. The combined transit queue therefore has a probability mass function that is proportional to a Poisson random variable with mean $\sum_{i,j} \lambda_{ij}/\mu_{ij}$. This is the stationary probability distribution of a $\cdot/M/\infty$ queue with arrival rate $\lambda = \sum_{i,j} \lambda_{ij}$ and service rate $\mu = \mu_{\text{avg}}$.

□

Our performance metrics do not depend on the distribution of vehicles

across specific transit queues, so we can focus on the simpler system with only 1 transit queue. This decreases the number of queues to $O(n)$ from $O(n^2)$. For the rest of this paper, we will refer to the 1 transit queue variation and assume that the state is given by $\mathbf{X}_{comb}$ with state space $S_{m,n}$ redefined appropriately.

Similar to the state aggregation over all transit queues, we can aggregate the queues for vehicles in regions. The system is then characterized by two quantities: the number of vehicles in transit and the number of vehicles in regions. Since there are a fixed number of vehicles, m, this yields a 1D representation of the steady-state distribution, where the state represents the number of vehicles in transit, X^t. This state evolves according to a birth-death process where the birth process represents new trips beginning and the death process represents trips ending. The principle underlying this aggregation is that the queues in each region each have an unnormalized stationary distribution that is geometrically distributed, so the aggregation of those queues has an unnormalized stationary distribution that is a (truncated) negative binomial distribution. Before introducing Theorem 1, which formalizes this reduction, we introduce some new notation to specify the 1-D birth-death process.

Let $A_{nt}(m)$ be the availability in a system with instantaneous travel (or "no transit" queues, hence the notation "nt") and m vehicles. To be more precise, consider a system of n queues with no transit times and equal relative throughputs $(r_i : i = 1, 2, \ldots, n)$. In this system, the availability is known to be $A_{nt}(m) = m/(m + n - 1)$ [Banerjee et al., 2022]. The rationale behind this result, which we need, is that the stationary distribution is uniform over all states. Therefore, the availability in a region is simply the fraction of states with at least 1 vehicle at that region. There are n queues and we are allocating m vehicles

across them. A standard balls and bins argument [Banerjee et al., 2022] applies — we are allocating $n-1$ dividing lines among $m+n-1$ spots. The number of empty spots between dividing lines corresponds to the number of vehicles at stations. Therefore, availability is the number of configurations where the first dividing line is not in the first spot ($n-1$ dividers chosen from $m+n-2$ spots) divided by the number of configurations ($n-1$ dividers among $m+n-1$ spots). This ratio yields the expression for $A_{nt}(m)$.

Let $D = \sum_i d_i$ be the arrival rate of customers that are willing to pay for a trip. The arrival (birth) rate in the state $X^t = k$ is $A_{nt}(m-k)D$ and the service (death) rate is $k\mu_{\text{avg}}$. Let the unnormalized stationary distribution of the 1D reduction be $\tilde{\pi}_{\text{1D}}(\cdot)$, where $\tilde{\pi}_{\text{1D}}(0) = 1$.

Theorem 1. *The unnormalized stationary distribution $\tilde{\pi}_{1D}(\cdot)$ coincides with the unnormalized stationary distribution of the transit queue, i.e., for all $k = 0, 1, \ldots, m$ representing the number of vehicles in transit, $\sum_{x \in S_{m,n} | x^t = k} \tilde{\pi}(x) = \tilde{\pi}_{1D}(k)$.*

Proof. Proof of Theorem 1 For ease of notation, let $\tilde{\pi}_{\text{agg}}(k) = \sum_{\mathbf{x} \in S_{m,n} | x^t = k} \tilde{\pi}_m(\mathbf{x})$. We use induction on k to show that $\tilde{\pi}_{\text{1D}}(k) = \tilde{\pi}_{\text{agg}}(k)$ for all $k \in \{0, 1, ..., m\}$. For the base case, $\tilde{\pi}_{\text{1D}}(0) = \tilde{\pi}_{\text{agg}}(0) = 1$. Now assume that $\tilde{\pi}_{\text{1D}}(k) = \tilde{\pi}_{\text{agg}}(k)$ for some $k < m$. We want to show that it holds true for $k+1$. The recursive expression for birth-death processes, $\tilde{\pi}_{\text{1D}}(k+1) = \tilde{\pi}_{\text{1D}}(k) A_{nt}(m-k)D/(k+1)\mu_{\text{avg}}$, along with the inductive hypothesis for $\tilde{\pi}_{\text{1D}}$, yields

$$\tilde{\pi}_{\text{1D}}(k+1) = \tilde{\pi}_{\text{1D}}(k) \frac{\frac{m-k}{m+n-k-1} D}{(k+1)\mu_{\text{avg}}} = \sum_{\mathbf{x} \in S_{m,n} | \mathbf{x}^t = k} \tilde{\pi}_m(\mathbf{x}) \left(\frac{\frac{m-k}{m+n-k-1} D}{(k+1)\mu_{\text{avg}}} \right).$$

The unnormalized stationary probability mass function for each non-transit queue i evaluated at $k \in \{0, 1, \ldots, m\}$ is of the form r_i^k, and recall that $r_i = r$ is equal over all regions i. Thus, $r = \lambda_i / d_i$ for all regions i and therefore

$r = \sum_i \lambda_i / \sum_i d_i = \Lambda/D$. In addition, $\tilde{\pi}^r_m(\mathbf{x}^r) = r^{m-k}$ for all states $\mathbf{x}$ with $m-k$ vehicles in regions. The number of states with $m-k$ vehicles in n regions is the number of ways to pick $n-1$ divisions among $m-k+n-1$ spots. This, with the facts that $\tilde{\pi}_t(k) = (\Lambda/\mu_{\text{avg}})^k/k!$ and $r = \Lambda/D$, yields

$$\begin{aligned}
\tilde{\pi}_{1\text{D}}(k+1) &= \binom{m-k+n-1}{n-1} r^{m-k} \left(\frac{\Lambda}{\mu_{\text{avg}}}\right)^k \frac{1}{k!} \frac{\frac{m-k}{m+n-k-1} D}{(k+1)\mu_{\text{avg}}} \\
&= \binom{m-(k+1)+n-1}{n-1} r^{m-(k+1)} \left(\frac{\Lambda}{\mu_{\text{avg}}}\right)^{k+1} \frac{1}{(k+1)!} \\
&= \sum_{x \in S_{m,n} | x^t = k+1} \tilde{\pi}(x_s) \tilde{\pi}_t(k+1) = \tilde{\pi}_{\text{agg}}(k+1).
\end{aligned}$$

The last equality follows since states x_s with $m-(k+1)$ vehicles in regions have $\tilde{\pi}(x_s) = r^{m-(k+1)}$ and there are $\binom{m-(k+1)+n-1}{n-1}$ such states. □

We have now aggregated a closed queueing network with n^2+n queues down to a single birth death process! From this single birth-death process we can compute the overall system availability (approximation ratio) in two ways as follows. The first way computes the expected rate at which trips begin divided by the customer arrival rate. This is equal to $\mathbb{E}[A_{nt}(m-X)]$, where the random variable $X \sim \pi_{1\text{D}}$, and so

$$A_m(\mathbf{q}) = \frac{1}{D} \sum_{k=0}^{m} A_{nt}(m-k) D \mathbb{P}[k \text{ vehicles in transit}] = \sum_{k=0}^{m} A_{nt}(m-k) \tilde{\pi}_{1\text{D}}(k).$$

The second way uses the fact that in steady-state, the rate at which trips begin is equal to the rate at which trips end. Therefore, we can also define availability by the expected rate at which trips end divided by the customer arrival rate, yielding

$$A_m(\mathbf{q}) = \frac{1}{D} \sum_{k=0}^{m} k\mu_{\text{avg}} \mathbb{P}[k \text{ vehicles in transit}] = \frac{\mu_{\text{avg}}}{D} \mathbb{E}_{X \sim \pi_{1\text{D}}}[X]. \tag{2.7}$$

By normalizing $\tilde{\pi}_{1\text{D}}$ and then computing the expected value above, the availability can be computed in $O(m)$ time, which matches the performance from

Benjaafar et al. [2022]. That paper extends a specific case of mean-value analysis from Reiser and Lavenberg [1980] which gives a recursive scheme for exactly computing properties of closed queueing networks [Benjaafar et al., 2022]. Mean value analysis requires that all queues have constant service rates, while Benjaafar et al. [2022] requires a constant relative throughput across all queues. Benjaafar et al. [2022] also extends this technique to allow for transit times and shows how it can be used to compute availability via a recursive expression. The key observation of the method is that when a vehicle arrives at a station, the remaining $m-1$ vehicles are distributed according to π_{m-1}. This provides an explicit way to relate a system with m vehicles to a system with $m-1$ units.

Our 1D reduction achieves the same complexity as Benjaafar et al. [2022] and in the next section we show it maintains the same complexity for all settings to which their methodology can be extended. Our reduction also works for new settings, critically those where service rates are not constant. Moreover, our approach provides a new method for *approximating* the performance of the system. Indeed, from (2.7), we simply require the mean of the birth-death process, $\mathbb{E}_{X\sim\pi_{1\mathrm{D}}}[X]$.

The 1D birth-death process has the same service rates as those of an $M/M/c/c$ queue. However, the arrival rates are not constant; instead they are decreasing in k, the number of vehicles in transit. We approximate $\mathbb{E}_{X\sim\pi_{1\mathrm{D}}}[X]$ by the point, b say, where the rates balance, i.e., where $A_{nt}(m-b) = b\mu_{\mathrm{avg}}$. This gives

$$\frac{m-b}{m-b+n-1}D = b\mu_{\mathrm{avg}} \implies b = \frac{\left(m+n-1+\frac{D}{\mu_{\mathrm{avg}}}\right) - \sqrt{\left(m+n-1+\frac{D}{\mu_{\mathrm{avg}}}\right)^2 - 4\left(m\frac{D}{\mu_{\mathrm{avg}}}\right)}}{2},$$

and then the system availability is approximated by $b\mu_{\mathrm{avg}}/D$. This approximation works very well because the form of the birth and death rates ensures that

the tails around the balance point for the 1D birth-death process are lighter than those in an $M/M/c/c$ queue, which are already very light since they are those of a truncated Poisson distribution. Indeed, the arrival rates are decreasing in k, so there are stronger restoring forces relative to a queue with constant arrival rates, as we formalize next under the simplifying assumption that the balance point is integral.

Lemma 9. *Assume for simplicity that b, the balance point of the rates in the 1D birth-death process, is integral and let $\pi_{M/M/c/c}$ be the stationary distribution of an $M/M/c/c$ queue with the same balance point b and service rates. Then, the tails of π_{1D} decay faster than $\pi_{M/M/c/c}$ away from b in the sense that $\pi_{1D}(l)/\pi_{1D}(b) \leq \pi_{M/M/c/c}(l)/\pi_{M/M/c/c}(b)$ for all $l \neq b$.*

Proof. Proof of Lemma 9 Choose the arrival rate λ to the $M/M/c/c$ queue to be $\lambda = b\mu_{\text{avg}}$. Since b is the balance point of π_{1D}, $\lambda = b\mu_{\text{avg}} = A_{nt}(m-b)D$. Then, for all $l = b+1, b+2, \ldots, m$,

$$\frac{\pi_{1D}(l)}{\pi_{1D}(b)} = \prod_{z=b}^{l-1} \frac{A_{nt}(m-z)D}{(z+1)\mu_{\text{avg}}} \leq \prod_{z=b}^{l-1} \frac{\lambda}{(z+1)\mu_{\text{avg}}} = \frac{\pi_{M/M/c/c}(l)}{\pi_{M/M/c/c}(b)},$$

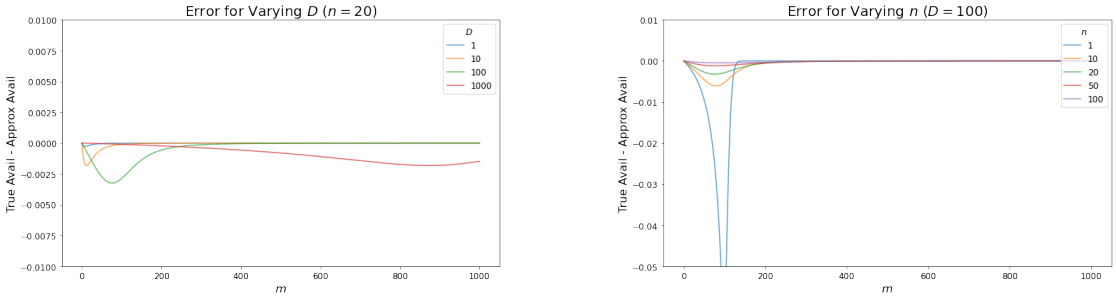
since $\lambda = A_{nt}(m-b)D$ and A_{nt} is decreasing. Similarly, for all $l = 0, 1, \ldots, b-1$,

$$\frac{\pi_{1D}(l)}{\pi_{1D}(b)} = \prod_{z=l+1}^{b} \frac{z\mu_{\text{avg}}}{A_{nt}(m-z-1)D} \leq \prod_{z=l+1}^{b} \frac{z\mu_{\text{avg}}}{\lambda} = \frac{\pi_{M/M/c/c}(l)}{\pi_{M/M/c/c}(k)}$$

since $\lambda = A_{nt}(m-b)D$ and A_{nt} is decreasing. □

This approximation for availability is further supported by computational evidence showing that the approximation closely matches system performance under a wide range of parameters as shown in Table 2.1.

This is the first closed-form approximation of availability given the number of vehicles for this class of systems. While Benjaafar et al. [2022] provided a



(a) Varying D (b) Varying n

Figure 2.1: Graphs showing the approximation error as m grows while varying other parameters.

Parameters		m^*	Approx Avail	True Avail	Difference	% Difference
$D = 10,\ \mu = 1$	n					
	1	100	1	1	0	0
	10	100	0.91	0.91	0.00008	0.01%
	20	100	0.828	0.828	0.0001	0.01%
	50	100	0.656	0.656	0.0001	0.02%
	100	100	0.49	0.49	0.00006	0.01%
$n = 20,\ \mu = 1$	D					
	1	100	0.839	0.839	9.7 e-6	0.00%
	10	100	0.828	0.828	0.0001	0.01%
	100	100	0.649	0.646	0.0029	0.45%
	1000	873	0.798	0.796	0.0018	0.23%
$D = 100,\ \mu = 1$	n					
	1	1000	1	0.975	0.0248	2.54%
	10	908	0.855	0.852	0.0028	0.33%
	20	873	0.798	0.796	0.0018	0.23%
	50	820	0.74	0.703	0.001	0.14%
	100	780	0.619	0.619	0.0006	0.09%
$D = 10n,\ \mu = 1$	n					
	1	100	1	1	0	0%
	10	100	0.742	0.736	0.005	0.71%
	20	157	0.626	0.623	0.003	0.47%
	50	391	0.621	0.62	0.0011	0.19%
	100	780	0.619	0.618	0.0006	0.09%

Table 2.1: Numerical Results for 1D Approximation

closed-form approximation for the number of vehicles required to meet availability thresholds, our approximation can be used in more general settings, including those with non-constant service rates and charging dynamics.

2.4.2 Electric Vehicles

We now consider electric vehicles with charging dynamics. In this section, we assume that charging rates are not state-dependent but we allow regions to have different relative throughputs and charging rates. In Section 2.5, we allow state-dependent charging-rates.

We begin by considering a setting with instantaneous travel and derive lower and upper bounds on performance. Then, we provide two techniques for analyzing the system with travel times, both of which build upon Mean-Value Analysis [Reiser and Lavenberg, 1980]. The first extends the methodology from Benjaafar et al. [2022] while the second extends the 1D-reduction previously introduced.

Bounding performance

We begin by analyzing the system with instantaneous travel times (so there are no transit queues) and focus on 2 cases: instantaneous charging (upper bound) and 'slow' charging (lower bound). With instantaneous charging, all uncharged vehicles immediately join the charged queue (as if the uncharged queues did not exist). Therefore, we recover the dynamics of the non-electric system that we previously analyzed. Without travel times, the availability in this setting is

$m/(m+n-1)$ [Banerjee et al., 2022], giving an upper bound for the performance of our system.

In the slow charging scenario, we let $r_i^u = r_i^c$ for all regions i since we require that charging is not the bottleneck. In the no-travel-time version of this system, we have a series of $\cdot/M/1$ queues, each with the same relative throughput. Therefore, the stationary distribution is the uniform distribution over all states and the availability is then $m/(m+2n-1)$ using the same counting argument we discussed in the last section and since we have $2n$ queues. Combining the bounds, $A_{nt}(m) \in [m/(m+2n-1), m/(m+n-1)]$. Later, in Lemma 13, we show that decreasing r_i^u, which results if we increase the charging rate at i, increases the availability of the system.

Mean-value analysis

We extend the mean-value analysis argument from Benjaafar et al. [2022] to our more complicated setting with charging dynamics. We develop a new recursive expression for availability that allows for queues with different relative throughputs. For each queue, the expected queue length is related to a system with 1 fewer vehicle. Similar to Benjaafar et al. [2022], the main tools are the random observer property [Reiser and Lavenberg, 1980] and Little's law. The random observer property states that in a stationary system with m units, at the moment of an arrival to a queue, the arriving unit will observe the stationary distribution of the system with one fewer unit ($m-1$ total units).

We now review some notation for this approach. Let $X_i^c(m)\left(X_i^u(m)\right)$ be a random variable denoting the steady-state number of charged (uncharged) vehicles

in region i with m total units in the system. Also let $X^t(m)$ be a random variable representing the steady-state total number of vehicles in transit (with m total vehicles in the system). Let $v_i(m)$ be the expected trip completion rate for trips starting from region i in steady-state. Then, the mean steady-state number of vehicles in transit is $\sum_i v_i(m)/\mu_{\text{avg}}$. When a unit arrives to the charged queue, it sees $\mathbb{E}\left[X_i^c(m-1)\right]$ vehicles before it in expectation. Since we have a constant trip demand rate (service rate of the queue), d_i, the expected time until the vehicle is used for another trip (leaves the queue) is $(1 + \mathbb{E}\left[X_i^c(m-1)\right])/d_i$. Using Little's law to relate the expected length of the queue (charged vehicles at i) to the long-term arrival rate ($v_i(m)$) and the expected time in the queue (calculated previously) yields

$$\mathbb{E}\left[X_i^c(m)\right] = v_i(m)\frac{1 + \mathbb{E}\left[X_i^c(m-1)\right]}{d_i}. \tag{2.8}$$

Let b_i be the charging rate in region i. Following the same argument yields

$$\mathbb{E}\left[X_i^u(m)\right] = pv_i(m)\frac{1 + \mathbb{E}\left[X_i^u(m-1)\right]}{b_i}. \tag{2.9}$$

Since all m vehicles must be in charged, uncharged, or transit queues, we know

$$\begin{aligned} m &= \mathbb{E}\left[X^t(m)\right] + \sum_i \mathbb{E}\left[X_i^c(m)\right] + \sum_i \mathbb{E}\left[X_i^u(m)\right] \\ &= \sum_i \frac{v_i(m)}{\mu_{\text{avg}}} + \sum_i \frac{v_i(m)}{d_i}\left[1 + \mathbb{E}\left[X_i^c(m-1)\right]\right] + \sum_i \frac{pv_i(m)}{b_i}\left[1 + \mathbb{E}\left[X_i^u(m-1)\right]\right]. \end{aligned}$$

Recall that D is the total rate that customers (who are willing to pay) request trips and the availability at i is given by $A_i(m) = v_i(m)/d_i = A(m)$ and is the same for all regions. In this setting, the relative throughput of the charged queues is r^c and the relative throughput of the uncharged queue in region i is r_i^u, so $r_i^u/r^c = pd_i/b_i$. Since we assume that charging is not the bottleneck, $r_i^u \leq r^c$.

Therefore, we get

$$m = \frac{A(m)D}{\mu_{\text{avg}}} + A(m)\left[n + \sum_i \mathbb{E}\left[X_i^c(m-1)\right]\right] + A(m)\sum_i \frac{r_i^u}{r^c}\left[1 + \mathbb{E}\left[X_i^u(m-1)\right]\right]$$

$$\implies A(m) = m\left(\frac{D}{\mu_{\text{avg}}} + n + \sum_i \mathbb{E}\left[X_i^c(m-1)\right] + \sum_i \frac{r_i^u}{r^c}\mathbb{E}\left[1 + X_i^u(m-1)\right]\right)^{-1} \tag{2.10}$$

which guides us to the following recursive scheme to exactly calculate the availability.

Algorithm 1: An algorithm for calculating availability

Initialize: $\mathbb{E}\left[X_i^c(0)\right] \leftarrow 0,\ \mathbb{E}\left[X_i^u(0)\right] \leftarrow 0\ \forall i$

for $k \leftarrow 1, ..., m$ **do**

  Calculate $A(k)$ using (2.10);

  For each region i, using $v_i(k) = A(k)d_i$, calculate $\mathbb{E}\left[X_i^c(k)\right]$ and $\mathbb{E}\left[X_i^u(k)\right]$ using (2.8) and (2.9);

end

Calculate expected vehicles in transit using $\mathbb{E}[X^t] = m - \sum_i \mathbb{E}[X_i^c(m)] - \mathbb{E}[X_i^u(k)]$ and availability using $\mathbb{E}[X^t]\mu_{\text{avg}}/D$;

To improve the speed of the above algorithm, we can recursively keep track of $\sum_i \mathbb{E}\left[X_i^c(m-1)\right]$ instead of each quantity individually. The update then becomes $\sum_i \mathbb{E}\left[X_i^c(m-1)\right] = A(m)\left(n + \sum_i \mathbb{E}[X_i^c(m-1)]\right)$. For uncharged queues, we only require one computation to update all queues with the same r_i^u. Indeed, $r_i^u/r^c = pd_i/b_i$ and $v_i(m) = A(m)d_i$. Therefore, all queues with the same r_i^u have the same values for $pv_i(m)/b_i$ and can be updated together according to (2.9). Therefore, the overall run-time is $O(mn_{\text{distinct}})$ where n_{distinct} is the number of distinct r_i^u values.

1D reduction extension

A major limitation of the argument in the previous section is that it does not allow state-dependent service rates. Instead, we can extend our previous 1D reduction to this setting with charging. The state space aggregation remains the same - the 1D state is the number of vehicles in transit. The collapse of transit queues still applies since the analysis only requires a product-form stationary distribution and is independent of the dynamics of the non-transit queues. The arrival rate of the birth-death process when there are k vehicles in transit remains $A_{nt}(m-k)D$ where $A_{nt}(k) = \mathbb{P}[X_i^c > 0 | k$ vehicles in the charged and uncharged queues]. A proof of the 1D reduction in this more general setting (with state-dependent service rates) is provided in Section 5.

Unlike the previous setting, we no longer have a closed form expression for A_{nt}. However, the system without transit times is just a network of $\cdot/M/1$ queues. We can use mean-value analysis to exactly compute availability in this simpler system [Reiser and Lavenberg, 1980].

Similar to the previous methodology, mean-value analysis gives recursive expressions for expected queue lengths. Combined with the total number of vehicles in the system, we can inductively solve for availability with k vehicles ($k \in \{0, 1, ..., m\}$). With the arrival and service rates for the 1D birth death process, the stationary distribution and expected vehicles in transit can be computed, which gives us the overall availability of the system. The run-time of this procedure is $O(mn_{\text{distinct}})$ which matches the extension of Benjaafar et al. [2022].

2.4.3 Scaling Analysis

In this section we derive results for two scaling regimes. First, we consider the setting where the number of vehicles m grows to infinity while demand is held fixed. We show linear convergence, meaning that availability is equal to $1 - O(1/m)$. This regime is a useful intermediate step to ease understanding of a second, more realistic regime in which both the number of vehicles and demand grow in proportion. The second regime exhibits interesting scaling behaviour. Since the number of vehicles and demand jointly grow, we have to consider a new static pricing policy parameterized by the number of vehicles. This new policy introduces a constraint that the expected vehicles in transit is bounded by m. If this constraint is not tight, then we observe linear convergence as in the first regime. Otherwise we observe slower convergence, namely that availability is $1 - O(1/\sqrt{m})$. These rates match the rates presented in Banerjee et al. [2022] in the simpler setting without charging dynamics.

Vehicle Scaling Limit

Lemma 10. *For any fixed, state-independent, demand-balanced pricing policy* $\mathbf{q}$*, when the number of vehicles in the system grows to infinity with demand held constant, the availability goes to 1 linearly, i.e.,* $A_m(\mathbf{q}) \geq 1 - O(1/m)$.

Proof. Proof of Lemma 10

For any fixed state-independent pricing policy, $\mathbf{q}$, there is a maximum expected vehicles in transit, $v_t = \sum_{ij} \phi_{ij} q_{ij} / \mu_{ij}$, that arises when all demand is satisfied by the platform.

Let T be a random variable representing the steady-state number of vehicles in transit. Then T is stochastically dominated by $\widetilde{T} \sim \text{Poisson}(v_t)$. Since $A_{nt}(m-k)$ is decreasing in k,

$$A_m(\mathbf{q}) = \mathbb{E}\left[A_{nt}(m-T)\right] \geq \mathbb{E}\left[A_{nt}(m-\widetilde{T})\right],$$

where we define $A_{nt}(k) = 0$ for $k \leq 0$. We can lower bound this expression by selecting a threshold for vehicles in stations, $\ell(m)$, to give

$$A_m(\mathbf{q}) \geq \mathbb{P}\left[\widetilde{T} \leq m - \ell(m)\right] A_{nt}(\ell(m)) \geq \mathbb{P}\left[\widetilde{T} \leq m - \ell(m)\right] \frac{\ell(m)}{\ell(m) + 2n - 1}.$$

Taking $\ell(m) = cm$ for some $c < 1$ and using Markov's inequality, $P(\widetilde{T} \geq \ell(m)) \leq E\widetilde{T}/\ell(m) = v_t/(cm)$. Then, the availability is bounded by

$$\left(1 - \frac{v_t}{cm}\right)\left(1 - \frac{2n-1}{cm + 2n - 1}\right) = 1 - O\left(\frac{1}{m}\right).$$

□

Since the throughput is the offered load times availability and the offered load is constant, the throughput scales in the same way as availability.

Vehicle and Demand Scaling Limit

In this section, we jointly scale both the demands and the number of vehicles at the same rate. More precisely, we consider a sequence of systems, indexed by m. In the mth system there are m vehicles, and we assume that the demands are given by $\phi_{ij}(m)$, where $\phi_{ij}(m) = m\phi_{ij}$. We begin with a tail bound on vehicles in transit and get a corresponding lower bound on availability. Then, we introduce a static pricing policy (parameterized by m) and prove an availability bound

for this new pricing policy. Finally, we analyze two regimes showing different convergence rates. The proofs in this section follow the corresponding proofs in Banerjee et al. [2022].

Lemma 9 of Banerjee et al. [2022] proves a tail bound on $\widetilde{T}$, namely that for $m - 2\sqrt{m\ln(m)} \geq v_t$ and $m \geq 100$,

$$\mathbb{P}\left[\widetilde{T} > m\left(1 - \sqrt{\frac{\ln m}{m}}\right)\right] \leq \frac{3}{\sqrt{m}}.$$

Therefore, we can bound availability by considering a threshold $\ell(m) = \sqrt{m\ln m}$ as in the proof of Lemma 10, giving

$$A_m(\mathbf{q}) \geq \mathbb{P}\left[\widetilde{T} \leq m - \ell(m)\right] A_{nt}(\ell(m)) \geq \left(1 - \frac{3}{\sqrt{m}}\right)\left(\frac{\sqrt{m\ln m}}{\sqrt{m\ln m} + 2n - 1}\right)$$

which corresponds to Lemma 9 in Banerjee et al. [2022].

Now, we introduce the parameterized static pricing policy that we use to analyze limiting availability and throughput. We begin with the modified fluid relaxation,

$$\begin{aligned}
\max_{q} \quad & \sum_{i,j} \phi_{ij}(m) q_{ij} \\
\text{s.t.} \quad & \sum_{k} \phi_{ki}(m) q_{ki} = \sum_{j} \phi_{ij}(m) q_{ij} && \forall i \\
& \sum_{i,j} \phi_{ij}(m) \tau_{ij} q_{ij} \leq m \\
& q_{ij} \in [0,1] && \forall i, j,
\end{aligned}$$

where we add a constraint that limits the expected vehicles in transit to be at most m. If the constraint is not satisfied, then availability will not go to 1 because the rate of customers willing to pay for a trip will be too high. Since our approximation ratio is the availability, we require that it goes to 1 as $m \to \infty$. The modified fluid relaxation simplifies because $\phi_{ij}(m) = m\phi_{ij}$ for all i, j and m,

yielding

$$\begin{aligned} \max_{q} \quad & \sum_{i,j} \phi_{ij} m q_{ij} \\ \text{s.t.} \quad & \sum_{k} \phi_{ki} q_{ki} = \sum_{j} \phi_{ij} q_{ij} \qquad \forall i \\ & \sum_{i,j} \phi_{ij} \tau_{ij} q_{ij} \leq 1 \\ & q_{ij} \in [0,1] \qquad \forall i, j, \end{aligned}$$

where the constraints are simplified by dividing both sides by m. The m in the objective function can be factored out of the sum and does not change the optimal pricing policy, only the optimal objective value. Therefore, the optimal solution to the new pricing problem, q' say, is constant in m. Let $\epsilon_m \in (0,1)$ be a quantity to be specified later. If $\sum_{i,j} \phi_{ij} \tau_{ij} q'_{ij} \geq (1 - \epsilon_m)$, then, for all i and j, let $\tilde{q}_{ij} = q'_{ij}(1 - \epsilon_m)$ and otherwise $\tilde{q}_{ij} = q'_{ij}$. Our new pricing policy is $\tilde{q}$. Now, we analyze the scaling limit under the static pricing policy $\tilde{q}$.

Lemma 11. *Let OPT_m be the throughput of the optimal state-dependent pricing policy in the system with m vehicles and $\tilde{q}$ be defined as above with $\epsilon_m := 2\sqrt{\ln m}/m$. Then, for $m \geq 100$,*

$$\frac{OBJ_m(\tilde{q})}{OPT_m} \geq (1 - \epsilon_m)\left(\frac{\sqrt{m \ln m}}{\sqrt{m \ln m} + 2n - 1} - \frac{3}{\sqrt{m}} \right).$$

Proof. Proof of Lemma 11 Fix an arbitrary $m \geq 100$, making the demand $\phi_{ij} m$ for all i and j and fixing the fluid relaxation (to this m). Let $\bar{q}$ be the average fraction of riders that complete their trips under the optimal state-dependent pricing policy (for the m-vehicle system). From Lemma 1, $\bar{q}$ satisfies the flow-balance constraint and since there are m total vehicles, the maximum expected vehicles in transit must be upper bounded by m, so $\bar{q}$ also satisfies the new constraint. Therefore, the optimal objective function value of the modified fluid relaxation is an upper bound on the throughput of the optimal state-dependent pricing

policy.

The state-independent pricing policy, q', has objective value in the modified fluid relaxation equal to $\widehat{OBJ}(q) = m \sum_{i,j} \phi_{ij} q'_{ij}$. Moreover, in a limit where we fix the demand at the level $\phi_{ij} m$, and send the number of vehicles $m' \to \infty$, the attained throughput in the stochastic system with m vehicles of the policy q' converges to $m \sum_{i,j} \phi_{ij} q'_{ij}$. From the definition of $\tilde{q}$, $\widehat{OBJ}(\tilde{q}) \geq (1 - \epsilon_m)\widehat{OBJ}(q')$, i.e., $\tilde{q}$ is nearly optimal in the modified fluid relaxation. We also have that $\sum_{i,j} \phi_{ij} \tilde{q}_{ij} \tau_{ij} \leq (1 - \epsilon_m)$ by definition. Let $OBJ_\infty(q)$ be the throughput of the system with demands $\phi_{ij} m$ and an infinite number of vehicles. Lemma 6 tells us that $OBJ_\infty(\tilde{q}) = m \sum_{i,j} \phi_{ij} \tilde{q}_{ij}$. Let $OBJ_m(q)$ be the throughput of the stochastic system with demands $\phi_{ij} m$ and m total vehicles. Then, by Lemma 7, $OBJ_m(\tilde{q}) = A_m(\tilde{q}) OBJ_\infty(\tilde{q})$. Combining these observations yields

$$\begin{aligned}
\frac{OBJ_m(\tilde{q})}{OPT_m} &\geq \frac{OBJ_m(\tilde{q})}{OBJ_\infty(q')} \\
&= \frac{A_m(\tilde{q}) OBJ_\infty(\tilde{q})}{OBJ_\infty(q')} \\
&\geq \left(1 - \frac{3}{\sqrt{m}}\right)\left(\frac{\sqrt{m \ln m}}{\sqrt{m \ln m} + 2n - 1}\right) \frac{OBJ_\infty(\tilde{q})}{OBJ_\infty(q')} \\
&\geq (1 - \epsilon_m)\left(1 - \frac{3}{\sqrt{m}}\right)\left(\frac{\sqrt{m \ln m}}{\sqrt{m \ln m} + 2n - 1}\right) \\
&\geq (1 - \epsilon_m)\left(\frac{\sqrt{m \ln m}}{\sqrt{m \ln m} + 2n - 1} - \frac{3}{\sqrt{m}}\right),
\end{aligned}$$

where the first inequality follows by the upper bound on OPT_{m_0} from the modified fluid relaxation, the second inequality uses the lower bound on availability, and the third inequality uses the relation between the objective function values of $\tilde{q}$ and q'. □

If the new constraint is tight, by Lemma 11, the convergence rate is $O(1/\sqrt{m})$. If the constraint is not tight, via Corollary 1 of Banerjee et al. [2022], the conver-

gence rate is $O(1/m)$.

2.5 State-dependent Service Rates

So far we have considered non-state-dependent rates in the charged and uncharged queues, but such rates are not ideal when modeling dockless systems. Indeed, we expect that a potential rider in a given region is more likely to find a nearby charged vehicle when many are available in the region, and thus the service rate for the charged queue should increase with the number of charged vehicles. And as we'll see next, state-dependent service rates in the uncharged queues allow us to better model different charging mechanisms.

2.5.1 Charging Mechanisms

We introduce and describe four different charging mechanisms and how they can be captured in our modeling framework. One could also use our framework to model combinations of charging mechanisms, e.g., different mechanisms in different regions or even multiple mechanisms used within the same region, but we do not explore such combinations here.

Crowd-sourcing

In crowd-sourced charging systems, platforms pay freelancers to pick up e-scooters and charge them before redeploying them on the street. As the number, and thus density, of uncharged vehicles increases, it should be more attractive

to freelancers to engage, since it is then easier to collect multiple scooters to recharge. We thus model this charging mechanism as an infinite-server queue, reflecting the expectation that more uncharged vehicles in a region will motivate more freelancers to recharge them. We assume that the charging rate per uncharged vehicle is $h_{\text{crowd}} = 1/5$ per hour, so that $\mu_i^{crowd}(k) = h_{crowd}k$. The charging time reflects the time for the scooter to be picked up, transported, recharged, and redeployed. Since we are modelling the crowd-sourced recharging as an infinite server queue, we can collapse the charging queues into the transit queues using Lemma 8.

Employees swapping batteries

Another common mechanism used by electric-micromobility providers is hiring employees to travel to vehicles to swap batteries. As the density of uncharged vehicles in a region increases, the average distance to the next vehicle decreases, increasing the rate of recharging. To motivate the form of charging rates, consider the setting where a travelling salesman problem (TSP) path is constructed to visit all k uncharged vehicles. The total length of this path scales according to $O(\sqrt{k})$ and therefore the time to travel to the next vehicle scales according to $O(1/\sqrt{k})$ [Beardwood et al., 1959]. If we assume that locating the vehicle and swapping the battery takes t_{switch}, the expected rate that vehicles get recharged is $\mu_i^{\text{employee}} = \left[1/\left(h_{\text{employee},i}\sqrt{k}\right) + t_{\text{switch}}\right]^{-1}$ and $h_{\text{employee},i}$ is a constant that takes into account the geography of the region, speed of the employee, and TSP constant for expected tour length.

Charging Stations

Several systems rely on charging stations to charge their fleet, mostly in the context of e-bikes. Stations are connected to the grid and charge bikes when bikes are parked. We model such systems as a birth-death process with a linearly increasing recharging rate until its threshold (station capacity) is reached. Let the charging rate per dock be h_{station} and the capacity of station i be C_i. Then the recharging rate is $\mu_i^{\text{station}}(k) = h_{\text{station}} \min\{k, C_i\}$.

We consider two variants of this charging mechanism: with perfect and imperfect compliance. In an ideal world, all system users are responsible and return bikes/scooters to charging stations. This is modelled as a single uncharged queue per region with recharging rates described above.

However, in reality, not all users are perfect and some vehicles may be returned away from stations or not returned properly. For these vehicles, the system requires another charging mechanism, e.g., crowd-sourcing or employees. Let p_{abandon} be the probability that an uncharged vehicle is not properly returned. We model this scenario by having two uncharged queues in each region, one for charging stations and one for the alternative mechanism. When an uncharged vehicle arrives to region i, it joins the charging-station queue with probability $(1-p_{\text{abandon}})$ and the other uncharged queue with probability p_{abandon}.

Vending Machines

In this charging mechanism, users ride up to a vending machine where they swap their battery with a fresh one. Vending machines can either be co-located with stations or independently distributed around the city. Operators may

prefer this mechanism relative to charging stations because they have a much smaller footprint and can store more batteries than bikes, allowing more reliable access to fully charged batteries. While vending machines have not been implemented in North America, this charging strategy has been rapidly growing in popularity around the world since 2021. Currently, vending machines are in use for non-shared e-scooters, moped and bicycles in several countries, including China, India, Taiwan, Indonesia, Rwanda, Uganda and Kenya [Inc., 2023a,b, Kuhudzai, 2023], and a pilot is under way in New York city to serve e-bike couriers [NYC Mayor Office, 2023]. To model this mechanism, we reduce the probability that a vehicle is uncharged after completing its trip. Since our model extends to the setting with heterogeneous p_{ij}, these probabilities can be individually adjusted to reflect the location and distribution of vending machines. Similar to imperfect charging stations, a backup charging mechanism must be specified.

2.5.2 Performance Metrics

To compute the availability of the system with state-dependent service rates, we use the 1D reduction introduced previously. Theorem 2 proves that we can reduce these state-dependent systems to a 1D birth-death process. To compute $A_{nt}(k)$, the availability in the corresponding system with instantaneous travel, we calculate the normalization constants of the instantaneous-travel system using the recursive scheme from Denning and Buzen [1978]. We denote these normalization constants of the system without transit queues as G_m^{nt} where "nt" indicates that there are no transit queues. For simplicity, the dependency on the number of regions, n, is suppressed.

Theorem 2. *The collapse of the closed queuing network to a 1D open queue with state representing the number of vehicles in transit holds for systems with charging dynamics and arbitrary state-dependent service rates. For the birth-death process, the arrival rates are $\lambda_i = A_{nt}(m-i)D$ and service rates are $\mu_i = i\mu_{avg}$.*

Proof. Proof of Theorem 2 We begin by noting that Lemma 8 holds in this setting as well. Therefore, we consider the setting with only 1 transit queue. Define $\tilde{\pi}_{1D}$ to be the unnormalized stationary distribution of the 1D birth-death process defined in the theorem statement with $\tilde{\pi}_{1D}(0) = 1$. Recall the notation for states in the full system, where we write $\mathbf{x} = (\mathbf{x}^r, \mathbf{x}^t)$ giving the breakdown into non-transit queues and transit queues respectively, and that we write $x^t = \sum_{i,j=1}^{n} \mathbf{x}^t_{ij}$ to be the corresponding total number of vehicles in transit. Let $\tilde{\pi}_{\text{agg}}(k) = \sum_{\mathbf{x} \in S_{m,n} | x^t = k} \tilde{\pi}_m(\mathbf{x})$ be the sum of the unnormalized steady-state probabilities of the full system over all states with k vehicles in transit. We show by induction that $\tilde{\pi}_{\text{agg}}(k) = \tilde{\pi}_{1D}(k)$ for all $k \in \{0, \ldots, m\}$.

For the base case, $\tilde{\pi}_{1\text{D}}(0) = \tilde{\pi}_{\text{agg}}(0) = 1$. Now, assume that $\tilde{\pi}_{1\text{D}}(k) = \tilde{\pi}_{\text{agg}}(k)$ for some $k < m$. We want to show that it holds true for $k+1$. The recursive expression for birth death processes, $\tilde{\pi}_{1D}(k+1) = \tilde{\pi}_{1D}(k) A_{nt}(m-k)D/((k+1)\mu_{\text{avg}})$ combined with the inductive hypothesis and the product form of the stationary distribution of the full system yields

$$\tilde{\pi}_{1\text{D}}(k+1) = \tilde{\pi}_{1\text{D}}(k)\frac{A_{nt}(m-k)D}{(k+1)\mu_{\text{avg}}} = \tilde{\pi}_{\text{agg}}(k)\frac{A_{nt}(m-k)D}{(k+1)\mu_{\text{avg}}} = \sum_{\mathbf{x} \in S_{m,n} | x^t = k} \tilde{\pi}^r(\mathbf{x}^r)\tilde{\pi}^t(k)\frac{A_{nt}(m-k)D}{(k+1)\mu_{\text{avg}}}.$$

Using $\tilde{\pi}^t(k) = \left(\Lambda/\mu_{\text{avg}}\right)^k / k!$ and rearranging yields

$$\begin{aligned}
\tilde{\pi}_{1\text{D}}(k+1) &= \sum_{\mathbf{x}\in S_{m,n}|x^t=k} \tilde{\pi}^r(\mathbf{x}^r)\left(\frac{\Lambda}{\mu_{\text{avg}}}\right)^k \frac{1}{k!}\frac{A_{nt}(m-k)D}{(k+1)\mu_{\text{avg}}} \\
&= \sum_{\mathbf{x}\in S_{m,n}|x^t=k} \tilde{\pi}^r(\mathbf{x}^r)A_{nt}(m-k)\frac{D}{\Lambda}\left(\frac{\Lambda}{\mu_{\text{avg}}}\right)^{k+1}\frac{1}{(k+1)!} \\
&= \sum_{\mathbf{x}\in S_{m,n}|x^t=k} \tilde{\pi}^r(\mathbf{x}^r)A_{nt}(m-k)\frac{D}{\Lambda}\tilde{\pi}^t(k+1) \\
&= \sum_{\mathbf{x}\in S_{m,n}|x^t=k} \tilde{\pi}^r(\mathbf{x}^r)\left(r\frac{G^{nt}_{m-k-1}}{G^{nt}_{m-k}}\right)\frac{1}{r}\tilde{\pi}_t(k+1) = G^{nt}_{m-k-1}\tilde{\pi}_t(k+1) \\
&= \sum_{\mathbf{x}\in S_{m,n}|x^t=k+1} \tilde{\pi}^r(\mathbf{x}^r)\tilde{\pi}_t(k+1) = \tilde{\pi}_{agg}(k+1)
\end{aligned}$$

where the fourth line uses the definition of availability, $A_{nt}(m-k) = rG^{nt}_{m-k-1}/G^{nt}_{m-k}$, and normalization constants, $G^{nt}_{m-k} = \sum_{\mathbf{x}\in S_{m,n}|x^t=k}\tilde{\pi}^r(\mathbf{x}^r)$. □

Therefore, by induction we have shown that the stationary distribution of the 1D birth-death process exactly matches the state aggregation of the complete queueing network.

The last component is to solve for the availability in the no travel time system, $A_{nt}(\ell)$ for $\ell = 0, 1, \ldots, m$. To do this, we enumerate all queues from 1 to n_{queues} and follow the recursive scheme from Denning and Buzen [1978], which is built upon the relationship $G_{m,n} = \sum_{\ell=0}^{m}\tilde{\pi}_n(\ell)G_{m-\ell,n-1}$. We begin by initializing $G_{\ell,1} = \tilde{\pi}_1(k)$ for all $k \in \{0, 1, \ldots, m\}$ and $G_{0,\ell} = 1$ for all $\ell \in \{1, 2, \ldots, n_{queues}\}$. Then, we calculate the normalization constants when adding the second queue $G_{k,2}$ for all $k \in \{0, 1, \ldots, m\}$. We continue adding one queue at a time until all queues have been added. After computing the normalization constants, we can calculate $A_{nt}(k)$ for all $k \in \{1, ..., m\}$ using $r_i G_{m-1}/G_m$. This algorithm takes $O(mn)$ time and suffers from some numerical stability challenges that we overcame using an arbitrary precision library.

2.5.3 Balance Point Approximation

Since the 1D reduction technique applies in this setting, we can extend the balance point approximation. However, without a closed form expression for A_{nt}, we cannot get a closed for expression for the approximation. Instead, we compute $A_{nt}(\ell)$ (via normalization constants) for iteratively larger ℓ. With this approximation, we stop when $A_{nt}(\ell)D > (m - \ell)\mu_{\text{avg}}$. At this point, increasing ℓ increases the left hand side while decreasing the right hand side, driving us farther from the balance point. The benefit of this, is reducing the amount of computation via early termination, which can reduce some of the challenges with numerical instability or amount of precision required.

2.5.4 Scaling Limits

This section analyzes the scaling limits for the system with state-dependent service rates. In order to do this, we prove a novel lower bound for the instantaneous-travel system, analogous to the $m/(m + n - 1)$ guarantee in Banerjee et al. [2022]. In fact, their guarantee can be recovered as a special case of our analysis. With this new lower bound, we can apply the same arguments used previously to characterize the scaling limits of the system with travel times. We consider the same regimes as in Section 2.4.3 : 1) vehicle scaling, 2) demand scaling linearly with the number of vehicles.

No-transit Availability

To prove a lower bound for the instantaneous-travel system, we begin by considering a more restrictive system. This system has two limitations: all queues have the same relative throughput and service rates converge exactly linearly. Under these assumptions, Lemma 13 below establishes that the availability in the system is $m/(m + c')$ for some positive constant c'. With this result, we then relax the two restrictions by showing that 1) increasing the throughput of uncharged queues, r_i^u, increases the throughput of the system and 2) increasing the service rate of any queue, $\mu_i(\ell)$, increases the throughput of the system. Intuitively, both of these changes increase the throughput of the system since they correspond to increasing the service rate of queues in the network. Therefore, the restrictions on the system decrease the throughput and $m/(m + c')$ is thus a lower bound on availability.

Before developing these results, we introduce a lemma that will be used.

Lemma 12. *Let $k,\ m,\ c,\ c'$ be non-negative integers. Then,*

$$\sum_{k=0}^{m} \binom{k+c'}{c'} \binom{m-k+c}{c} = \binom{m+c+c'+1}{m}$$

Proof. Proof of Lemma 12 We begin with property 5.26 from Graham et al. [1994] which states

$$\sum_{a=0}^{b} \binom{b-a}{d} \binom{e+a}{g} = \binom{b+e+1}{d+f+1}$$

for integers $b,\ d \geq 0$ and integers $g \geq e \geq 0$. Let $b = m + c,\ d = c,\ e = c'$, and $g = c'$, yielding

$$\begin{aligned}\binom{m+c+c'+1}{m} = \binom{m+c+c'+1}{c+c'+1} &= \sum_{k=0}^{m+c} \binom{k+c'}{c'} \binom{m+c-k}{m} \\ &= \sum_{k=0}^{m} \binom{k+c'}{c'} \binom{m+c-k}{m} + \sum_{k=m+1}^{m+c} \binom{k+c'}{c'} \binom{m+c-k}{m}.\end{aligned}$$

Now, we consider the second summation where $k > m$. Therefore, $m + c - k < m$ and $\binom{m+c-k}{m} = 0$ for all terms in the second summation, proving the lemma. □

Now, we prove the closed form availability expression for the restricted system.

Lemma 13. *In a balanced closed queueing network where all queues have the same relative throughput, and service rates converge exactly linearly, i.e., $\mu_i(k) = \mu_i(\infty)k/(k + c_i)$ for some positive integer c_i for all queues i and $k = 1, 2, \ldots, m$, the availability of the system is $m/(m + c')$ where $c' = n - 1 + \sum_{i=1}^{n} c_i$.*

Proof. Proof of Lemma 13 Let λ_i be the (relative) arrival rate to each queue and $\mu_i(k)$ be the service rate of queue i with k vehicles. Since the queueing network is balanced, there is a common value $r = \lambda_i/\mu_i(\infty)$ for all queues i.

We prove the linear convergence of availability by inductively adding queues. The base case is when there is 1 queue and then $G_{m,1} = \tilde{\pi}_1(m)$, where

$$\tilde{\pi}_1(m) = \prod_{j=1}^{m} \frac{\lambda_1}{\mu_1(j)} = \prod_{j=1}^{m} \frac{\lambda_1(j + c_1)}{j\mu_1(\infty)} = \frac{r^m(m + c_1)!}{m!c_1!} = r^m \binom{m + c_1}{c_1}.$$

Then, the availability is

$$A_1(m) = r\frac{G_{m-1,1}}{G_{m,1}} = r\frac{r^{m-1}\binom{m-1+c_1}{c_1}}{r^m\binom{m+c_1}{c_1}} = \frac{m}{(m + c_1)},$$

where $c' = c_1$ as required.

Our inductive hypothesis is that the availability of the system with the first n queues is $A_n(m) = rG_{m-1,n}/G_{m,n} = m/(m + c)$ for $c = n - 1 + \sum_{i=1}^{n} c_i$ Therefore, c depends on n but not on m. Now, we want to show that when we add the $n + 1^{\text{st}}$ queue, the availability has the form $A_{n+1}(m) = m/(m + c')$ where $c' = c + 1 + c_{n+1}$.

We begin by considering the normalization constant of the n-queue system,

$$G_{m,n} = \frac{r}{A_n(m)} G_{m-1,n} = \frac{r}{A_n(m)} \frac{r}{A_n(m-1)} G_{m-2,n} = \frac{r}{A_n(m)} \frac{r}{A_n(m-1)} \cdots \frac{r}{A_n(1)}$$
$$= r^m \left(\frac{m+c}{m}\right)\left(\frac{m-1+c}{m-1}\right)\cdots\left(\frac{1+c}{1}\right) = r^m \frac{(m+c)!}{c!m!} = r^m \binom{m+c}{c}.$$

Now, we use the recursive expression for the normalization constant,

$$G_{m,n+1} = \sum_{x \in S_{m,n+1}} \tilde{\pi}(x) = \sum_{k=0}^{m} \tilde{\pi}_{n+1}(k) G_{m-k,n} = \sum_{k=0}^{m} r^k \binom{k+c_{n+1}}{c_{n+1}} r^{m-k} \binom{m-k+c}{c}$$
$$= r^m \sum_{k=0}^{m} \binom{k+c_{n+1}}{c_{n+1}} \binom{m-k+c}{c} = r^m \binom{m+c+c_{n+1}+1}{m}$$

where the final equality comes from Lemma 12. Lastly, we consider the availability of the $n+1$-queue system,

$$A_{n+1}(m) = r \frac{G_{m-1,n+1}}{G_{m,n+1}} = r \left(r^{m-1} \binom{m+c+c_{n+1}}{m-1}\right)\left(r^m \binom{m+c+c_{n+1}+1}{m}\right)^{-1} = \frac{m}{m+c+c_{n+1}+1}$$

Letting $c' = c + c_{n+1} + 1$ shows that the inductive hypothesis is true. □

Now that we have shown the restricted system's availability is $m/(m+c')$, we turn to Theorem 3 to relax the two restrictions.

Theorem 3. *Consider the instantaneous-travel system with state-dependent service, where all service rates converge linearly, i.e., for each queue, i, there exists a positive integer c_i, such that $\mu_i(k) \geq \mu_i(\infty)k/(k+c_i)$, and the relative throughput of uncharged queues is at most the relative throughput of the charged queues ($r_i^{\mu} \leq r^c$). Then, the availability of the system converges to 1 linearly as the number of vehicles $m \to \infty$.*

Proof. Proof of Theorem 3 As with Lemma 13, we prove this claim by inductively adding queues. For the base case, we begin with a charged queue. In a 1 queue system, $G_{m,1} = \tilde{\pi}_1(m)$. Therefore, using the fact that $\mu_1(\infty) = d_1$,

$$A_1(m) = r^c G_{m-1,1}/G_{m,1} = r^c \frac{\tilde{\pi}_1(m-1)}{\tilde{\pi}_1(m-1)\left(\lambda_1/\mu_1^c(m)\right)} = \frac{\lambda_1}{d_1} \frac{\mu_1^c(m)}{\lambda_1} \geq \frac{d_1(m/(m+c_1))}{d_1} = \frac{m}{m+c_1}$$

and $A_1(m)$ converges to 1 at least linearly quickly.

Now, we consider the inductive hypothesis where $A_n(m) = r^c G_{m-1,n}/G_{m,n} \geq m/(m+c)$ for some integer c and we want to show that $A_{n+1}(m) \geq m/(m+c')$ for some integer c'. We prove this by establishing a connection between the computational perspective, $A_{n+1}(m) = r^c G_{m-1,n+1}/G_{m,n+1}$ and a specific birth-death process. Then, we show that the throughput of the birth-death process is lower bounded when we replace the arrival rate and service rate with their respective lower bounds. Since the throughput of the queue is proportional to the throughput of the system, this gives us a lower bound on availability of the queue, and that lower bound was seen to converge linearly to 1 as $m \to \infty$ in Lemma 13.

For establishing the connection between the computational perspective and the birth-death process, we consider two cases.

Case 1: Adding a charged queue to the network. The relative throughput of the new queue is r^c, with relative arrival rate λ_{n+1} and limiting service rate $\mu_{n+1}(\infty) = d_{n+1}$.

We begin with the expression for the (marginal) stationary distribution of the number of vehicles in the $(n+1)^{\text{st}}$ queue, π^{comp}, which is a function of normalization constants derived in Denning and Buzen [1978]. We then construct a birth-death process whose stationary distribution, π^{bd} matches π^{comp}.

The marginal distribution, as derived in Denning and Buzen [1978], is

$$\pi^{comp}(k) = \frac{(r^c)^k}{\prod_{l=1}^{k} \mu_{n+1}(l)/d_{n+1}} \frac{G_{m-k,n}}{G_{m,n}}$$

for $k = 0, 1, \ldots, m$, which implies

$$\frac{\pi^{comp}(k+1)}{\pi^{comp}(k)} = \frac{r^c}{\mu_{n+1}(k+1)/d_{n+1}} \frac{G_{m-k-1,n}}{G_{m-k,n}} = \frac{d_{n+1}}{\mu_{n+1}(k+1)} r^c \frac{G_{m-k-1,n}}{G_{m-k,n}} = \frac{A_n(m-k)d_{n+1}}{\mu_{n+1}(k+1)}.$$

Now, consider a birth death process with state k corresponding to the number of vehicles in queue $n+1$. The arrival rate is $A_n(m-k)d_{n+1}$ and the service rate is $\mu_{n+1}(k)$. Intuitively, the arrival rate corresponds to the rate at which vehicles would enter the charged queue for the new region conditioned on $m-k$ vehicles being in the first n queues and the service rate remains unchanged. Therefore, the marginal distribution has the same distribution as this birth-death process,

$$\pi^{bd}(k+1) = \pi^{bd}(k)\frac{A_n(m-k)d_{n+1}}{\mu_{n+1}(k+1)}.$$

This establishes an explicit relationship between the computational perspective and the birth-death process reduction.

Case 2: Adding an uncharged queue to the network. Let the relative arrival rate of the corresponding charged queue be λ_{n+1} and therefore the relative arrival rate of the new queue is $p\lambda_{n+1}$. Also, let the maximum recharging rate be $b_{n+1} = \mu_{n+1}(\infty)$. Then, the relative throughput of the new queue is $r^u_{n+1} = p\lambda_{n+1}/b_{n+1}$.

The marginal distribution is

$$\pi^{\text{comp}}(k) = \frac{r^k_{n+1}}{\prod_{l=1}^{k}\mu_{n+1}(l)/b_{n+1}}\frac{G_{m-k,n}}{G_{m,n}}$$

which implies that

$$\begin{aligned}\frac{\pi^{\text{comp}}(k+1)}{\pi^{\text{comp}}(k)} &= \frac{r^u_{n+1}}{\mu_{n+1}(k+1)/b_{n+1}}\frac{G_{m-k-1,n}}{G_{m-k,n}} = \frac{b_{n+1}}{\mu_{n+1}(k+1)}r^u_{n+1}\frac{G_{m-k-1,n}}{G_{m-k,n}} \\ &= \frac{b_{n+1}}{\mu_{n+1}(k+1)}\frac{pd_{n+1}}{b_{n+1}}r^c\frac{G_{m-k-1,n}}{G_{m-k,n}} = \frac{pd_{n+1}}{\mu_{n+1}(k+1)}A_n(m-k).\end{aligned}$$

Now, consider a birth-death process with state k corresponding to the number of vehicles in the uncharged queue we are adding. The arrival rate in State k is $pd_{n+1}A_n(m-k)$ and the service rate is $\mu_{n+1}(k)$. Intuitively, the arrival rate corresponds to the expected rate at which vehicles would enter the uncharged queue,

conditioned on their being $m - k$ vehicles across the first n queues. Therefore, the marginal distribution has the same stationary distribution as this birth-death process,

$$\pi^{bd}(k+1) = \pi^{bd}(k)\frac{pd_{n+1}A_n(m-k)}{\mu_{n+1}(k+1)}.$$

This completes the link between the computational perspective and certain birth-death processes. Now, we show that decreasing the arrival and service rates of the birth-death process reduces the queue's throughput.

First, we reduce the arrival rate to the lower bound provided by the inductive hypothesis. For charged queues in State k this is $\lambda^{lb}(k) = d_{n+1}(m-k)/(m-k+c)$ and for uncharged queues in State k there is an extra factor of p. Let this new birth-death process with reduced arrival rates have a stationary distribution π'. Then, $X' \sim \pi'$ is stochastically dominated by $X \sim \pi^{bd}$. Since $\mu_{n+1}(k)$ is increasing in k, then the throughput of π^{bd}, $\mathbb{E}[\mu_{n+1}(X^{bd})] \geq \mathbb{E}[\mu_{n+1}(X')]$, the throughput of π'.

Second, we reduce the service rates and show that this reduces the mean service rate, i.e., the throughput, of the birth-death process. For charged queues, we use the linear convergence assumption to give the lower bound $\mu^{lb}_{n+1}(k) = d_{n+1}k/(k+c_{n+1})$. For uncharged queues, we use the lower bound that occurs when the maximum charging rate is minimized, $b_i = pd_i$, corresponding to $r^{u}_{n+1} = r^{c}$. Specifically, $\mu^{lb}_{n+1}(k) = pd_{n+1}k/(k+c_{n+1})$. Let π^{lb} be the stationary distribution of the birth death process with arrival rates λ^{lb} and service rates μ^{lb}. Now, we compare π', the stationary distribution just with reduced arrival rates and π^{lb} with reduced arrival rates and reduced service rates. Since π' has faster service rates than π^{lb}, X' is stochastically dominated by $X^{lb} \sim \pi^{lb}$. Moreover, in steady state the throughput as measured through the mean service rate is equal to the throughput as measured through the mean arrival rate. Since the arrival rates

$\lambda^{lb}(k)$ are decreasing in k, we get

$$\mathbb{E}[\mu_{n+1}(X')] = \mathbb{E}[\lambda^{lb}(X')] \geq \mathbb{E}[\lambda^{lb}(X^{lb})] = \mathbb{E}[\mu^{lb}_{n+1}(X^{lb})],$$

From Lemma 13, we know that the availability of π^{lb} is $m/(m + c')$ for some integer c', and therefore $\mathbb{E}[\mu_{n+1}(X^{bd})] \geq \mathbb{E}[\mu^{lb}_{n+1}(X^{lb})] = d_{n+1}m/(m + c')$. Therefore, $A_{n+1} = \mathbb{E}[\mu_{n+1}(X^{bd})]/d_{n+1} \geq m/(m + c')$ as desired. □

Now that we have a closed form lower bound for the availability of the system with instantaneous travel, we can use the exact same arguments as in Section 2.4.3 while replacing $m/(m + 2n - 1)$ with $m/(m + c')$.

2.6 Computational Results

This section will begin with an exploratory data analysis of e-scooter usage in Minneapolis. The latter part of this section will present results from the closed queuing model built upon Minneapolis data.

2.6.1 Exploratory Data Analysis

This section will begin by exploring the temporal distribution of micromobility trips completed across various cities. Then, basic properties of trips in Minneapolis will be explored before focusing on the origin-destination distribution of trips. We conclude by showing the impact of available scooters on trips completed.

Temporal Trip Demand

Previous literature has established that bike-sharing systems often see time-heterogeneous demand with peaks corresponding to morning and evening commutes. To explore this further, in Figure 2.2, we graph the distribution of trip start times for public micromobility programs in Austin, Chicago, Minneapolis, and New York City (NYC). Austin, Chicago, and Minneapolis data reflects only e-scooter trips while the NYC data corresponds to all citibike trips. The distributions in Figure 2.2 are from July 2021 or 2022 (depending on available data) and show the fraction of overall trips completed in each hour of the day. For reference, NYC July 2018 data is also included.

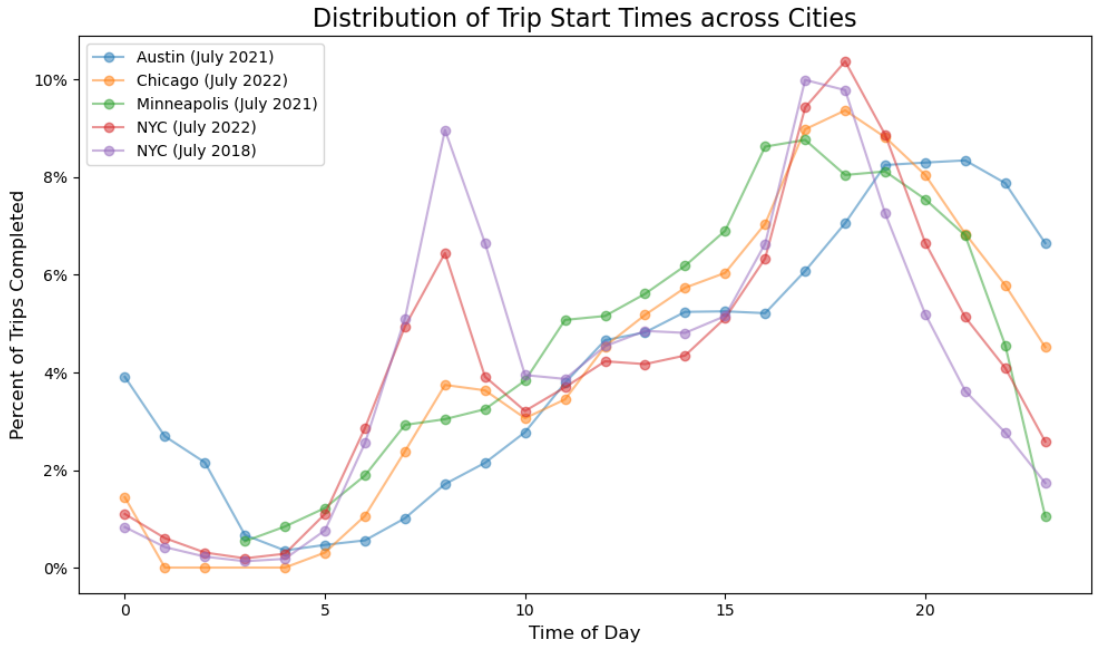


Figure 2.2: The distribution of micromobility trip start times in Austin, Chicago, Minneapolis and New York City in July. Minneapolis has no data between midnight and 4-am because the scooters are disabled at this time.

A few interesting insights emerge from Figure 2.2. The bimodal demand peaks corresponding to the morning and evening commute are most pro-

nounced in NYC and the morning commute has a comparatively small impact in the other three cities. The peak for the morning commute in NYC has decreased significantly between 2018 and 2022. The evening usage seems very similar across all systems. These distributions suggest similar e-scooter demand profiles across cities (demand slowly ramping up during the day before peaking between 5-8 pm) that may be very different from classic bike-sharing systems such as citibike.

Minneapolis Trip Properties

Now, we shift the focus to Minneapolis e-scooter trips. Two data sources are used, trip data and system snapshots [Minneapolis, 2022a,b]. The trip data has a record of every trip completed that includes key information such as trip duration, distance, start time, end time, start location and end location. The system snapshots (provided every four hours) identify the number of available scooters by CenterlineID (approximately 7500 total) that are grouped in 87 zones within the city. We combine zones until we have 11 neighborhoods as identified in this map of Minneapolis [Wikipedia, the free encyclopedia, 2024a]. Unless otherwise specified, trips only reflect those during weekdays since weekends typically have very different demand profiles.

Figure 2.3 is a histogram of e-scooter trip times that has been truncated to only show trips under an hour. The majority of trips are under 10 minutes with the peak of the distribution occurring around 5 minutes. Despite e-scooters charging per minute, there is still a significant right tail to the distribution.

Figure 2.3 is a histogram of e-scooter trip distances that has been truncated to

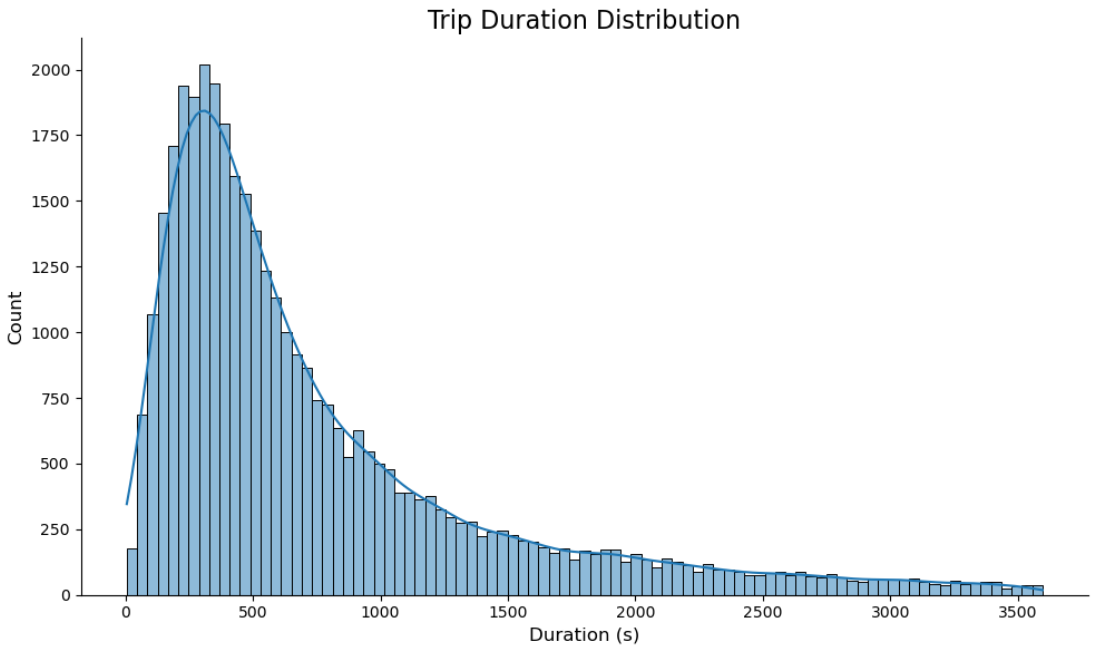


Figure 2.3: The distribution of trip duration (truncated at 1 hour) for Minneapolis e-scooter trips on weekdays in July 2021.

only show trips under 5km. There is a noticeable spike of trips with distance under 125m without a corresponding spike of very short trips. Several trips begin and end in very close locations, potentially corresponding to joy rides or running errands as opposed to transportation. Without these trips, the distribution peaks at just over 1km with a significant right tail.

Figure 2.5a is a times series of the number of trips initiated per day by region. This reveals a couple key insights about e-scooter demand in Minneapolis. First, there is a significant shift in demand around mid-September when the trips completed in the University neighborhood increase significantly. This roughly corresponds with the beginning of the semester at the University of Minnesota. Second, the demand is dominated by two neighborhoods: Central and University with all other neighborhoods contributing significantly fewer trips. Third, the trips initiated appears to be periodic on a weekly cadence with

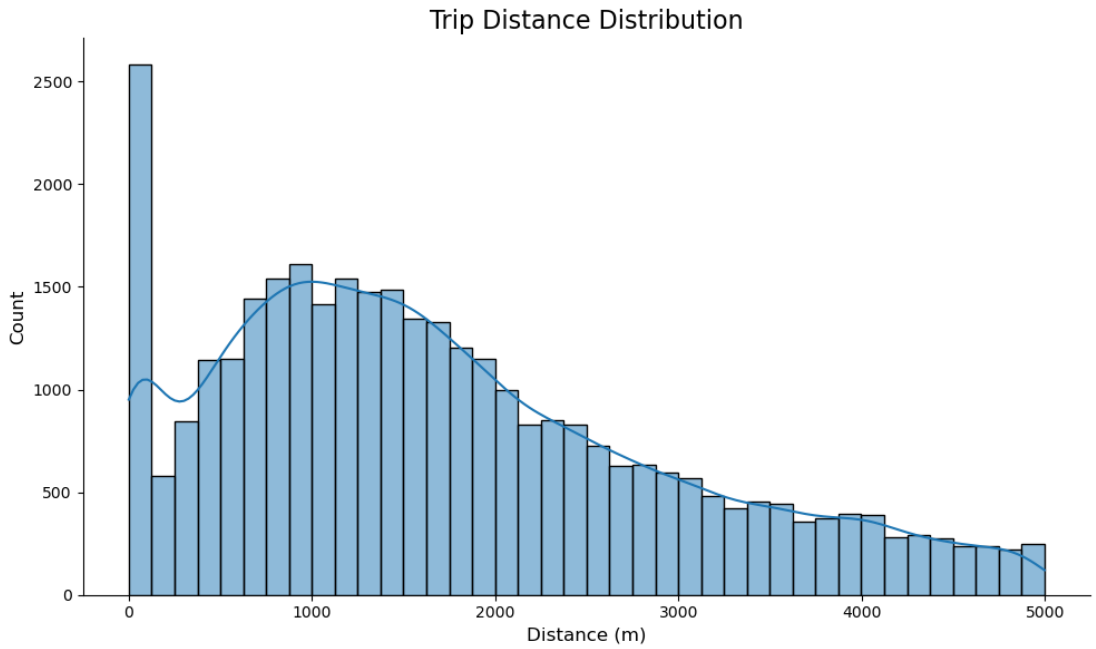


Figure 2.4: The distribution of trip distance (truncated at 5km) for Minneapolis e-scooter trips in July 2021.

the busiest days being weekends. All five of the busiest days in 2021 were weekends: September 25th (Sat), July 17th (Sat), September 11th (Sat), August 14th (Sat), and July 25th (Sun). Last, there is a significant decline in the number of trips completed towards the end of the year (November and December), likely a result of colder weather.

We are also able to track the number of available e-scooters in each region throughout the year, as illustrated by Figure 2.5b. This figure shows that significantly more e-scooters are positioned in the University neighborhood coinciding with the beginning of the semester. As expected, most vehicles are in one of the two regions responsible for the majority of trips. However, there are also many scooters in the Near North neighborhood despite this neighborhood not contributing many trips. This may reflect service level requirements for e-scooter providers operating in Minneapolis. As winter approaches, e-scooters

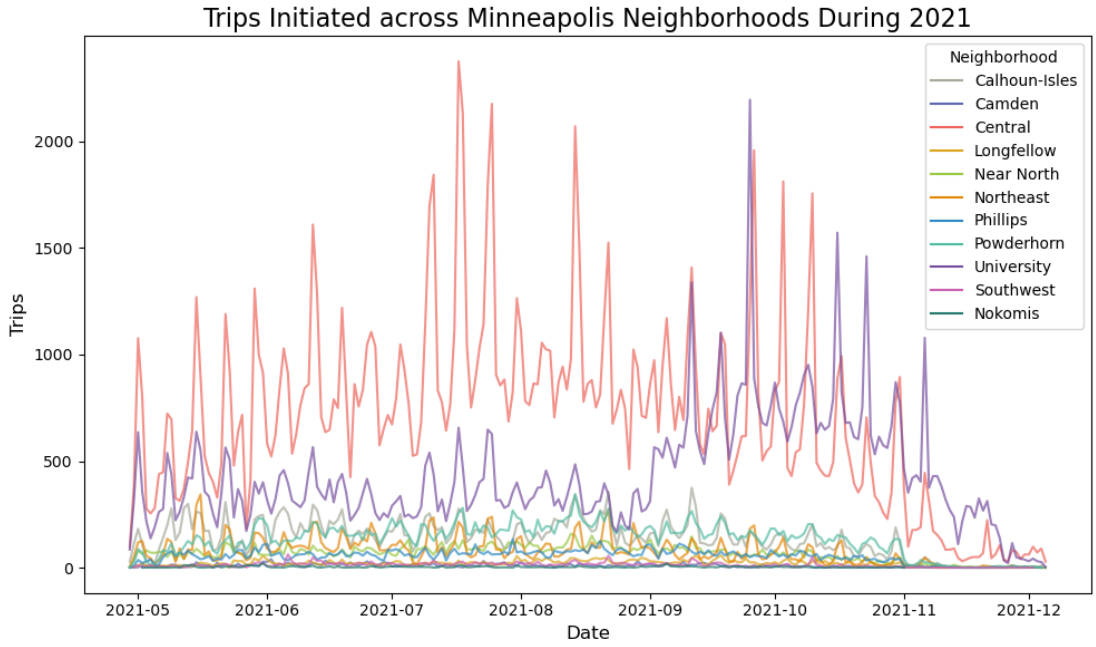
are removed from service.

Origin-Destination Demand

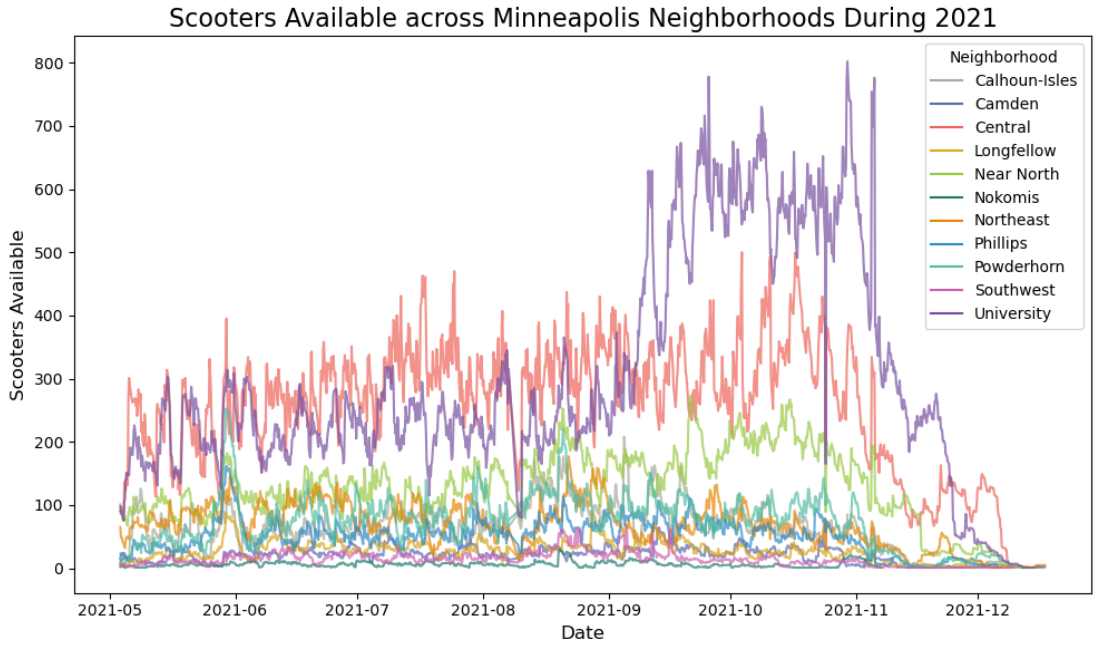
Now, we focus on understanding the distribution of trips according to their origin and destination zones. We begin by analyzing all trips in July before partitioning by time of day (morning, midday, and evening). After, we repeat these figures for October, after the university semester has started. Figure 2.6a shows the distribution of all weekday July trips by origin and destination. The shade of each entry in the matrix corresponds to the number of trips observed (on a log scale) between the origin (row) and destination (column) zone. All zones within a neighborhood are grouped together and the squares along the diagonal highlight trips that occur within a neighborhood. Zones without either a trip origination or destination are omitted.

When looking at all trips, we notice that most trips begin and end within the same neighborhood, with several beginning and ending within the same zone (on the diagonal). This supports Figure 2.4 which shows the distribution of trip length, including several trips shorter than 125m and otherwise a peak at just over 1km. The trip demand looks largely symmetric about the diagonal line, implying that trips are relatively balanced between zones or neighborhoods. However, any system that does not rely heavily on rebalancing would show balanced trips over long time scales such as a month. Neighborhoods show a very heterogeneous distribution of trips with respect to both frequency and destination.

Now, we consider the trip distribution during different times of the day



(a) Time series of the number of e-scooter trips initiated in each neighborhood.



(b) Time series of the number of e-scooters available in each neighborhood.

Figure 2.5: The trips initiated and e-scooter allocation by neighborhood in Minneapolis in 2021.

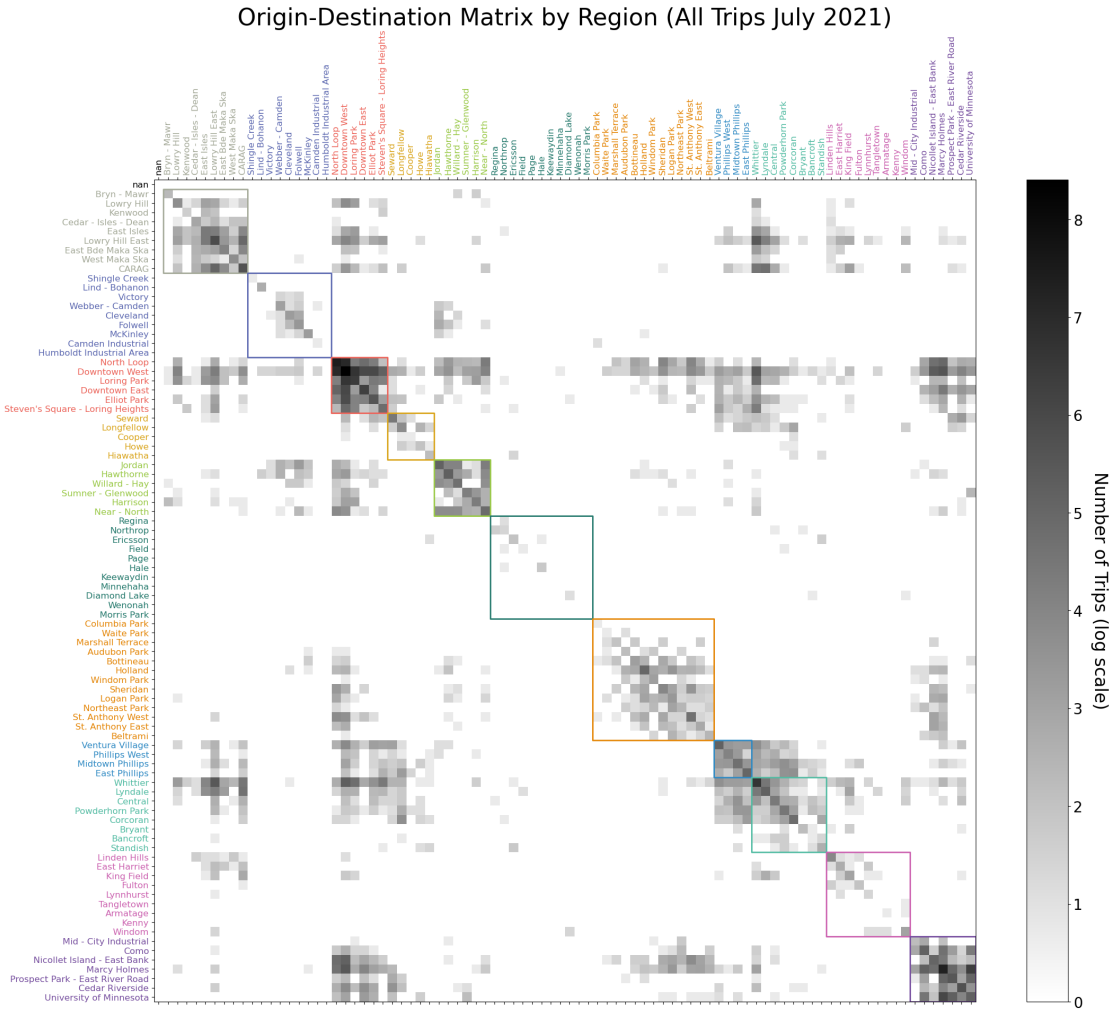
(morning, midday, and evening). As discussed previously, morning is the least busy time for the system with the majority of trips occurring within the Central neighborhood. This suggests that the e-scooters are rarely used for the morning commute outside Central.

As we move to midday and evening, the system becomes busier and we observe trips covering more parts of the city. The trip distribution observed during midday and evening appear similar with the evening having more overall trips. Importantly, the symmetry about the diagonal appears to be present during both time periods, implying that we are not seeing heavy directional demand dictated by morning or evening commutes. This is an important distinction from bike-sharing programs in large cities that have significant commuting flows that dominate their dynamics. The model assumption of constant demand rates is therefore more appropriate in the Minneapolis setting.

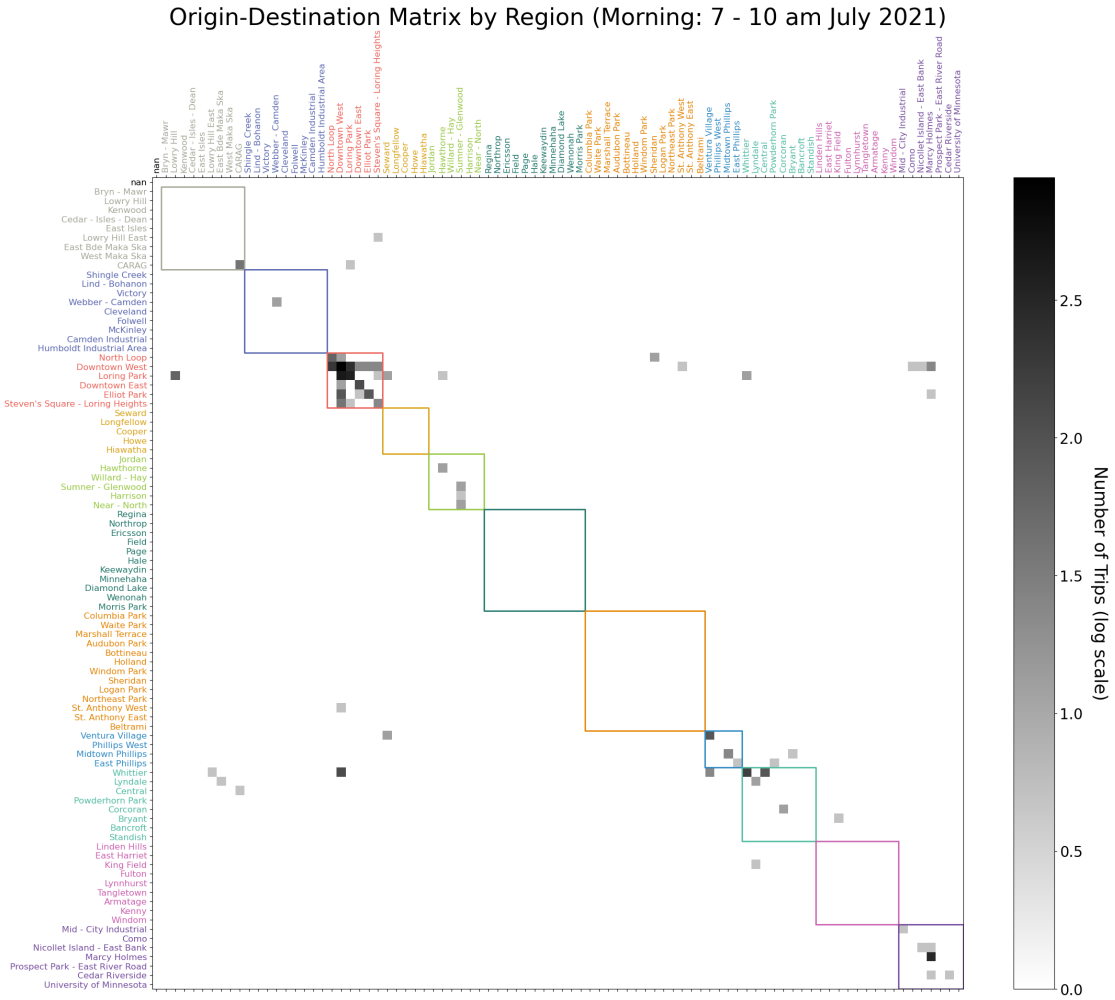
The insights previously noted appear to extend to October. Despite the significant increase in trips from the University region, the trip distribution is visually very similar between July and October.

Spatial Demand

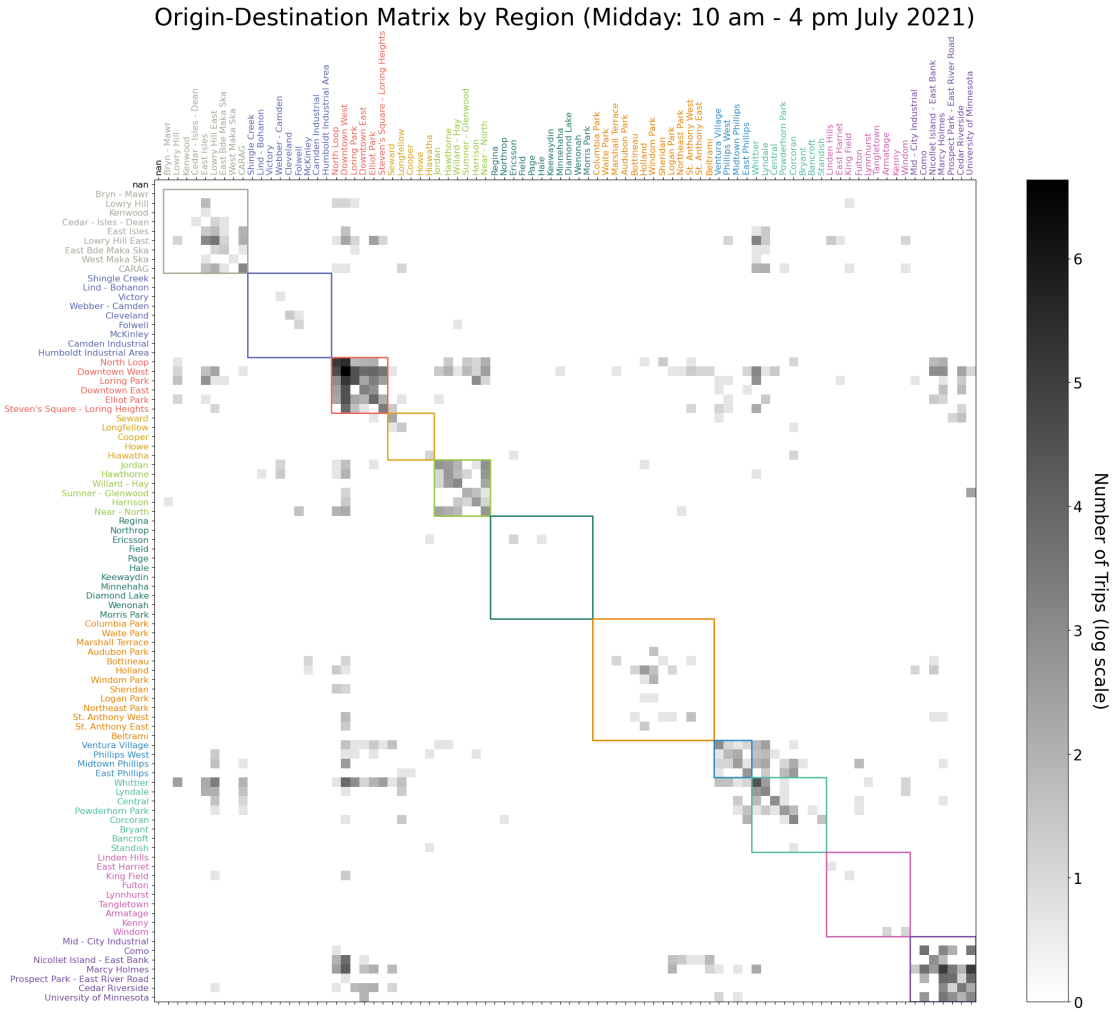
The last part of the exploratory data analysis considers the impact of the number of available vehicles on trips initiated. Potential users of the system first locate the closest available (and charged) vehicle before deciding whether to walk to it and begin their trip. As more vehicles are available in a region, it becomes more likely that an e-scooter is sufficiently close to the potential user. Therefore, we expect the trip initiation rate increase with the number of available vehicles in



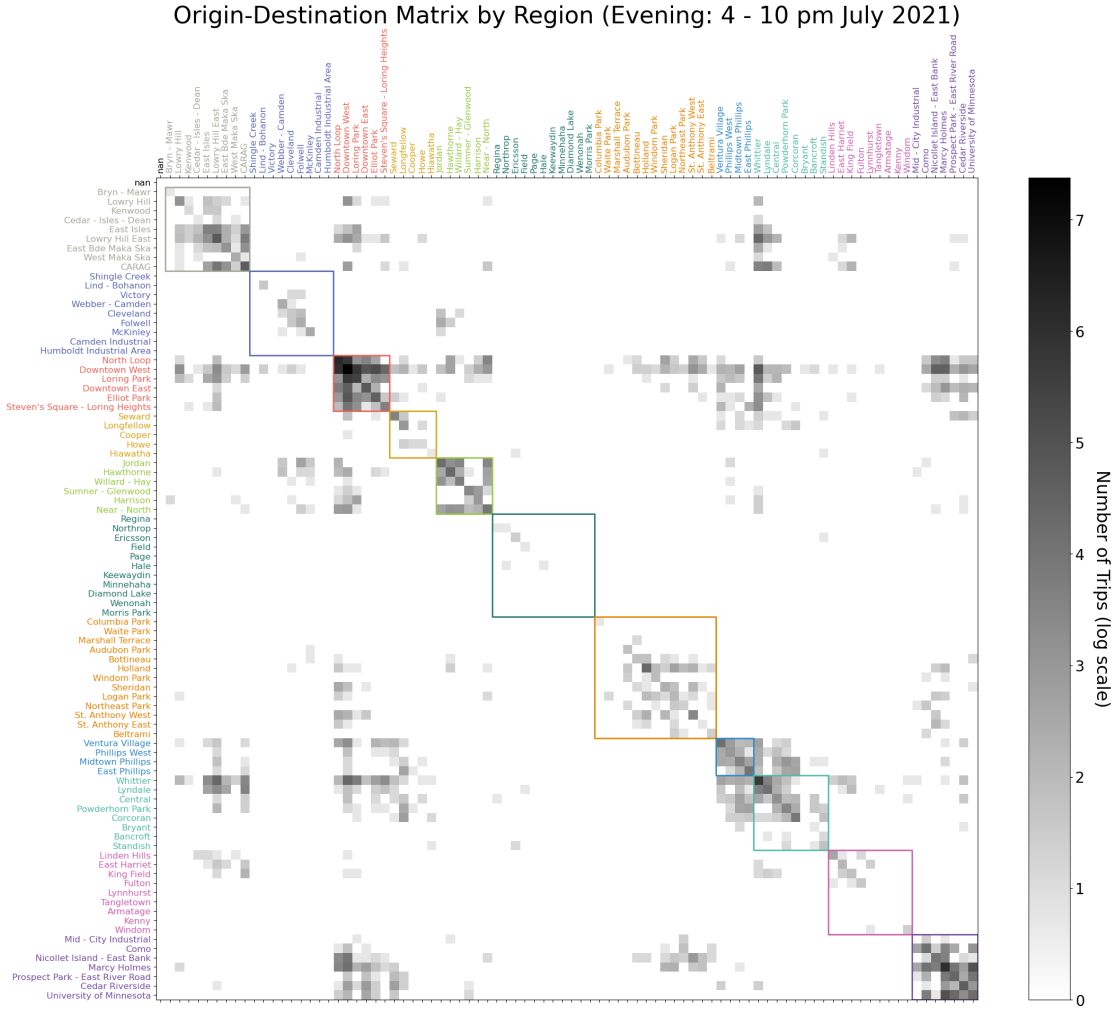
(a) All trips from July 2021



(b) Morning trips (7 - 10 am) from July 2021

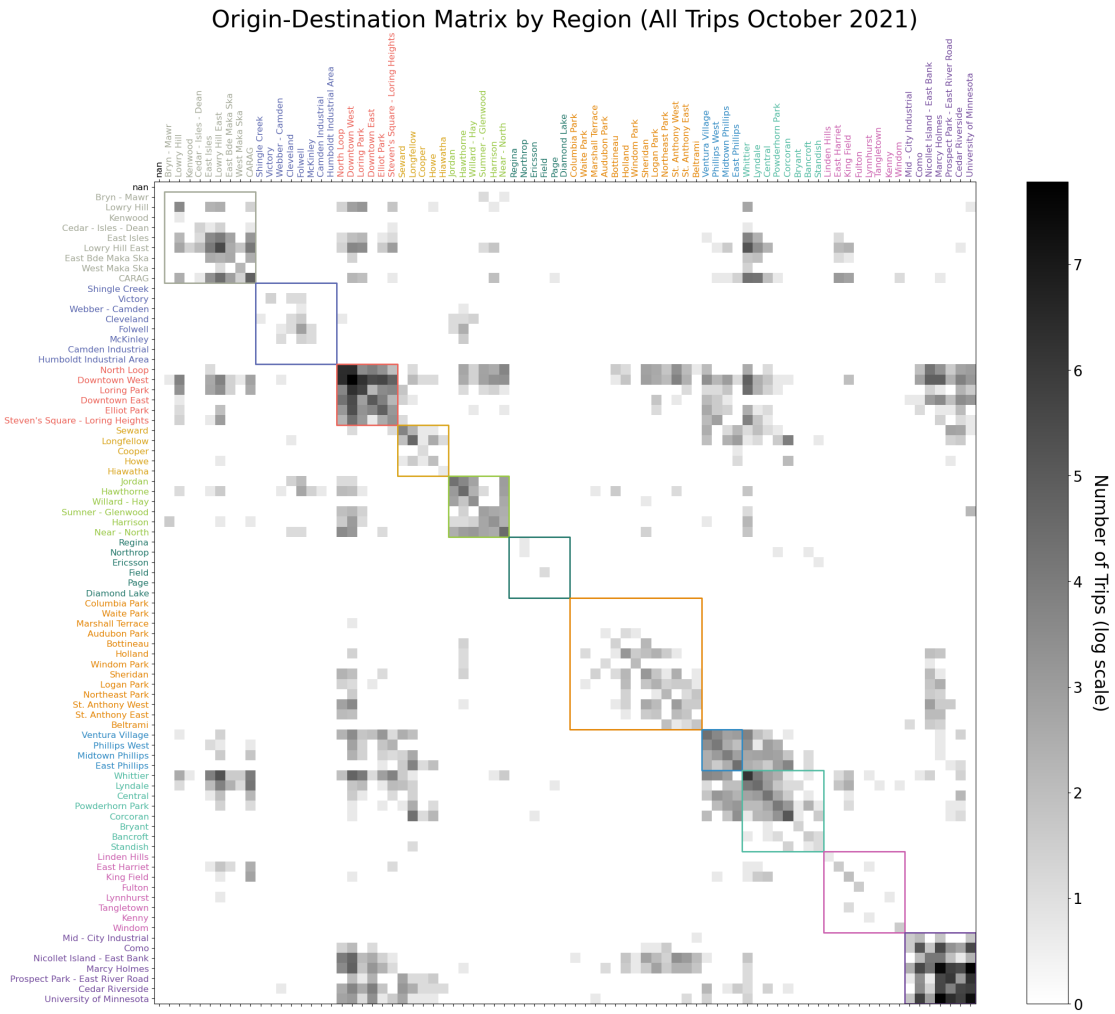


(c) Midday trips (10 am - 4 pm) from July 2021

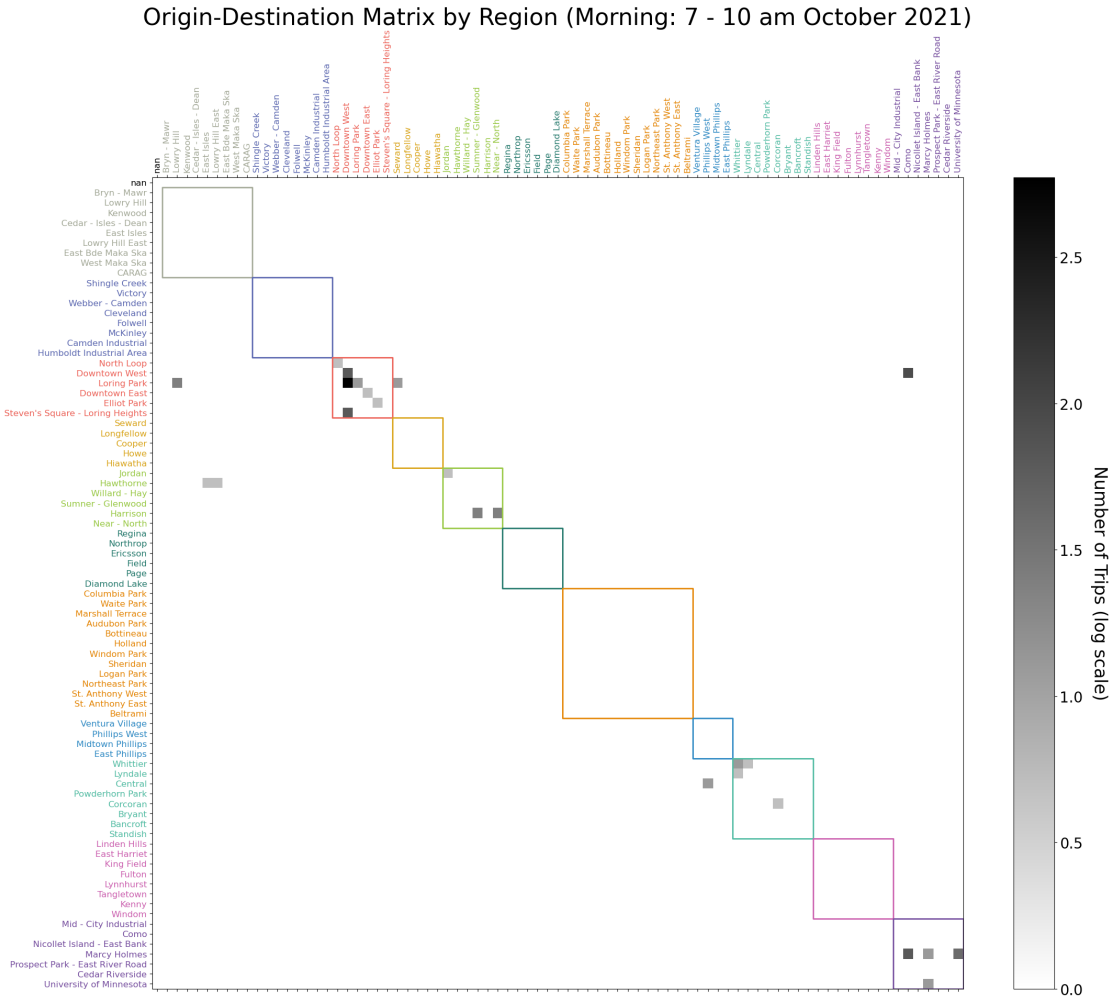


(d) Evening trips (4 - 10 pm) from July 2021

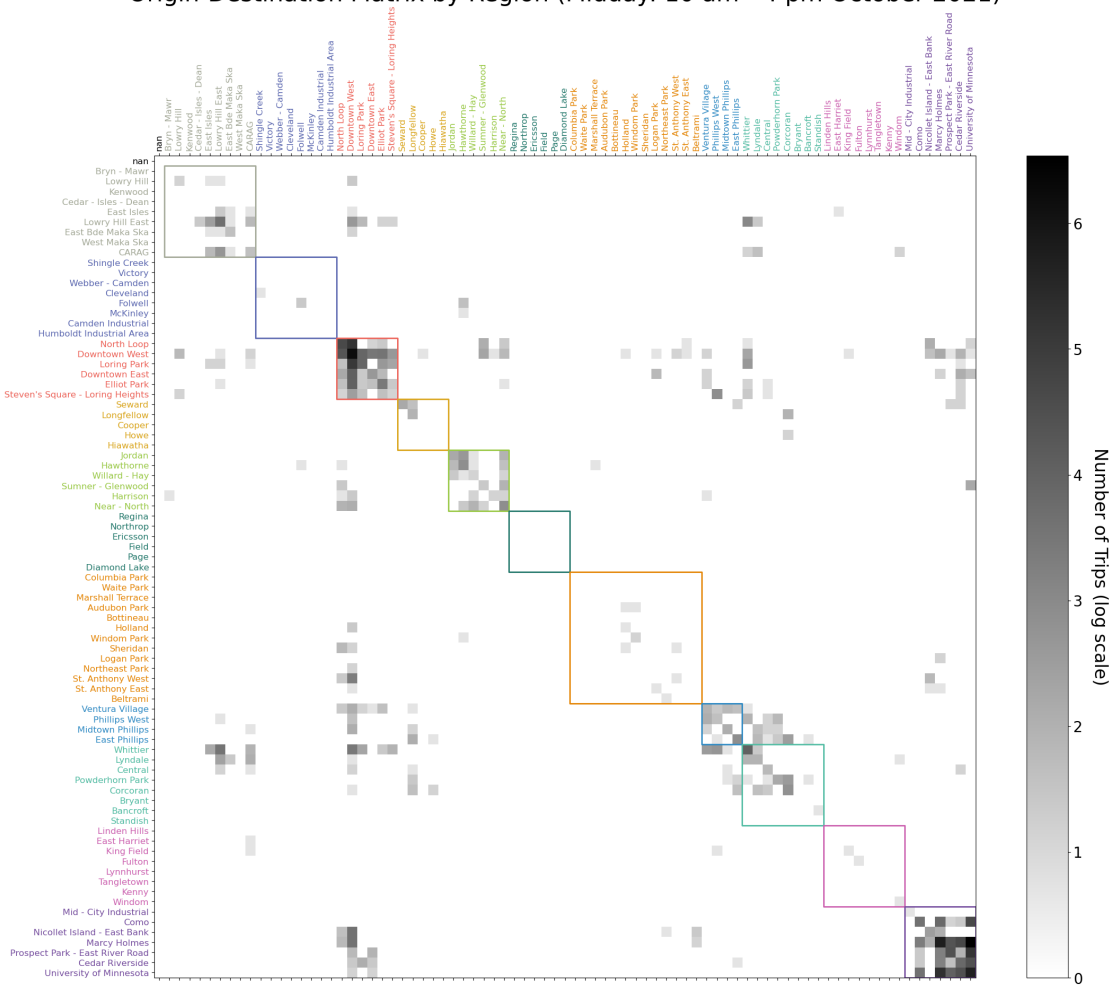
Figure 2.6: The number of e-scooter trips by each origin (row) and destination (column) zone in Minneapolis at different times in July.



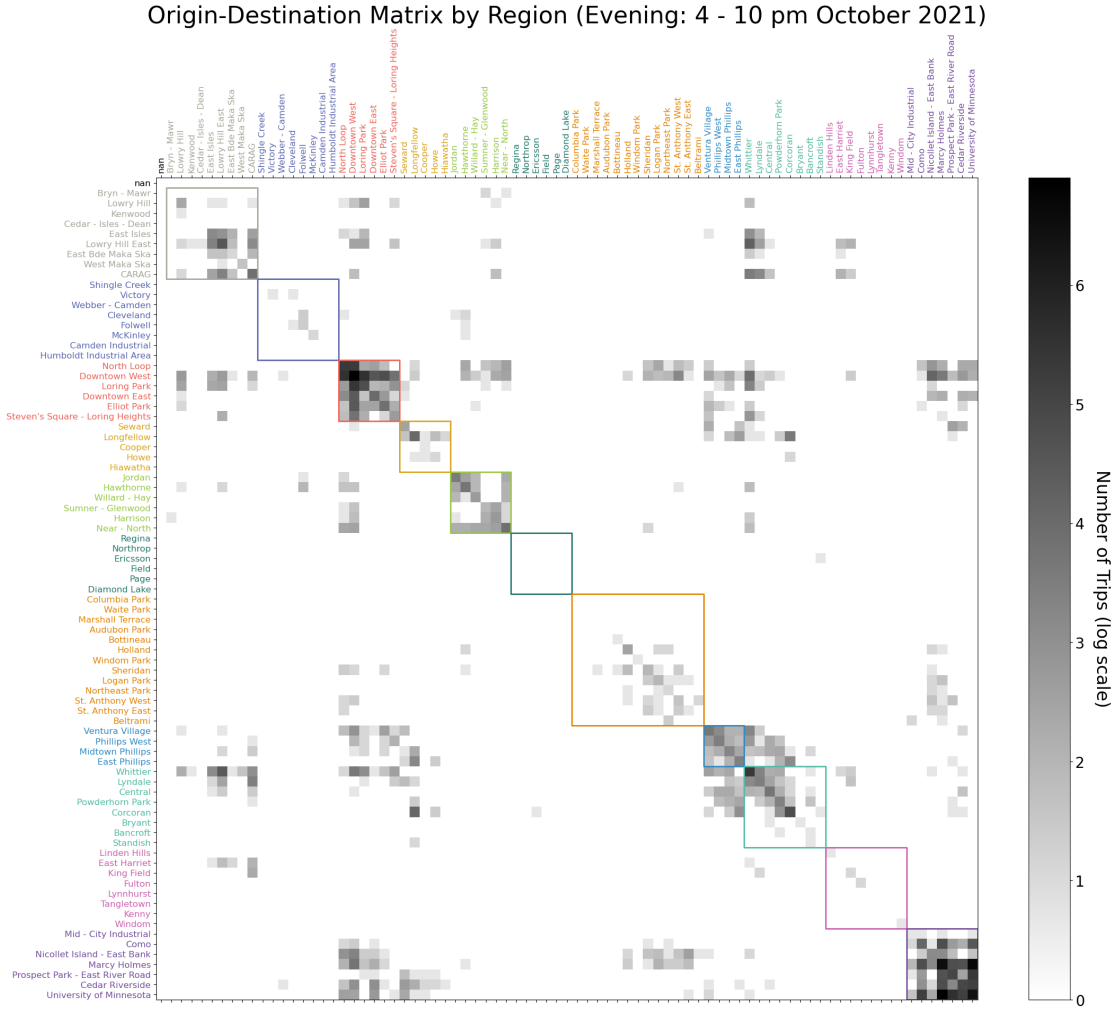
(a) All trips from October 2021



(b) Morning trips (7 - 10 am) from October 2021



(c) Midday trips (4 pm - 10 pm) from October 2021



(d) Evening trips (4 - 10 pm) from October 2021

Figure 2.7: The number of e-scooter trips by each origin (row) and destination (column) zone in Minneapolis at different times in October.

the neighborhood. In our model, this is reflected by the state-dependent service rates of the charged queues.

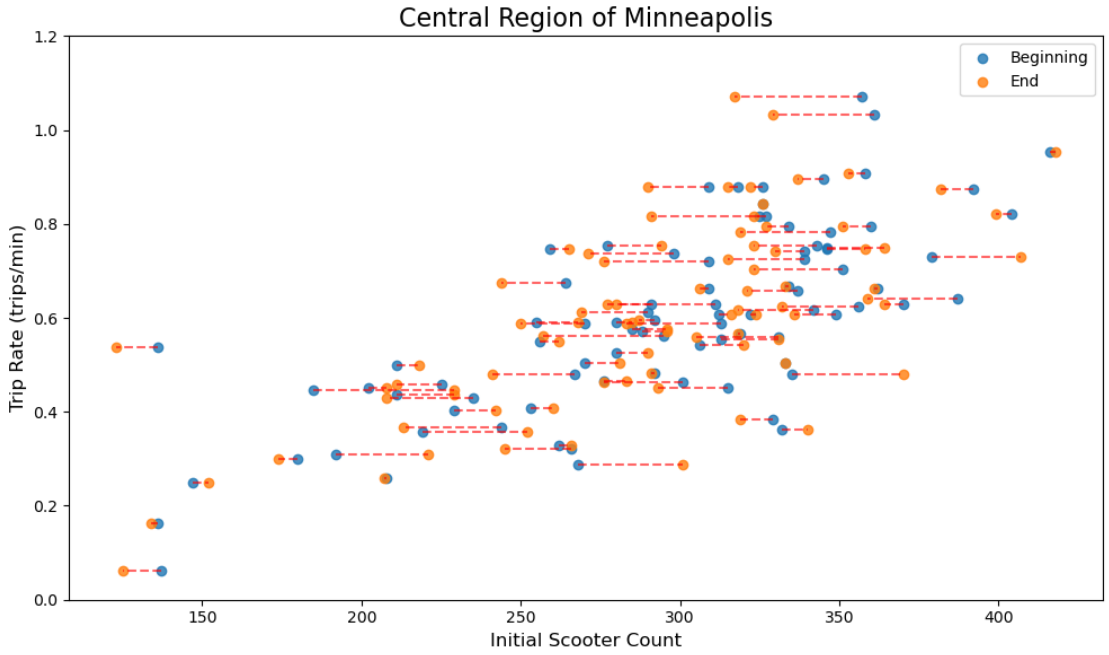
This intuition is supported by Figure 2.8 which illustrates that more vehicles result in more trips initiated. The x-axis is the number of scooters available and the y-axis corresponds to the observed trip initiation rate. For every weekday in the time period, the number of available vehicles at 4pm and 8pm is recorded, along with the observed trip initiation rate for the specified neighborhood. Each day's data is recorded as a blue point (e-scooters at 4pm), an orange point (e-scooters at 8pm) connected by a horizontal dashed red line. The y-coordinate of both points is the observed trip initiation rate for that day.

For all three scenarios in Figure 2.8 (Central: May - August, University: May - August, University: September and October), there is a positive correlation between the observed trip rate and the initial scooter count. A linear approximation appears to be a reasonable method to capture this relationship.

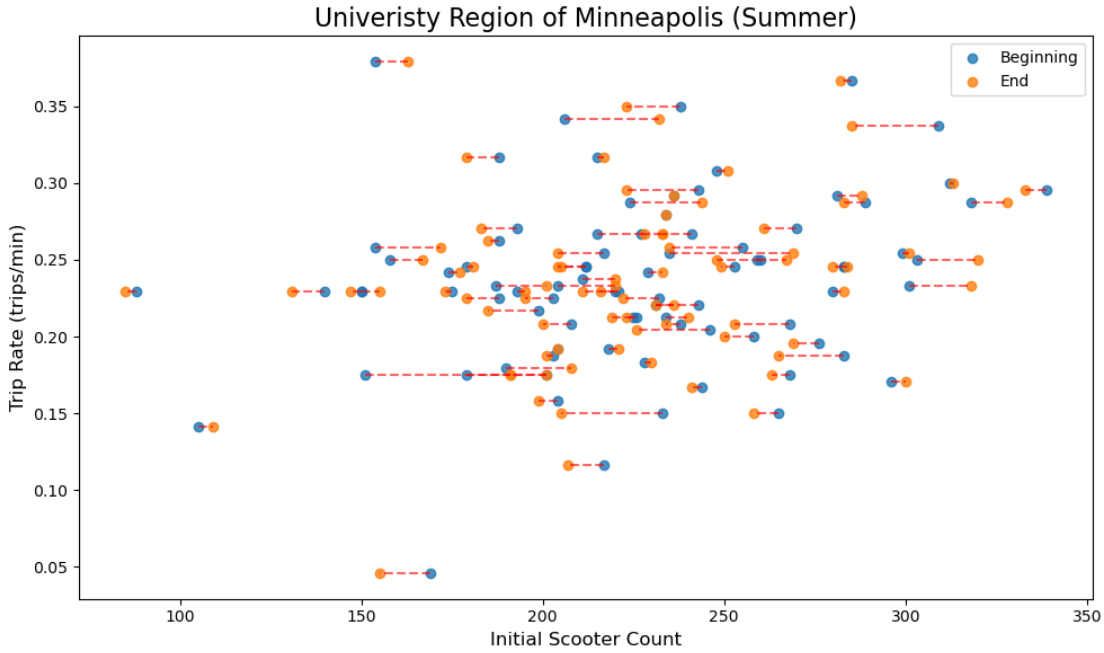
2.6.2 Case Study

In this subsection, we illustrate how to use the closed queueing model to estimate the tradeoffs between different charging mechanisms in Minneapolis. We begin by describing how the inputs to the problem are derived from Minneapolis' open data.

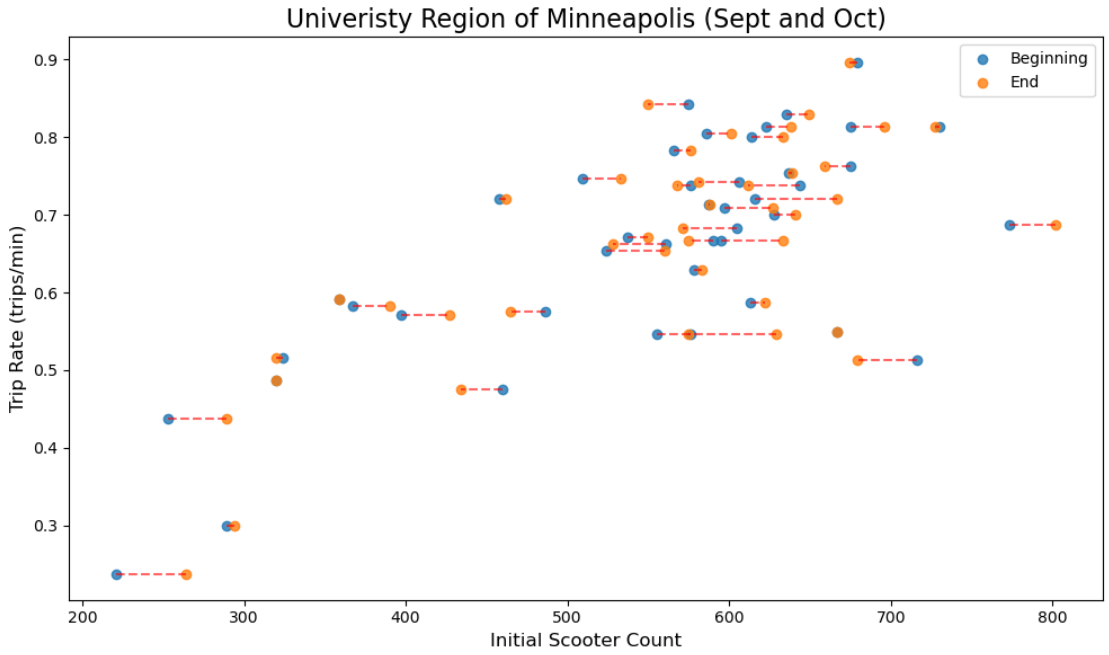
1. The regions in our model are the 11 neighborhoods identified in Minneapolis.
2. τ_{ij} represents the average trip time for a trip from neighborhood i to neigh-



(a) Observed trip initiation rate vs. available vehicles for the Central region between May and August 2021.



(b) Observed trip initiation rate vs. available vehicles for the University region between May and August 2021.



(c) Observed trip initiation rate vs. available vehicles for the University region between September and October 2021.

Figure 2.8: Plots of the observed trip rate (between 4-8pm) vs the number of available vehicles at 4pm and 8 pm. Each daily observation corresponds to a blue dot (vehicles available at 4pm), an orange dot (vehicles available at 8pm), and a dashed horizontal red line connecting the two. The y-coordinate corresponds to observed trip rate.

borhood j. This is computed using the Minneapolis trip data.

3. p_{ij} represents the probability that a vehicle is out of charge after a trip is modeled as depending on both the origin and destination. We assume that e-scooters have a battery life of 2 hours and set $p_{ij} = \tau_{ij}/2$ hours.

4. ϕ_{ij} represents the maximum demand for trips from location i to j. A system operator may not observe this directly, but can use information such as app openings to estimate this quantity. Unfortunately, we do not have access to this data and therefore use the following estimate

$$\frac{2 \times \text{trips between } i \text{ and } j}{8 \times \#\text{ dates}}$$

where trips between i and j are the total observed weekday trips between

May and September and dates is the number of weekdays in this time period. The factor of 2 on the numerator addresses our underestimate from only considering completed trips (and not observing latent demand). The 8 on the denominator normalizes the trip rate by assuming that all trips occur within 8 hours.

5. $\zeta_i(m_i)$ represents the fraction of customers that will find a vehicle sufficiently close if there are m_i charged vehicles in neighborhood i. $\zeta_i(m_i) \times \phi_{ij}$ is estimated by fitting a linear regression model to the midpoint of the red dashed lines from Figure 2.8 for each neighborhood. For the service rates of the charged queues (given by $\zeta_i(m_i)\phi_{ij}$), the minimum is $0.0001/\phi_{ij}$ (ensuring a rate of at least 0.0001 trips/second) and the maximum is ϕ_{ij}.

By fitting the linear function to the observed trip rates between 4pm - 8pm and normalizing ϕ_{ij} by assuming an 8 hour time window for all trips, we are intentionally biasing our input parameters to reflect the busier times of the system.

Now, we describe how the service rates of the charging mechanisms are determined.

1. Crowd-sourcing: each bike takes an average of 5 hours to be collected, brought home, charged, and re-deployed. System operators will have access to better information for estimating this quantity.
2. Employees: There are 11 total employees, each assigned to one neighborhood. The area of each neighborhood is found from Wikipedia and the speed of the employee is assumed to be 18km/hour (5m/s) [Wikipedia, the free encyclopedia, 2024b,c,d,e,f,g,h,i,j,k,l]. The employees speed is slower, reflecting that we are using Euclidean distances (where they can

travel in any direction) which would be optimistic for a city setting. We use a TSP tour constant of 1 and assume it takes t_{switch} =1 minute to find each vehicle and swap the battery.

3. Charging Stations (perfect): We assume a total of 175 charging stations are setup across the neighborhoods and allocated proportional to demand. We round to ensure each region has an integer number of stations and at least 1 per neighborhood. The recharge time is 2 hours per vehicle
4. Charging Stations (imperfect): We keep the same 175 charging stations and set $p_{\text{abandon}} = 0.5$. We use crowd-sourcing as the back-up mechanism for abandoned vehicles.
5. Vending Machines: We assume that vending machines are distributed around the city and reduce the probability of a vehicle being uncharged after every trip by 20%. Depending on the allocation of these stations, the reduction in probability can be origin and destination dependent. We also use crowd-sourcing as a back up mechanism since bikes will arrive uncharged.

Using the assumptions detailed above, we are able to compare each mechanism. Figure 2.9 shows the availability of each mechanism as the fleet size (m) grows. Recall, the availability is our performance measure and all systems have the same fluid relaxation (upper bound) that they are being compared to. Therefore, this represents a direct comparison of the mechanisms.

This graph presents some interesting findings. First, employees are the most efficient mechanism and crowd-sourced is the least efficient. Second, incorporating people not returning e-scooters to charging station significantly hurts the performance of the system. Last, there is an ordering of the performance of the

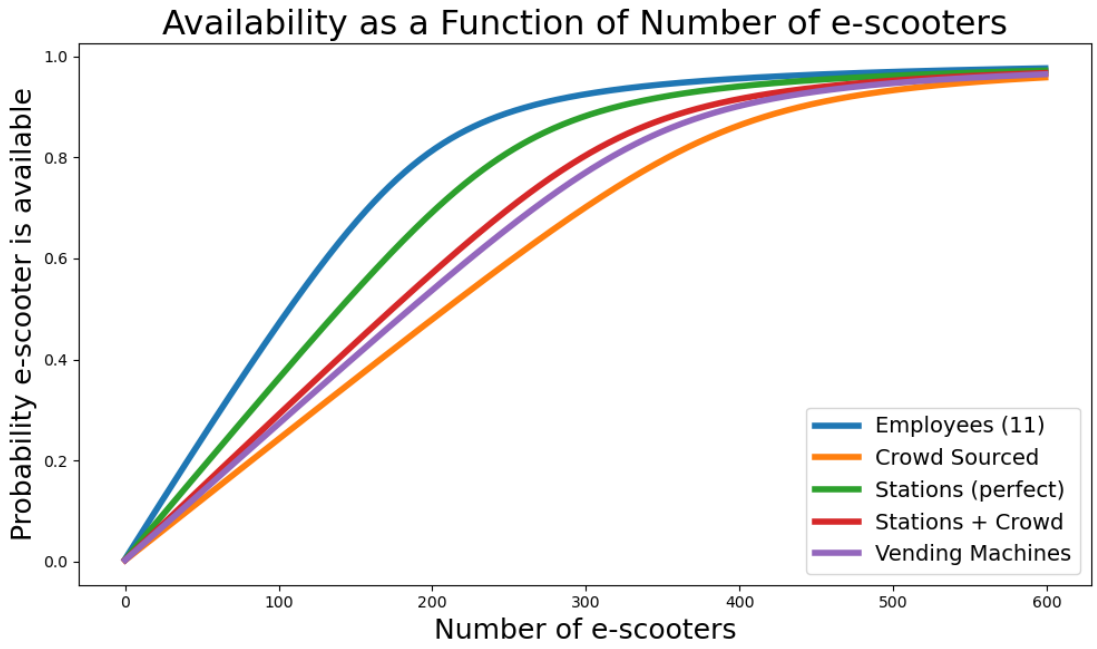


Figure 2.9: The availability for each charging mechanism as the number of e-scooters increases.

systems across all values of m. This does not always have to be true, but is likely a consequence of an ordering of charging rates for the states frequently visited.

Despite employees being the most efficient, this does not mean it is the best choice for charging mechanisms. The decision of charging mechanism should also consider fleet size and cost-structure. In some settings, it may be better to use a less efficient charging mechanism, but deploy more vehicles to get the same performance. The cost structure of these different operating models is also a very important consideration. For example, charging stations are expensive to build and connect to the grid but have a relatively lower recurring cost. Employees and crowd-sourcing are the opposite, with higher recurring costs but lower initial costs. As operators and cities negotiate the system design when granting permission to operate, this model quantifies the trade-offs between key design decisions such as the business model and fleet size.

CHAPTER 3

PLANNING FOR ON-DEMAND MULTIMODAL TRANSIT SYSTEMS

3.1 Introduction

On-Demand Multimodal Transit Systems (ODMTS) are transit systems that pair on-demand shuttles with high-frequency and high-capacity fixed transit services (e.g. subway, rail, or bus line). The shuttles typically serve as the first and last-mile solution, bringing passengers to and from stations which are connected via fixed transit. It aims to improve upon the inefficiencies of using low-ridership fixed bus lines to serve as a first and last-mile solution.

Quantitative modelling illustrates that ODMTS can simultaneously reduce operating costs and transit time [Maheo et al., 2019, Guan et al., 2022, Yan et al., 2021]. This inspired a 6-month pilot in Atlanta, where on-demand shuttles were deployed in specific regions. Most of the trips served connected with another form of public transportation and many of the replaced trips would have been completed with either a personal vehicle, ride-hailing service, or taxi [Van Hentenryck et al., 2023]. Lastly, surveys of low-resource communities have also supported a broad preference for mobility on-demand over fixed route transit [Yan et al., 2021].

The goal of this paper is to provide a simple, extensible model that helps cities explore the feasibility of ODMTS. To aid cities with this exploration, it is critical that a variety of real world considerations can be easily evaluated, such as equity between regions, latent demand, and peak vs. off-peak performance. Previous work in this area focuses on a more operational perspective, by formu-

lating IPs to optimize routing decisions [Maheo et al., 2019, Guan et al., 2022, Yan et al., 2021]. These large IP models are complex to build and solve since the methodologies proposed include bender's decomposition, trip filtering, cut bundling, and more [Maheo et al., 2019]. They also make strong assumptions such as all demand is known apriori and passenger trips can be sampled from a known prior distribution. In addition, it typically requires research and new algorithms to extend them for real-world considerations such as capturing latent demand [Guan et al., 2022]. Due to these complexities, it is challenging for these models to address all the design considerations cities may have.

This paper introduces a new modelling framework that consists of two components. First, we model the dynamics of individual regions as a discrete-time Markov chain (DTMC) with 2 classes of customers: first and last-mile. The DTMC can be decomposed into 2 independent DTMCs if it takes constant time for a shuttle to return to service each queue. This relates the number of shuttles in service to performance metrics such as waiting time. Then, an IP is formulated to find the optimal deployment of shuttles to minimize the expected waiting time of all customers under time-varying (piecewise-constant) demand. This can be solved efficiently by a network flow relaxation if there are diminishing returns for additional shuttles in service.

These components are modular. For instance, instead of using the DTMCs to compute performance metrics, either simulation or flow based methods could be used. These alternative performance metrics can easily be plugged into the second stage, allowing diversity in the modelling 'physics' used. In addition, key considerations such as equity, latent demand, and peak vs. off-peak performance can be easily integrated into this modelling framework by adjusting

either the first or second component.

The primary contributions of this paper are:

1. A simple model 'physics' for the performance of shuttles in a region with a mass transit station. The simple dynamics allow the considerations of transit authorities to be easily incorporated when computing relevant performance metrics.
2. An efficient way to solve for the optimal deployment of shuttles throughout the day under time-varying demand. This can be applied even if a different model 'physics' is selected for the first stage, as long as there are diminishing returns for additional shuttles.
3. A city scale experiment on Washington DC that shows how our method can illustrate the potential of ODMTS.

3.2 Literature Review

We being with discussing literature on the design of On-Demand Multimodal Transit Systems (ODMTS), first proposed by Maheo et al. [2019]. Motivated by improving transit service during off-peak hours, the paper proposes a system where stations are connected by high-frequency bus lines and shuttles bring passengers to and from the stations. They formulate a Mixed Integer Program and present solution techniques. Future work relaxes some of the assumptions in this work. Dalmeijer and Van Hentenryck [2020] relaxes three constraints to better reflect transit operations: 1) lines can have different frequencies, 2) other modes of transit (e.g. rail) are added, and 3) a limit on the number of transfers

per passenger is introduced. Auad-Perez and Van Hentenryck [2022] considers fleet-sizing and integrates ride-sharing into the model while Périvier et al. [2021] considers line planning. Basciftci and Van Hentenryck [2023] and Guan et al. [2022] integrate latent demand by modelling consumer choosing whether or not to participate in the system. These models solve for game theoretic equilibrium. Another approach to consumer choice is presented by Banerjee et al. [2023] which focuses on the transit operator's perspective and decouples the network design and pricing problems. This work does not consider the network design element of the problem, allowing us to use a simpler model 'physics' that is easier to extend to other questions.

Another common approach to evaluating the potential of microtransit is detailed agent-based simulation studies. Archetti et al. [2018] conducts a study where potential passengers consider alternatives (e.g. walking, private car) and shows that the system can satisfy many users and substantially reduce private car usage. Oh et al. [2020] finds similar results when evaluating the potential for autonomous on-demand shuttles in Singapore.

Some other notable approaches include Steiner and Irnich [2020] which develops a planning model for integrating shuttles with bus lines. Given a demand for trips, the model decides which lines should be used and how shuttles will integrate with the new bus system. Martínez Mori et al. [2023] quantifies the benefit of allowing recourse decisions as uncertainty is revealed throughout the day. Lastly, Wang and Odoni [2016] studies a stochastic last-mile system where users arrive in batches to the service station and on-demand shuttles bring them to their destinations.

The majority of academic literature in this area focuses on operational ques-

tions and network design formulations. Typically, papers consider jointly optimizing the public transit lines and the on-demand shuttle operations. These lead to large mixed integer programs that can be challenging to solve. It also makes it challenging to extend these models to important considerations such as latent demand or robustness to mis-specified parameters. Our work simplifies the model dynamics by not considering the network design problem, allowing for easier extensions to key considerations.

3.3 Individual Region Model

The individual region model is the more detailed 'physics' of the system which we use for computing performance metrics, such as waiting time. We first introduce the individual region model under the assumption that only 1 shuttle is serving the region. Later, we extend the model to multiple shuttles./

The city is divided into non-overlapping regions such that each region has exactly one mass transit station (e.g. subway, rail or high-capacity bus line). We model the number of customers waiting to be picked up as two discrete-time Markov chains (DTMCs), one for the station (last-mile) and one for the neighborhood (first-mile), that can be viewed, together, as a single Markov chain.

In this model, we abstract away detailed operational decisions such as matching and routing. This makes it easier to analyze, solve, and extend. However, without detailed operational decisions, this model is only suited for understanding the impact and trade-off between key design decisions.

3.3.1 Customer Arrivals

There are two types of customers: first-mile (from the neighborhood to the station) and last-mile (from the station to the neighborhood). Upon arriving, customers wait in their respective queues (station or neighborhood) until they are picked up by the shuttle. Customers arrive to the system according to independent Poisson processes with rates λ_s and λ_n for the station and neighborhood respectively.

3.3.2 Shuttle Service

Each shuttle has a capacity of C customers and operates on a predefined loop: picking up passengers from the station and dropping them at their destinations in the neighborhood, then picking up customers from the neighborhood and dropping them at the station. When picking up customers, shuttles will pick up as many as possible up to their capacity C according to first-come first-served at each location. If no customers are available for pickup, the shuttle departs and maintains the original schedule. Let τ_{sn} (τ_{ns}) be the *deterministic* time required to pick up and drop off up to C customers when leaving from the station (pick up and drop off up to C customers and return to the station). Define $\tau = \tau_{sn} + \tau_{ns}$.

For simplicity, we assume that the shuttle picks up all passengers in the neighborhood simultaneously. The additional waiting time (while the shuttle is picking up other passengers in the neighborhood) is a function of the region geography and number of passengers to pick up. Since it is not a function of the queueing dynamics, this quantity can be estimated separately.

3.3.3 DTMC Formulation

We can formulate the process described above as a Discrete-time Markov Chain (DTMC). The Markov Chain tracks three quantities: 1) the number of customers waiting at the station (n_s), 2) the number of customers waiting in the neighborhood (n_n) and 3) the location of the shuttle. The state is updated immediately before service (customers picked up at the station or in the neighborhood). Therefore, the transition dynamics of the DTMC can be described as

$$(n_s, n_n, \text{station}) \rightarrow ((n_s - C)^+ + Poisson(\lambda_s \tau_{sn}), n_n + Poisson(\lambda_n \tau_{sn}), \text{ neighborhood})$$

$$(n_s, n_n, \text{neighborhood}) \rightarrow (n_s + Poisson(\lambda_s \tau_{ns}), (n_n - C)^+ + Poisson(\lambda_n \tau_{ns}), \text{ station}).$$

Now, if we consider 2 time steps we get the transition dynamics

$$\begin{aligned}(n_s, n_n, \text{station}) \rightarrow (\ & (n_s - C)^+ + Poisson(\lambda_s \tau), \\ & (n_n + Poisson(\lambda_n \tau_{sn}) - C)^+ + Poisson(\lambda_n \tau_{ns}), \\ & \text{station}) \\ (n_s, n_n, \text{neighborhood}) \rightarrow (\ & (n_s + Poisson(\lambda_s \tau_{ns}) - C)^+ + Poisson(\lambda_s \tau_{sn}), \\ & (n_n - C)^+ + Poisson(\lambda_n \tau), \\ & \text{neighborhood}).\end{aligned}$$

Consider the two step transition probability when the shuttle starts at the station. The number of customers at the station (n_s), between each consecutive visit to the station, does not depend on n_n. Similarly, for the successive visits to the neighborhood, the update of n_n does not depend on n_s. Therefore, the DTMC that describes both locations can be decomposed into two independent DTMCs, each tracking one location. Each independent DTMC is modeled as a deterministic batch-service scheduled queue with Poisson arrivals. The transi-

tion dynamics are given by

$$\mathbb{P}(x_{t+1}|x_t) = \begin{cases} Pois(x_{t+1}) \text{ if } x_t \leq C, \\ Pois(x_{t+1} - (x_t - C)) \text{ if } x_t > C \end{cases}$$

where $Pois(x)$ is the pmf of a Poisson random variable with mean $\lambda_s\tau$ or $\lambda_n\tau$ for the station and neighborhood respectively.

The key property that allows this independence is the constant cycle time. Intuitively, if the shuttle always returns in exactly τ time units, regardless of the state of the other location, it operates independently of the other location's dynamics.

3.3.4 Stability and Performance Metrics

We are interested in getting steady-state performance metrics (e.g. mean waiting time) for the DTMCs. Therefore, we begin with the conditions for which a stationary distribution exists. Specifically, the maximum service rate must exceed the arrival rate of customers. Therefore, $C/\tau > \lambda_s$ and $C/\tau > \lambda_n$.

Proposition 1. *Consider a queue with Poisson arrivals (rate λ) and batch scheduled service (up to C jobs every τ time units). If $\lambda < C/\tau$, then the stationary distribution of queue length immediately before service, π say, exists, is unique, and the mean customer wait time (not counting the time riding in the shuttle) is $\tau/2 + \mathbb{E}_{X\sim\pi}(X - C)^+/\lambda$.*

Proof. Proof of Proposition 1 The Markov chain is readily seen to be irreducible and aperiodic. Moreover, positive recurrence can be verified using a Lyapunov function argument with the Lyapunov function $V(x) = x$. Thus, the stationary

distribution π exists and is unique. We assume customers are served in first-in-first-out order and the first shuttle services the location at time 0, beginning cycle 1. Let N_k be the number of arrivals in the k^{th} cycle where $N_k \sim Poisson(\lambda\tau)$ are iid for $k \geq 1$. Then $M_k = N_1 + \cdots + N_k$ is the number of arrivals by the end of the k^{th} cycle, where we assume for simplicity that there are no customers waiting for the shuttle at time 0. (This is easily relaxed.) Let X_k be the number of customers present immediately before the start of cycle $k+1$, for $k \geq 0$. Then, $X_0 = 0$ is the initial number of customers and the recursive relationship is $X_{k+1} = (X_k - C)^+ + N_k$. Let $\tau_k = (k-1)\tau$ be the start time of the k^{th} cycle. Let W_j be the waiting time of the j^{th} customer, from the time they arrive until they depart on the shuttle. Let $A_{k,i}$ be the arrival time of the i^{th} customer in cycle k.

Consider the sum of the waiting times of all customers that arrive in the first n cycles. The waiting time for each customer can be decomposed into two quantities: the time until they first see a shuttle arrive and the time spent waiting for shuttles filled with previously arriving customers. For the first quantity, customers are separated into their arrival periods. For the i^{th} arrival in the k^{th} arrival period, the time until the next shuttle is $\tau_{k+1} - A_{k,i}$. For the second quantity, we count the total waiting time for full shuttles by cycle. Specifically, each customer remaining after a shuttle service must wait an additional τ time units until the next shuttle service. Therefore, the total shuttle waiting time can be decomposed by cycle where, after each cycle k, $(X_k - C)^+$ customers must all wait τ additional time units until the next service. Combining these quantities together, the total waiting time for all customers arriving in the first n cycles is

$$\sum_{j=1}^{M_n} W_j = \sum_{k=1}^{n} \left(\left[\sum_{i=1}^{N_k} (\tau_{k+1} - A_{k,i}) \right] + (X_k - C)^+ \tau \right) + R_n,$$

where R_n corresponds to the sum of the residual waiting times after time τ_n of

the remaining customers after the n^{th} cycle. Up to C of these customers will be picked up immediately after the nth cycle so have 0 waiting time. The remaining $\lceil (X_n - C)^+ \rceil$ customers will each have to wait up to $\lceil (X_n - C)^+ / C \rceil$ cycles to be picked up by the shuttle. Thus $0 \le R_n \le (X_n - C)^+ \lceil (X_n - C)^+ / C \rceil \tau$.

Thus, the mean steady-state customer waiting time is

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{M_n} \sum_{j=1}^{M_n} W_j &= \lim_{n\to\infty} \frac{n}{M_n} \left[\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{N_k} (\tau_{k+1} - A_{k,i}) + \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} (X_k - C)^+ \tau + \frac{R_n}{n} \right] \\ &= \frac{1}{\mathbb{E}[N_1]} \left[\mathbb{E}\left[\sum_{i=1}^{N_1} (\tau_2 - A_{1,i}) \right] + \tau \mathbb{E}_\pi (X_0 - C)^+ + \lim_{n\to\infty} \frac{R_n}{n} \right].\end{aligned}$$

By the order statistic property, the first term is

$$\mathbb{E}\left[\sum_{i=1}^{N_1} (\tau_2 - A_{i,1}) \right] = \frac{\tau}{2} \mathbb{E}[N_1]$$

where $\mathbb{E}[N_1] = \lambda\tau$. Also, $R_n/n \to 0$ almost surely if $\mathbb{E}_\pi[X_0^2] < \infty$, since

$$\begin{aligned}\mathbb{E}_\pi[X_0^2] &= \lim_{n\to\infty} \frac{1}{n} \left(X_1^2 + X_2^2 + \cdots + X_n^2 \right) \\ &= \lim_{n\to\infty} \frac{n-1}{n} \frac{1}{n-1} \left(X_1^2 + \cdots + X_{n-1}^2 \right) + \lim_{n\to\infty} \frac{X_n^2}{n} \\ &= \mathbb{E}_\pi[X_0^2] + \lim_{n\to\infty} \frac{X_n^2}{n},\end{aligned}$$

where equality holds throughout almost surely. Therefore, $\lim_{n\to\infty} X_n^2/n = 0$ almost surely and $R_n/n \to 0$ almost surely. Therefore, the mean customer wait is $\tau/2 + \mathbb{E}_\pi(X_0 - C)^+/\lambda$ provided that $E_\pi X_0^2 < \infty$. But this follows from a standard Lyapunov argument with the Lyapunov function $V(x) = x^3$. □

3.3.5 Multiple Shuttles

We now extend the previous model to the setting with multiple shuttles. We assume shuttles take a fixed amount of time, τ, to complete each cycle and all

shuttles are equally spaced in time around this loop. For example, if there are n shuttles in service in the region (notice that we are redefining the notation n to refer to the number of shuttles, rather than the number of cycles as in the previous single-shuttle model), shuttles will pick up new customers from the station every τ/n time units. Under these assumptions, the two DTMCs remain independent, allowing us to use the analysis developed previously. Therefore, our new stability conditions are $Cn/\tau > \lambda_s$ and $Cn/\tau > \lambda_n$. The mean customer wait time becomes $\tau/2n + \mathbb{E}_\pi(X_0 - C)^+/\lambda$.

This expression applies to the station or neighborhood queues only. Let w_s (w_n) be the mean waiting time for the station (neighborhood) queue. Then, the mean customer waiting time for all customers serviced in that region is $(\lambda_s w_s + \lambda_n w_n)/(\lambda_s + \lambda_n)$.

We conjecture that the DTMCs examined have decreasing marginal benefit of adding shuttles. If the conjecture holds, then the mean waiting time is concave in the number of shuttles, which then allows tractable optimization of allocation of shuttles over regions, as we explore next.

Lastly, this model can be extended to consider important features such as latent demand by allowing the arrival rates λ_s and λ_n to be functions of the expected waiting times at each queue.

3.4 Optimizing Fleet Deployment

Now, we consider how to optimally deploy the fleet of shuttles during the day to minimize the expected waiting over all customers. Since public transit sys-

tems have highly time-varying demand, it is essential to determine the optimal fleet deployment over the course of the day. We begin with the simplest setting where there are a fixed number of shuttles available throughout the day. Later, we extend the model to allow for the number of available shuttles to change throughout the day.

3.4.1 Problem Input and Assumptions

Let L be the set of all locations (regions), indexed by $\ell \in L$ and K be the total number of shuttles available. We discretize the day into T time periods indexed by $t \in \{1, 2, \ldots, T\}$. Each time period is modelled as having constant arrival rates, $\lambda^s_{\ell t}$ and $\lambda^n_{\ell t}$ for the station and neighborhood respectively. Let $w_{\ell t}(j)$ be the expected waiting time of a passenger in location ℓ during time period t if there are j shuttles operating, averaged over both neighborhood and station queues. This can be computed from the queuing model introduced in Section 3.3 or through some other methodology, e.g., simulation. From Section 3.3, let b^t_ℓ be the minimum number of vehicles that must be in region ℓ during time period t to ensure stability. Then, let $v^t_{\ell j}$ be the marginal value of adding the j^{th} shuttle to region ℓ during time period t, i.e., the reduction in mean steady-state waiting time (adjusted for arrival rates) from adding the j^{th} shuttle to location ℓ for time period t. More precisely, we define $v^t_{\ell j} = 0$ for all $j < b^t_\ell$, $v^t_{\ell j} = -(\lambda^s_{\ell t} + \lambda^n_{\ell t})(w_{\ell t}(j))$ for $j = b^t_\ell$ and $v^t_{\ell j} = (\lambda^s_{\ell t} + \lambda^n_{\ell t})(w_{\ell t}(j) - w_{\ell t}(j-1))$ for all $j > b^t_\ell$. Since waiting times are decreasing as a function of shuttles, $v^t_{\ell j} < 0$ for $j > b^t_\ell$. The key assumption we require is that the marginal benefit of adding shuttles is non-increasing.

3.4.2 Integer Programming Formulation

We use a path-based formulation on a time-expanded network to optimize the deployment of shuttles throughout the day. Let all potential paths that a shuttle can take during the day be P, indexed by $p \in P$. Specifically, $p = \{(\ell, t)\}$ for all locations ℓ where the shuttle would be at the beginning of time period t. If a shuttle is in the same location at time t and $t+1$, it represents being in service at location ℓ for time period t. A shuttle being in different locations in successive time periods represents the shuttle rebalancing during the day. For notational simplicity, we assume rebalancing always requires 1 time period.

Let y_p be the number of shuttles that take path $p \in P$. Let $x^t_{\ell j}$ be a binary variable that equals 1 if at least j vehicles are assigned to region ℓ in time period t and is otherwise 0. It is defined for all $\ell \in L,\ j \in \{1, \dots, K\},\ t \in \{1, \dots, T\}$. The optimal allocation of shuttles throughout the day can be found by solving

$$\min_{x,y} \quad \sum_{t=1}^{T} \sum_{\ell \in L} \sum_{j=1}^{K} x^t_{\ell j} v^t_{\ell j} \tag{3.1}$$

$$\text{s.t.} \quad x^t_{\ell j} \geq \mathbb{1}\{j \leq b^t_\ell\} \qquad \forall \ell \in L, j \in \{1, \dots, K\}, t \in \{1, \dots, T\} \tag{3.2}$$

$$\sum_{p \in P} y_p = K \tag{3.3}$$

$$\sum_{j=1}^{K} x^t_{\ell j} = \sum_{p \in P} y_p \mathbb{1}\{(\ell, t) \in p\} \mathbb{1}\{(\ell, t+1) \in p\} \qquad \forall t \in \{1, \dots, T\}, \ell \in L \tag{3.4}$$

$$x^t_{\ell j} \in \{0, 1\} \qquad \forall \ell \in L, j \in \{1, \dots, T\}, t \in \{1, \dots, T\} \tag{3.5}$$

$$y_p \in \mathbb{Z}^+ \qquad \forall p \in P \tag{3.6}$$

which we refer to as the IP formulation. We analyze the LP relaxation that corresponds to replacing (3.5) with $x^t_{\ell j} \in [0, 1]$ and (3.6) with $y_p \geq 0$. The objective

function, 3.13, corresponds to minimizing the waiting time across the system. If there are $z_{\ell t}$ shuttles servicing location ℓ during time period t, the average waiting time is

$$\frac{\sum_{\ell \in L} \sum_{t=1}^{T} (\lambda_{\ell t}^{s} + \lambda_{\ell t}^{n})(w_{\ell t}(z_{\ell t}))}{\sum_{\ell \in L} \sum_{t=1}^{T} (\lambda_{\ell t}^{s} + \lambda_{\ell t}^{n})}.$$

Now, consider the objective function. First, $\sum_{j=1}^{k} x_{\ell j}^{t} v_{\ell j}^{t} = (\lambda_{\ell t}^{s} + \lambda_{\ell t}^{n}) \sum_{j=1}^{k} v_{\ell j}^{t} = (\lambda_{\ell t}^{s} + \lambda_{\ell t}^{n}) w_{\ell t}(z_{\ell t})$. Therefore, the objective function corresponds to the numerator of the mean waiting time expression. Since the denominator is constant, it optimizes the same quantity. Constraint (3.2) ensures that the minimum shuttles in service requirement is met for all locations and time periods. Constraint (3.3) constrains the total number of shuttles. Constraint (3.4) computes the number of shuttles in service at each location ℓ over each time period. Since there are diminishing returns for allocating shuttles, $v_{\ell j}^{t} < v_{\ell m}^{t}$ for all $m > j \geq b_{\ell}^{t}$. Therefore the optimal solution allocates all shuttles to the smallest available j in order to minimize the objective.

The objective function can also be decomposed into

$$\sum_{t=1}^{T} \sum_{\ell \in L} \sum_{j=1}^{K} x_{\ell j}^{t} v_{\ell j}^{t} = \sum_{t=1}^{T} \sum_{\ell \in L} \sum_{j=1}^{b_{\ell}^{t}} x_{\ell j}^{t} v_{\ell j}^{t} + \sum_{t=1}^{T} \sum_{\ell \in L} \sum_{j=b_{\ell}^{t}+1}^{K} x_{\ell j}^{t} v_{\ell j}^{t}$$

where the first sum is a constant due to Constraint (3.2) that forces all the $x_{\ell j}^{t} = 1$ in the sum. The objective function could be replaced with just the second sum, corresponding to only focusing on the impact of additional shuttles allocated throughout the day.

3.4.3 Min Cost Circulation Reduction

The integer programming model can be solved efficiently by reformulating it as a min cost circulation problem on a time-expanded network. Let $G = (N, E)$ be a graph with three types of nodes: source (s), sink (s'), and (location, time) nodes (ℓ, t) for all $\ell \in L$ and $t \in \{1, \ldots, T+1\}$. The graph has five types of edges:

1. Allocation arcs ($\mathcal{E}_1 = \{A(\ell) | \ell \in L\}$) correspond to the initial deployment of shuttles where the edge $A(\ell) = (s, (\ell, 1))$. The lower bound is $\underline{u}(e) = 0$, the upper bound is $\overline{u}(e) = \infty$, and the objective contribution is $o(e) = 0$ for all $e \in \mathcal{E}_1$. The amount of flow on an arc is denoted by $f_a(\ell)$.
2. Collection arcs ($\mathcal{E}_2 = \{C(\ell) | \ell \in L\}$) correspond to the final position of shuttles where the edge $C(\ell) = ((\ell, T+1), s')$. The lower bound is $\underline{u}(e) = 0$, the upper bound is $\overline{u}(e) = \infty$, and the objective contribution is $o(e) = 0$ for all $e \in \mathcal{E}_2$. The amount of flow on an arc is denoted by $f_c(\ell)$
3. Circulation arc ($\mathcal{E}_3 = \{(s', s)\}$) corresponds to an arc that connects the sink to the source transforming this into a circulation problem. The lower bound is $\underline{u}(e) = K$, the upper bound is $\overline{u}(e) = K$ and the objective contribution is $o(e) = 0$ for $e = (s', s)$.
4. Service arcs ($\mathcal{E}_4 = \{S(\ell, t, j) | \ell \in L, t \in \{1, \ldots, T\}, j \in \{1, \ldots, K\}\}$) correspond to a shuttle being assigned to location ℓ for time period t where $S(\ell, t, j) = ((\ell, t), (\ell, t+1))$. Note, there are multiple arcs connecting the same nodes and they are indexed by $j = \{1, \ldots, K\}$. The lower bound is $\underline{u}(e) = \mathbb{1}\{e_i \leq b_\ell^t\}$, the upper bound is $\overline{u}(e) = 1$, and the objective contribution is $o(e) = v_{\ell j}^t$. The amount of flow on an arc is denoted by $f_s(\ell, t, j)$.
5. Repositioning arcs ($\mathcal{E}_5 = \{R(\ell_1, \ell_2, t) | \ell_1, \ell_2 \in L, \ell_1 \neq \ell_2, t \in \{1, \ldots, T\}\}$)

correspond to a shuttle relocating from one location to another where $R(\ell_1, \ell_2, t) = ((\ell_1, t), (\ell_2, t+1))$. The lower bound is $\underline{u}(e) = 0$, the upper bound is $\overline{u}(e) = \infty$ and the objective contribution is $o(e) = 0$. The amount of flow on an arc is denoted by $f_r(\ell_1, \ell_2, t)$.

Then, the set of all edges is $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3 \cup \mathcal{E}_4 \cup \mathcal{E}_5$. Let $f(e)$ be the amount of flow on edge $e \in \mathcal{E}$. The min cost circulation formulation is

$$\min_f \quad \sum_{e \in E} o(e) f(e) \tag{3.7}$$

$$\text{s.t.} \quad \sum_{e \in \delta^-(i)} f(e) = \sum_{e \in \delta^+(i)} f(e) \qquad \forall i \in N \tag{3.8}$$

$$\underline{u}_e \leq f(e) \leq \overline{u}_e \qquad \forall e \in \mathcal{E} \tag{3.9}$$

where $\delta^+(i)(\delta^+(i))$ is the set of all arcs whose origin (endpoint) is the node i.

Since only $\mathcal{E}_4$ contributes to the objective, we can rewrite the optimization problem as

$$\min_f \quad \sum_{t=1}^{T} \sum_{j=1}^{k} \sum_{\ell \in L} f_s(\ell, t, j) v^t_{\ell j} \tag{3.10}$$

$$\text{s.t.} \quad \sum_{e \in \delta^-(i)} f(e) = \sum_{e \in \delta^+(i)} f(e) \qquad \forall i \in N \tag{3.11}$$

$$\underline{u}_e \leq f(e) \leq \overline{u}_e \qquad \forall e \in \mathcal{E} \tag{3.12}$$

which we refer to as the flow LP.

Proposition 2. *There exists an integer optimal solution to the LP relaxation of the integer programming formulation.*

Proof. Proof of Proposition 2 Our proof consists of 3 claims. First, we show that the flow LP has an integral optimal solution. Second, we show that each feasible

solution to the LP relaxation corresponds to a feasible solution to the flow LP with the same objective value. Last, we show that each integral flow solution corresponds to an integral feasible solution to the LP relaxation with the same objective value.

Claim 1: The flow LP has an integral optimal solution.

Proof: The problem is a min cost circulation problem with integral upper and lower bounds for all arcs. Therefore, by property 5.5 from Williamson [2019], the optimal flow is integral.

Claim 2: Each feasible solution to the LP relaxation corresponds to a feasible solution to the flow LP with the same objective value

Proof: Given an LP solution (x, y), consider the following flow solution,

$$\begin{aligned} f_a(\ell) &= \sum_{p \in P} y_p \mathbb{1}\{(\ell, 1) \in p\} \\ f_c(\ell) &= \sum_{p \in P} y_p \mathbb{1}\{(\ell, T+1) \in p\} \\ f_s(\ell, t, j) &= x^t_{\ell j} \\ f_r(\ell_1, \ell_2, t) &= \sum_{p \in P} y_p \mathbb{1}\{(\ell_1, t) \in p\} \mathbb{1}\{(\ell_2, t+1) \in p\} \end{aligned}$$

where for relocation, $\ell_1 \neq \ell_2$. The objective values are the same by construction and the upper and lower bounds are automatically satisfied by construction. Therefore, we only need to show that the flow variables satisfy conservation of flow. We begin with the source node, where because the circulation arc is the only arc ending at the source, the flow entering the source is K. Therefore, starting from constraint 3.3,

$$K = \sum_{p \in P} y_p = \sum_{p \in P} \sum_{\ell \in L} y_p \mathbb{1}\{(\ell, 1) \in p\} = \sum_{\ell \in L} \sum_{p \in P} y_p \mathbb{1}\{(\ell, 1) \in p\} = \sum_{\ell} f_a(\ell).$$

Because allocation arcs are the only arcs leaving the source node, the total flow out is $\sum_\ell f_a(\ell) = K$ and conservation of flow is satisfied at the source. Similarly, for the sink node, the only arc leaving the node is the circulation arc with flow K. Therefore,

$$K = \sum_{p \in P} y_p = \sum_{p \in P} \sum_{\ell \in L} y_p \mathbb{1}\{(\ell, T+1) \in p\} = \sum_{\ell \in L} \sum_{p \in P} y_p \mathbb{1}\{(\ell, T+1) \in p\} = \sum_{\ell} f_c(\ell).$$

Because the collection arcs are the only arcs ending at the sink, conservation of flow is satisfied at the sink. For all other nodes, (ℓ, t), the conservation of flow is automatically satisfied since the flow is a superposition of paths which satisfy conservation of flow.

Claim 3: Each integral flow solution corresponds to an integral feasible solution to the LP relaxation with the same objective value.

Proof: Run a standard flow decomposition on the integral flow solution to get an integral y for the LP relaxation, satisfying (3.6). Since there are K units of flow in the solution, Constraint (3.3) is satisfied. To get an integral x, let $x^t_{\ell j} = 1$ if at least j vehicles are in service at location ℓ during time period t according to y and otherwise 0. This definition of x satisfies Constraint (3.4). The lower bounds on the service arcs ensures that at least b^t_ℓ vehicles are in service as necessary (Constraint (3.2)). The monotonicity of the objective terms ($v^t_{\ell j}$) ensures that the appropriate $x^t_{\ell j}$ terms are set to 1. Therefore, the LP relaxation solution has the same objective value as the integral flow solution. □

3.4.4 Extending to time-varying shuttle availability

The formulation introduced previously assumes a constant number of shuttles available during the service period. However, transit systems typically adjust the number of vehicles in service in response to anticipated demand. In this subsection, we extend the previous formulation to allow for a varying number of shuttles available for service.

IP Formulation

We extend the formulation by introducing a hub 'location' h that removes a minimum number of shuttles from service. The T discrete time periods are partitioned into $T_s \leq T$ service periods, each with K_r shuttles available, where $r \in \{1, \ldots, T_s\}$. Let service period r run from time period $s(r)$ to $e(r)$ inclusive. Then, we can define a new set of plausible paths. If a shuttle is not active during service period r, then it is neither in service nor relocating during $s(r)$ to $e(r)$ inclusive. For defining a path, $p \in P$, a shuttle is not active during service period r if $(h, r) \in p$ and by construction, $(\ell, t) \notin p$ for all $\ell \in L$ and $t \in \{s(r), \ldots, e(r)\}$. The

revised IP formulation is

$$\min_{x,y} \quad \sum_{t=1}^{T} \sum_{\ell \in L} \sum_{j=1}^{K} x_{\ell j}^{t} v_{\ell j}^{t} \tag{3.13}$$

$$\text{s.t.} \quad x_{\ell j}^{t} \geq \mathbb{1}\{j \leq b_{\ell}^{t}\} \qquad \forall \ell \in L, j \in \{1, \ldots, K\}, t \in \{1, \ldots, T\} \tag{3.14}$$

$$\sum_{p \in P} y_p = K \tag{3.15}$$

$$\sum_{p \in P} y_p (1 - \mathbb{1}\{(h, r) \in p\}) \leq K_r \qquad \forall r \in \{1, \ldots, T_s\} \tag{3.16}$$

$$\sum_{j=1}^{K} x_{\ell j}^{t} = \sum_{p \in P} y_p \mathbb{1}\{(\ell, t) \in p\} \mathbb{1}\{(\ell, t+1) \in p\} \qquad \forall t \in \{1, \ldots, T\}, \ell \in L \tag{3.17}$$

$$x_{\ell j}^{t} \in \{0, 1\} \qquad \forall \ell \in L, j \in \{1, \ldots, T\}, t \in \{1, \ldots, T\} \tag{3.18}$$

$$y_p \in \mathbb{Z}^{+} \qquad \forall p \in P \tag{3.19}$$

where we have added constraint 3.16 which ensures that at most K_r shuttles are in service during service period r.

Network Flow Reduction

The new IP Formulation can be solved as a network flow problem. We add a hub node, (h, r), for each service period $r \in \{1, \ldots, T_s\}$. The additional edges are detailed below. Let $\beta(t)$ be the service period number corresponding to time period t.

1. We add a single arc from the source node, s, to the first hub node, $(h, 1)$, representing shuttles that are not used at the beginning of the day. For this arc e, the lower bound is $\underline{u}(e) = 0$, the upper bound is $\overline{u}(e) = \infty$, and the objective contribution is $o(e) = 0$. The flow on this arc is denoted by $f_a(h)$.

2. Similarly, we add a single arc from the last hub node, (h, T_s), to the sink, s', representing shuttles not used during the last service period. The arc, e, has a lower bound $\underline{u}(e) = 0$, upper bound $\overline{u}(e) = \infty$, and objective contribution $o(e) = 0$. The flow on this arc is denoted by $f_c(h)$.

3. Into hub arcs ($\mathcal{E}_5 = \{H_i(\ell, t) | \ell \in L, t \in \{1, \ldots, e(T_s - 1)\}\}$) correspond to an arc from location ℓ and time period t to the hub location for the next service period where edge $H_i(\ell, t) = ((\ell, t), (h, \beta(t)+1))$. The lower bound is $\underline{u}(e) = 0$, the upper bound is $\overline{u}(e) = \infty$, and the objective contribution is $o(e) = 0$ for all $e \in \mathcal{E}_5$. The flow on these arcs is denoted by $f_i(\ell, t)$.

4. Out of hub arcs ($\mathcal{E}_6 = \{H_e(\ell, t) | \ell \in L, t \in \{s(2), \ldots, T\}\}$) correspond to a shuttle rejoining service from the hub node where edge $H_e(\ell, t) = ((h, \beta(t) - 1), (\ell, t))$. Vehicles that are not used during the first service period must join service starting at $t \geq s(2)$. The lower bound is $\underline{u}(e) = 0$, the upper bound is $\overline{u}(e) = \infty$, and the objective contribution is $o(e) = 0$ for all $e \in \mathcal{E}_6$. The flow on these arcs is denoted by $f_o(\ell, t)$.

5. Hub to hub arcs ($\mathcal{E}_7 = \{H_s(r) | r \in \{1, 2, \ldots, T_s - 1\}\}$) correspond to a shuttle remaining out of service for service periods j and $j + 1$ where edge $H_s(r) = ((h, r), (h, r + 1))$. The lower bound is $\underline{u}(e) = 0$, the upper bound is $\overline{u}(e) = \infty$, and the objective contribution is $o(e) = 0$ for all $e \in \mathcal{E}_7$. The flow on these arcs is denoted by $f_h(r)$.

We can extend the previous network flow formulation

$$\min_{f} \quad \sum_{t=1}^{T} \sum_{j=1}^{K} \sum_{\ell \in L} f_s(\ell, t, j) v_{\ell j}^t \tag{3.20}$$

$$\text{s.t.} \quad \sum_{e \in \delta^-((h,r))} f(e) \geq K - K_r \qquad \forall r \in \{1, \ldots, T_s\} \tag{3.21}$$

$$\sum_{e \in \delta^-(i)} f(e) = \sum_{e \in \delta^+(i)} f(e) \qquad \forall i \in N \setminus \{s, s'\} \tag{3.22}$$

$$\underline{u}_e \leq f(e) \leq \overline{u}_e \qquad \forall e \in \mathcal{E} \tag{3.23}$$

by adding constraint 3.21 which ensures that at least $K - K_r$ vehicles enter (h, r) and are out of service for service period r. Lastly, we want to show that the LP relaxation still has integer optimal solutions.

Proposition 3. *There exists an integer optimal solution to the LP relaxation of the modified integer programming formulation.*

Proof. Proof of Proposition 3 Our proof follows the same three steps as Proposition 2.

Claim 1: The flow LP has an integral optimal solution.

Proof: The flow LP still has an integer optimal solution since the new constraint (3.21) can be implemented via node splitting.

Claim 2: Each feasible solution to the LP relaxation corresponds to a feasible solution to the flow LP with the same objective value.

Proof: Given an LP solution (x, y), consider the following flow solution

where f_a, f_c, f_s, and f_r are defined the same as previously. Let

$$f_i(\ell, t) = \sum_{p \in P} y_p \mathbb{1}\{(\ell, t) \in p\} \mathbb{1}\{(h, \beta(t) + 1) \in p\}$$

$$f_o(\ell, t) = \sum_{p \in P} y_p \mathbb{1}\{(h, \beta(t) - 1 \in p\} \mathbb{1}\{(\ell, t) \in p\}$$

$$f_h(r) = \sum_{p \in P} y_p \mathbb{1}\{(h, r) \in p\} \mathbb{1}\{(h, r + 1) \in p\}.$$

Just like the previous proof, the objective values are the same by construction and the upper and lower bounds are also automatically satisfied by construction. Conservation of flow is automatically satisfied for all nodes that are not the source or sink since the flow is a superposition of paths which individually obey conservation of flow. For the source and sink nodes, the previous proof can easily be adapted by considering $f_a(h)$ and $f_c(h)$ respectively. Lastly, we must show that constraint 3.21 is satisfied. By combining constraints 3.15 and 3.16, $\sum_{p \in P} y_p \mathbb{1}\{(h, r) \in p\} \geq K - K_r$, implying that the total flow entering hub node r is at least $K - K_r$, as desired.

Claim 3: Each integral flow solution corresponds to an integral feasible solution to the LP relaxation with the same objective value.

Proof: The proof is the same as in Proposition 2. □

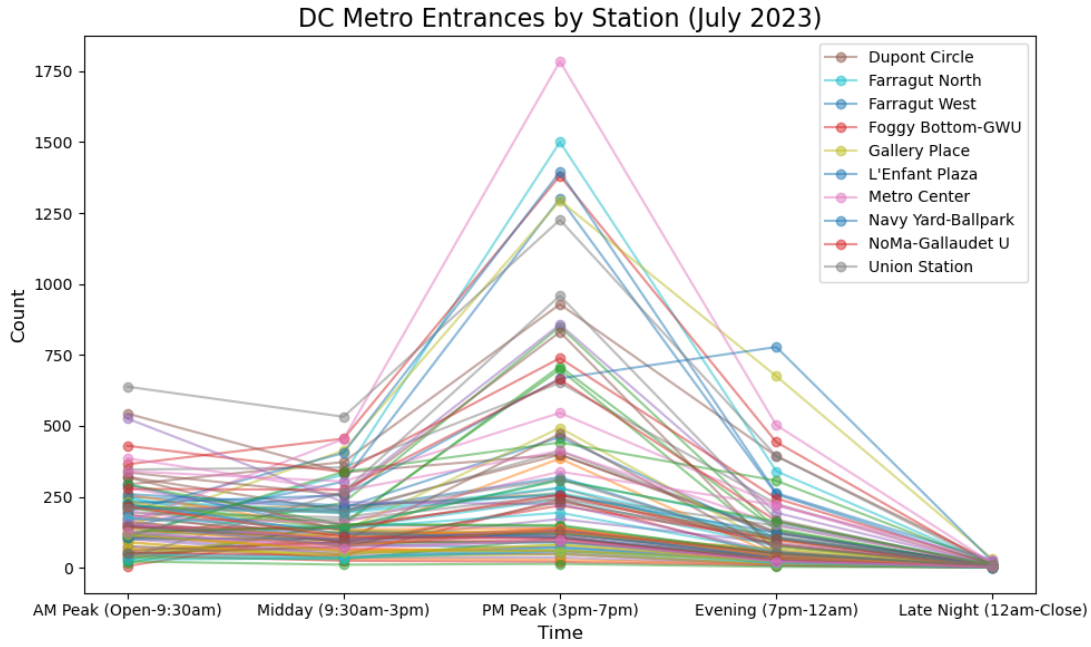
Therefore, we have shown that we can extend the model to consider varying fleet sizes while preserving the integrality of the LP relaxation. Lastly, this optimization model can be extended to consider features such as equity requirements for regions. This can be accomplished by changing how the minimum number of shuttles in service is calculated to ensure that performance targets are met.

3.5 Case Study

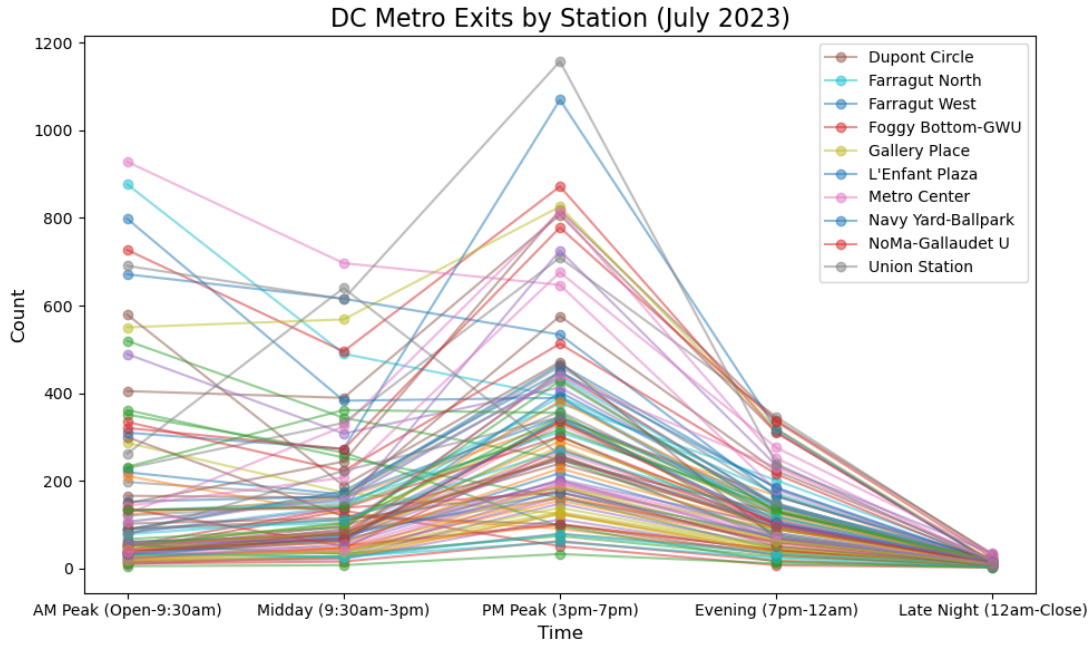
We use Washington DC as a case study to show how the model can be used to evaluate the potential of an ODMTS. This subsection first describes how the input parameters for the model are derived, followed by presenting the model results.

Our model uses all all 98 subway stations in the Washington Metropolitan Area Transit Authority (WMATA) as transit hubs that users are trying to get to and from. Each station corresponds to its own region and later we describe how the service areas for each station are found. This case study only consider data from non-holiday weekdays in July 2023. The optimization problem is solved over the time horizon of one day, from 5 am - midnight, discretized into 38 half-hour periods. We assume each shuttle has a capacity of 6.

We begin by estimating the arrival rate of customers. WMATA data provides entry and exit counts for each station across 5 time windows: AM Peak (Open-9:30am), Midday (9:30am-3pm), PM Peak (3pm-7pm), Evening (7pm-12am), and Late Night (12am-Close) [Authority, 2024b]. We ignores the last time window because it only applies to Friday evenings. A visualization of the entrance and exit counts by station is shown in Figure 3.1 We assume that 20% of all entering (exiting) customers will use a shuttle to get to the station (their destination). Chicago bike-sharing data is then used to determine the distribution of arrivals within each of the 4 time windows [Bikeshare, 2024]. We assume that the arrival rate of customers is proportional to the rate at which bike-sharing trips begin. The temporal distribution of bike-sharing trips is shown in Figure 3.2.



(a) Average daily entrances for each station across all 5 time windows.



(b) Average daily exits for each station across all 5 time windows.

Figure 3.1: The average daily entrances and exits for each station over the 5 time windows.

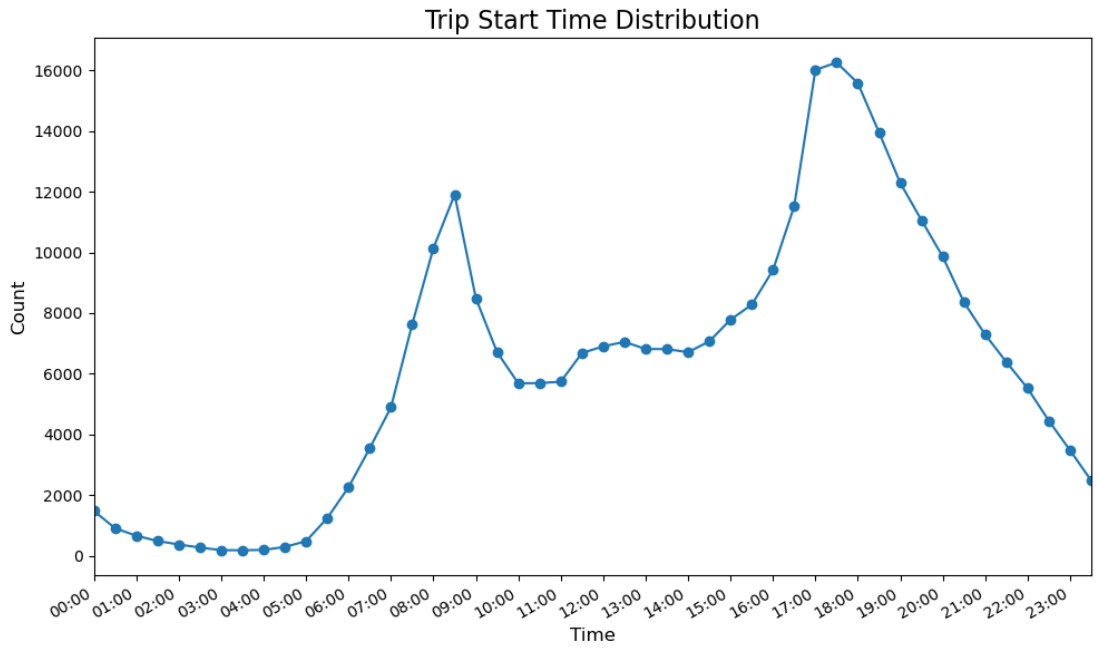


Figure 3.2: The distribution of trip start times for Chicago Bikeshare in July 2023.

Now, we describe how the cycle length is calculated. We begin by defining the catchment area around each station using a Voronoi diagram (shown in Figure 3.3). This assigns every point in the plane to its closest station by Euclidean distance. The locations for each station were found via the WMATA developer portal. Using this diagram, we find the catchment area for all enclosed regions. All unbounded regions are assumed to have a catchment area of 81 km^2. We assume the cycle length is twice the time required for a TSP tour in the region, using TSP tour constant $\beta = 0.72$ and the average speed of a WMATA bus, 9.8 mph [Nguyen, 2024]. These estimates are pessimistic for a couple reasons. First, shuttles would have to stop less frequently, and therefore should travel faster. Second, doubling the TSP tour length overestimates the expected distance the shuttle must travel.

For simplicity, we will assume that there are a constant number of shuttles available throughout the day. Now that we have finished specifying the inputs

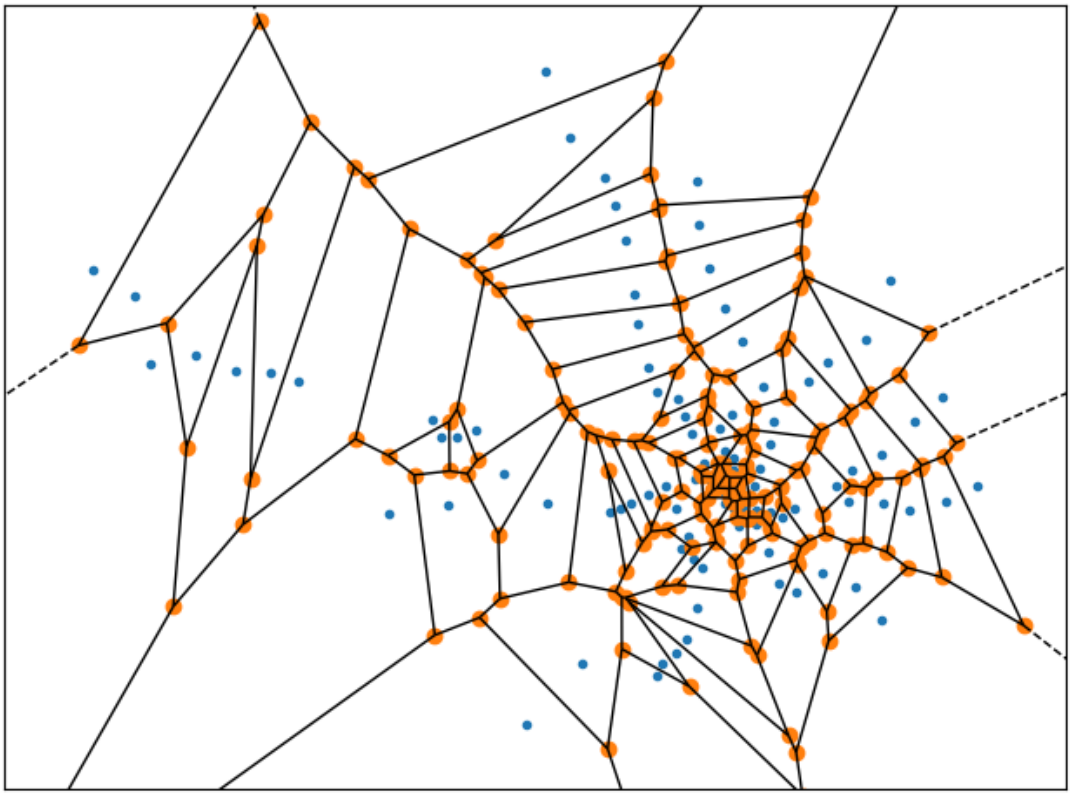


Figure 3.3: The service regions for each subway station, drawn as a Voronoi diagram. Each blue point denotes a station and each orange point represents an intersection of the partitioning lines.

to the model, we can solve the model and analyze the results.

The minimum number of shuttles required (to meet the stability requirements) is approximately 1375. This is a significant number of shuttles, approaching the current size of WMATA's bus fleet [Authority, 2024a]. However, this fleet sizing corresponds to the system's busiest time and assumes that all stations and regions are switching to an on-demand model. There may still be a significant benefit to on-demand shuttles by restricting either the times or locations that they are used.

As we increase the number of shuttles in the system, the objective value im-

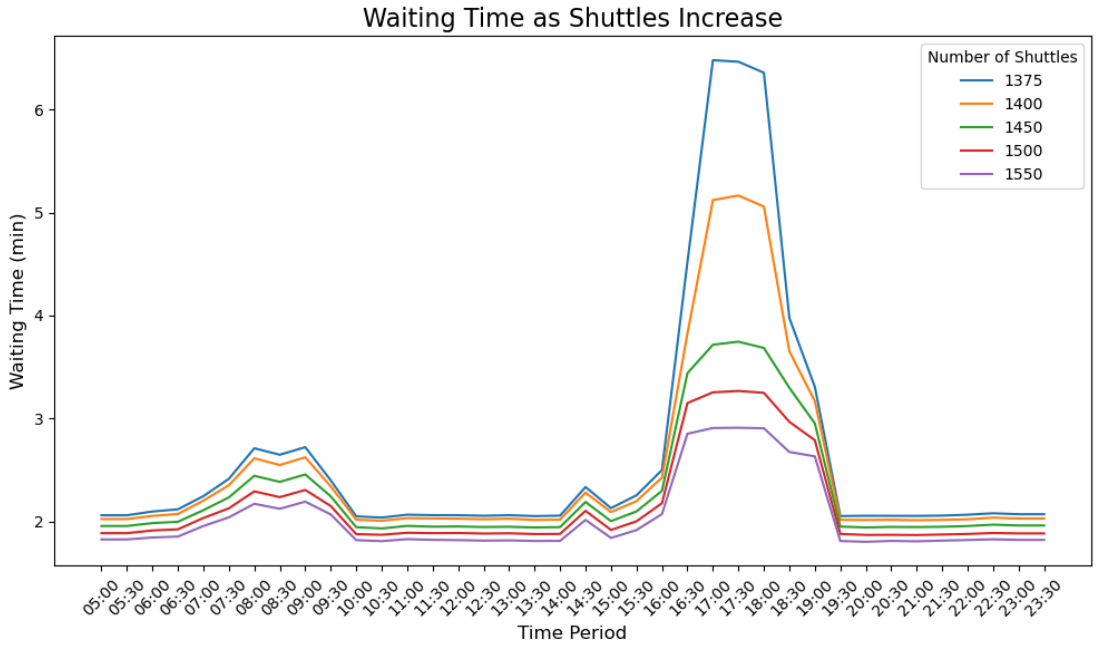
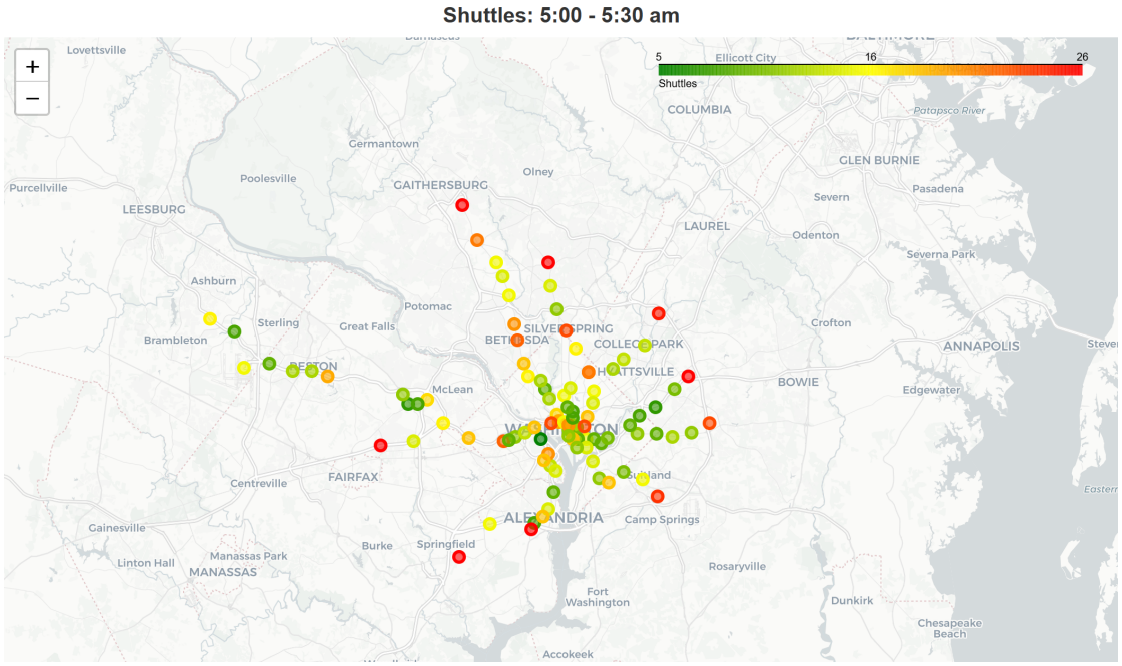


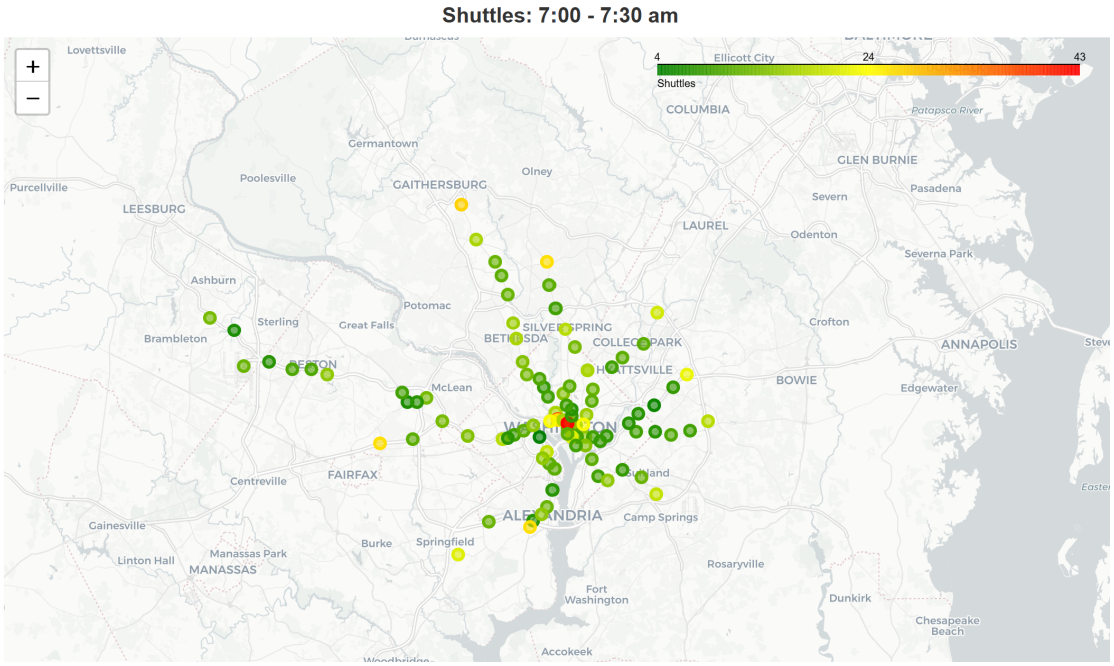
Figure 3.4: The service regions for each subway station, drawn as a Voronoi diagram. Each blue point denotes a station and each orange point represents an intersection of the partitioning lines.

proves monotonically with diminishing returns. We can also examine how the performance metrics change throughout the day as illustrated in Figure 3.4. As expected, increasing the number of shuttles improves system performance. The most challenging time for the system is the evening rush, between 4:30 - 7:00 pm. The demand in this time period drives the significant number of shuttles required. While waiting times during the morning commute increase, it appears to be less intense than the evening rush. This is consistent with Figures 3.1. For a system operator, this gives a very clear signal about the source of the challenges of an ODMTS and can help shape its successful implementation. Lastly, in Figure 3.5, we visualize the deployment of solution with 1450 shuttles throughout the day. The scale is not consistent and varies significantly across figures for different times of day. This figure illustrates that during off-peak hours, the shuttles are distributed across the city. During the morning commute, shuttles

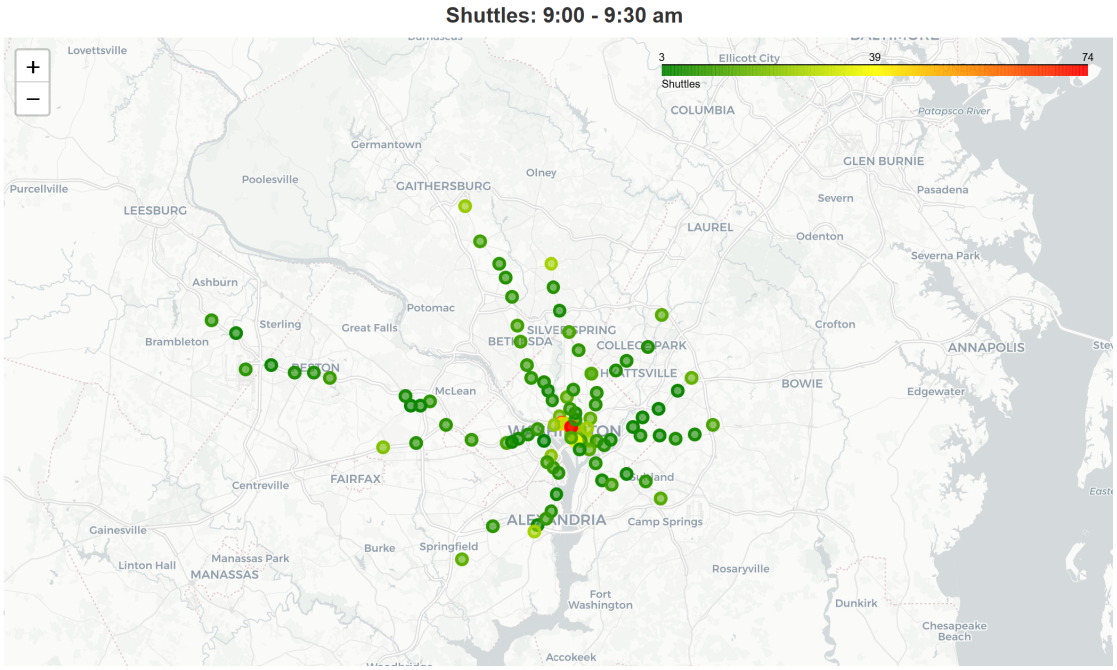
concentrate near downtown and at a handful out endpoint stations while during the evening commute, they mostly concentrate downtown.



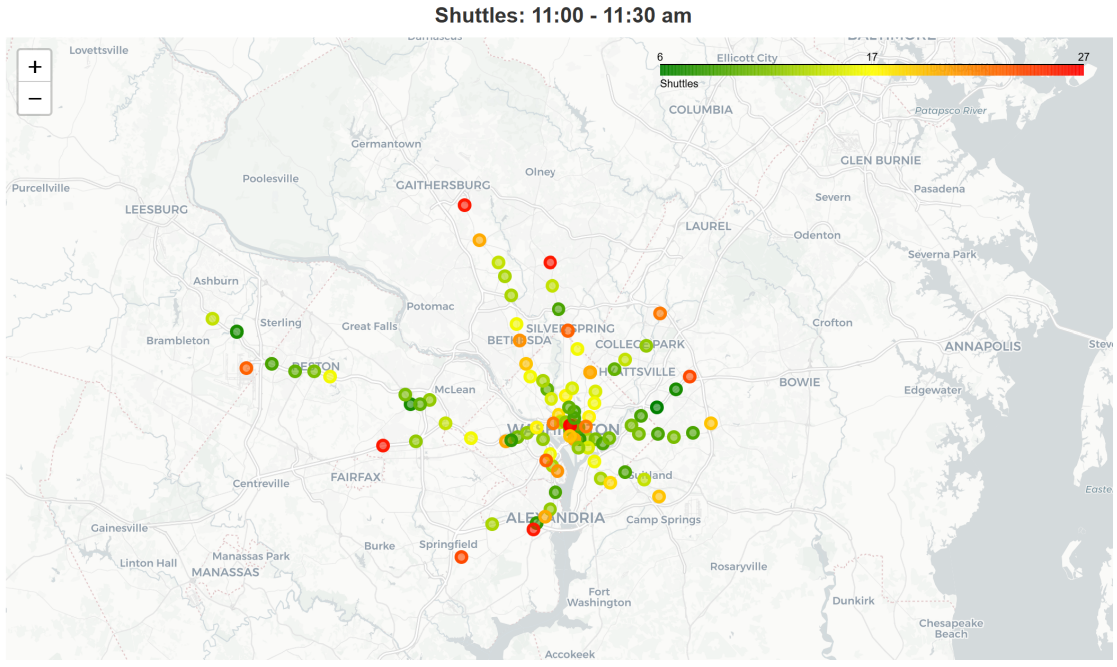
(a) Distribution of shuttles from 5 - 5:30 am.



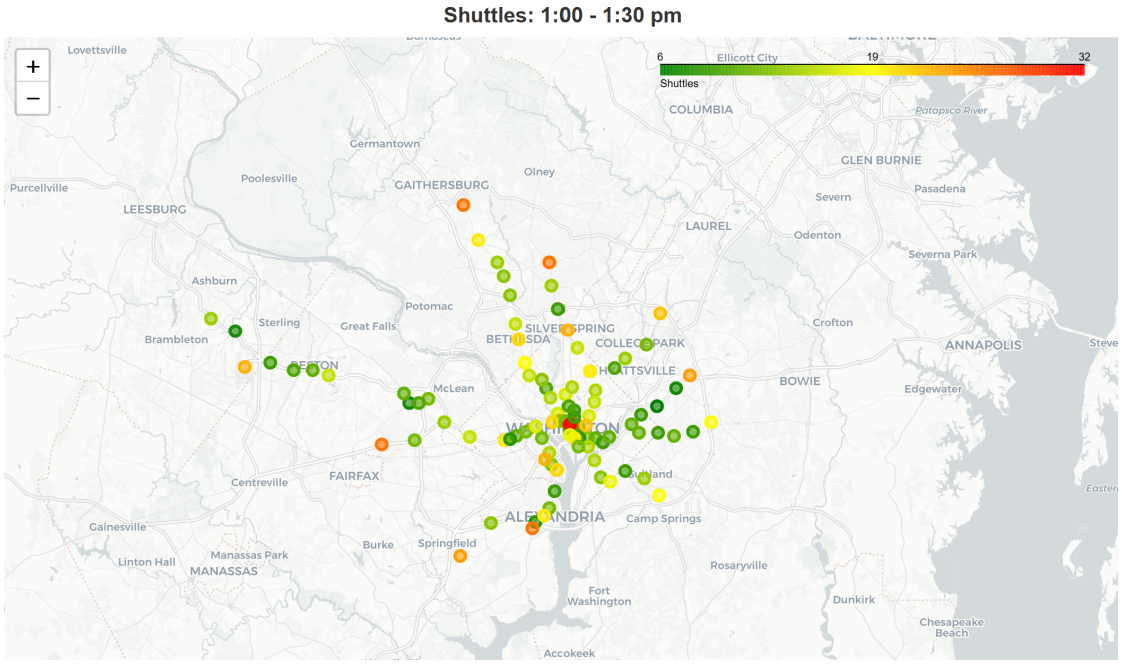
(b) Distribution of shuttles from 7 - 7:30 am.



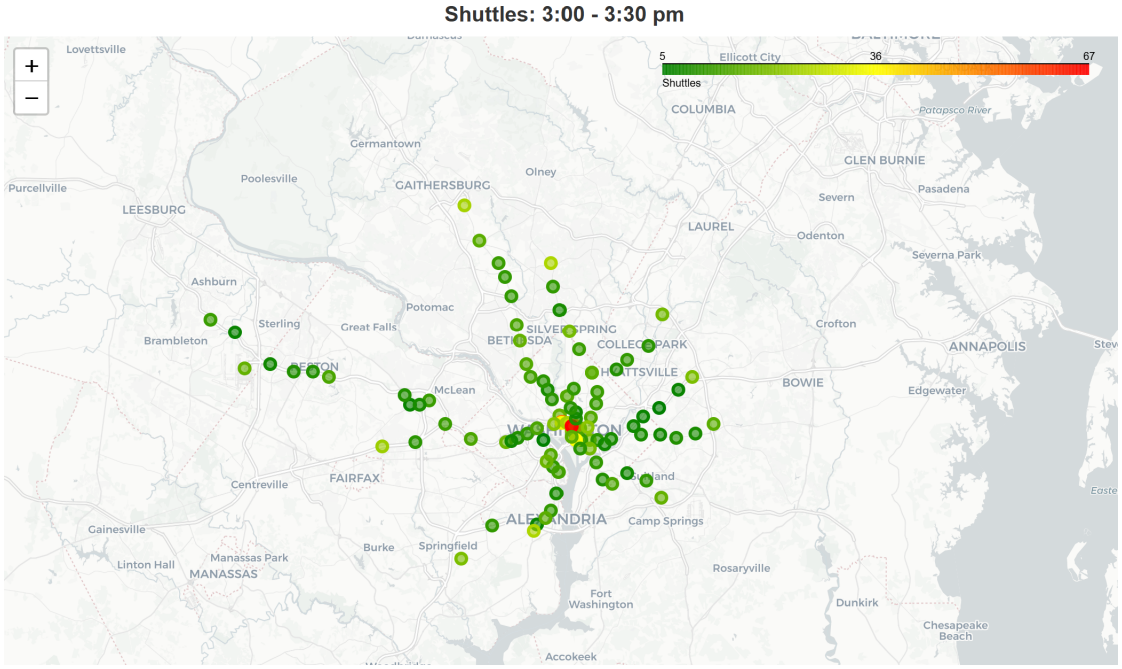
(c) Distribution of shuttles from 9 - 9:30 am.



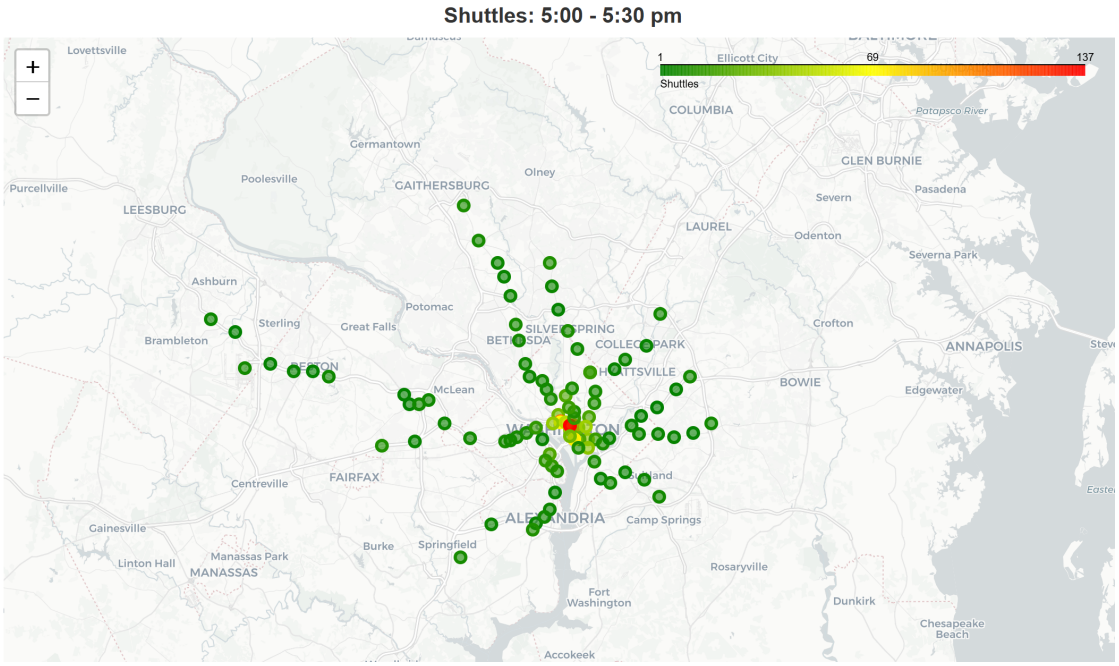
(d) Distribution of shuttles from 11 - 11:30 am.



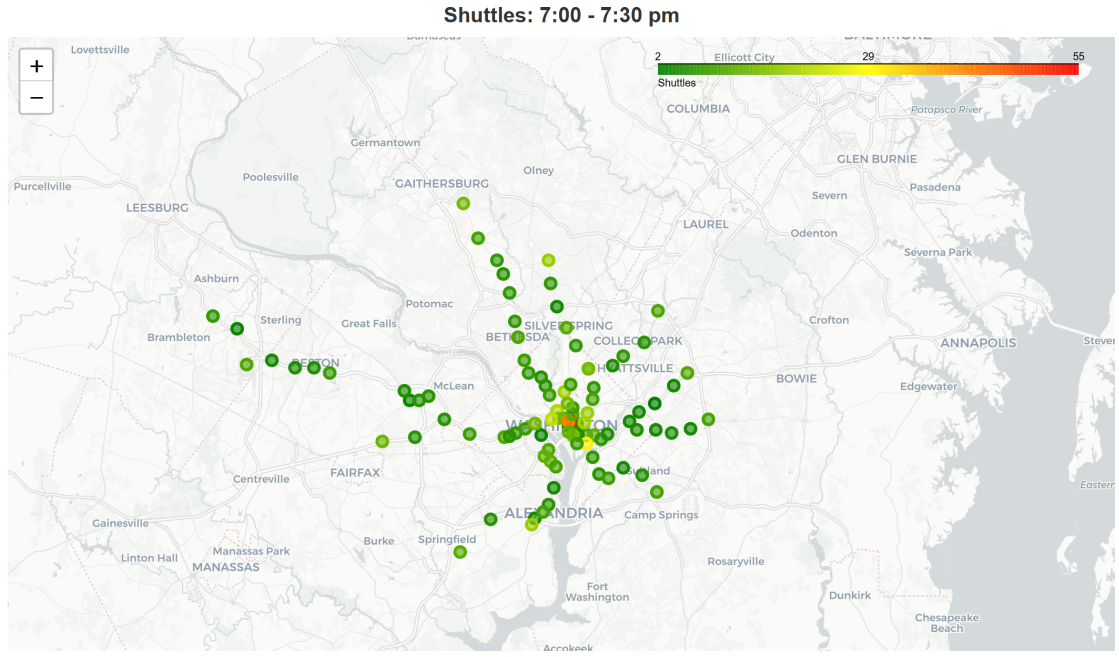
(e) Distribution of shuttles from 1 - 1:30 pm.



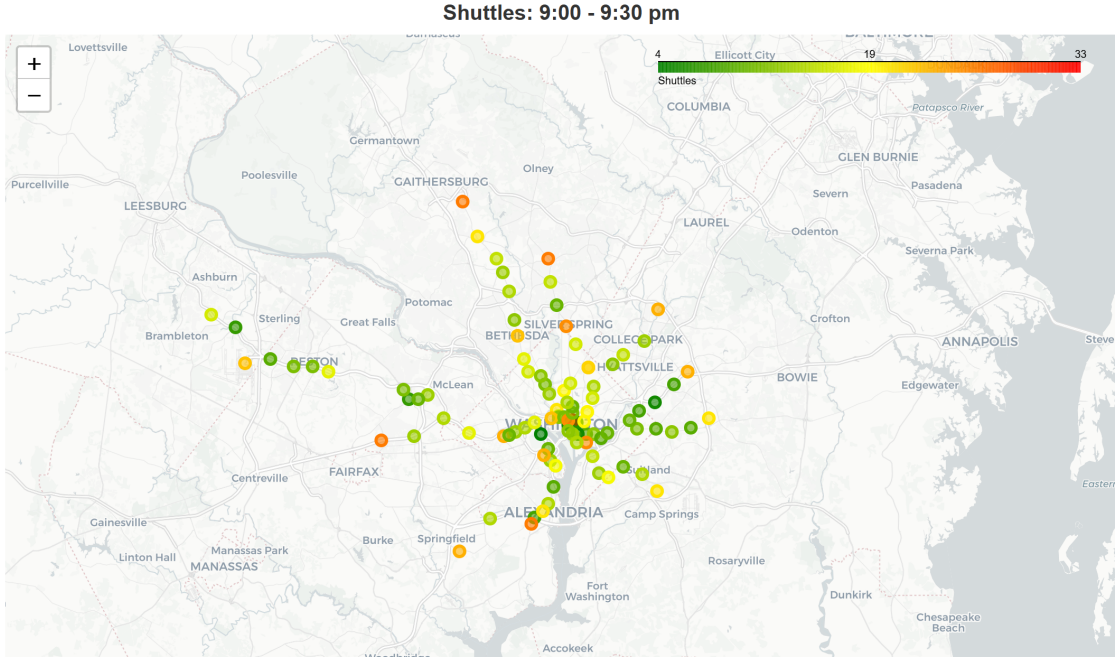
(f) Distribution of shuttles from 3 - 3:30 pm.



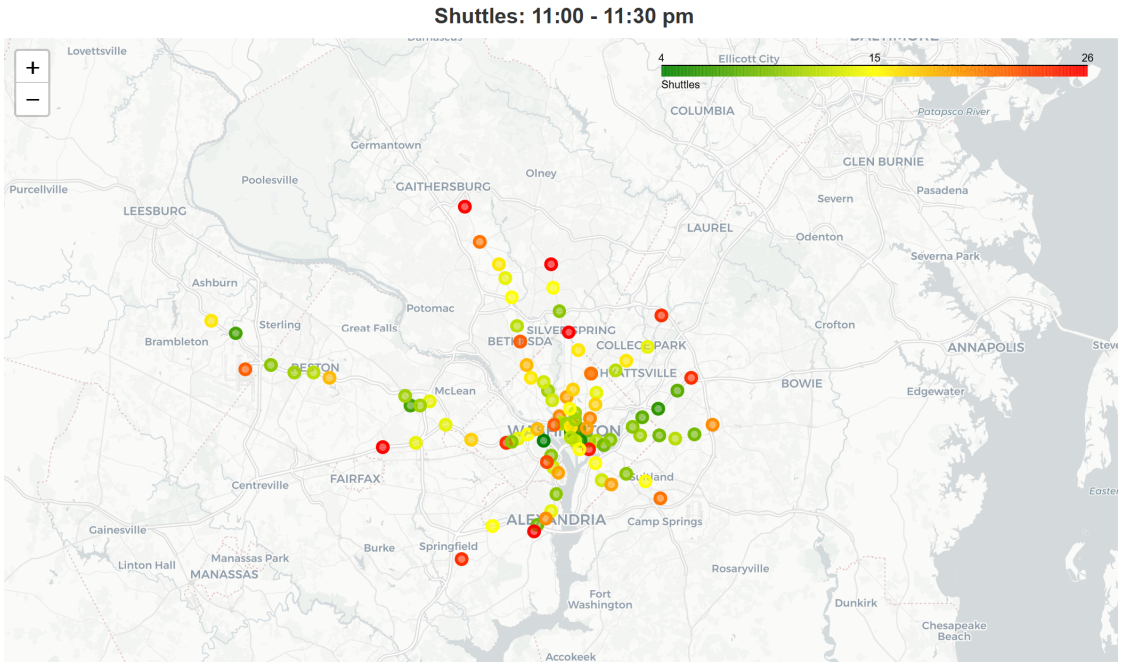
(g) Distribution of shuttles from 5 - 5:30 pm.



(h) Distribution of shuttles from 7 - 7:30 pm.



(i) Distribution of shuttles from 9 - 9:30 pm.



(j) Distribution of shuttles from 11 - 11:30 pm.

Figure 3.5: The distribution of shuttles throughout the day according to solution for 1450 shuttles.

BIBLIOGRAPHY

C. Archetti, M. G. Speranza, and D. Weyland. A simulation study of an on-demand transportation system. *International Transactions in Operational Research*, 25(4):1137–1161, 2018.

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

W. M. A. T. Authority. Bus, bus facilities and paratransit investments, 2024a. URL https://www.wmata.com/initiatives/capital-improvement-program/investments/Bus-and-Paratransit-Investments.cfm.

W. M. A. T. Authority. Daily ridership dashboards (bus and rail), 2024b. URL https://www.wmata.com/initiatives/ridership-portal/daily-summary.cfm.

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

S. Banerjee, C. Hssaine, Q. Luo, and S. Samaranayake. Plan your system and price for free: Fast algorithms for multimodal transit operations. *Available at SSRN 3972423*, 2023.

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

F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *Journal of the ACM (JACM)*, 22(2):248–260, 1975.

J. Beardwood, J. H. Halton, and J. M. Hammersley. The shortest path through many points. In *Mathematical proceedings of the Cambridge philosophical society*, volume 55, pages 299–327. Cambridge University Press, 1959.

S. Benjaafar and X. Shen. Pricing in on-demand and one-way vehicle-sharing networks. *Operations Research*, 71(5):1596–1609, 2023.

S. Benjaafar, S. Wu, H. Liu, and E. B. Gunnarsson. Dimensioning on-demand vehicle sharing systems. *Management Science*, 68(2):1218–1232, 2022.

C. Bikeshare. System data, 2024. URL `https://s3.amazonaws.com/capitalbikeshare-data/index.html`.

A. Braverman, J. G. Dai, X. Liu, and L. Ying. Empty-car routing in ridesharing systems. *Operations Research*, 67(5):1437–1452, 2019.

K. Dalmeijer and P. Van Hentenryck. Transfer-expanded graphs for on-demand multimodal transit systems. In *International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pages 167–175. Springer, 2020.

P. J. Denning and J. P. Buzen. The operational analysis of queueing network models. *ACM Comput. Surv.*, 10(3):225–261, sep 1978. ISSN 0360-0300. doi: 10.1145/356733.356735. URL `https://doi.org/10.1145/356733.356735`.

D. K. George and C. H. Xia. Fleet-sizing and service availability for a vehicle

rental system via closed queueing networks. *European journal of operational research*, 211(1):198–207, 2011.

R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley Professional, 1994.

H. Guan, B. Basciftci, and P. Van Hentenryck. Heuristic algorithms for integrating latent demand into the design of large-scale on-demand multimodal transit systems. *arXiv preprint arXiv:2212.03460*, 2022.

L. He, H.-Y. Mak, Y. Rong, and Z.-J. M. Shen. Service region design for urban electric vehicle sharing systems. *Manufacturing & Service Operations Management*, 19(2):309–327, 2017.

K. Heineke, B. Kloss, D. Scurtu, and C. Wiemuth. Electric kickscooters have come of age. regulators have taken notice, 2023. URL https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/electric-kickscooters-have-come-of-age-regulators-have-taken-notice

B. I. Inc. Bounce Infinity home page, 2023a. URL https://bounceinfinity.com/home.html.

G. Inc. Gogoro home page, 2023b. URL https://www.gogoro.com/.

A. Kabra, E. Belavina, and K. Girotra. Bike-share systems: Accessibility and availability. *Management Science*, 66(9):3803–3824, 2020.

R. J. Kuhudzai. Spiro targets 1 million electric motorcycles & 3,000 battery charging & swapping stations in kenya, 2023. URL https://cleantechnica.com/2023/09/11/

spiro-targets-1-million-electric-motorcycles-3000-battery-charging-#:~:text=Here%20is%20a%20summary%20of,Togo%2C%20Rwanda%2C%20Uganda%2C%20Kenya.

A. Maheo, P. Kilby, and P. Van Hentenryck. Benders decomposition for the design of a hub and shuttle public transit system. *Transportation Science*, 53(1): 77–88, 2019.

J. C. Martínez Mori, M. G. Speranza, and S. Samaranayake. On the value of dynamism in transit networks. *Transportation Science*, 57(3):578–593, 2023.

M. Minneapolis. Scooter availability 2021, March 2022a. URL https://opendata.minneapolismn.gov/datasets/cityoflakes::scooter-availability-2021/about.

M. Minneapolis. Motorized foot scooter trips 2021, March 2022b. URL https://opendata.minneapolismn.gov/datasets/cityoflakes::motorized-foot-scooter-trips-2021/about.

D. Nguyen. D.c. will begin fining drivers in bus-only lanes, January 2024.

NYC Mayor Office. Charge safe, ride safe, 2023. URL https://www.nyc.gov/assets/home/downloads/pdf/office-of-the-mayor/2023/micromobility-action-plan.pdf.

S. Oh, R. Seshadri, D.-T. Le, P. C. Zegras, and M. E. Ben-Akiva. Evaluating automated demand responsive transit using microsimulation. *IEEE Access*, 8: 82551–82561, 2020.

E. Özkan and A. R. Ward. Dynamic matching for real-time ride sharing. *Stochastic Systems*, 10(1):29–70, 2020.

N. Périvier, C. Hssaine, S. Samaranayake, and S. Banerjee. Real-time approximate routing for smart transit systems. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 5(2):1–30, 2021.

P. Pollett. Modelling congestion in closed queueing networks. *International Transactions in Operational Research*, 7(4-5):319–330, 2000.

M. Reiser and S. S. Lavenberg. Mean-value analysis of closed multichain queuing networks. *Journal of the ACM (JACM)*, 27(2):313–322, 1980.

D. Schellong, P. Sadek, C. Schaetzberger, and T. Barrack. The promise and pitfalls of e-scooter sharing, 2023. URL `https://www.bcg.com/publications/2019/promise-pitfalls-e-scooter-sharing`.

K. Steiner and S. Irnich. Strategic planning for integrated mobility-on-demand and urban public bus networks. *Transportation Science*, 54(6):1616–1639, 2020.

P. Van Hentenryck, C. Riley, A. Trasatti, H. Guan, T. Santanam, J. A. Huertas, K. Dalmeijer, K. Watkins, J. Drake, and S. Baskin. Marta reach: Piloting an on-demand multimodal transit system in atlanta. *arXiv preprint arXiv:2308.02681*, 2023.

H. Wang and A. Odoni. Approximating the performance of a "last mile" transportation system. *Transportation Science*, 50(2):659–675, 2016.

A. Waserhole and V. Jost. Pricing in vehicle sharing systems: Optimization in queuing networks with product forms. *EURO Journal on Transportation and Logistics*, 5(3):293–320, 2016.

Wikipedia, the free encyclopedia. Communities and neighborhoods of minneapolis, 2024a. URL `https://upload.wikimedia.org/`

wikipedia/commons/c/c3/Communities_and_neighborhoods_of_Minneapolis.svg.

Wikipedia, the free encyclopedia. Calhoun isles, minneapolis, 2024b. URL https://en.wikipedia.org/wiki/Calhoun_Isles,_Minneapolis.

Wikipedia, the free encyclopedia. Camden, minneapolis, 2024c. URL https://en.wikipedia.org/wiki/Camden,_Minneapolis.

Wikipedia, the free encyclopedia. Central, minneapolis, 2024d. URL https://en.wikipedia.org/wiki/Central,_Minneapolis.

Wikipedia, the free encyclopedia. Longfellow, minneapolis, 2024e. URL https://en.wikipedia.org/wiki/Longfellow,_Minneapolis.

Wikipedia, the free encyclopedia. Near north, minneapolis, 2024f. URL https://en.wikipedia.org/wiki/Near_North,_Minneapolis.

Wikipedia, the free encyclopedia. Nokomis, minneapolis, 2024g. URL https://en.wikipedia.org/wiki/Nokomis,_Minneapolis.

Wikipedia, the free encyclopedia. Northeast, minneapolis, 2024h. URL https://en.wikipedia.org/wiki/Northeast,_Minneapolis.

Wikipedia, the free encyclopedia. Phillips, minneapolis, 2024i. URL https://en.wikipedia.org/wiki/Phillips,_Minneapolis.

Wikipedia, the free encyclopedia. Powderhorn, minneapolis, 2024j. URL https://en.wikipedia.org/wiki/Powderhorn,_Minneapolis.

Wikipedia, the free encyclopedia. Southwest, minneapolis, 2024k. URL https://en.wikipedia.org/wiki/Southwest,_Minneapolis.

Wikipedia, the free encyclopedia. University, minneapolis, 2024l. URL https://en.wikipedia.org/wiki/University,_Minneapolis.

D. P. Williamson. *Network flow algorithms*. Cambridge University Press, 2019.

X. Yan, X. Zhao, Y. Han, P. Van Hentenryck, and T. Dillahunt. Mobility-on-demand versus fixed-route transit systems: An evaluation of traveler preferences in low-income communities. *Transportation Research Part A: Policy and Practice*, 148:481–495, 2021.

R. Zhang, F. Rossi, and M. Pavone. Analysis, control, and evaluation of mobility-on-demand systems: A queueing-theoretical approach. *IEEE Transactions on Control of Network Systems*, 6(1):115–126, 2018.