Randomized FIFO Mechanisms*

Francisco Castro† Hongyao Ma‡ Hamid Nazerzadeh§ Chiwei Yan¶

August 15, 2021

Abstract

We study the matching of jobs to workers in a queue, for example a ridesharing platform dispatching drivers to pick up riders at an airport. Under FIFO dispatching, the heterogeneity in earnings from different trips incentivizes drivers to cherrypick, increasing riders' waiting times for a match, and resulting in a loss of efficiency and reliability. We first propose *the direct FIFO mechanism*, which offers lower-earning trips to drivers further from the head of the queue. The option to skip the rest of the line incentivizes drivers to accept all dispatches, and the equilibrium outcome is envy-free and achieves the first best throughput and the second best revenue. To preserve fairness of the dispatch rule in the sense that drivers closer to the head of the queue have higher priority for trips to all destinations, we introduce a family of *randomized FIFO mechanisms*. By sending declined trips gradually down the queue in a randomized manner, a randomized FIFO mechanism achieves the same optimal outcome in equilibrium, while incurring a small variance in driver earnings which diminishes as riders' patience increases. Extensive counterfactual simulations using data from the City of Chicago demonstrate substantial improvements of revenue and throughput compared to the status quo FIFO dispatching, and highlight the effectiveness of using waiting time to reduce the variability of driver earnings.

1 Introduction

Matching marketplaces play an instrumental role in economic exchanges and the allocation of public and private resources. Over the past decade, the rise of online platforms connecting people with gig workers has also radically changed many aspects of our daily lives— ridesharing platforms such as Uber and Lyft provide billions of rides each quarter,[1] and the grocery sales processed by Instacart was projected to exceed 35 billion dollars in 2020.[2]

*The authors would like to thank Nick Arnosti, Omar Besbes, Peter Cohen, Jose Correa, John Dickerson, Sergey Gitlin, Yash Kanoria, Cinar Kilcioglu, Jake Marcinek, Eoin O'Mahony, David Parkes, Scott Rodilitz, Garrett van Ryzin, Lior Seeman, Carmen Wang, Adam Wierman, Zhixi Wan, and participants at Uber Marketplace Matching Science Deep Dive, INFORMS 2020, Simons Institute Online and Matching-Based Market Design Reunion, for helpful comments and discussions.

†UCLA Anderson School of Management. 110 Westwood Plaza, room B505, Los Angeles, CA 90095, USA. Email: francisco.castro@anderson.ucla.edu.

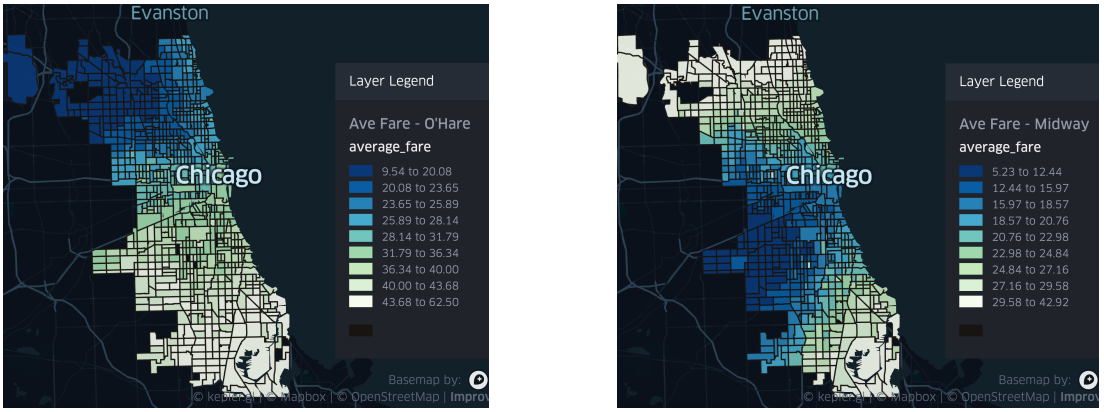
‡Decision, Risk, and Operations Division, Columbia Business School. 423 Uris Hall, 3022 Broadway, New York, NY, 10027, USA. Email: hongyao.ma@columbia.edu.

§USC Marshall School of Business. Bridge Memorial Hall, University of Southern California, Los Angeles, CA 90089 USA. Email: nazerzad@usc.edu.

¶Department of Industrial and Systems Engineering, University of Washington. 3900 E Stevens Way NE, Seattle, WA 98195, USA. Email: chiwei@uw.edu.

[1] `https://www.statista.com/statistics/946298/uber-ridership-worldwide/`, accessed 02/23/2021; `https://www.statista.com/statistics/916456/lyft-number-of-rides/`, accessed 02/23/2021.

[2] `https://www.bloomberg.com/news/features/2020-05-06/instacart-was-overwhelmed-by-coronavirus-overnight`, accessed 02/23/2021.



(a) Trips from O'Hare.

(b) Trips from Midway.

Figure 1: Average trip fare by destination Census Tract in Chicago, for trips originating from the O'Hare International Airport and the Midway International Airport. See Section 5 for more details.

To improve efficiency and reduce waiting times, platforms often aim to match rider or grocery delivery trips with the closest available drivers. When requests are concentrated in space, however, matching by proximity has unintended consequences. As an example, Amazon drivers have been reportedly hanging their smartphones in trees near Amazon delivery stations and Whole Foods stores, in order to appear even closer and gain higher priority for job offers.[3] A similar problem existed for Uber and Lyft at airports and event venues. Matching riders to the closest drivers incentivizes drivers to get as close to the terminal or venue as possible, leading to traffic congestions.[4][5]

Many ridesharing platforms now maintain *virtual queues* at airports for drivers who are waiting in designated areas, and dispatch drivers from the queue in a first-in-first-out (FIFO) manner.[6] This resolves the congestion issues and is also considered more fair by many since drivers who have waited the longest in the queue are now the first in line to receive trip offers. At major U.S. airports, however, a driver at the head of the queue will receive the next trip offer in a few seconds under FIFO dispatching if she declines an offer from the platform (see Figure 12). As we shall see, this lowered cost of cherrypicking substantially exacerbates existing problems on incentive alignment.

Figure 1 shows the average trip fare by destination Census Tract in Chicago, for trips originating from the O'Hare and Midway airports. A short trip from O'Hare to a nearby area pays an average of \$10-\$20, but a long trip can pay an average of \$60. During busy hours, instead of accepting an average trip, drivers who are close enough to the head of the queue are better off declining most trip offers, and waiting for only the highest earning trips instead. Riders, however, have finite patience, despite being willing to wait for some time for a match. When each driver decline takes ten seconds on average, two minutes had passed after a trip with low or moderate earnings was offered to and declined by the top twelve drivers in the queue.[7] At this point, it is very likely that

[3] `https://www.bloomberg.com/news/articles/2020-09-01/amazon-drivers-are-hanging-smartphones-in-trees-to-get-more-work`, accessed 09/07/2020.

[4] `https://www.vice.com/en/article/gvy357/the-new-system-uber-is-implementing-at-airports-has-some-drivers-worried`, accessed 02/23/2021.

[5] Airport trips account for 15% of Uber's gross bookings. See Form S-1 of Uber's IPO filing: `https://www.sec.gov/Archives/edgar/data/1543151/000119312519103850/d647752ds1.htm`.

[6] `https://help.lyft.com/hc/en-us/articles/115012922787-Receiving-Airport-FIFO-pickup-requests`, `https://www.uber.com/us/en/drive/dayton/airports/day/`, accessed 02/18/2021.

[7] When offered a trip, Uber and Lyft drivers have 15 seconds to decide whether to accept. `https://help.uber.com/driving-and-delivering/article/getting-a-trip-request?nodeId=e7228ac8-7c7f-4ad6-b120-086d39f2c94c`, `https://help.lyft.com/hc/en-/articles/115013080028-How-to-give-a-Lyft-ride`, accessed 02/24/2021.

the rider cancels her trip request, not knowing when a driver will be assigned, if at all.

What we have seen is that in the presence of heterogeneous earnings and finite rider patience, trips with moderate or lower earnings never reach drivers in the queue who are willing to accept them. This undercuts the platforms' mission of providing reliable transportation for riders, and leads to low revenue and trip throughput for the platform. Moreover, fulfilling only the small number of high earning trips is also a poor outcome for the drivers, since many drivers who just dropped off a rider at the airport will have to relocate back to the city with an empty car, and those who do join the queue need to wait for a very long time for a ride.

Simple fixes by limiting dispatching transparency or drivers' flexibility are not desirable— in recent years, ridesharing platforms are moving towards sharing trip destination and earnings estimation upfront, as well as providing drivers the options to accept or decline any trips without penalties.[8] Hiding information or imposing penalties are not fully effective either. For example, experienced drivers often call riders to ask about trip details when destinations are hidden before the pick-up [Cook et al., 2018]. Forcing drivers to accept every dispatch improves reliability in the short run, but also imposes a lottery (with possible outcomes ranging from $9 to over $60) on drivers who might have waited for two hours in line. Such high variance in earnings discourages future engagement, and leads to drivers' churning from the platform in the long run.

Recognizing the inefficiencies under FIFO dispatching, alternative allocation mechanisms have been studied in the literature. In particular, last-in-first-out (LIFO) dispatching is shown to be optimal in various settings [Hassin, 1985, Su and Zenios, 2004], since drivers lose (instead of gain) priority over time, thus have strong incentives to accept any trip they're dispatched. LIFO dispatching is, however, perceived as unfair by many [Breinbjerg et al., 2016], and is also easy to manipulate as participants may rejoin the queue at the end to (re)gain priority. This renders LIFO unsuitable and ineffective in ridesharing, as drivers have the option to go offline and online again to rejoin the end of the virtual queue at any time.

Ideally, platforms may properly price trips by destination and eliminate drivers' incentives to cherrypick. In recent work, Ma et al. [2019] propose the spatio-temporal pricing mechanism, which is welfare optimal, incentive aligned, and guarantees that drivers at the same origin are indifferent towards trips to all destinations. This remains an idealized target for our current setting, but is hard to achieve in practice. Consider, again, the O'Hare example as shown in Figure 1a. Tripling the fares of the short trips to match the earnings from the long trips is suboptimal, and the platform is unable to decrease driver payouts below some pre-determined per-minute and per-mile rates, thus earnings from long trips cannot be effectively reduced.

1.1 Our Results

In this work, we study the dispatching of trips to drivers who are waiting in a virtual queue, where some trips are necessarily more lucrative than the others due to operational constraints. Without the power to adjust trip prices, the mechanisms we design make use of drivers' waiting times in the queue to effectively align incentives, improve reliability and efficiency, and reduce the variability in drivers' total payoffs.

[8] The specific policies and their implementations vary across companies and geographical regions. As an example, see `https://www.uber.com/blog/california/keeping-you-in-the-drivers-seat-1/` (accessed 02/21/2021). This is in part due to regulatory requirements for categorizing drivers as independent contractors. For context, we cite an excerpt from California Proposition 22: "The network company does not require the appbased driver to accept any specific rideshare service or delivery service request as a condition of maintaining access to the network company's online-enabled application or platform."

The model. We study a continuous time, non-atomic model, with one origin (e.g. the airport) where trips are dispatched to drivers in a virtual queue. Riders request trips to a finite number of destinations with heterogeneous earnings for drivers. Upon the arrival of each rider, and after a rider's trip request is declined, the platform offers the rider's trip to a driver in the queue. Riders are willing to wait for some time for a match, but have finite patience and will cancel their requests and leave after a certain number of declines from drivers. Drivers' waiting in the queue (as opposed to driving elsewhere in the city, for example) is costly for both drivers and the platform. Drivers are strategic, aiming to optimize their earnings from trips minus the waiting costs they incur.

We study mechanisms that are fully *transparent* and *flexible* (see Footnote 8). At any point in time time, drivers know about the supply, demand, the length of the queue and their positions in their queue. When offered trip requests, drivers are provided trip destinations and earnings *upfront*, so that they can decide whether to accept based on this information. Moreover, drivers are not penalized for actions they take, and have the freedom to (i) decline dispatches they do not want to accept (ii) rejoin the virtual queue at the tail at any point of time, and (iii) decide to not join the queue upon arrival, or leave the queue at any point of time to perhaps relocate back to the city without a rider.

Main results. To optimize trip throughput and net revenue (i.e. total earnings from completed trips, minus the opportunity costs the platform incurs due to drivers' waiting in the queue), a platform should keep no driver in the queue, and dispatch all drivers upon arrival to destinations in decreasing order of trip earnings. However, under the status quo *strict FIFO dispatching* where trips are dispatched to each and every driver starting from the head of the queue, drivers close to the head of the queue are incentivized to excessively decline trip offers and wait for higher-earning trips. We analyze the equilibrium outcome under strict FIFO, and show that with finite rider patience, most trips except for the highest earning ones become unfulfilled. Driver earnings and the net revenue of the platform are further reduced due to drivers' excessive waiting in the queue.

Recognizing that the moderate and lower earning trips never reach drivers in the queue who are willing to accept them, we propose the *direct FIFO mechanism*, which offers these trips directly to drivers further down the queue. We prove that it is a subgame perfect equilibrium for drivers to accept all dispatches, and that the equilibrium outcome achieves the first best trip throughput, and *the second best net revenue*, i.e. the highest net revenue achievable by any mechanism that is flexible and transparent. The direct FIFO mechanism, however, suffers from fairness concerns and is ill-suited for practice. Consider, for example, the Chicago Midway airport (Figure 1b). A driver close enough to the head of the queue will no longer receive any trip back to downtown Chicago, since direct FIFO skips drivers at the head of the queue when dispatching lower-earning trips, and all high-earning trips the driver may receive will be heading to the suburbs.

To achieve optimal throughput and revenue without sacrificing fairness, we propose a family of *randomized FIFO mechanisms*. A randomized FIFO mechanism first offers each rider's trip request to a driver close to the head of the queue, and then sends the trip gradually down the queue in a randomized manner after each time the trip is declined by a driver. The dispatch rule is fair in the sense that when drivers are non-strategic, a driver closer to the head of the queue receives trips to every destination at a higher rate than any driver further down the queue. We prove that the randomized FIFO mechanisms appropriately align incentives using waiting times, achieving the first best throughput and second best net revenue in equilibrium— the option to skip the rest of the line incentivizes drivers further down the queue to accept trips with lower earnings; randomizing each dispatch among a small group of drivers increases each individual drivers' waiting times for the next dispatch, allowing the mechanism to prioritize drivers closer to the head of the queue for

trips to every destination without creating incentives for excessive cherrypicking.

Extensive counterfactual simulations using ridesharing data from the City of Chicago suggest that in comparison to the *status quo* strict FIFO dispatching, the randomized FIFO mechanism achieves substantial improvements in revenue, throughput, and driver earnings. Moreover, the variance in drivers' total payoffs is small, and diminishes rapidly as riders' patience increases since the randomized FIFO mechanism can more effectively match higher-earning trips with drivers who have incurred higher waiting costs in the queue. This demonstrates the desirable balance achieved by the randomized FIFO mechanisms between efficiency, fairness, and the variability in driver earnings, and highlights the effectiveness of using waiting times to align incentives and reduce inequity when the flexibility to set prices is limited due to operational constraints.

1.2 Related Work

Ridesharing Markets. The literature on pricing and matching in ridesharing platforms is rapidly growing. Castillo et al. [2017] and Yan et al. [2020] establish the importance of dynamic pricing in maintaining the spatial density of open driver supply, which reduces waiting times and improves operational efficiency. In the presence of spatial imbalance and temporal variation of supply and demand, Bimpikis et al. [2019] and Besbes et al. [2020] study revenue-optimal pricing for the platform, Ma et al. [2019] propose origin-destination based pricing that is welfare optimal and incentive compatible, and Garg and Nazerzadeh [2020] show that additive instead of multiplicative "surge" pricing is more incentive aligned for drivers when prices need to be origin-based only. Considering the online arrival of supply and demand and their distribution in space, Kanoria and Qian [2020], Qin et al. [2020] and Özkan and Ward [2020] study dynamic matching policies that dispatch drivers from areas with relatively abundant supply, and Ashlagi et al. [2018], Dickerson et al. [2018] and Aouad and Saritaç [2020] focus on the online matching between riders and drivers and the pooling of shared rides. In this work, we focus on a single origin where the optimal destination-based pricing is infeasible due to operation constraints, so that some trips are necessarily more lucrative than the others. This leads to the need of using drivers' waiting times to align incentives and to reduce the inequity in driver earnings.

The operation of ridesharing platforms is also studied using queueing-theoretic models. Banerjee et al. [2015] compare optimal dynamic and static pricing policies; Banerjee et al. [2018] propose state-dependent dispatching policies to minimize unfulfilled demand; Afeche et al. [2018] study the impact of platform control on platform revenue and driver income; Besbes et al. [2019] show that in comparison to traditional service settings, higher capacity is needed when spatial density affects operational efficiency; Castro et al. [2020] study practical dispatching policies when drivers have heterogeneous compatibility over trips. These works use queueing-theoretic framework to model and analyze the availability of supply, but for settings where drivers are spread out in space at different locations. In contrast, we focus on the matching of trips to drivers who are waiting in a *virtual queue*, addressing the problem of dispatching heterogeneous trips to drivers who have incurred different waiting costs in a way that is reliable, efficient and fair.

Various empirical studies analyze the Uber platform as a two-sided marketplace, focusing on the labor market of Uber drivers [Hall et al., 2017], the longer-term labor market equilibration [Hall and Krueger, 2016], the value of flexible work arrangements [Chen et al., 2019], learning-by-doing and the gender earnings gap [Cook et al., 2018], and the surplus of consumers [Cohen et al., 2016]. In regard to the dynamic "surge" pricing, Hall et al. [2015], Chen and Sheldon [2015], and Lu et al. [2018] demonstrate its effectiveness in improving reliability and efficiency, increasing driver supply during high-demand times, as well as incentivizing drivers to relocate to higher surge areas. In contrast, we use data from ridesharing platforms (including Uber and Lyft, made public by

the City of Chicago) to understand the heterogeneity in driver earnings by trip destination. We also demonstrate via counterfactual simulations the inefficiencies of the equilibrium outcome when drivers are strategic, as well as the substantial improvements achieved by our proposed mechanisms.

Queue-Based Mechanisms. Designing incentive-aligned and efficient mechanisms to allocate jobs/resources to a queue of strategic agents with heterogeneous or homogeneous preferences has been studied in various settings.

When agent types are homogeneous, pioneering work such as Naor [1969] suggests to levy tolls to alleviate the negative externalities caused by the waiting agents. In this realm, papers have studied how to address the inefficiencies that arise in queueing settings with self-interested customers using transfers (see Hassin [2016] for a comprehensive review). However, in many important practical settings including ours, the use of monetary incentives is often restricted, by regulation or business constraints. When monetary transfers are not used, Hassin [1985] and Su and Zenios [2004] showed that LIFO dispatching is socially optimal, but as we have discussed earlier in the introduction it suffers from fairness concerns and can be easy to manipulate in practice. Assuming a constant queue length, Bloch and Cantala [2017] compare probabilistic queueing disciplines which assign every types of jobs to agents with higher queue positions with higher probabilities, similar to our fairness consideration, and conclude that FIFO is always preferred by all agents (who joined the queue) while random allocation reduces waste caused by agent rejections. In our setting, agents can decide to join or not based on queue length and, therefore, the queue size is an equilibrium outcome. As a consequence, FIFO could lead to excessive loss of jobs due to limited rider patience, which results in poor outcomes for riders, drivers and the platform; we also show that a completely random allocation can cause large variation in driver payoffs.

When agents have private and heterogeneous preference over jobs, various queue-based mechanisms and information designs have been proposed to ensure efficiency and incentivize truthful reporting. A closely related paper to ours is Ashlagi et al. [2020]. They consider the problem of how to match heterogeneous objects to agents with private types. They identify the welfare maximizing mechanism and establish that it can be implemented by a FIFO queue together with an appropriate information disclosure policy — adjacent job types are pooled together and only information about the interval a job type belongs to is shown to the agent, who can decide to accept or reject. By hiding the exact job information, the highest value for a match might not be attained; however, the mechanism still achieves the maximum welfare by incentivizing agents to accept objects which leads to lower wait times. A strict FIFO mechanism leads to poor outcomes in our setting due to limited job patience. Importantly, we assume that the platform must show the complete trip information to drivers in the queue. As we explain below, these are unique aspects of our setting that lead to a different class of dispatching rules. Other related work in this line of literature include Su and Zenios [2006], Leshno [2019] and Arnosti and Shi [2020]. Su and Zenios [2006] consider a model with finite number of agent types and a continuum of job types. The mechanism they propose is a monotone disjoint queue mechanism where jobs types are partitioned into subsets and each subset of jobs are separately allocated to different queues. Leshno [2019] focuses on a special case of two agent and job types under an overloaded setting without waiting cost and shows that a buffer queue mechanism for each of the job types is optimal. We note that implementing a disjoint queue mechanism could be practically challenging in our setting as a large number of separate queues of trip destinations might need to be organized.[9] Arnosti and Shi [2020] establish equivalence between certain common procedures used in the practice of public housing

[9] Ashlagi et al. [2020] show that a maximum of $2n + 1$ number of queues are needed for n types of jobs.

allocation including independent and common lotteries and waiting lists with or without deferral option.

In contrast to the aforementioned papers, our proposed randomized FIFO mechanism is unique in the novel ways waiting time is used to balance different dimensions of the equilibrium outcome: (1) offering drivers the option to skip the line incentivizes them to accept trips with lower earnings, which reduces cherry pickings and balances each driver's waiting cost and trip earning; (2) randomizing the dispatch incentivizes drivers to accept a wider range of trip types, which is critical to maintain reliability without having to skip a long list of drivers, as the opportunity for allocating a trip is limited due to rider patience; (3) finally, carefully controlling the dispatch randomness to be confined within a small list of drivers reduces the variance of driver payoffs. Such novel combination of randomization and targeting, to the best of our knowledge, is new and leads to a practical and efficient dispatching mechanism can be used in major ridesharing platforms to achieve desirable outcomes.

2 Preliminaries

We study a continuous time model, with one origin (e.g. an airport) where trips are dispatched to drivers who are waiting in a virtual queue. $\mathcal{L} = \{1, 2, \ldots, \ell\}$ is the set of $\ell \in \mathbb{Z}_{>0}$ discrete trip types (e.g. trips to different destinations). Rider demand and driver supply are infinitesimal, and are stationary over time. For each location $i \in \mathcal{L}$, $\mu_i > 0$ denotes the arrival rate of riders requesting trips to location i (i.e. the mass of riders arriving per unit of time). Upon arrival, riders' trip requests need to be dispatched to the drivers. All riders have a *patience level* of $P \in \mathbb{Z}_{>0}$, meaning that a rider may be willing to wait for a while for a driver to accept her trip request, but she will leave (e.g. cancel her request) after the P^{th} time that her trip is declined by the drivers. Each driver can drive any rider to her destination, and riders do not have preferences over drivers.

Let $\lambda > 0$ be the arrival rate of drivers at the origin. Upon arrival, the driver may decide whether to join the queue. The *net earnings* of a trip to each location $i \in \mathcal{L}$ is w_i. A driver who completes a rider trip to location i gets a payoff of w_i from the trip, and the payoff of a driver who does not join the queue, or leaves the queue without a rider trip, is normalized to be zero. While waiting in the queue, a driver incurs an opportunity cost of $c > 0$ per unit of time. This cost captures, for example, the potential earnings a driver can make from driving elsewhere instead of waiting in the queue. Drivers are strategic, aim to optimize their earnings from trips minus their waiting costs, and do not have preferences over riders or destinations.

For each unit of time a driver spends waiting in the queue, the platform incurs an opportunity cost of $c_p \in [0, c]$. This models the revenue the driver could have made for the platform, when the driver is driving elsewhere in the city instead of waiting in the queue. Drivers may not always drive for the platform, and the market might also be oversupplied As a result, the opportunity cost for the platform c_p may be lower than that for the drivers.

An informal timeline of a dispatching mechanisms is as follows (see Section 3 for a formal definition). Upon the arrival of each rider, the mechanism may dispatch the rider's trip request to a driver in the queue. If the driver accepts the dispatch, she leaves the queue and picks up the rider. Otherwise, the trip may be dispatched again, until (i) some driver accepts the trip, or (ii) the rider cancels her request when her patience runs out (after the trip is declined for P times), or (iii) the mechanism decides to not dispatch the trip again.

We consider a setting where the platform has complete information about demand, supply, opportunity costs, and earnings from trips to different destinations. We assume drivers have the same information, and that this is common knowledge amongst the drivers. We study mechanisms

that are fully *transparent* and *flexible*. At any point in time, all drivers know the total length of the queue and their positions in the queue. When offered trip dispatches, drivers are provided trip destinations and earnings *upfront*, so that they can decide whether to accept a dispatch based on this information. Moreover, drivers are not penalized for actions they take, and have the options to (i) decline dispatches they do not want to accept without losing their position in the queue, (ii) rejoin the virtual queue at the tail at any point of time, and (iii) decide to not join the queue upon arrival, or leave the queue at any point of time to perhaps relocate back to the city without a rider.

A platform's *trip throughput* is the mass of trips completed per unit of time by all drivers in the queue. A platform's *net revenue* is the sum of the net earnings from trips made by all drivers in the queue per unit of time, minus the opportunity cost the platform incurs due to drivers' waiting in the queue (this opportunity cost models the platform's loss of revenue elsewhere in the city, due to driver supply being tied-up in the queue). When drivers are non-strategic and always follow the platform's dispatches, we refer to the highest possible trip throughput and net revenue a platform can achieve the *first best*.

For simplicity of notation, we assume the destinations are ordered such that $w_1 > w_2 > \cdots > w_\ell \geq 0$.[10] With stationary and infinitesimal demand and supply, a platform does not need a non-zero driver queue. In steady state, a platform that aims to optimize its net revenue should keep no driver in the queue, but dispatch drivers upon their arrival to destinations in decreasing order of w_i until either all drivers are dispatched or all riders are picked-up. Denote the lowest-earning trip that is (partially) completed as

$$i^* = \max\{i \in \mathcal{L} \mid \lambda > \sum_{j=1}^{i-1} \mu_j\}. \tag{1}$$

Proposition 1 (The first best)**.** The steady state first best outcome has zero drivers in the queue. Upon arrival, drivers are dispatched to pick up arriving riders in decreasing order of w_i. The remaining drivers (if any) are suggested to leave without joining the queue. The *first best trip throughput* is

$$T_{\text{FB}} = \min\{\lambda,\ \sum_{i\in\mathcal{L}} \mu_i\}, \tag{2}$$

and the *first best net revenue* is

$$R_{\text{FB}} = \sum_{i=1}^{i^*-1} w_i\mu_i + w_{i^*} \min\{\lambda - \sum_{j=1}^{i^*-1} \mu_i,\ \mu_{i^*}\}. \tag{3}$$

2.1 Strict FIFO Dispatching

The first-in-first-out (FIFO) queue discipline is considered "fair" by most, and is the default discipline in many everyday situations [Platz and Østerdal, 2017]. When drivers have the flexibility to decline undesirable trips, however, we show that dispatching each trip to every driver starting from the head of the queue leads to poor outcomes for the riders, drivers, and the platform. To avoid ambiguity, we refer to this mechanism as *strict FIFO dispatching.*

[10] Combining destinations with the same net earnings does not affect the equilibrium outcome of any mechanism. For drivers who are free to decline dispatches based on trip destinations, no trip with $w_i < 0$ will be accepted, since completing one such trip is worse than declining the dispatch and leave the queue immediately. Thus we assume for now that $w_i \geq 0$ for all $i \in \mathcal{L}$.

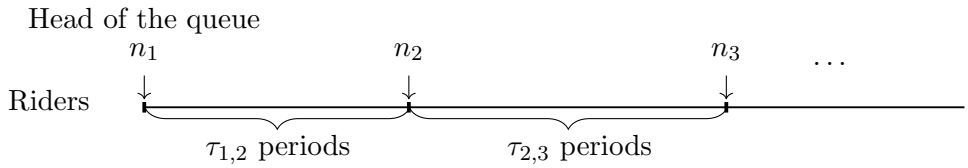


Figure 2: The equilibrium outcome under strict FIFO dispatching, assuming infinite rider patience.

We start by analyzing the equilibrium outcome under strict FIFO dispatching. Consider a rider request for a trip to location 1. Under strict FIFO, the trip will be accepted by the driver at the head of the queue, since a trip to location 1 has the highest earnings among all trips the driver may receive in the queue. Moreover, the (infinitesimal) driver at the head of the queue will only be willing to accept trips to location 1, since she is the first in line to receive all incoming trip dispatches, thus she does not have to wait any time for a trip dispatch to location 1.

Similar reasoning holds for drivers who are close to the head of the queue, and a driver is willing to accept a trip to location 2 only if the additional waiting cost for a trip to location 1 outweighs the increase in earnings $w_1 - w_2$. Let $\tau_{1,2}$ be the maximum additional time a driver is willing to wait for a trip to location 1, in comparison to immediately taking a trip to location 2. We know

$$\tau_{1,2}c = w_1 - w_2 \Rightarrow \tau_{1,2} = \frac{w_1 - w_2}{c}. \tag{4}$$

By Little's Law, the first position in the queue where the driver is willing to accept a location 2 trip is $n_2 \triangleq \mu_1\tau_{1,2} = \mu_1(w_1 - w_2)/c$, since the waiting time from this position to the head of the queue is $\tau_{1,2}$ when all drivers ahead of this position accept only trips to location 1. For a driver at position n_2, her *continuation payoff* (i.e. net earnings from the trip the driver accepts, minus the waiting costs the driver incurs from this point of time onward) is w_2, regardless of whether the driver accepted a trip to location 2, or if the driver continued to wait for a trip to location 1.

Similarly, in comparison to accepting a trip to location $i + 1$, a driver is willing to wait an additional $\tau_{i,i+1} = (w_i - w_{i+1})/c$ units of time for a trip to location i. We can compute the first positions in the queue where drivers are willing to accept trips to each location $i \in \mathcal{L}$, assuming that riders have infinite patience and will not cancel their trip requests regardless of how many times their trips have been declined by the drivers (see Figure 2).

Lemma 1 (informal)**.** Assume that riders have infinite patience. Under strict FIFO dispatching, it is an equilibrium for a driver to accept trip dispatches to each location $i \in \mathcal{L}$ only if the driver is at position $q \geq n_i$ in the queue, where $n_1 \triangleq 0$ and

$$n_i \triangleq \sum_{j=1}^{i-1} \left(\frac{w_j - w_{j+1}}{c} \sum_{k=1}^{j} \mu_k \right), \ \forall i \geq 2. \tag{5}$$

We provide the formal statement and the proof of the equilibrium outcome under strict FIFO dispatching in Appendix A.1. Briefly, we prove by induction that assuming infinite rider patience, for each location $i \in \mathcal{L}$, a driver at position n_i in the queue gets a continuation payoff of w_i, regardless of if she accepted a trip to location i or not. Drivers at positions earlier than n_i in the queue are, however, better off waiting for trips with higher earnings.

When riders have a finite patience level P, trip requests to locations $i \in \mathcal{L}$ with $n_i > P$ will not reach a driver who is willing to accept this trip before the rider's patience runs out. As a result, trips to these destinations become unfulfilled, leading to poor efficiency and reliability.

Example 1. Consider an airport queue, where riders request trips to three destinations $\mathcal{L} = \{1, 2, 3\}$. The arrival rate of riders to each destination is $\mu_1 = 1$, $\mu_2 = 6$, and $\mu_3 = 3$, and the net earnings from these trips are $w_1 = 75$, $w_2 = 25$, $w_3 = 15$. Intuitively, trips to location 1 represent the rare but high-earning long trips from the airport. Location 2 can be considered as the downtown area with high trip volumes and medium earnings, and think about location 3 trips as short rides to the hotels and towns surrounding the airport with low earnings.

Drivers arrive at a rate of $\lambda = 5$ per unit of time, and the opportunity costs for the drivers and the platform are $c = c_p = 1/3$. Considering each unit of time as one minute, this corresponds to a scenario where a driver driving for the platform elsewhere in the city will make \$20 per hour. Riders have a patience level of $P = 12$. When it takes an average of 10 seconds for each driver to decline a dispatch, this corresponds to the riders' being willing to wait for two minutes before canceling their trip requests.

The first best. The first best outcome accepts all trips to location 1, and dispatches the remaining 4 units of driver to trips to location 2. No drivers wait in the line. The first best trip throughput is $T_{\text{FB}} = 5$, and the first best net revenue is $R_{\text{FB}} = w_1 + 4w_2 = 175$ per minute. This outcome can be implemented, for example, by forcing each driver to always accept the trip she receives. This introduces a high variance in drivers' total payoffs (net earnings from trip minus waiting costs): the average total payoff of a driver who arrived at the queue is 35, and the variance is 400.

Strict FIFO dispatching. Under strict FIFO dispatching, when drivers has the flexibility to decide which trips to take, the driver at the head of the queue is only willing to accept trips to location 1. A driver with a location 2 trip in hand is willing to wait an additional $\tau_{1,2} = (w_1 - w_2)/c = 150$ minutes for a trip to location 1. Therefore, the first position in the queue where the driver is willing to go to location 2 is $n_2 = \tau_{1,2}\mu_1 = 150$. With a patience level of 12, riders requesting trips to location 2 will cancel their trip requests and leave after their trips are declined by the 12^{th} driver in the queue. This is poor service reliability for the riders. Location 3 trips are similarly unfulfilled, thus strict FIFO achieves only a trip throughput of $T_{\text{strict}} = 1$ per unit of time.

The remaining 4 units of drivers will need to leave the queue without a rider trip. The drivers, however, will not leave the queue if the payoff from joining the queue is better than that from leaving. Drivers are willing to wait for $w_1/c = 225$ minutes for a trip to location 1, thus the steady state queue length will be $\mu_1 w_1/c = 225$. In equilibrium, drivers get a payoff of zero regardless of whether they joined the queue, or left without a rider. The large number of drivers waiting in the queue is also very costly for the platform, which achieves in this example a net revenue of zero: $R_{\text{strict}} = w_1\mu_1 - c_p(\mu_1 w_1/c) = 0$. □

Strict FIFO dispatching is *fair*, in the sense that drivers who are closer to the head of the queue has higher priority for receiving trips to every destination. As we have seen in the example, however, dispatching each trip to each and every driver in the queue leads to poor reliability for the riders, low trip throughput and net revenue for the platform, and zero earnings for the drivers themselves, despite their strategizing for better earnings. In next section, we will see that by deprioritizing drivers at the head of the queue for trips to certain destinations (thus violating what is typically perceived as fair dispatching), we are able to substantially improve the outcome for the riders, drivers, and the platform, even without the power to adjust trip prices.

3 The Direct FIFO Mechanism

In this section, we introduce *the direct FIFO mechanism*. The mechanism is based largely on FIFO dispatching, but sends lower-earning trips starting from positions further down the queue where

drivers are willing to accept the dispatches for the option to skip the rest of the line. Accepting all trip dispatches forms a *subgame perfect equilibrium* among drivers, and the mechanism achieves the highest revenue and throughput possible under any mechanism that is flexible and transparent.

3.1 A Dispatching Mechanism

We first formally define a dispatching mechanism. Let $Q \geq 0$ denote the length of the virtual queue, and let $q \in [0, Q]$ be a particular position in the queue. $q = 0$ and $q = Q$ are the *head* and the *tail* of the queue, respectively— positions where the drivers have waited the longest and the shortest time in line. Let h denote the past dispatching history of a particular rider trip, representing for how many times was the trip offered (if any), and to drivers at which positions in the queue. Finally, we use ϕ to denote the decision to not dispatch a rider's trip request to any driver.

Definition 1 (Dispatching mechanism)**.** Given the queue length Q, the past dispatching history h of a trip, and the trip's destination, a dispatching mechanism x determines a probability distribution over $[0, Q] \cup \{\phi\}$. Upon the arrival of a rider, or after a rider's trip is declined by some driver, the mechanism uses x to (i) dispatch the trip to a driver at some position $q \in [0, Q]$ in the queue, or (ii) decide to not dispatch the trip (which we denote as ϕ).

The queue length Q represents the *state* of the queue. The dispatching mechanisms we study make dispatch decisions for each trip based on the state and the past dispatch history of this particular trip, but not on other factors such as how the state had evolved over time, or what actions the drivers had taken in the past.[11] Similarly, we focus on driver strategies that depend on the queue length and a driver's position in the queue, and we denote a strategy as a tuple $\sigma = (\alpha, \beta, \gamma)$. For any queue length $Q \geq 0$, and at any position $q \in [0, Q]$ in the queue,

(i) $\alpha(q, Q, i) \in [0, 1]$ for each location $i \in \mathcal{L}$ is the probability with which the driver at position q in the queue accepts the trip dispatch, if she is dispatched a trip to location i,

(ii) $\beta(q, Q) \in [0, 1]$ determines the probability with which the driver at position q in the queue re-joins the queue at the tail (by going offline and online again, for example), and

(iii) $\gamma(q, Q) \in [0, 1]$ is the probability for the driver at q to leave the queue without a rider.

Let $U(q, Q, \sigma, \sigma')$ denote the random variable representing the *continuation payoff* of the driver at position q in the queue, when the current length of the queue is Q, when this driver employs strategy σ, and when all other drivers employ strategy σ' (including those drivers who will arrive in the future). This includes the net earnings from the trip the driver may complete in the future, minus the total opportunity cost she incurs from this point of time onward waiting in the queue. Denote $\pi(q, Q, \sigma, \sigma') \triangleq \mathbb{E}\left[U(q, Q, \sigma, \sigma')\right]$ as the driver's *expected continuation payoff* from position q onward, where the expectation is taken over randomness in both the mechanism's decisions and the strategies of drivers. $\pi(Q, Q, \sigma, \sigma')$ thus represents the expected payoff of a driver with strategy σ, who joins the queue when the queue length is Q, and when all other drivers employ strategy σ'.

We define the following properties.

Definition 2 (Subgame-perfect equilibrium)**.** A strategy σ^* forms a *subgame perfect equilibrium* (SPE) among drivers under a mechanism if for any economy and any feasible strategy σ,

$$\pi(q, Q, \sigma^*, \sigma^*) \geq \pi(q, Q, \sigma, \sigma^*), \ \forall Q \geq 0, \ \forall q \in [0, Q]. \tag{6}$$

[11] When a mechanism is allowed to make dispatch decisions based on drivers' actions in the past, the mechanism may easily align incentives by no longer sending any trip offers to a driver who had declined a dispatch, for example.

Definition 3 (Individual rationality)**.** A mechanism is *individually rational in SPE* if under a strategy σ^* that forms an SPE among the drivers, for any economy,

$$\pi(q, Q, \sigma^*, \sigma^*) \geq 0, \ \forall Q \geq 0, \ \forall q \in [0, Q]. \tag{7}$$

Definition 4 (Envy-freeness)**.** A mechanism is *envy-free in SPE* if under a strategy σ^* that forms an SPE among the drivers, for any economy,

$$\pi(q_1, Q, \sigma^*, \sigma^*) \geq \pi(q_2, Q, \sigma^*, \sigma^*), \ \forall Q \geq 0, \ \forall q_1, q_2 \in [0, Q] \text{ s.t. } q_1 \leq q_2. \tag{8}$$

Intuitively, under a mechanism that is individually rational and envy-free in SPE, a driver anywhere in the queue always gets non-negative continuation payoff, and does not envy the future earnings of any driver who arrived later in time.

Given a mechanism $\mathcal{M}$ and a strategy σ^* that forms an SPE under $\mathcal{M}$, let Q^* be length of the queue under σ^* in steady state. This is the case if the number of drivers joining the queue per unit of time is equal to the number of drivers dispatched from the queue, when (i) the length of the queue is Q^*, and (ii) all drivers adopt strategy σ^*. Moreover, let $z_i(\sigma^*)$ denote the fraction of trips to location $i \in \mathcal{L}$ that are completed under σ^* in steady state.

The *trip throughput* of mechanism $\mathcal{M}$ is the amount of trips completed per unit of time under σ^* in steady state:

$$T_{\mathcal{M}}(\sigma^*) \triangleq \sum_{i \in \mathcal{L}} z_i(\sigma^*) \mu_i. \tag{9}$$

The *net revenue* achieved by mechanism $\mathcal{M}$ is the total net earnings all drivers made from trips per unit of time under σ^* in steady state, minus the total opportunity costs the platform incurs due to drivers' waiting in the queue:

$$R_{\mathcal{M}}(\sigma^*) \triangleq \sum_{i \in \mathcal{L}} z_i(\sigma^*) \mu_i w_i - Q^* c_p. \tag{10}$$

When $c_p = c$, the net revenue of the platform is $R_{\mathcal{M}}(\sigma^*) = \sum_{i \in \mathcal{L}} z_i(\sigma^*) \mu_i w_i - Q^* c$, i.e. the total net payoffs to all drivers who arrive at the queue.

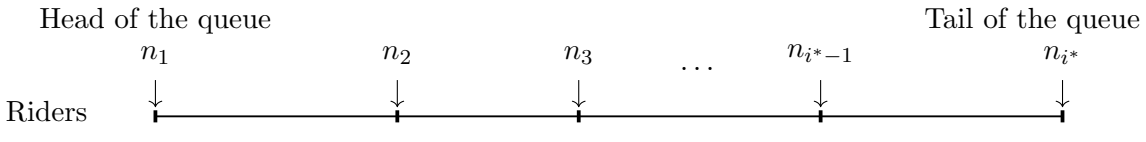
The objective of a mechanism is to optimize trip throughput and net revenue achieved under σ^* in steady state.[12] We say a mechanism is *optimal* if in equilibrium in steady state (i) the mechanism achieves the first best trip throughput, and (ii) the mechanism achieves the *second best net revenue* i.e., the highest steady sate equilibrium net revenue that is achievable by any dispatching mechanism that is flexible and transparent, providing trip information to drivers upfront and not penalizing drivers for any actions they take.[13]

3.2 Optimality of Direct FIFO

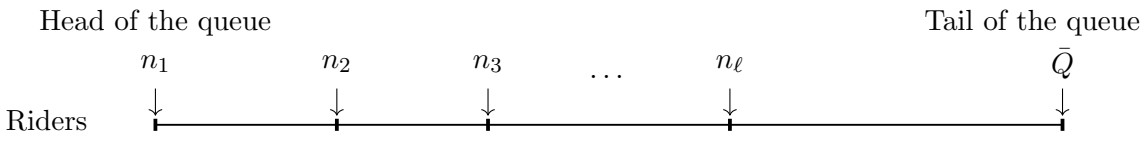
Definition 5 (Direct FIFO)**.** Under *the direct FIFO mechanism*, trips to each location $i \in \mathcal{L}$ are dispatched in a FIFO manner to drivers starting from position n_i (as defined in (5)) in the queue, when the length of the queue is $Q \geq n_i$. When $Q < n_i$, trips to location i are not dispatched.

[12] When the platform takes as commission a fixed fraction of the earnings made by the drivers (both from the queue and also from driving elsewhere in the city), the problem of maximizing a platform's total commission is equivalent to that of maximizing the net revenue as defined in (10).

[13] As we shall see later in this section, we may not be able to achieve the first best net revenue in certain settings, despite achieving the first best trip throughput. This is the case when a mechanism completes the same set of trips as those under the first best outcome, but drivers' strategically waiting for higher earning trips leads to a non-zero queue length, reducing the net revenue for the platform.



(a) The under-supplied scenario, with $\lambda \leq \sum_{i \in \mathcal{L}} \mu_i$.



(b) The over-supplied scenario, with $\lambda > \sum_{i \in \mathcal{L}} \mu_i$.

Figure 3: The steady-state equilibrium outcome under the direct FIFO mechanism.

Under the direct FIFO mechanism, the highest earning trips to location 1 are dispatched to the head of the queue, where the driver have waited for the longest time (thus have incurred the highest waiting costs). For a trip to location $i > 1$ with lower earnings, the mechanism skips drivers close to the head of the queue who will be unwilling to accept, and dispatches the trip starting from the n_i^{th} position— the first position in the queue where the driver is willing to accept a trip to location i under strict FIFO dispatching. The following theorem proves that this option to "skip the rest of the line" incentivizes drivers to accept all dispatches they receive.

Theorem 1 (Incentive compatibility of direct FIFO)**.** *It is a subgame-perfect equilibrium for drivers to accept al dispatches from the direct FIFO mechanism, and to join the queue if and only if the length of the queue is at most*

$$\bar{Q} \triangleq n_\ell + \frac{w_\ell}{c} \sum_{i \in \mathcal{L}} \mu_i. \tag{11}$$

Moreover, the equilibrium outcome is individually rational and envy-free.

The proof is via induction on queue positions, and is provided in Appendix A.2. Intuitively, this is a "direct implementation" of the equilibrium outcome under strict FIFO dispatching when riders have infinite patience (see Lemma 1). Trips are dispatched starting from the positions in the queue where the drivers are indifferent towards accepting the trip or continuing to wait, and the equilibrium payoff from joining the queue is non-negative when the queue length is at most $\bar{Q}$.

When there are more drivers than needed to complete all trips to location 1, the direct FIFO mechanism does not achieve the first best net revenue— drivers are willing to spend time waiting for trips with higher earnings, leading to a queue of non-zero length and lowering the net revenue of the platform. This kind of "strategic waiting" is, however, not avoidable. We prove in the following theorem that the direct FIFO mechanism achieves the *second best net revenue* in equilibrium in steady state, i.e. the highest equilibrium net revenue achievable by any dispatching mechanism that provides trip destinations upfront and does not penalize drivers for declining dispatches.

Theorem 2 (Optimality of direct FIFO)**.** *For every economy, the direct FIFO mechanism achieves in SPE the first best steady state trip throughput. Moreover, the equilibrium outcome achieves the first best net revenue when $c_p = 0$, and the second best net revenue when $c_p \in (0, c]$.*

We prove this theorem in Appendix A.2. Briefly, we first show that the steady state equilibrium outcome under direct FIFO is as illustrated in Figure 3. When $\lambda \leq \sum_{i \in \mathcal{L}} \mu_i$, i^* as defined in (1) is

the lowest-earning trip that is (partially) completed in equilibrium. Drivers will line up for a trip to location $j < i^*$ (which has higher earnings than w_{i^*}), but the equilibrium queue length is $Q^* = n_{i^*}$ and there is no wait for a trip to location i^*. See Figure 3a. Every driver gets a total payoff of w_{i^*}, regardless of which trip they take, and all trips that are completed under the first-best outcome are also completed. When $\lambda > \sum_{i \in \mathcal{L}} \mu_i$, the queue is over-supplied such that all trips are accepted and completed, and the equilibrium queue length is $Q^* = \bar{Q}$ (see Figure 3b). At this point, the drivers are indifferent between joining the queue and leaving, and all drivers get a zero payoff.

Given that all trips completed under the first best outcome are also completed, direct FIFO achieves the first best trip throughput, and also the first best net revenue if $c_p = 0$, (i.e. when the platform does not incur any opportunity cost due to drivers' waiting in the queue). To prove that the mechanism achieves the second best revenue when $c_p > 0$, we first establish that no mechanism can achieve in equilibrium a strictly higher total payoff for drivers. This implies that if the same set of trips are completed, reducing the equilibrium queue length (thereby reducing the opportunity costs for the platform) is not possible, and we also show that completing a different set of trips in return for a shorter queue is not an improvement.

Example 1 (Continued)**.** Consider the economy in Example 1 in Section 2, for which strict FIFO dispatching achieves trip throughput $T_{\text{strict}} = 1$ and net revenue $R_{\text{strict}} = 0$. Under direct FIFO, trips to each location will be dispatched to drivers in the queue starting at positions $n_1 = 0$, $n_2 = 150$, and $n_3 = 360$, respectively. With $\lambda = 5$, $\mu_1 = 1$ and $\mu_2 = 6$, the lowest earning trip accepted will be trips to location $i^* = 2$, and the equilibrium queue length is $Q^* = n_2 = 150$.

Upon arrival at the tail of the queue at n_2, one unit of driver moves on to wait for trips to location 1, and the remaining 4 units of drivers immediately accept trips to location 2 and leave the queue. In equilibrium, all drivers get a total payoff of $w_2 = 25$, thus there is no variance in drivers' total payoffs. The platform achieves a trip throughput of $T_{\text{direct}} = 5$ and a net revenue of $R_{\text{direct}} = 1 \cdot w_1 + 4 \cdot w_2 - Q^* c_p = 125$ per unit of time. Since $c = c_p$ this net revenue is also the total payoff for all drivers combined. □

In comparison to strict FIFO dispatching, the direct FIFO mechanism substantially improves driver earnings, trip throughput, and the net revenue of the platform. The mechanism is, however, not fair because once a driver had moved close enough to the head of the queue, she is no longer eligible for dispatch for most of the trips. Take the Midway airport as an example. A driver who has waited long enough in the queue will never receive a trip back to downtown Chicago again—as we can see from Figure 1b, all high-earning trips direct FIFO dispatches to her will be heading to the suburbs. This renders the direct FIFO mechanism ill-suited for practice.

4 The Randomized FIFO Mechanism

In this section, we introduce a family of *randomized FIFO mechanisms*, which achieve the same optimal trip throughput and net revenue as the equilibrium outcome under direct FIFO. At the same time, the randomized FIFO mechanisms retain the fairness property that when drivers are straightforward and accept all dispatches, a driver closer to the head of the queue receives trip dispatches to *every* destination at a (weakly) higher rate than any driver further down the queue.

Before introducing the randomized FIFO mechanisms, we first analyze the steady state Nash equilibrium under *random dispatching*, where every trip request is simply dispatched to a driver in the queue uniformly at random. This demonstrates how optimal throughput and revenue can be achieved in a fair manner via randomization.

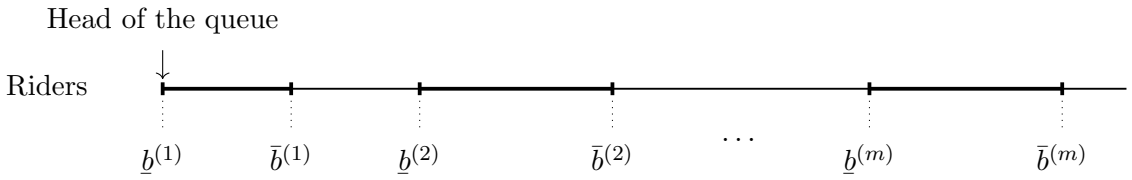


Figure 4: Illustration of a randomized FIFO mechanism.

Definition 6 (Nash equilibrium in steady state)**.** A strategy σ^* forms a Nash equilibrium among drivers in steady state under a mechanism if there exists a queue length $Q^* \geq 0$ such that

(i) for any feasible strategy σ and any position in the queue $q \in [0, Q^*]$,

$$\pi(q, Q^*, \sigma^*, \sigma^*) \geq \pi(q, Q^*, \sigma, \sigma^*), \tag{12}$$

(ii) when all drivers employ strategy σ^*, the steady state queue length is Q^*.

Proposition 2 (Optimality of random dispatching)**.** In Nash equilibrium in steady sate, dispatching every trip to all drivers in the queue uniformly at random achieves the first best trip throughput and the second best net revenue. When $c_p = 0$, the equilibrium net revenue is also the first best.

See Appendix A.3 for the proof of this result. Briefly, we show that under random dispatching, every driver in the queue is willing to accept a trip to location i^* (the lowest earning trip accepted in equilibrium under direct FIFO) despite the fact that the driver may still receive trips with higher earnings later. This is different from the outcome under strict FIFO, because in comparison to the driver at the head of the queue under strict FIFO, a driver who declines a dispatch under random dispatching will need to wait for a much longer time to receive her next dispatch.

This additional waiting time introduced by randomization increases drivers' costs of cherrypicking, and allows random dispatching to align incentives without skipping drivers at the head of the queue while dispatching trips to any location. Naively randomizing over all drivers in the queue, however, introduces substantial variability in drivers' waiting times, on top of the variability in the earnings drivers get from the trips themselves. This is in stark contrast to direct FIFO, which sends lower-earning trips to drivers who have waited less time in the queue, thereby reducing the variation in drivers' total payoffs.

The randomized FIFO mechanisms we now propose make proper use of drivers' waiting times in the queue in both ways. By gradually sending declined trips down the queue in a randomized manner, a randomized FIFO mechanism aligns incentives, and also guarantees that drivers in earlier segments in the queue (who have incurred higher waiting costs) take trips with higher earnings.

Definition 7 (Randomized FIFO)**.** A *randomized FIFO mechanism* is specified by $m \geq 1$ "bins" in the queue $([\underline{b}^{(1)}, \bar{b}^{(1)}],\ [\underline{b}^{(2)}, \bar{b}^{(2)}],\ \ldots, [\underline{b}^{(m)},\ \bar{b}^{(m)}])$. For the k^{th} time a trip is dispatched, the mechanism dispatches the trip to a driver in the k^{th} bin $[\underline{b}^{(k)}, \bar{b}^{(k)}]$ uniformly at random.

See Figure 4 for an illustration. Intuitively, all rider requests are first dispatched to drivers in the first bin $[\underline{b}^{(1)}, \bar{b}^{(1)}]$ uniformly at random. If a dispatch is declined, the mechanism will then dispatch the trip to a random driver in the next bin. When all drivers are straightforward and do not decline trip dispatches, only drivers closest to the head of the queue will receive trip dispatches from the mechanism.

Recall that i^* as defined in (1) is the lowest-earning trip that is (partially) completed under the first best outcome. With a rider patience level of P, each trip may be dispatched a maximum of P times before the rider cancels the trip request. Given any economy, let $(\mathcal{L}^{(1)}, \mathcal{L}^{(2)}, \ldots, \mathcal{L}^{(m)})$ for some $m \leq \min\{i^*,\ P\}$ be an *ordered partition* of the top i^* destinations $\{1, 2, \ldots, i^*\} \subseteq \mathcal{L}$, i.e.

(i) for all $k = 1, \ldots, m$, $\mathcal{L}^{(k)} \subseteq \{1, 2, \ldots, i^*\}$ and $\mathcal{L}^{(k)} \neq \emptyset$,

(ii) for all $k_1, k_2 \leq m$ s.t. $k_1 \neq k_2$, $\mathcal{L}^{(k_1)} \cap \mathcal{L}^{(k_2)} = \emptyset$, and

(iii) for all $k_1, k_2 \leq m$ s.t. $k_1 < k_2$, we have $i < j$ for all $i \in \mathcal{L}^{(k_1)}$ and all $j \in \mathcal{L}^{(k_2)}$.

Condition (iii) implies that trips in an earlier partition has higher earnings than those in a later partition. Given an economy and any ordered partition $(\mathcal{L}^{(1)}, \ldots, \mathcal{L}^{(m)})$ of the top i^* destinations, we construct a corresponding set of m bins in the queue $([\underline{b}^{(1)}, \bar{b}^{(1)}],\ \ldots, [\underline{b}^{(m)},\ \bar{b}^{(m)}])$ as follows:

$$\underline{b}^{(k)} \triangleq \sum_{i \in \cup_{k'<k} \mathcal{L}^{(k')}} \left(w_i - \min_{i' \in \mathcal{L}^{(k)}} \{w_{i'}\} \right) \mu_i / c, \tag{13}$$

$$\bar{b}^{(k)} \triangleq \sum_{i \in \cup_{k' \leq k} \mathcal{L}^{(k')}} \left(w_i - \min_{i' \in \mathcal{L}^{(k)}} \{w_{i'}\} \right) \mu_i / c. \tag{14}$$

In Lemma 4 in Appendix A.4 we show that $\underline{b}^{(1)} = 0$, $\bar{b}^{(k)} \geq \underline{b}^{(k)} \geq 0$ for each $k \leq m$, and $\underline{b}^{(k+1)} \geq \bar{b}^{(k)}$ for all $k \leq m-1$. This guarantees that the bins start from the head of the queue, are well defined, and do not overlap with each other.

We now present the main result of this paper, that the family of randomized FIFO mechanisms constructed in this way achieves the optimal steady state outcome in Nash equilibrium.

Theorem 3 (Optimality of randomized FIFO)**.** *For any economy, and any ordered partition of the top i^* destinations $(\mathcal{L}^{(1)}, \ldots, \mathcal{L}^{(m)})$ with $m = \{i^*, P\}$, a randomized FIFO mechanism corresponding to (13) and (14) is optimal. It achieves the first best trip throughput and the second best net revenue in Nash equilibrium in steady state, and when $c_p = 0$, the net revenue is also the first best.*

We provide the proof of this theorem in Appendix A.4. Briefly, let Q^* denote the steady state equilibrium queue length under the direct FIFO mechanism. We first show that under a randomized FIFO mechanism, it is a Nash equilibrium in steady state that (i) all drivers in the k^{th} bin in the queue accept all and only trips in the top k partitions $\cup_{k'=1}^{k} \mathcal{L}^{(k')}$ (with some potential randomization over trips to location i^*), (ii) after joining the queue, no driver leaves the queue without a rider trip, or rejoins the queue at its tail, (iii) drivers join the queue with probability $\min\{1,\ \sum_{i \in \mathcal{L}} \mu_i / \lambda\}$ upon arrival, and (iv) the length of the queue remains constant at Q^*. In this equilibrium outcome, all trips that are completed under the first best are also completed, so that the randomized FIFO and direct FIFO mechanisms complete the same set of trips in steady sate. The queue lengths being the same then implies that randomized FIFO also achieves the optimal net revenue, given the optimality of the direct FIFO mechanism we established in Theorem 2.

More formally, let $\pi^*(q) \triangleq \pi(q, Q^*, \sigma^*, \sigma^*)$ be the continuation payoff of a driver at position $q \in [0, Q^*]$ in the queue, when the queue length is Q^* and when all drivers adopt the strategy σ^* described above. We show that $\pi^*(q)$ is non-negative, piece-wise linear, and monotonically non-increasing in q. Moreover, $\pi^*(q) = \min_{i \in \mathcal{L}^{(k)}} \{w_i\}$ for all $q \in [\underline{b}^{(k)}, \bar{b}^{(k)}]$ for each $k \leq m$, i.e. the continuation earning of any driver in the k^{th} bin is equal to the net earning of the lowest earning trip in the k^{th} partition. This allows us to prove by induction on k that σ^* forms a Nash equilibrium.

The non-negativity and monotonicity of $\pi^*(q)$ also imply that the randomized FIFO mechanism is *individually rational* and *envy free in Nash equilibrium in steady state.*

We now revisit the economy in Example 1, and demonstrate that the randomized FIFO mechanism achieves a substantial reduction of the variability in drivers' earnings in comparison to dispatching trips to all drivers in the queue uniformly at random.

Example 2. Consider an economy with three destinations $\mathcal{L} = \{1, 2, 3\}$, rider arrival rates $\mu_1 = 1$, $\mu_2 = 6$, $\mu_3 = 3$, and net earnings from trips $w_1 = 75$, $w_2 = 25$, $w_3 = 15$. The opportunity costs per minute are $c = c_p = 1/3$, and drivers arrive at a rate of $\lambda = 8$ per minute, higher than the arrival rate of drivers in Example 1. Moreover, we assume that riders have a patience level of $P = 2$, so that each trip can be dispatched only twice before riders cancel their requests.

Strict FIFO dispatching. Trips to locations 2 and 3 cannot reach drivers in the queue who are willing to accept them. $T_{\text{strict}} = 1$ as a result. Moreover, $R_{\text{strict}} = 0$ since queue is long enough that drivers get a payoff of zero regardless of whether they had joined the queue, and when $c_p = c$ the platform's net revenue is equal to the total payoff of all drivers who arrived at the queue.

Random dispatching. When the length of the queue is $Q^* = 360$, it is an equilibrium for all drivers to accept all trips to locations 1 and 2, and randomize on trips to location 3. In steady state, all location 1 and 2 trips, and a third of location 3 trips are completed. The throughput is $T_{\text{rand}} = 8$, and the platform achieves a net revenue of $R_{\text{rand}} = 120$ per minute. The average waiting time in the queue is $Q^*/T_{\text{rand}} = 45$ minutes, and the drivers get an average total payoff of 15. However, due to the high level of uncertainty in (i) a driver's waiting times for a trip, and (ii) the net earnings from the trip a driver accepts, the variance of drivers' total payoffs is 500 (see Appendix B.4 for the computation of earning variance).

Randomized FIFO. Consider a randomized FIFO mechanism corresponding to the ordered partition $(\mathcal{L}^{(1)}, \mathcal{L}^{(2)}) = (\{1\}, \{2, 3\})$. The corresponding bins are given by $\underline{b}^{(1)} = \bar{b}^{(1)} = 0$, $\underline{b}^{(2)} = 180$, and $\bar{b}^{(2)} = 360$. All trips are first sent to drivers in $[\underline{b}^{(1)}, \bar{b}^{(1)}] = \{0\}$, i.e. the head of the queue. In equilibrium, trips to location 1 will be accepted immediately, and the declined trips to locations 2 and 3 will then be randomly dispatched to drivers at positions 180 to 360 in the queue.

In equilibrium, the length of the queue is $Q^* = 360$, and randomized FIFO achieves the same trip throughput, net revenue, average driver waiting time, and average driver payoff, as those under random dispatching. In contrast, the variance of the total payoffs of the drivers is only 75 (see Appendix B.5 for more details). By matching higher earning trips with drivers in earlier bins who have incurred a higher waiting cost, the randomized FIFO mechanism is able to substantially reduces the variability in drivers' total payoffs. □

For an economy where riders' patience level is higher than the number of trips types completed, (i.e. when $i^* \leq P$), the randomized FIFO mechanism corresponding to $m = i^*$ partitions has a single trip in each partition: $\mathcal{L}^{(k)} = \{k\}$ for each $k \in \mathcal{L}$ and $k \leq i^*$. In this case, $\bar{b}^{(k)} = \underline{b}^{(k)} = n_k$ for each $k \leq i^*$, meaning that for the k^{th} time each trip is dispatched, the trip will be offered to the driver at position n_k in the queue. In equilibrium, trips to each location $k \leq i^*$ are accepted by the drivers at n_k, and the equilibrium outcome reduces to that under direct FIFO. This implies that when riders are sufficiently patient, the randomized FIFO mechanism achieves zero variance in drivers' total payoffs.

4.1 Discussion

Under the randomized FIFO mechanisms, the small variance in drivers' total payoffs and the envy-freeness of the equilibrium outcome (which guarantees that no driver will want to swap positions

with any other driver who joined the queue later in time) can both be considered as fairness properties for *the equilibrium outcome* (see Avi-Itzhak and Levy [2004], Platz and Østerdal [2017], and Wierman [2011]). These are important considerations in practice, and complement what we adopt as the fairness of *the dispatch rule*, i.e. when drivers are non-strategic, drivers who arrived earlier in time will receive trips to every destination at a higher rate.

With strategic drivers, in equilibrium, it is possible for drivers in later bins to receive certain low-earning trips at a higher rate than drivers in an earlier bin, after these trips are first dispatched to and declined by drivers in the earlier bins— improving reliability and efficiency requires that the mechanism dispatch lower-earning trips to drivers further down the queue who are willing to accept them before riders' patience runs out.[14] In practice, it is appealing that the fairness of a dispatch rule depends only on its design, but not the analysis of the equilibrium outcome in every economy.

Our mechanisms are also appealing for practice since drivers are provided trip destination and earnings information upfront, as well as the flexibility to accept or decline any trip dispatches. It is tempting to think that a mechanism that imposes penalties could easily achieve a better outcome. However, the proof of Theorem 2 implies that even if a mechanism is allowed to disincentivize declines such that drivers would lose their positions in the queue after declining trip offers,[15] no such mechanism can achieve a better throughput or revenue than the randomized FIFO mechanisms. Intuitively, all trips that are accepted under the first best outcome are also accepted under randomized FIFO, and the only inefficiency arises due to drivers' strategic waiting in the queue. Reducing the length of the driver queue (relative to the equilibrium queue length under randomized FIFO) leads to a smaller opportunity cost for the platform, but at the same time reduces drivers' cost of being moved to the tail of the queue. This renders such penalties less effective. All things considered, the randomized FIFO mechanism achieves a desirable balance between flexibility, efficiency, fairness and variability in drivers' earnings.

5 Simulation Results

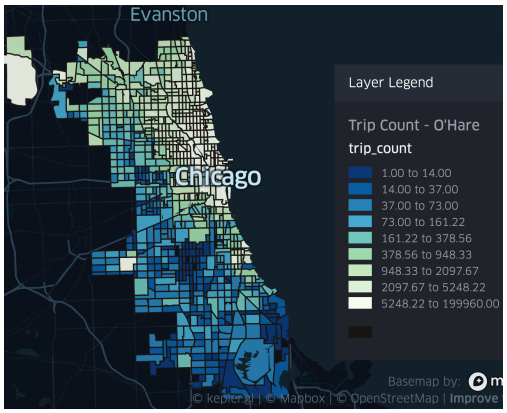
In this section, we present counterfactual simulation results for the Chicago O'Hare International Airport. As we vary the level of driver supply or rider patience, we compare various mechanisms and benchmarks in the equilibrium net revenue, trip throughput, queue length, and drivers' average waiting time, average earnings, and earning variance. Additional simulation results for O'Hare and for the Chicago Midway International Airport are provided in Appendix D.

To estimate the distribution of trips and the net earnings from trips by destination, we make use of trip-level data from ridesharing platforms (including Uber and Lyft) made public by the City of Chicago.[16] The dataset provides the fare for each trip (rounded to the nearest $2.50), the origin

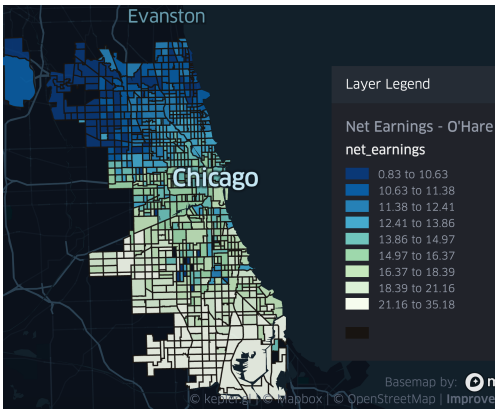
[14] There may also be segments in the queue where the drivers do not receive any dispatches under the randomized FIFO mechanisms. It is possible for $\bar{b}^{(k-1)} < \underline{b}^{(k)}$, meaning that a driver who have just moved past the k^{th} bin may need to wait for some time before she reaches the $k-1^{\text{th}}$ bin, and in the mean time will not receive any dispatches. The existence of such segments in between bins is to guarantee that a driver who is about to reach $\underline{b}^{(k)}$ is not getting a continuation payoff that is too high such that the driver will decline the lower-earning trips in $\mathcal{L}^{(k)}$.

[15] In practice, Uber and Lyft drivers may lose their positions in line and move back to the tail of the queue after declining (multiple) trip dispatches. See `https://help.lyft.com/hc/en-us/articles/115012922787-Receiving-Airport-FIFO-pickup-requests` and `https://www.uber.com/us/en/drive/san-francisco/airports/san-francisco-international/`, accessed 09/27/2020. Such penalties can improve the outcome under strict FIFO dispatching.

[16] `https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p`, accessed December 12, 2020. This dataset contains all trips in Chicago from November 2018 through September 2020, but we use data up to mid March of 2020, before the COVID-19 pandemic substantially changed the dynamics



(a) Trip count by destination.



(b) Net earnings by destination.

Figure 5: Trip volume and estimated net earnings (assuming $c = 1/3$) by destination Census Tract in Chicago, for trips originating from the Chicago O'Hare International Airport.

and destination of each trip on the Census Tract level, as well as the timestamps at the beginning and the end of each trip (rounded to the nearest 15 minutes). There are a total of 801 census tracts within the City of Chicago, which we consider as the set of destinations.[17]

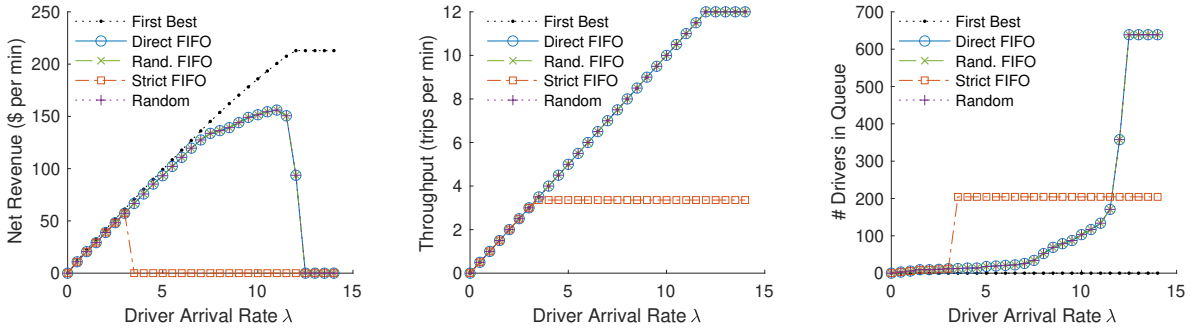
From November 1, 2018 to March 11, 2020, there are a total of 4.53 million trips originating from O'Hare (see Figure 11 in Appendix D.1). The number of trips by destination census tract is as shown in Figure 5a, and the average trip fare by destination is presented in Figure 1a in Section 1. Without driver identifiers, we are unable to estimate the average earnings of a driver in Chicago. We assume throughout this section that the opportunity cost of a driver is $c = 1/3$, representing the scenario that an average driver driving in the city makes \$20 per hour.[18] Combining the average fare, trip duration (see Figure 13a), and the opportunity cost, we estimate the net earnings by destination as shown in Figure 5b. See Appendix C.1 for more details.

Throughout this section, we fix the total arrival rate of riders at $\sum_{i\in\mathcal{L}} \mu_i = 12$ per minute. This is roughly equal to the rate of completed trips during early evening hours on weekdays (see Figure 12 in Appendix D.1 for the average number of completed trips by hour-of-week). We assume that the platform's opportunity cost of drivers' time is $c_p = c = 1/3$ per minute, which corresponds to the scenario where (i) the net revenue of the platform is equal to the total payoff of all drivers who arrived at the queue, and (ii) the gap between the first best and the second best net revenue is the largest. Finally, the randomized FIFO mechanism we evaluate in this section corresponds to an ordered partition of the set of completed trips into at most P subsets, each containing roughly the same number of destination census tracts.

of the market. During this time of consideration, drivers in Chicago do not have trip destinations up front.

[17] `https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Census-Tracts-2010/5jrd-6zik`. Accessed September 14, 2020. The destination census tracts are not available for trips ending outside of the City of Chicago, and may also be hidden to protect riders' privacy when trips are sparse. Overall, 42.6% of trips originating from O'Hare do not have a destination census tract, thus we cannot take these trips into consideration for our simulations. We do not expect any qualitative change in the simulation results, if trips to all destinations are included. In fact, incorporating the long trips to the suburbs with very high earnings (which are currently missing) will likely lead to a worse outcome under strict FIFO, since these trips provide strong incentives for drivers at the head of the queue to wait and cherrypick.

[18] The simulation results are not sensitive to the choice of c. A proposal from Uber in 2019 (see `https://p2a.co/H9gttWA`, accessed September 14, 2020) discussed ensuring drivers are paid an average of \$21 per hour while *on trip*, the earnings per hour online could be currently slightly lower, depending on the average utilization level.



(a) Net revenue. (b) Trip throughput. (c) Equilibrium queue length.

Figure 6: Equilibrium net revenue, trip throughput, and length of the queue in steady state, as the arrival rate of drivers varies. Chicago O'Hare.

Varying Driver Supply. We first compare the different mechanisms and benchmarks as the arrival rate of drivers λ varies from zero to twenty percent over the total rider arrival rates. We fix the rider patience level at $P = 12$, representing the scenario where each driver decline takes 10 seconds on average, and where riders are willing to wait for 2 minutes for a match. Figure 6 presents the equilibrium net revenue, trip throughput, and queue length achieved in steady state.

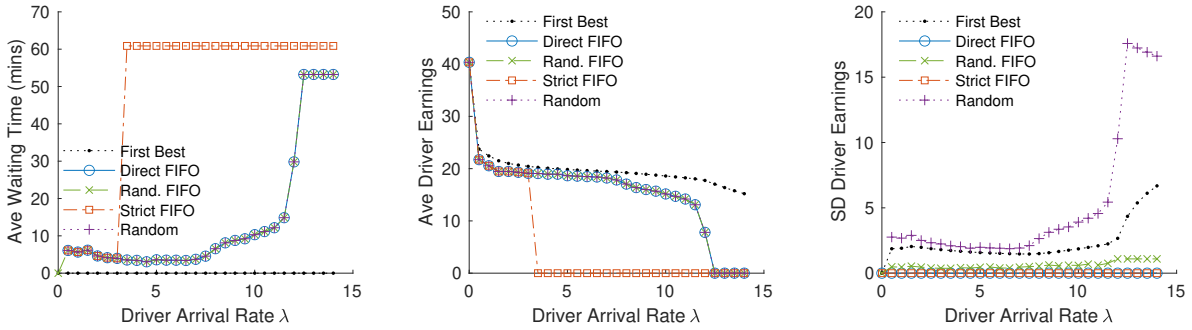
When the arrival rate of drivers is very low, the outcome under direct FIFO, randomized FIFO, strict FIFO and random dispatching coincide, and all mechanisms achieve a net revenue very close to that under the first best outcome. This is because all drivers are able to accept trips with high earnings, and do not spend much time lining up in the queue. As the arrival rate of drivers increases, the length of the queue increases, and so does the gap between the first best and the second best net revenue (which is achieved by direct FIFO, randomized FIFO, and random dispatching).

In contrast to the other mechanisms, the trip throughput under strict FIFO dispatching quickly plateaus despite the increasing driver supply, since rider requests for lower earning trips can no longer reach drivers in the queue who are willing to accept them. These trips become unfulfilled, and at the same time, some drivers will have to deadhead back to the city with an empty car. As a result, the net revenue under strict FIFO drops to zero— drivers will continue to join the queue until the queue is so long that joining is no better than leaving without a rider, thus in equilibrium all drivers who arrived at the queue get a zero total payoff.

Once the queue is over-supplied, i.e. when driver arrival rate exceeds total rider demand, the net revenue under the rest of the mechanisms also drop to zero. This is inevitable, since no driver is willing to leave the airport as long as continuing to wait has a strictly positive payoff. Nevertheless, as is shown in Figure 7a, the average waiting time under the randomized FIFO mechanism is still shorter than that under strict FIFO dispatching, since the trip throughout is substantially higher.

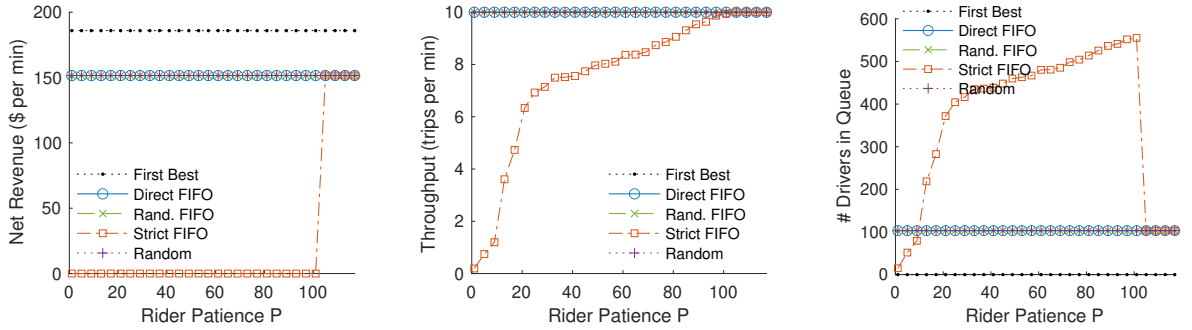
In Figures 7b and 7c, we compare the average payoff (i.e. the net earnings from trips, if any, minus the waiting costs) of all drivers who arrived at the airport, and also the standard deviation of drivers' payoffs. As expected, random dispatching introduces substantial uncertainty in drivers' payoffs. In contrast, by matching higher-earning trips with drivers who have waited longer in the queue, the randomized FIFO mechanism achieves a much smaller variance in drivers' payoffs in comparison to the first best, which effectively forces drivers to accept every trip they receive.

Varying Rider Patience. Fixing the arrival rate of drivers at $\lambda = 10$, we now compare the equilibrium, steady state outcomes under different mechanisms when riders' patience level increases from $P = 1$ to $P = 120$. Figure 8 presents the net revenue, trip throughput, and the length of



(a) Average waiting time. (b) Average driver payoff. (c) SD of driver payoffs.

Figure 7: Drivers' average waiting times, average total payoff, and the standard deviation (SD) in drivers' total payoff in equilibrium in steady state, as the arrival rate of drivers varies. Chicago O'Hare.



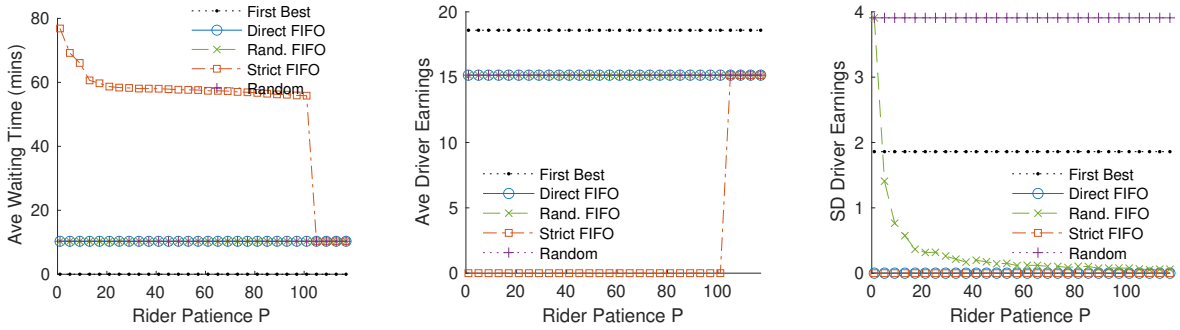
(a) Net revenue. (b) Trip throughput. (c) Equilibrium queue length.

Figure 8: Equilibrium net revenue, trip throughput, and length of the queue in steady state, as riders' patience level varies. Chicago O'Hare.

the queue, and Figure 9 shows drivers' average waiting times in queue, drivers' average payoff after arriving at the queue, and the standard deviation in drivers' payoffs.

The outcomes under the direct FIFO mechanism and random dispatching are not affected by riders' patience levels. Both mechanisms achieve the first best trip throughput, a high net revenue for the platform, and a low waiting time for the drivers. The randomized FIFO mechanism achieves the same throughput, revenue, and driver waiting times. Moreover, we see from Figure 9c that (i) the variance in drivers' total payoffs is substantially lower than that under random dispatching, and (ii) this variance diminished rapidly as riders' patience level increases. Intuitively, riders' patience level P limits the number of times a trip can be dispatched, hence the number of bins a randomized FIFO mechanism may employ. As P increases, the mechanism is able to better match trips with higher earnings with drivers who had waiting the longer time in the queue.

The strict FIFO mechanism, on the other hand, performs poorly. As the patience level increases, more rider trips are able to reach drivers in the queue who are willing to accept them, thus the trip throughput increases. The net revenue of the platform and the average driver payoff remain at zero, however, because drivers continue to join the queue until the payoff from joining is no better than that from leaving without a rider. After P exceeded 100, strict FIFO finally is able to dispatch all drivers that arrive at the airport and join the other mechanisms in achieving the second best net revenue. This is not possible in practice, since even when each driver decline takes only 10 seconds,



(a) Average waiting time. (b) Average driver payoff. (c) SD of driver payoffs.

Figure 9: Drivers' average waiting times, average total payoff, and the standard deviation (SD) in drivers' total payoff in equilibrium in steady state, as riders' patience level varies. Chicago O'Hare.

$P > 100$ requires that riders wait for over fifteen minutes to get matched to a driver.

6 Conclusion

We study the dispatching of trips to drivers in a queue, where some trips necessarily have higher earnings than the others due to operational constraints. We propose a family of randomized FIFO mechanisms, which fairly prioritize drivers who have waited longer in the queue when dispatching trips to all destinations, and provably achieve the highest revenue and throughput achievable by any mechanism that is transparent and flexible. Extensive counterfactual simulations demonstrate substantial improvements of throughput and revenue in comparison to the *status quo* strict FIFO dispatching, and highlight the effectiveness of using drivers' waiting times in the queue to align incentives, improve efficiency and reliability, and reduce the variability in driver earnings.

From a technical perspective, our setting generalizes existing work in the literature by modeling rider impatience, and endogenizing drivers' decision to join, leave, or re-join the platform. The randomized FIFO mechanisms we propose are also appealing for practice since drivers are provided trip destination and earnings information upfront, as well as the flexibility to freely accept or decline any dispatches. Even when a mechanism is allowed to disincentivize rejection such that drivers would lose their position in the queue after declining a dispatch (i.e., moving back to the tail of the queue), no such mechanism can achieve a better throughput or revenue. All things considered, the randomized FIFO mechanism achieves a desirable balance between efficiency, flexibility, fairness and the inequity in drivers' earnings.

References

Philipp Afeche, Zhe Liu, and Costis Maglaras. Ride-hailing networks with strategic drivers: The impact of platform control capabilities on performance. *Columbia Business School Research Paper*, (18-19):18–19, 2018.

Ali Aouad and Ömer Saritaç. Dynamic stochastic matching under limited time. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 789–790, 2020.

Nick Arnosti and Peng Shi. Design of lotteries and wait-lists for affordable housing allocation. *Management Science*, 66(6):2291–2307, 2020.

Itai Ashlagi, Maximilien Burq, Chinmoy Dutta, Patrick Jaillet, Amin Saberi, and Chris Sholley. Maximum weight online matching with deadlines. *arXiv preprint arXiv:1808.03526*, 2018.

Itai Ashlagi, Faidra Monachou, and Afshin Nikzad. Optimal dynamic allocation: Simplicity through information design. *Available at SSRN*, 2020.

Benjamin Avi-Itzhak and Hanoch Levy. On measuring fairness in queues. *Advances in applied probability*, pages 919–936, 2004.

Siddhartha Banerjee, Ramesh Johari, and Carlos Riquelme. Pricing in ride-sharing platforms: A queueing-theoretic approach. In *ACM Conference on Economics and Computation (EC)*, 2015.

Siddhartha Banerjee, Yash Kanoria, and Pengyu Qian. State dependent control of closed queueing networks. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, 2018.

Omar Besbes, Francisco Castro, and Ilan Lobel. Spatial capacity planning. *Proceedings of the 2019 ACM Conference on Economics and Computation*, 2019.

Omar Besbes, Francisco Castro, and Ilan Lobel. Surge pricing and its spatial supply response. *Management Science*, 2020.

Kostas Bimpikis, Ozan Candogan, and Daniela Saban. Spatial pricing in ride-sharing networks. *Operations Research*, 67(3):744–769, 2019.

Francis Bloch and David Cantala. Dynamic assignment of objects to queuing agents. *American Economic Journal: Microeconomics*, 9(1):88–122, 2017.

Jesper Breinbjerg, Alexander Sebald, and Lars Peter Østerdal. Strategic behavior and social outcomes in a bottleneck queue: experimental evidence. *Review of Economic Design*, 20(3):207–236, 2016.

Juan Camilo Castillo, Dan Knoepfle, and Glen Weyl. Surge pricing solves the wild goose chase. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 241–242, 2017.

Francisco Castro, Peter Frazier, Hongyao Ma, Hamid Nazerzadeh, and Chiwei Yan. Matching queues, flexibility and incentives. *arXiv preprint arXiv:2006.08863*, 2020.

M Keith Chen and Michael Sheldon. Dynamic pricing in a labor market: Surge pricing and the supply of uber driver-partners. *University of California (Los Angeles) Working Paper URL http://citeseerx. ist. psu. edu/viewdoc/download*, 2015.

M Keith Chen, Peter E Rossi, Judith A Chevalier, and Emily Oehlsen. The value of flexible work: Evidence from uber drivers. *Journal of Political Economy*, 127(6):2735–2794, 2019.

Peter Cohen, Robert Hahn, Jonathan Hall, Steven Levitt, and Robert Metcalfe. Using big data to estimate consumer surplus: The case of uber. *NBER Working Paper No. 22627*, 2016.

Cody Cook, Rebecca Diamond, Jonathan Hall, John A List, and Paul Oyer. The gender earnings gap in the gig economy: Evidence from over a million rideshare drivers. Technical report, National Bureau of Economic Research, 2018.

John Dickerson, Karthik Sankararaman, Aravind Srinivasan, and Pan Xu. Allocation problems in ride-sharing platforms: Online matching with offline reusable resources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Nikhil Garg and Hamid Nazerzadeh. Driver surge pricing. *Proceedings of the 21st ACM Conference on Economics and Computation*, 2020.

Jonathan Hall, Cory Kendrick, and Chris Nosko. The effects of uber's surge pricing: A case study. Technical report, The University of Chicago Booth School of Business, 2015.

Jonathan V Hall and Alan B Krueger. An analysis of the labor market for uber's driver-partners in the united states. *NBER Working Paper No. 22843*, 2016.

Jonathan V Hall, John J Horton, and Daniel T Knoepfle. Labor market equilibration: Evidence from uber. Technical report, New York University Stern School of Business, 2017.

Refael Hassin. On the optimality of first come last served queues. *Econometrica*, 53(1):201–202, 1985.

Refael Hassin. *Rational queueing*. CRC press, 2016.

Yash Kanoria and Pengyu Qian. Blind dynamic resource allocation in closed networks via mirror backpressure. In *EC'20: Proceedings of the 21st ACM Conference on Economics and Computation*, 2020.

Jacob Leshno. Dynamic matching in overloaded waiting lists. *Available at SSRN 2967011*, 2019.

Alice Lu, Peter I Frazier, and Oren Kislev. Surge pricing moves uber's driver-partners. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 3–3, 2018.

Hongyao Ma, Fei Fang, and David C Parkes. Spatio-temporal pricing for ridesharing platforms. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 583–583, 2019.

Pinhas Naor. The regulation of queue size by levying tolls. *Econometrica: journal of the Econometric Society*, pages 15–24, 1969.

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

Trine Tornøe Platz and Lars Peter Østerdal. The curse of the first-in–first-out queue discipline. *Games and Economic Behavior*, 104:165–176, 2017.

Zhiwei Qin, Xiaocheng Tang, Yan Jiao, Fan Zhang, Zhe Xu, Hongtu Zhu, and Jieping Ye. Ride-hailing order dispatching at DiDi via reinforcement learning. *INFORMS Journal on Applied Analytics*, 50(5):272–286, 2020.

Xuanming Su and Stefanos Zenios. Patient choice in kidney allocation: The role of the queueing discipline. *Manufacturing & Service Operations Management*, 6(4):280–301, 2004.

Xuanming Su and Stefanos A Zenios. Recipient choice can address the efficiency-equity trade-off in kidney transplantation: A mechanism design model. *Management science*, 52(11):1647–1660, 2006.

Adam Wierman. Fairness and scheduling in single server queues. *Surveys in Operations Research and Management Science*, 16(1):39–48, 2011.

Chiwei Yan, Helin Zhu, Nikita Korolko, and Dawn Woodard. Dynamic pricing and matching in ride-hailing platforms. *Naval Research Logistics (NRL)*, 67(8):705–724, 2020.

Appendix

Appendix A provides proofs omitted from the body of the paper. We derive the equilibrium, steady state outcome under various mechanisms and benchmarks in Appendix B. Additional examples and discussions are provided in Appendix C, and we include in Appendix D more detailed description of the data from the City of Chicago, additional simulation results for O'Hare, as well as simulations for trips originating from the Chicago Midway International Airport.

A Proofs

A.1 Equilibrium Outcome Under Strict FIFO

Before formally stating and proving the equilibrium outcome under strict FIFO dispatching, we first provide the following result on necessary conditions of best-response strategies.

Recall from Section 3.1 that given a mechanism, $\pi(q, Q, \sigma, \sigma')$ denotes the expected continuation payoff (net earnings from trip minus waiting costs) of a driver at position $q \geq 0$ in the queue, when the length of the queue is $Q \geq q$, and when this driver adopts strategy σ while all other drivers employ strategy σ'. Moreover, under a strategy $\sigma = (\alpha, \beta, \gamma)$, $\alpha(q, Q, i)$, $\beta(q, Q)$ and $\gamma(q, Q)$ denote the probability for a driver to (i) accept a trip to location $i \in \mathcal{L}$, (ii) re-joins the queue at the tail, and (iii) leave the queue without a rider, when the length of the queue is $Q \geq 0$ and when the driver is at some position $q \in [0, Q]$.

Lemma 2. Fix a strategy σ^* adopted by the rest of the drivers. $\sigma = (\alpha, \beta, \gamma)$ is a *best-response strategy* only if for any queue length $Q \geq 0$ and at any position in the queue $q \leq Q$,

(i) the driver accepts (or declines) with probability one trips for which the net earnings is strictly above (or below) the continuation payoff, i.e. for all $i \in \mathcal{L}$, $w_i > \pi(q, Q, \sigma, \sigma^*) \Rightarrow \alpha(q, Q, i) = 1$, and $w_i < \pi(q, Q, \sigma, \sigma^*) \Rightarrow \alpha(q, Q, i) = 0$,

(ii) the driver rejoins at the tail of the queue with probability one (or zero) if the continuation payoff at the tail of the queue is strictly higher (or lower), i.e. $\pi(q, Q, \sigma, \sigma^*) < \pi(Q, Q, \sigma, \sigma^*) \Rightarrow \beta(q, Q) = 1$ and $\pi(q, Q, \sigma, \sigma^*) > \pi(Q, Q, \sigma, \sigma^*) \Rightarrow \beta(q, Q) = 0$, and

(iii) the driver leaves the queue without a rider trip with probability one (or zero) if the continuation payoff is strictly negative (or positive), i.e. $\pi(q, Q, \sigma, \sigma^*) < 0 \Rightarrow \gamma(q, Q) = 1$ and $\pi(q, Q, \sigma, \sigma^*) > 0 \Rightarrow \gamma(q, Q) = 0$.

When the length of the queue is Q, a driver at location q who is dispatched a trip to location $i \in \mathcal{L}$ faces the decision of whether to accept the trip and get a continuation payoff of w_i, or to decline the trip and remain in the queue. The continuation payoff from remaining in the queue given strategy σ is $\pi(q, Q, \sigma, \sigma^*)$, thus a best response must satisfy condition (i) in Lemma 2. Similarly, it is easy to see that a violation of either condition (ii) or (iii) leads to a useful deviation that improves the driver's payoff, contradicting the assumption that σ is a best-response strategy.

Condition (i) also implies that an optimal acceptance strategy α must have a cut-off structure, such that for any $Q \geq 0$ and any $q \in [0, Q]$, $\alpha(q, Q, i) > 0$ for location $i \in \mathcal{L}$ implies $\alpha(q, Q, j) = 1$ for all $j < i$, since trips to these destinations have higher net earnings.

Recall that $n_0 \triangleq 0$, and observe that for each $i \geq 2$, n_i as defined in (5) can be rewritten as:

$$n_i \triangleq \sum_{j=1}^{i-1} \left(\frac{w_j - w_{j+1}}{c} \sum_{k=1}^{j} \mu_k \right) = \sum_{j=1}^{i-1} \frac{w_j - w_i}{c} \mu_j. \tag{15}$$

Similarly, $\bar{Q}$ as defined in (11) can be rewritten as:

$$\bar{Q} \triangleq n_L + \frac{w_L}{c} \sum_{i=1}^{L} \mu_i = \sum_{i \in \mathcal{L}} w_i \mu_i / c, \tag{16}$$

and it is straightforward to see that $n_i \leq \bar{Q}$ for all $i \in \mathcal{L}$. We now formally state and prove Lemma 1, on the equilibrium outcome of strict FIFO dispatching when riders have infinite patience and never cancel their trip requests.

Lemma 1 (SPE under strict FIFO with infinite rider patience)**.** Assume that riders are infinitely patient. Under strict FIFO dispatching, it is a subgame-perfect equilibrium for drivers to:[19]

- accept trips to each location $i \in \mathcal{L}$ if and only if the driver is at position $q \geq n_i$ in the queue, and
- join the queue if and only if the length of the queue is weakly below $\bar{Q}$, and never leave the queue or move to the tail after joining.

Proof. Let $\sigma^* = (\alpha^*, \beta^*, \gamma^*)$ be the strategy specified by the lemma, i.e. for any queue length $Q \geq 0$ and any position in the queue $q \in [0, Q]$,

$$\begin{aligned} \alpha^*(q, Q, i) =& \mathbb{1}\{q \geq n_i\}, \\ \beta^*(q, Q) =& 0, \\ \gamma^*(q, Q) =& \mathbb{1}\{q > \bar{Q}\}. \end{aligned}$$

Here, $\mathbb{1}\{\cdot\}$ is the indicator function. $\gamma^*(q, Q) = \mathbb{1}\{q > \bar{Q}\}$ means that the driver will leave (or not join) the queue if and only if the driver's position is (or will be) later than $\bar{Q}$. What we need to show is that starting from any queue length $Q \geq 0$, assuming that the rest of the drivers all adopt strategy σ^*, it is a best response for a driver at any position $q \in [0, Q]$ in the queue to also employ strategy σ^*. We prove this by induction on (segments of) positions in the queue, starting from the head of the queue.

The base case. First, consider the driver at the head of the queue (i.e. at position $q = 0$). The (infinitesimal) driver does not have to wait any time for a dispatch to location 1. The continuation payoff for the driver at $q = 0$ under σ^* is therefore

$$\pi(0, Q, \sigma^*, \sigma^*) = w_1. \tag{17}$$

This is the highest net earnings a driver may get from any trip, thus no other strategy may achieve a higher payoff, and σ^* is a best response for the driver at the head of the queue.

The induction step. Now assume that for some $i \geq 2$, it is a best response for drivers at positions $q \leq n_{i-1}$ in the queue to employ strategy σ^*, and that a driver's optimal continuation payoff starting from position $q = n_{i-1}$ onward is $\pi(n_{i-1}, Q, \sigma^*, \sigma^*) = w_{i-1}$ for any $Q \geq n_{i-1}$. We prove the induction step by showing that:

(i) σ^* is a best response for a driver at any position $q \in (n_{i-1}, n_i]$ in the queue, and

(ii) the optimal continuation payoff from position n_i onward is $\pi(n_i, Q, \sigma^*, \sigma^*) = w_i$.

[19] Note that the strategy prescribed by this lemma is a particular SPE. There exist other strategies that may form an SPE among the drivers, depending on how drivers break ties between alternatives with equal continuation payoffs.

We first compute the continuation payoff for drivers at positions $(n_{i-1}, n_i]$ in the queue, assuming that all drivers adopt strategy σ^*. First, consider a driver at some position $q \in (n_{i-1}, n_i)$. Under σ^*, the driver accepts trips to locations $j < i$, does not leave the queue without a rider trip, or move to the tail of the queue. When all drivers adopt σ^*, trips to locations 1 through $i-1$ are accepted by drivers at or ahead of n_{i-1}, thus the driver at q will not accept any trip she receives from the platform. By Little's Law, a driver at q will wait $(q - n_{i-1})/\sum_{j\le i-1}\mu_j$ units of time before she reaches position n_{i-1} in the queue. By the induction assumption, the driver gets an optimal continuation payoff of w_{i-1} starting from n_{i-1}. As a result, for any $q \in (n_{i-1}, n_i)$,

$$\pi(q, Q, \sigma^*, \sigma^*) = \pi(n_{i-1}, Q, \sigma^*, \sigma^*) - c(q - n_{i-1})/\sum_{j\le i-1}\mu_j = w_{i-1} - c(q - n_{i-1})/\sum_{j\le i-1}\mu_j.$$

Now consider a driver at position $q = n_i$ in the queue. If the driver is dispatched and accepts a trip to location i, she gets w_i. If not, the driver moves forward in the queue and her continuation payoff is again

$$\lim_{q\to n_i-}\pi(q, Q, \sigma^*, \sigma^*) = w_{i-1} - c(n_i - n_{i-1})/\sum_{j\le i-1}\mu_j = w_i.$$

Combining the two cases, we know

$$\pi(q, Q, \sigma^*, \sigma^*) = w_{i-1} - c(q - n_{i-1})/\sum_{j\le i-1}\mu_j,\ \forall q \in (n_{i-1}, n_i], \tag{18}$$

and we have $\pi(q, Q, \sigma^*, \sigma^*) > w_i$ when $q < n_i$ and $\pi(n_i, Q, \sigma^*, \sigma^*) = w_i$.

We now prove that σ^* is a best response for drivers at $(n_{i-1}, n_i]$ in the queue. Assume towards a contradiction, that there exists some strategy σ, and some $q \in (n_{i-1}, n_i]$, s.t. $\pi(q, Q, \sigma, \sigma^*) > \pi(q, Q, \sigma^*, \sigma^*)$ for some $Q \ge q$. Note that the driver does not get dispatched any trip with net earnings higher than w_i until the driver reaches position n_{i-1} in the queue. Consider the following two scenarios:

- If the driver left the queue (with or without a rider) under σ before she reaches n_{i-1}, the driver's payoff is upper-bounded by $w_i \le \pi(q, Q, \sigma, \sigma^*)$.
- When the driver did reach n_{i-1} under σ, her highest continuation payoff from n_{i-1} onward is at most w_{i-1} given the induction assumption. Moreover, the driver will incur a waiting cost of at least $c(q - n_{i-1})/\sum_{j\le i-1}\mu_j$ before reaching n_{i-1} thus driver's continuation payoff starting from q is again upper bounded by $\pi(q, Q, \sigma^*, \sigma^*)$.

Combining the two cases, we know that $\pi(q, Q, \sigma, \sigma^*) > \pi(q, Q, \sigma^*, \sigma^*)$ is not achievable for any $q \in (n_{i-1}, n_i]$ under any strategy σ, thus σ^* is a best response. This completes the proof of the induction step.

End of the queue. What we have proved by induction is that σ^* is a best response for any driver at positions $q \in [0, n_\ell]$ in the queue, and that the optimal continuation payoff from n_ℓ onward under any strategy is w_ℓ. Now consider drivers at positions $q \in (n_\ell, \bar{Q}]$ in the queue. Following σ^* implies waiting until reaching n_ℓ in the queue, thus the continuation payoff is:

$$\pi(q, Q, \sigma^*, \sigma^*) = w_\ell - c(q - n_\ell)/\sum_{j\le \ell}\mu_j \ge w_\ell - c(\bar{Q} - n_\ell)/\sum_{j\le \ell}\mu_j = 0,\ \forall q \in (n_\ell, \bar{Q}]. \tag{19}$$

An argument very similar to the proof of the induction step shows that regardless of whether a driver reached n_ℓ in the queue or not, achieving continuation payoff higher than $\pi(q, Q, \sigma^*, \sigma^*)$ is not possible and σ^* is a best response. Moreover, $\pi(\bar{Q}, \bar{Q}, \sigma^*, \sigma^*) = 0$ holds, thus it is a best response to leave (or not join) the queue when at position $q > \bar{Q}$ (or when the length of the queue is longer than $\bar{Q}$). This completes the proof of this lemma. □

A.2 Incentive Compatibility and Optimality of Direct FIFO

In this section, we provide proofs for the incentive compatibility and optimality of the direct FIFO mechanism.

Theorem 1 (Incentive compatibility of direct FIFO)**.** *It is a subgame-perfect equilibrium for drivers to accept al dispatches from the direct FIFO mechanism, and to join the queue if and only if the length of the queue is at most*

$$\bar{Q} \triangleq n_\ell + \frac{w_\ell}{c} \sum_{i \in \mathcal{L}} \mu_i. \tag{11}$$

Moreover, the equilibrium outcome is individually rational and envy-free.

Proof. Let σ^* denote the strategy of (i) always accepting trip dispatches from the direct FIFO mechanism, (ii) join the queue if and only if the length of the queue is $Q \le \bar{Q}$, and (iii) once in the queue, never leave the queue without a rider or move to the tail of the queue.

To establish the incentive compatibility of the direct FIFO mechanism, we need to show that starting from any queue length $Q \ge 0$, and assuming the rest of the drivers all adopt strategy σ^*, it is a best response for a driver to also employ strategy σ^*. This can be established using a very similar (and in fact, slightly simpler) argument as in the proof of Lemma 1. We do not repeat the same arguments here but refer the readers to Appendix A.1, where we established the SPE under strict FIFO (assuming infinite rider patience) by induction on segments of the queue.

It is also straightforward to show that the equilibrium continuation payoff under direct FIFO is also identical to that under strict FIFO dispatching where riders are infinitely patient. Combining equations (17), (18) and (19), we have the following expression for the equilibrium continuation payoff of a driver at position $q \in [0, Q]$ in the queue for any queue length $Q \ge 0$ under the direct FIFO mechanism:

$$\pi(q, Q, \sigma^*, \sigma^*) = \begin{cases} w_1, & \text{if } q = 0, \\ w_{i-1} - c(q - n_{i-1}) / \sum_{j \le i-1} \mu_j, & \text{if } q \in (n_{i-1}, n_i],\ \forall i \ge 2, \\ w_\ell - c(q - n_\ell) / \sum_{j \in \mathcal{L}} \mu_j, & \text{if } q \in (n_\ell, \bar{Q}], \\ 0, & \text{if } q > \bar{Q}. \end{cases} \tag{20}$$

For any $q \le \bar{Q}$, $\pi(q, Q, \sigma^*, \sigma^*)$ is non-negative, continuous, and monotonically decreasing in q, thus the direct FIFO mechanism is individually rational and envy-free (i.e., no drivers envies the other drivers in positions behind her in the queue).

Also observe that there is no randomness at all in a driver's continuation payoff starting from any position in the queue, since at each point n_i, the driver gets precisely w_i regardless of whether the driver accepted a trip to location i and left, or if the driver moved forward in the queue. As a result, the individual rationality and envy-freeness properties also hold *ex post*. This completes the proof of the theorem. □

Theorem 2 (Optimality of direct FIFO)**.** *For every economy, the direct FIFO mechanism achieves in SPE the first best steady state trip throughput. Moreover, the equilibrium outcome achieves the first best net revenue when $c_p = 0$, and the second best net revenue when $c_p \in (0, c]$.*

Proof. Let σ^* denote the equilibrium strategy of accepting all dispatches from the direct FIFO mechanism, and joining the queue if and only if the length of the queue is at most $\bar{Q}$ (see Theorem 1). We prove the optimality of direct FIFO with the following three steps:

Step 1. Establish the steady state equilibrium outcome under when all drivers adopt strategy σ^*, and prove that the same set of trips that are completed under the first best outcome are also completed under direct FIFO.

Step 2. Show that in equilibrium, no transparent and flexible mechanism is able to achieve a higher total payoff (per unit of time) for all drivers who arrive at the queue, and

Step 3. Complete the proof that no mechanism is able to achieve a better net revenue.

We start from Step 1.

Step 1. We first establish the steady state equilibrium outcome under the direct FIFO mechanism. There are two cases, depending on whether the platform is over or under-supplied.

Step 1.1: $\lambda > \sum_{i\in\mathcal{L}} \mu_i$. We first show that in the over-supplied case, when all drivers adopt strategy σ^*, $Q^* = \bar{Q}$ is a steady-state queue length. To prove this, first observe that with $Q^* = \bar{Q} \geq n_i$ for all $i \in \mathcal{L}$, all rider trip requests are accepted. The rate at which drivers are dispatched from the queue is $\sum_{i\in\mathcal{L}} \mu_i < \lambda$, thus drivers effectively join the queue with probability $\sum_{i\in\mathcal{L}} \mu_i/\lambda$ and the length of the queue remains constant at $Q^* = \bar{Q}$. Observe that the total payoff achieved by all drivers who arrive at the queue is zero, because a driver gets a zero payoff regardless of whether she join the queue upon arrival, or left immediately without joining.

We also show that $Q^* = \bar{Q}$ is the unique steady queue length, and that starting from any queue length $Q \neq \bar{Q}$, the length of the queue will converge to $\bar{Q}$ within a finite amount of time. First, we know from (20), we know that the equilibrium continuation payoff of a driver at any position $q < \bar{Q}$ is strictly positive. If the length of the queue Q is strictly shorter than $\bar{Q}$, a driver strictly prefers to join the queue upon arrival, and drivers join the queue at a rate of λ under σ^*. This cannot be the steady state outcome, since the rate at which drivers are dispatched from the queue is at most $\sum_{i\in\mathcal{L}} \mu_i$, and even lower when $Q < n_\ell$. As a result, the queue length will grow at a rate of at least $\lambda - \sum_{i\in\mathcal{L}} \mu_i$, whenever $Q < \bar{Q}$. Moreover, any queue length $Q > \bar{Q}$ cannot be an steady sate either, since (20) implies that the drivers at positions $q > \bar{Q}$ have negative payoffs, thus will leave the queue immediately.

Step 1.2: $\lambda \leq \sum_{i\in\mathcal{L}} \mu_i$. Recall that when a platform is not over-supplied, $i^* \in \mathcal{L}$ as defined in (1) denotes the lowest-earning and highest index trip that is (partially) completed under the first best outcome, when the λ units of drivers are dispatched to destinations in decreasing order of w_i.

We first show that $Q^* = n_{i^*}$ is a steady state equilibrium queue length. When the length of the queue is n_{i^*}, all trips to locations $i < i^*$ will be dispatched and accepted by drivers in the queue. $\sum_{i<i^*} \mu_i$ out of the λ drivers move forward in the queue upon arrival, and the remaining $\lambda - \sum_{i<i^*} \mu_i$ drivers leave the queue immediately with trips to location i^* that are dispatched to the tail of the queue $Q^* = n_{i^*}$. In this way, the length of the queue remains at n_{i^*}.

Observe that the set of trips completed in steady state under direct FIFO is the same as those completed under the first best outcome. Moreover, a driver gets a payoff of w_{i^*} regardless of whether the driver accepted a trip to location i^* immediately after arrival. As a result, the total payoff of all drivers combined is λw_{i^*} per unit of time.

We also show that $Q^* = n_{i^*}$ is the unique steady state queue length for all non-degenerate economies, meaning that $\lambda \neq \sum_{j=1}^{i} \mu_i$ for any $i \in \mathcal{L}$. Consider the following two scenarios:

- When the length of the queue is $Q < n_{i^*}$, trips to locations $j \geq i^*$ are not dispatched under the direct FIFO mechanism. The excess drivers, however, will still join the queue under σ^*, since at $Q < \bar{Q}$ the payoff from joining is strictly positive. As a result, the length of the queue will grow at a rate at least $\lambda - \sum_{j<i^*} \mu_j$, as long as it is strictly below n_{i^*}.
- When $Q > n_{i^*}$, all trips to locations $j \leq i^*$ are dispatched and accepted under direct FIFO. As a result, the length of the queue will decrease at a rate of $\sum_{j\leq i^*} \mu_j - \lambda$ when $\lambda < \sum_{j\leq i^*} \mu_j$, until it reaches $Q = n_{i^*}$. In the degenerate case where $\lambda = \sum_{j\leq i^*} \mu_j$, any queue length between n_{i^*} and n_{i^*+1} may be a steady state queue length, and we break ties in favor of shorter queues under the direct FIFO mechanism.

Combining the two settings in Step 1.1 and 1.2, we know that the same set of trips that are completed under the first best outcome are also completed under direct FIFO. This implies that the direct FIFO mechanism achieves in equilibrium the first best steady state trip throughput of $T_{\text{direct}} = \min\{\sum_{i\in\mathcal{L}} \mu_i, \lambda\}$. Moreover, when $c_p = 0$, the outcome under direct FIFO also achieves the first best revenue, since the total net earnings from trips is the same as that under the first best, and drivers' lining up in the queue is not costly for the platform.

Step 2. We now prove that it is not possible to improve the total payoff of all drivers who had arrived at the queue, when drivers have access to trip destinations upfront, and can freely decline trips or re-join the queue at the tail at any point of time. Again we discuss the under-supplied and the over-supplied case s separately.

Step 2.1: $\lambda > \sum_{i\in\mathcal{L}} \mu_i$. We need to prove that in equilibrium, drivers always get an average payoff of zero after arriving at the queue. To show this, consider any mechanism $\mathcal{M}$ that is flexible and transparent. It cannot be a steady state equilibrium under $\mathcal{M}$ for every driver to leave the queue with a rider trip, because with $\lambda > \sum_{i\in\mathcal{L}} \mu_i$, the length of the queue will continue to grow. As a result, some driver must willingly leave without a rider trip, and the net payoff of such drivers is non-positive.

Assume towards a contradiction that $\mathcal{M}$ achieves a strictly positive average driver payoff, and let σ' and Q' denote the equilibrium strategy under $\mathcal{M}$, and the steady state queue length under $\mathcal{M}$, respectively. The expected continuation payoff of a driver who joined the queue at the tail must be strictly positive: $\pi_{\mathcal{M}}(Q', Q', \sigma', \sigma') > 0$. This is because the drivers who did not join the queue upon arrival (if any) have zero net earnings thus if $\pi_{\mathcal{M}}(Q', Q', \sigma', \sigma') \leq 0$, the average payoff of all drivers who arrived at the queue will be non-positive. $\pi_{\mathcal{M}}(Q', Q', \sigma', \sigma') > 0$, however, contradicts the assumption that the outcome forms an equilibrium, since in this case, no driver will be willing to leave the queue without a rider trip— it is a useful deviation to rejoin the queue at the tail instead, and get a strictly positive payoff.

Step 2.2: $\lambda \leq \sum_{i\in\mathcal{L}} \mu_i$. As we've shown in Step 1.2, in this case drivers have an average payoff of w_{i^*} after arriving at the queue, where i^* is the lowest earning trip that is (partially) completed in equilibrium. What we need to prove is that under any mechanism $\mathcal{M}$ that does not penalize drivers for declining dispatches or rejoining the queue at the tail, the average payoff of a driver who arrived at the queue cannot exceed w_{i^*}.

First, by definition of i^*, it cannot be a steady state equilibrium under $\mathcal{M}$ for every driver who arrive at the virtual queue to leave the queue with a rider trip to a location $j < i^*$. As a result, some driver must leave with a trip to some location $j \geq i^*$, or leave without a rider. In both cases, the driver's continuation payoff after accepting a trip or leaving the queue is upper bounded by w_{i^*}. This cannot form an equilibrium when $\pi_{\mathcal{M}}(Q', Q', \sigma', \sigma') > w_{i^*}$ (since a driver is better off re-joining the queue at the tail instead), therefore $\pi_{\mathcal{M}}(Q', Q', \sigma', \sigma') \leq w_{i^*}$ must hold.

This completes the proof of Step 2.

Step 3. We now prove that no mechanism can achieve a higher net revenue in equilibrium than that under the direct FIFO mechanism. The case of $c_p = 0$ was discussed in Step 1. The case where $c_p = c$ is also straightforward, since in this case the net revenue of the platform is equal to the total payoff of all drivers combined (see discussions in Section 3). As a result, Step 2 implies that no mechanism can achieve a higher net revenue.

What is left to prove is the case where $c_p \in (0, c)$. Consider an alternative mechanism $\mathcal{M}$, and let $\{\tilde{\mu}_i\}_{i\in\mathcal{L}}$ be the rate at which mechanism $\mathcal{M}$ completes trips to each destination. We are going to prove that the net revenue under $\mathcal{M}$ is optimized when the outcome under $\mathcal{M}$ is the same as that under direct FIFO, and we again discuss the over and under-supplied cases separately.

Step 3.1: $\lambda > \sum_{i\in\mathcal{L}} \mu_i$. Given Step 2, drivers get an total payoff of zero under $\mathcal{M}$. Assuming that the equilibrium queue length is $Q^*_{\mathcal{M}}$, we have:

$$\sum_{i\in\mathcal{L}} \tilde{\mu}_i w_i - cQ^*_{\mathcal{M}} = 0. \tag{21}$$

The platform, however, may still get a non-zero net revenue

$$R_{\mathcal{M}} = \sum_{i\in\mathcal{L}} \tilde{\mu}_i w_i - c_p Q^*_{\mathcal{M}} = (c - c_p)Q^*_{\mathcal{M}} \geq 0,$$

and it is straightforward to see that $R_{\mathcal{M}}$ is optimized when $Q^*_{\mathcal{M}}$ is the maximized. With (21), $Q^*_{\mathcal{M}}$ is maximized when $\tilde{\mu}_i = \mu_i$ for all $i \in \mathcal{L}$. This is the same outcome as that under the direct FIFO mechanism, thus no mechanism can achieve a better net revenue.

Step 3.2: $\lambda \leq \sum_{i\in\mathcal{L}} \mu_i$. In this case, drivers get an average payoff of w_{i^*} under direct FIFO, and the equilibrium queue length is $Q^*_{\text{direct}} = n_{i^*}$. Let $T_{\mathcal{M}} \triangleq \sum_{i\in\mathcal{L}} \tilde{\mu}_i$ denote the trip throughput under mechanism $\mathcal{M}$, and let $u^*_{\mathcal{M}}$ be the average equilibrium payoff of drivers achieved under $\mathcal{M}$. Consider the following two cases:

- $T_{\mathcal{M}} < \lambda$, in which case not all drivers are dispatched rider trips in equilibrium under $\mathcal{M}$. An argument very similar to that in Step 2 shows that in this case, the average payoff of a driver who joined the queue upon arrival must be zero, therefore $u^*_{\mathcal{M}} = 0$ must hold. Similar to the over-suppllied setting, we have

  $$\sum_{i\in\mathcal{L}} \tilde{\mu}_i w_i - cQ^*_{\mathcal{M}} = 0,$$

  which implies

  $$R_{\mathcal{M}} = \sum_{i\in\mathcal{L}} \tilde{\mu}_i w_i - c_p Q^*_{\mathcal{M}} = (c - c_p)Q^*_{\mathcal{M}}.$$

  $R_{\mathcal{M}}$ is again optimized when $Q^*_{\mathcal{M}}$ is the longest. For any fixed throughput $T_{\mathcal{M}} = \sum_{i\in\mathcal{L}} \tilde{\mu}_i < \lambda$, the queue length $Q^*_{\mathcal{M}} = \sum_{i\in\mathcal{L}} \tilde{\mu}_i w_i/c$ is maximized when the $T_{\mathcal{M}}$ units of drivers are dispatched to trips in decreasing order of w_i, and this implies that the net revenue $R_{\mathcal{M}} = (c - c_p)Q^*_{\mathcal{M}}$ is upper bounded by:

$$\begin{aligned} R_{\mathcal{M}} \leq & (c - c_p)\left(\sum_{i<i^*} \mu_i w_i/c + (\lambda - \sum_{i<i^*} \mu_i)w_{i^*}/c\right) \\ = & \sum_{i<i^*} \mu_i w_i + (\lambda - \sum_{i<i^*} \mu_i)w_{i^*} - \frac{c_p}{c}\left(\sum_{i<i^*} \mu_i w_i + (\lambda - \sum_{i<i^*} \mu_i)w_{i^*}\right) \end{aligned}$$

This is weakly below the net revenue under direct FIFO, which can be written as:

$$\begin{aligned} R_{\text{direct}} =& \sum_{i<i^*} \mu_i w_i + (\lambda - \sum_{i<i^*} \mu_i) w_{i^*} - c_p n_{i^*} \\ =& \sum_{i<i^*} \mu_i w_i + (\lambda - \sum_{i<i^*} \mu_i) w_{i^*} - \frac{c_p}{c} \left(\sum_{i<i^*} \mu_i (w_i - w_{i^*}) + (\lambda - \sum_{i<i^*} \mu_i)(w_{i^*} - w_{i^*}) \right). \end{aligned}$$

- Consider now the case where $T_{\mathcal{M}} = \lambda$. Drivers' getting an average payoff of $u^*_{\mathcal{M}}$ implies:

$$\sum_{i \in \mathcal{L}} \tilde{\mu}_i w_i - c Q^*_{\mathcal{M}} = \lambda u^*_{\mathcal{M}}. \tag{22}$$

  For each $i \in \mathcal{L}$, denote $\Delta_i \triangleq (w_i - w_{i^*})/c$. The equilibrium queue length can be written as:

$$Q^*_{\mathcal{M}} = \frac{1}{c} \left(\sum_{i \in \mathcal{L}} \tilde{\mu}_i w_i - \sum_{i \in \mathcal{L}} \tilde{\mu}_i u^*_{\mathcal{M}} \right) = \sum_{i \in \mathcal{L}} \tilde{\mu}_i \left(\Delta_i + (w_{i^*} - u^*_{\mathcal{M}})/c \right). \tag{23}$$

  The net revenue under $\mathcal{M}$ is therefore of the form:

$$R_{\mathcal{M}} = \sum_{i \in \mathcal{L}} \tilde{\mu}_i w_i - c_p Q^*_{\mathcal{M}} = \sum_{i \in \mathcal{L}} \tilde{\mu}_i \left(w_i - c_p \Delta_i \right) - \lambda (w_{i^*} - u^*_{\mathcal{M}}) c_p / c. \tag{24}$$

  For the first term in (24), $w_i - c_p \Delta_i = w_i - (w_i - w_{i^*}) c_p / c$ is highest for smaller i with higher w_i. The second term $\lambda (w_{i^*} - u^*_{\mathcal{M}}) c_p / c$ achieves its minimum at $u^*_{\mathcal{M}} = w_{i^*}$ given Step 2. As a result, we know that $R_{\mathcal{M}}$ is optimized when $u^*_{\mathcal{M}} = w_{i^*}$, and when $\tilde{\mu}_i$ is maximized for smallest $i \in \mathcal{L}$ first. This is, again, the same outcome as that under direct FIFO.

This completes the proof of Step 3, and concludes the proof of the optimality of direct FIFO. □

A.3 Optimality of Random Dispatching

Before proving the optimality of random dispatching, we first provide the following lemma on the best response strategy of a driver in a stationary environment.

Lemma 3. Consider a driver in a stationary environment, where she receives trip offers to each location $i \in \mathcal{L}$ at a rate of $\eta_i \geq 0$. The highest achievable net payoff from any feasible strategy is $\max\{\max_{j \in \mathcal{L}} \rho_j,\ 0\}$, where

$$\rho_j \triangleq \left(\sum_{i=1}^{j} w_i \eta_i - c \right) \Big/ \sum_{i=1}^{j} \eta_i. \tag{25}$$

Moreover, j^* is a maximizer of ρ_j if and only if $\rho_{j^*} \leq w_{j^*}$ and $\rho_{j^*} \geq w_{j^*+1}$.

Proof. Lemma 2 implies that any best response strategy on acceptance in this setting must have a cutoff structure, meaning that if the driver accepts a trip to some location $j \in \mathcal{L}$ with non-zero probability, then she must accept any trip to locations $i < j$ with probability 1. Moreover, the driver will decide to leave the queue only if the expected continuation payoff from the optimal acceptance strategy is non-positive. We now show that the highest achievable net payoff under any best-response strategy is this stationary environment is $\max\{\max_{j \in \mathcal{L}} \rho_j,\ 0\}$.

Consider for now a deterministic strategy such that the driver stays in the queue, and accepts all trips to locations 1 through j if offered. We denote this strategy as σ_j. The average net earnings the driver gets from the an average trip she accepts is $\sum_{i=1}^{j} w_i\eta_i / \sum_{i=1}^{j} \eta_i$, and in expectation, the driver will wait $1/\sum_{i=1}^{j} \eta_i$ units of time to receive a trip dispatch she will accept. Therefore, the expected net payoff (i.e. the net earnings from trip a driver accepts minus her expected waiting cost) under strategy σ_j is

$$\sum_{i=1}^{j} w_i\eta_i \Big/ \sum_{i=1}^{j} \eta_i - c \Big/ \sum_{i=1}^{j} \eta_i = \left(\sum_{i=1}^{j} w_i\eta_i - c\right) \Big/ \sum_{i=1}^{j} \eta_i = \rho_j.$$

Among all deterministic strategies s.t. the driver does not leave, the highest achievable net payoff is therefore $\max_{j\in\mathcal{L}} \rho_j$.

The cutoff structure proved by Lemma 2 also implies that the only potentially useful randomization in a driver's acceptance strategy is on the lowest earning trip that is accepted. Consider a strategy where the driver accepts all trips to locations 1 through $j-1$, but accepts location j trips with probability $\theta \in [0,1]$. The expected net payoff in this setting is:

$$\left(\sum_{i=1}^{j-1} w_i\eta_i + w_j\theta\eta_j - c\right) \Big/ \left(\sum_{i=1}^{j-1} \eta_i + \theta\eta_j\right) = \left(\rho_{j-1}\sum_{i=1}^{j-1} \eta_i + w_j\theta\eta_j\right) \Big/ \left(\sum_{i=1}^{j-1} \eta_i + \theta\eta_j\right).$$

This is a weighted average of ρ_{j-1} and w_j, thus can be optimized at $\theta = 0$ (or $\theta = 1$) if $\rho_{j-1} \geq w_j$ (or if $\rho_{j-1} \leq w_j$). Therefore, for a driver who does not choose to immediately leave the queue, the highest achievable net payoff can be achieved by a deterministic acceptance strategy, and the optimal payoff under any acceptance strategy is equal to $\max_{j\in\mathcal{L}} \rho_j$. When this payoff is negative, the driver is better off leaving the queue instead of waiting for any trip dispatches. As a result, a driver's highest possible payoff a driver may achieve in this stationary environment is $\max\{\max_{j\in\mathcal{L}} \rho_j,\ 0\}$.

What is left to show is that j^* is a maximizer of ρ_j if and only $\rho_{j^*} \leq w_{j^*}$ and $\rho_{j^*} \geq w_{j^*+1}$. To prove this, first observe that for any $j > 1$, ρ_j is a weighted average of ρ_{j-1} and w_j:

$$\rho_j = \left(\rho_{j-1}\sum_{i=1}^{j-1} \eta_i + w_j\eta_j\right) \Big/ \sum_{i=1}^{j} \eta_i. \tag{26}$$

This implies (i) when $\rho_j \geq \rho_{j-1}$, it must be the case that $w_j \geq \rho_j \geq \rho_{j-1}$, and (ii) $\rho_j \geq \rho_{j+1} \Rightarrow \rho_j \geq w_{j+1}$. Therefore, if j^* is a maximizer of ρ_j, we must have $\rho_{j^*} \geq \rho_{j^*-1} \Rightarrow w_{j^*} \geq \rho_{j^*}$, and $\rho_{j^*} \geq \rho_{j^*+1} \Rightarrow \rho_{j^*} \geq w_{j^*+1}$.

On the other hand, if $\rho_{j^*} \leq w_{j^*}$ and $\rho_{j^*} \geq w_{j^*+1}$ both hold, we now prove that j^* must be a maximizer of ρ_j. Denote $\hat{j} \in \mathcal{L}$ as the first location for which $\rho_j > w_{j+1}$, i.e.

$$\hat{j} \triangleq \min\{j \in \mathcal{L} \mid \rho_j > w_{j+1}\}. \tag{27}$$

We first claim that ρ_j must be monotonically non-decreasing when $j \leq \hat{j}$, i.e. for all $j < \hat{j}$, $\rho_j \leq \rho_{j+1}$. This is because for any $j < \hat{j}$, $\rho_j \leq w_{j+1}$ holds by definition of $\hat{j}$, thus by (26) we have $\rho_j \leq \rho_{j+1}$. Moreover, given (26) and the fact that w_j is monotonically decreasing, we can prove by a simple induction ($\rho_{\hat{j}} > w_{\hat{j}+1} \Rightarrow \rho_{\hat{j}} > \rho_{\hat{j}+1} > w_{\hat{j}+1} > w_{\hat{j}+2}$ and so on) that (i) ρ_j must be monotonically decreasing for all $j \geq \hat{j}$, i.e. $\forall j \geq \hat{j}$, $\rho_j \geq \rho_{j+1}$, and (ii) $\rho_j > w_{j+1}$ for all $j \geq \hat{j}$. Combining the two cases, we know that $\hat{j}$ is a maximizer of ρ_j.

For j^*, we know from (26) that $\rho_{j^*} \leq w_{j^*} \Rightarrow w_{j^*} \geq \rho_{j^*} \geq \rho_{j^*-1}$, therefore $j^* - 1 < \hat{j}$. Given $\rho_{j^*} \geq w_{j^*+1}$, consider the two possible scenarios.

- If $\rho_{j^*} > w_{j^*+1}$, we must have $j^* \geq \hat{j}$, thus $j^* = \hat{j}$ holds and j^* is a maximizer of ρ_j.
- If $\rho_{j^*} = w_{j^*+1}$, we have $j^* < \hat{j}$. Moreover, (26) implies $\rho_{j^*+1} = \rho_{j^*} = w_{j^*+1} > w_{j^*+2}$, which means $j^*+1 \geq \hat{j}$. As a result, $j^* = \hat{j}-1$, and j^* is still a maximizer of ρ_j because $\rho_{j^*} = \rho_{j^*+1} = \rho_{\hat{j}}$.

This completes the proof of this lemma. □

With Lemma 3 at hand, we now prove the result on the equilibrium outcome under a mechanism that dispatches every trip request to all drivers in the queue, uniformly at random.

Proposition 2 (Optimality of random dispatching)**.** In Nash equilibrium in steady sate, dispatching every trip to all drivers in the queue uniformly at random achieves the first best trip throughput and the second best net revenue. When $c_p = 0$, the equilibrium net revenue is also the first best.

Proof. We prove this result by showing that the equilibrium outcome under random dispatching has the same queue length Q^* as that under direct FIFO, and that the same set of trips that are completed under direct FIFO is also completed under random dispatching. Theorem 2 then implies the same optimality results for random dispatching.

We discuss the over-supplied and under-supplied settings separately.

Case 1: $\lambda > \sum_{i\in\mathcal{L}} \mu_i$. When the platform is over-supplied, we have proved in Theorem 2 that all rider trips are completed under direct FIFO, and that the equilibrium queue length is $Q^* = \bar{Q}$ (as defined in (11)). We now prove that under random dispatching, when the queue length is $\bar{Q}$, it is a Nash equilibrium for drivers to (i) join the queue with probability $\sum_{i\in\mathcal{L}} \mu_i/\lambda$ upon arrival, (ii) accept all trip dispatches while in the queue, and (iii) never move to the tail of the queue or leave the queue after joining.

More formally, we prove that the strategy $\sigma^* = (\alpha^*, \beta^*, \gamma^*)$ defined as follows forms a Nash equilibrium among the drivers when the queue length is $\bar{Q}$:

$$\begin{aligned}
\alpha^*(q, \bar{Q}, i) &= 1, \ \forall i \in \mathcal{L}, \ \forall q \in [0, \ \bar{Q}], \\
\beta^*(q, \bar{Q}) &= 0, \ \forall q \in [0, \ \bar{Q}], \\
\gamma^*(q, \bar{Q}) &= \begin{cases} 0, & \text{if } q < \bar{Q} \\ 1 - \sum_{i\in\mathcal{L}} \mu_i/\lambda, & \text{if } q = \bar{Q}. \end{cases}
\end{aligned}$$

When all drivers adopt strategy σ^*, the length of the queue remains at $\bar{Q}$, since the numbers of drivers who join the queue and who are dispatched from the queue are both $\sum_{i\in\mathcal{L}} \mu_i$ per unit of time. All rider trips are completed, implying the same steady state revenue and trip throughput as those under direct FIFO.

We now prove that σ^* forms a Nash equilibrium among the drivers under random dispatching when the queue length is $\bar{Q}$. First, observe that when the queue length is $\bar{Q}$ and when the rest of the driver adopts σ^*, (i) each rider trip is dispatched once since the probability of declines is zero, and (ii) a driver's position in the queue has no impact on the rate at which she receives dispatches to each destination. This is therefore a stationary setting we have analyzed in Lemma 3. For a driver anywhere in the queue, the rate at which she receives dispatches to each location $i \in \mathcal{L}$ is:

$$\eta_i = \mu_i/\bar{Q}.$$

Recall from (16) that $\bar{Q}$ can be rewritten as $\bar{Q} = \left(\sum_{i=1}^{\ell} w_i\mu_i\right)/c$. Therefore, the expected payoff ρ_j from accepting only trips to locations 1 through j (as defined in (25)) is of the form:

$$\rho_j = \left(\sum_{i=1}^{j} w_i\eta_i - c\right) \Big/ \sum_{i=1}^{j} \eta_i = \left(\sum_{i=1}^{j} w_i\mu_i - c\bar{Q}\right) \Big/ \sum_{i=1}^{j} \mu_i = \left(\sum_{i=1}^{j} w_i\mu_i - \sum_{i=1}^{\ell} w_i\mu_i\right) \Big/ \sum_{i=1}^{j} \mu_i.$$

This implies $\rho_\ell = 0$, and $\rho_j < 0$ for all $j < \ell$. By Lemma 3, we know that the best acceptance strategy is to accept all dispatches, which is aligned with α^*. This also implies that when all drivers adopt σ^*, the continuation payoff of a driver anywhere in the queue is $\pi(q, \bar{Q}, \sigma^*, \sigma^*) = \rho_\ell = 0$.

The drivers' being indifferent towards being in the queue and leaving the queue means that there is no useful deviation from joining the queue with probability $\sum_{i\in\mathcal{L}} \mu_i/\lambda$ (hence the probability of not joining the queue is $\gamma^*(\bar{Q}, \bar{Q}) = 1 - \sum_{i\in\mathcal{L}} \mu_i/\lambda$). Moreover, re-joining the queue at the tail is not useful since a driver's position in the queue has no impact on the rate at which the driver receives trip dispatches. This completes the proof that σ^* forms a Nash equilibrium among the drivers, thus concludes the discussion for Case 1, the over-supplied setting.

Case 2: $\lambda \le \sum_{i\in\mathcal{L}} \mu_i$. In the case without excess drivers, i^* as defined in (1) denotes the index of the lowest-earning trip that is (partially) completed in equilibrium under the direct FIFO mechanism and the first best. We know from Theorem 2 that the equilibrium queue length under direct FIFO is $Q^* = n_{i^*}$, and the drivers complete all trips to locations $j < i^*$, and in each unit of time the drivers also complete $\lambda - \sum_{i=1}^{i^*-1} \mu_i$ out of the μ_{i^*} trips to location i^*.

We now prove that random dispatching achieves the same equilibrium outcome (queue length and set of trips completed). Fix the length of the queue at $Q^* = n_{i^*}$, and consider the strategy $\sigma^* = (\alpha^*, \beta^*, \gamma^*)$ such that for all $q \in [0,\ \bar{Q}]$,

$$\alpha^*(q, \bar{Q}, i) = \begin{cases} 1, & \text{if } i < i^*, \\ 1 - \left(1 - (\lambda - \sum_{i=1}^{i^*-1} \mu_i)/\mu_{i^*}\right)^{1/P}, & \text{if } i = i^*, \\ 0, & \text{if } i < i^*, \end{cases} \tag{28}$$

$$\beta^*(q, \bar{Q}) = 0, \tag{29}$$

$$\gamma^*(q, \bar{Q}) = 0. \tag{30}$$

For simplicity of notation, let $\theta_{i^*} \triangleq 1 - \left(1 - (\lambda - \sum_{i=1}^{i^*-1} \mu_i)/\mu_{i^*}\right)^{1/P}$. When every driver adopts strategy σ^*, each trip to locations $i < i^*$ is dispatched once, the trip to location i^* is dispatched $\sum_{k=1}^{P}(1-\theta_{i^*})^{(k-1)}\theta_{i^*}k + (1-\theta_{i^*})^P P = (1-(1-\theta_{i^*})^P)/\theta_{i^*}$ times, and each trip to locations $i > i^*$ is dispatched P times. Given the queue length $Q^* = n_{i^*}$, the rate at which a driver anywhere in the queue receives trip dispatches to each location is:

$$\eta_i = \begin{cases} \mu_i/n_{i^*}, & \text{if } i < i^*, \\ \mu_i(1-(1-\theta_{i^*})^P)/(\theta_{i^*} n_{i^*}), & \text{if } i = i^*, \\ \mu_i P/n_{i^*}, & \text{if } i > i^*. \end{cases}$$

As we observed in (15), $n_{i^*} = \sum_{i=1}^{i^*-1}(w_i - w_{i^*})\mu_i/c$. For each $j < i^*$, the expected payoff from accepting only the top j trips can be written as:

$$\rho_j = \left(\sum_{i=1}^{j} w_i\mu_i - cn_{i^*}\right) \Big/ \sum_{i=1}^{j} \mu_i = \left(\sum_{i=1}^{j} w_i\mu_i - \sum_{i=1}^{i^*-1}(w_i - w_{i^*})\mu_i\right) \Big/ \sum_{i=1}^{j} \mu_i.$$

This implies that:

$$\rho_{i^*-1} = \left(\sum_{i=1}^{i^*-1} w_i\mu_i - \sum_{i=1}^{i^*-1}(w_i - w_{i^*})\mu_i\right) \Big/ \sum_{i=1}^{i^*-1} \mu_i = w_{i^*}.$$

We know from (26) that $\rho_{i^*} = w_{i^*}$ must hold as well since ρ_{i^*} is a weighted average of ρ_{i^*-1} and w_{i^*}.Moreover, since w_i is strictly decreasing in i, we have $\rho_{i^*-1} < w_{i^*-1}$. Applying Lemma 3,

we know that the highest possible expected payoff a driver may receive in this stationary setting is w_{i^*}, and this can be achieved by accepting all trips to location $i < i^*$, and accepting trips to location i^* with any probability in $[0, 1]$. α^* is therefore an optimal acceptance strategy. It is also straightforward to see that no strategy that involves not joining the queue the queue, and moving to the tail of the queue, or leave the queue without a rider trip, could achieve a higher expected payoff than w_{i^*}, thus σ^* forms a Nash equilibrium when the queue length is $Q^* = n_{i^*}$.

What is left to prove is that the length of the queue remains at $Q^* = n_{i^*}$ when all drivers adopt σ^*. To show this, we prove that the rate at which drivers are dispatched from the queue is equal to λ, the rate at which drivers join the queue. First, all trips to locations $i \le i^* - 1$ are accepted, so we only need to prove that $\lambda - \sum_{i=1}^{i^*-1} \mu_i$ drivers accept trips to location i^* per unit of time. For trips to location i^*, each time a trip is dispatched, it is *not* accepted with probability $\left(1 - (\lambda - \sum_{i=1}^{i^*-1} \mu_i)/\mu_{i^*}\right)^{1/P}$. Thus the probability for the trip to be unfulfilled after P dispatches is $1-(\lambda-\sum_{i=1}^{i^*-1} \mu_i)/\mu_{i^*}$. This implies that the probability for a trip to location i^* to be completed is $(\lambda-\sum_{i=1}^{i^*-1} \mu_i)/\mu_{i^*}$, so that $\lambda-\sum_{i=1}^{i^*-1} \mu_i$ drivers accept trips to location i^* per unit of time. This completes the proof of the under-supplied case, and concludes the proof of this proposition. □

A.4 Optimality of Randomized FIFO

In this section, we prove the optimality of the randomized FIFO mechanisms. We first provide the following lemma, which shows that the bins constructed as in (13) and (14) given any ordered partition are well-defined and not overlapping.

Lemma 4. For any ordered partition $(\mathcal{L}^{(1)}, \mathcal{L}^{(2)}, \ldots, \mathcal{L}^{(m)})$ of the top i^* trips $\{1, 2, \ldots, i^*\}$, the corresponding set of bins satisfies:

(i) $0 \le \underline{b}^{(k)} \le \bar{b}^{(k)}$ for each $k = 1, \ldots, m$, and $\bar{b}^{(k)} = \underline{b}^{(k)}$ if $|\mathcal{L}^{(k)}| = 1$,

(ii) $\bar{b}^{(k-1)} < \underline{b}^{(k)}$ for all $k = 2, 3, \ldots, m$.

Proof. For part (i), $\bar{b}^{(1)} \ge \underline{b}^{(1)} = 0$ trivially holds. For all $k = 2, 3, \ldots, m$, we have

$$\begin{aligned}
&\bar{b}^{(k)} - \underline{b}^{(k)} \\
=&\frac{1}{c}\left(\sum_{i\in\cup_{k'\le k}\mathcal{L}^{(k')}} (w_i - \min_{i'\in\mathcal{L}^{(k)}}\{w_{i'}\})\mu_i\right) - \frac{1}{c}\left(\sum_{i\in\cup_{k'<k}\mathcal{L}^{(k')}} \left(w_i - \min_{i'\in\mathcal{L}^{(k)}}\{w_{i'}\}\right)\mu_i\right) \\
=&\frac{1}{c}\left(\sum_{i\in\mathcal{L}^{(k)}} \left(w_i - \min_{i'\in\mathcal{L}^{(k)}}\{w_{i'}\}\right)\mu_i\right) \\
\ge&0.
\end{aligned}$$

Note that when $\mathcal{L}^{(k)}$ contains a single location, $\underline{b}^{(k)} = \bar{b}^{(k)}$ holds, meaning that for the k^{th} time each trip is dispatched, the trip will be offered to the driver at position $q = \underline{b}^{(k)}$ in the queue. This completes the proof of part (i).

For part (ii), first observe that for any $k > 1$, $\min_{i\in\mathcal{L}^{(k-1)}}\{w_i\} > \min_{i\in\mathcal{L}^{(k)}}\{w_i\}$, since the partition is ordered thus $w_i > w_j$ for all $i \in \mathcal{L}^{(k-1)}$ and all $j \in \mathcal{L}^{(k)}$. As a result,

$$\bar{b}^{(k-1)} = \frac{1}{c}\left(\sum_{i\in\cup_{k'<k}\mathcal{L}^{(k')}} \left(w_i - \min_{i'\in\mathcal{L}^{(k-1)}}\{w_{i'}\}\right)\mu_i\right) < \frac{1}{c}\left(\sum_{i\in\cup_{k'<k}\mathcal{L}^{(k')}} \left(w_i - \min_{i'\in\mathcal{L}^{(k)}}\{w_{i'}\}\right)\mu_i\right) = \underline{b}^{(k)}.$$

This completes the proof of this lemma. □

We now prove the main result of our paper on the optimality of randomized FIFO.

Theorem 3 (Optimality of randomized FIFO)**.** *For any economy, and any ordered partition of the top i^* destinations $(\mathcal{L}^{(1)}, \ldots, \mathcal{L}^{(m)})$ with $m = \{i^*, P\}$, a randomized FIFO mechanism corresponding to (13) and (14) is optimal. It achieves the first best trip throughput and the second best net revenue in Nash equilibrium in steady state, and when $c_p = 0$, the net revenue is also the first best.*

Proof. We first show that given a randomized FIFO mechanism corresponding to an ordered partition of the top i^* locations, under the Nash equilibrium in steady state, (i) the length of the queue is equal to the equilibrium queue length under the direct FIFO mechanism, and (ii) the same set of trips completed under direct FIFO are also completed. Theorem 2 then implies that the equilibrium outcome under randomized FIFO is optimal. We then establish individual rationality and envy-freeness under randomized FIFO by showing that a driver's continuation payoff as a function of the driver's position in the queue is non-negative and monotonically non-increasing.

Recall that i^* (defined in (1)) is the index of the lowest-earning trip that is (partially) completed in equilibrium under direct FIFO. We discuss the following cases:

Case 1. The total number of bins $m = \min\{i^*,\ P\} = 1$, in which case all trips are dispatched to drivers in the first bin. There are again two scenarios:

- Case 1.1 $i^* = 1$, and in which case only trips to location 1 are (partially) completed under the direct FIFO mechanism.
- Case 1.2 $i^* > 1$, but $P = 1$, meaning that riders are impatient, and will cancel their trip request after any driver decline.

Case 2. The number of bins $m = \min\{i^*,\ P\} > 1$, in which case trips may be dispatched multiple times, and we establish the equilibrium result by induction.

Case 1.1: $m = i^* = 1$**.** In this case, there is a single partition under randomized FIFO: $\mathcal{L}^{(1)} = \{1\}$, and we have $\underline{b}^{(1)} = \bar{b}^{(1)} = 0$. As a result, all trips are dispatched (only once) to the driver at the head of the queue. Recall that no driver will decline a trip to location 1, since there is no other trip with better earnings that the driver would like to wait for. Consider two cases:

- When $\lambda \leq \mu_1$, it is straightforward to verify that (i) the length of the queue is zero, and (ii) all drivers accept a trip to location 1 immediately upon arrival, forms a Nash equilibrium among drivers in steady state. This is the same outcome as that under direct FIFO.
- When $\lambda > \mu_1$ but $i^* = 1$, the number of locations must be $\ell = 1$, and all trips are accepted at the head of the queue. The equilibrium outcome is again the same as that under direct FIFO, where it is also the case that all trips are dispatched to and accepted by the driver at the head of the queue. In steady state, drivers join the queue with probability μ_1/λ, all trips are completed, and the length of the queue is $Q^* = \mu_1 w_1/c = \bar{Q}$ (at which point a driver is indifferent toward joining the queue and leaving without a rider trip).

Combining the two cases, we know that when $i^* = 1$, randomized FIFO achieves the same optimal outcome achieved by direct FIFO in equilibrium. Every driver gets a payoff of w_1 when $\lambda \leq \mu_1$, and when $\lambda > \mu_1$, the continuation payoff decreases linearly in the driver's position in the queue at takes value zero at $\bar{Q}$. The equilibrium outcome is, therefore, individually rational and envy-free.

Case 1.2: $i^* > 1$**,** $m = P = 1$**.** With $P = 1$, riders cancel their trip requests after a single driver decline, and all trips are dispatched by the randomized FIFO mechanism to the first (and only) bin of drivers uniformly at random. $\mathcal{L}^{(1)} = \{1, \ldots, i^*$, and the first bin is given by $\underline{b}^{(1)} = 0$ and $\bar{b}^{(1)} = n_{i^*} > 0$. There are two cases, depending on whether the queue is under or over-supplied.

Case 1.2.1: $\lambda \leq \sum_{i\in\mathcal{L}} \mu_i$. Assume that the length of the queue is $Q^* = n_{i^*}$. Under the randomized FIFO mechanism, all trips are randomly dispatched to drivers in $[\underline{b}^{(1)},\ \bar{b}^{(1)}] = [0, n_{i^*}]$, i.e. all drivers in the queue. This is the same scenario as the under-supplied setting under random dispatching, which we analyzed in Case 2 in the proof of Proposition 2. It is straightforward to verify that the same strategy (specified by (28), (29) and (30)) forms a Nash equilibrium in steady state under randomized FIFO, and the equilibrium queue length remains constant at $Q^* = n_{i^*}$. We refer the readers to the proof of Proposition 2, and do not repeat the same arguments here. Drivers at any position $q \in [0, Q^*]$ in the queue has the same continuation payoff $w_{i^*} \geq 0$, thus the equilibrium outcome is individually rational and envy-free.

Case 1.2.2: $\lambda > \sum_{i\in\mathcal{L}} \mu_i$. When the queue is over-supplied, all trips are completed under direct FIFO, and $i^* = \ell$. The randomized FIFO mechanism dispatches each trip once (since $P = 1$) to drivers in $[\underline{b}^{(1)}, \bar{b}^{(1)}] = [0, n_\ell]$ uniformly at random.[20] Consider the strategy $\sigma^* = (\alpha^*, \beta^*, \gamma^*)$:

$$\begin{aligned}
\alpha^*(q, \bar{Q}, i) &= 1,\ \forall i \in \mathcal{L}, \forall q \in [0, \bar{Q}] \\
\beta^*(q, \bar{Q}) &= 0, \forall q \in [0, \bar{Q}] \\
\gamma^*(q, \bar{Q}) &= \begin{cases} 0, & \text{if } q < \bar{Q}, \\ 1 - \sum_{i\in\mathcal{L}} \mu_i/\lambda, & \text{if } q = \bar{Q}. \end{cases}
\end{aligned}$$

Here, $\gamma^*(\bar{Q}, \bar{Q}) = 1 - \sum_{i\in\mathcal{L}} \mu_i/\lambda$ means that the drivers join the queue at the tail $q = \bar{Q}$ with probability $\sum_{i\in\mathcal{L}} \mu_i/\lambda$. It is clear that when σ^* is adopted by all drivers, the length of the queue remains at $\bar{Q}$. We now prove that with $Q^* = \bar{Q}$, it is a Nash equilibrium for all drivers to adopt strategy σ^*. In other words, when the length of the queue is $\bar{Q}$ and when σ^* is adopted by the rest of the drivers, σ^* is a best response strategy for a driver at any $q \in [0, \bar{Q}]$

Let us first consider a driver at some position $q \in [0, n_\ell]$ in the queue. If the driver does not leave the queue or move to the tail of the queue, this again is a stationary environment analyzed in Lemma 3. When every other driver adopts σ^*, every trip is dispatched only once, thus the rate at which a driver receives trip dispatches to each location $i \in \mathcal{L}$ is $\eta_i = \mu_i/n_\ell$. Since $n_\ell = \sum_{i=1}^{\ell-1}(w_i - w_\ell)\mu_i/c = \sum_{i=1}^{\ell}(w_i - w_\ell)\mu_i/c$ (see (15)), a driver's expected payoff from accepting only the top j trips ρ_j can therefore be rewritten as:

$$\begin{aligned}
\rho_j &= \left(\sum_{i=1}^{j} w_i\mu_i - cn_\ell\right) \Big/ \sum_{i=1}^{j} \mu_i = \left(\sum_{i=1}^{j} w_i\mu_i - \sum_{i=1}^{\ell}(w_i - w_\ell)\mu_i\right) \Big/ \sum_{i=1}^{j} \mu_i \\
&= \left(\sum_{i=j+1}^{\ell} (w_\ell - w_i)\mu_i + w_\ell \sum_{i=1}^{j} \mu_i\right) \Big/ \sum_{i=1}^{j} \mu_i = w_\ell + \left(\sum_{i=j+1}^{\ell} (w_\ell - w_i)\mu_i\right) \Big/ \sum_{i=1}^{j} \mu_i.
\end{aligned}$$

Since $w_\ell - w_i \leq 0$ for all $i \in \mathcal{L}$, ρ_j is maximized at $j = \ell$, and we also have $\rho_\ell = w_\ell$. Lemma 3 implies that σ^*, i.e. accepting all trips, is the optimal acceptance strategy for a driver at $q \in [0, \bar{b}^{(1)}]$. Under σ^*, the continuation payoff is of the form:

$$\pi(q, \bar{Q}, \sigma^*, \sigma^*) = \rho_\ell = w_\ell,\ \forall q \in [0, n_\ell]. \tag{31}$$

[20]The equilibrium queue length is $Q^* = \bar{Q} > \bar{b}^{(1)}$, as a result, drivers at positions $q \in (\bar{b}^{(1)}, \bar{Q}]$ do not receive any dispatches under randomized FIFO. This is, therefore, a different setting from the over-supplied setting under random dispatching (Case 1 of Proposition 2), where trips are dispatched to all drivers in the queue at random.

Since $w_\ell \geq 0$, there is no incentive for a driver to leave the queue without a rider trip, hence there is no useful deviation from $\gamma(q, \bar{Q}) = 0$. To see why moving to the tail of the queue is not a useful deviation either, observe that a driver will not receive any trip dispatch until she moves back to position $\bar{b}^{(1)} = n_\ell$ in the queue. Once the driver moves here (after incurring a non-negative waiting cost), the driver is in the exact same position as she is before, achieving an optimal payoff of w_ℓ when accepting all trips from the platform. This completes the proof that σ^* is a best response for a driver at some $q \in [0, \bar{b}^{(1)}]$ in the queue, when σ^* is adopted by the rest of the drivers.

Now consider any driver at some position $q \in (n_\ell, \bar{Q}]$ in the queue. The driver will not receive any trip dispatches until she reaches position n_ℓ in the queue, thus there is no useful deviation from the acceptance strategy σ^*. From $\bar{b}^{(1)}$ onward, the driver gets an optimal continuation payoff of w_ℓ as we have shown above. As a result, the driver's continuation payoff under σ^* is of the form:

$$\pi(q, \bar{Q}, \sigma^*, \sigma^*) = w_\ell - c(q - n_\ell) / \sum_{i \in \mathcal{L}} \mu_i, \ \forall q \in (\bar{b}^{(1)}, \bar{Q}]. \tag{32}$$

We can verify that $\pi(q, \bar{Q}, \sigma^*, \sigma^*) > 0$ for all $q < \bar{Q}$ and that $\pi(q, \bar{Q}, \sigma^*, \sigma^*) = 0$. Leaving the queue without a rider (and get zero) is therefore not a useful deviation. Moreover, at $\bar{Q}$ the drivers are indifferent towards being in the queue or leaving without a rider, thus a randomized joining decision is a best response. Individual rationality and envy-freeness both hold, since $\pi(q, \bar{Q}, \sigma^*, \sigma^*)$ is non-negative and monotonically non-increasing for all $q \in \bar{Q}$. This completes Case 1.2.

Case 2: $m > 1$**.** In this case we prove by induction on k (the index of bins, starting from the first bin) that in Nash equilibrium given the steady state queue length Q^*, drivers in the k^{th} bin accept all dispatches for trips in the first k partitions $\cup_{k'=1}^{k} \mathcal{L}^{(k')}$, and decline all lower earning trips in $\cup_{k'>k} \mathcal{L}^{(k')}$. We then establish individual rationality and envy-freeness by checking that the continuation payoff is non-negative and monotonically non-increasing.

Before analyze the base case of the induction, we first provide some notations. Denote $\underline{j}^{(k)} \in \mathcal{L}$ and $\bar{j}^{(k)} \in \mathcal{L}$ as the lowest and highest indices of trips in the k^{th} partition $\mathcal{L}^{(k)}$:

$$\underline{j}^{(k)} \triangleq \min\{i \in \mathcal{L} | i \in \mathcal{L}^{(k)}\}, \tag{33}$$

$$\bar{j}^{(k)} \triangleq \max\{i \in \mathcal{L} | i \in \mathcal{L}^{(k)}\}. \tag{34}$$

We know that a trip to location $\underline{j}^{(k)}$ (or $\bar{j}^{(k)}$) is the highest (or lowest) paying trip in $\mathcal{L}^{(k)}$. Let Q^* denote the equilibrium queue length under direct FIFO, i.e. $Q^* = n_{i^*}$ when $\lambda \leq \sum_{i \in \mathcal{L}} \mu_j$, and $Q^* = \bar{Q}$ when $\lambda > \sum_{i \in \mathcal{L}} \mu_j$. Let $\sigma^* = (\alpha^*, \beta^*, \gamma^*)$ be a strategy given by:

- accepting all trips in the top k partitions while in the k^{th} bin in the queue, and randomize only on trips to location i^* while in the last bin:

$$\alpha^*(q, Q^*, i) = \begin{cases} \mathbb{1}\{i \in \cup_{k'=1}^{k} \mathcal{L}^{(k')}\}, & \text{if } q \in [\underline{b}^{(k)}, \bar{b}^{(k)}] \text{ for some } k \leq m, \text{ and } i \neq i^*, \\ \min\{(\lambda - \sum_{i<i^*} \mu_i), \ \mu_{i^*}\}/\mu_{i^*}, & \text{if } q \in [\underline{b}^{(m)}, \bar{b}^{(m)}] \text{ and } i = i^*, \end{cases} \tag{35}$$

- never move to the tail of the queue:

$$\beta^*(q, Q^*) = 0, \ \forall q \in [0, Q^*], \tag{36}$$

- never leave the queue without a trip after joining the queue, and join the queue with probability $\min\{\sum_{i \in \mathcal{L}} \mu_i / \lambda, \ 1\}$, i.e.

$$\gamma^*(q, Q^*) = \begin{cases} 0, & \text{if } q < Q^*, \\ 1 - \min\{\sum_{i \in \mathcal{L}} \mu_i / \lambda, \ 1\}, & \text{if } q = Q^*. \end{cases} \tag{37}$$

We now prove by induction that σ^* forms a Nash equilibrium among the drivers in steady state with queue length Q^*.

Step 1: the base case with $k = 1$. We first prove that when the length of the queue is Q^* and when every other driver adopts strategy σ^*, it is a best response for any driver in the first bin $[\underline{b}^{(1)}, \bar{b}^{(1)}]$ to also adopt strategy σ^*.

The first bin consists of the top $\bar{b}^{(1)}$ drivers at the head of the queue. When $\mathcal{L}^{(1)} = \{1\}$, $\bar{b}^{(1)} = \underline{b}^{(1)} = 0$, thus all trips are first dispatched to the driver at the head of the queue. In this case, it is clear that accepting only trips to location 1 is the best response for a driver at $q = 0$, and this is aligned with σ^*. Now consider the case where $|\mathcal{L}^{(1)}| > 1$, s.t. $\bar{b}^{(1)} > 0$. With $m > 1$, the equilibrium length of the queue Q^* is above $\bar{b}^{(1)}$, thus the first bin is "full". For a driver at any position $q \in [0, \bar{b}^{(1)}]$, the rate at which she receives dispatches to each location $i \in \mathcal{L}$ is

$$\eta_i^{(1)} = \mu_i / \bar{b}^{(1)}.$$

Note that the rates $\{\eta_i^{(1)}\}_{i\in\mathcal{L}}$ are independent to both the strategies adopted by the rest of the drivers in the first bin, and also the strategies employed by all drivers later in the queue.

We first prove that if a driver does not leave the queue or move to the tail of the queue, then there is no useful deviation from $\alpha^*(q, Q^*, i) = \mathbb{1}\{i \in \mathcal{L}^{(1)}\}$, i.e. accepting all trips in $\mathcal{L}^{(1)}$. This is a stationary setting that we have analyzed in Lemma 3. Given (14), we know that $\bar{b}^{(1)}$ is of the form: $\bar{b}^{(1)} = \frac{1}{c}\left(\sum_{i\in\mathcal{L}^{(1)}}(w_i - \min_{i'\in\mathcal{L}^{(1)}}\{w_{i'}\})\mu_i\right)$. The utility for a driver in the first bin from accepting only the top $\bar{j}^{(1)}$ trips (as defined in (25)) can therefore be written as:

$$\begin{aligned}
\rho^{(1)}_{\bar{j}^{(1)}} &= \left(\sum_{i\in\mathcal{L}^{(1)}} w_i\eta_i^{(1)} - c\right) \Big/ \sum_{i\in\mathcal{L}^{(1)}} \eta_i^{(1)} = \left(\sum_{i\in\mathcal{L}^{(1)}} w_i\mu_i - c\bar{b}^{(1)}\right) \Big/ \sum_{i\in\mathcal{L}^{(1)}} \mu_i^{(1)} \\
&= \left(\sum_{i\in\mathcal{L}^{(1)}} w_i - \left(\sum_{i\in\mathcal{L}^{(1)}} \left(w_i - \min_{i'\in\mathcal{L}^{(1)}}\{w_{i'}\}\right)\mu_i\right)\right) \Big/ \sum_{i\in\mathcal{L}^{(1)}} \mu_i^{(1)} \\
&= \min_{i\in\mathcal{L}^{(1)}}\{w_i\}.
\end{aligned}$$

This implies $\rho^{(1)}_{\bar{j}^{(1)}} \le w_i$ for all $i \in \mathcal{L}^{(1)}$, and $\rho^{(1)}_{\bar{j}^{(1)}} > w_i$ for all $i \notin \mathcal{L}^{(1)}$ (recall that the partitions are ordered). Lemma 3 then implies that an optimal acceptance strategy is to accept all trips to locations 1 through $\bar{j}^{(1)}$, and this is aligned with σ^*. Lemma 3 also implies that the continuation payoff of any driver in the first bin given strategy σ^* is:

$$\pi(q, Q^*, \sigma^*, \sigma^*) = \min_{i\in\mathcal{L}^{(1)}}\{w_i\},\ \forall q \in [0, \bar{b}^{(1)}]. \tag{38}$$

Since $\min_{i\in\mathcal{L}^{(1)}}\{w_i\} \ge 0$, deviating from $\gamma^*(q, Q^*) = 0$ and leaving the is not a useful deviation. Moreover, by moving to the tail of the queue, a driver will not receive any trip with net earnings higher than $\min_{i\in\mathcal{L}^{(1)}}\{w_i\}$, if the driver does not move all the way back to the first bin again. Once a driver is back to the first bin (after incurring some non-negative waiting costs), the driver is in the exact same situation as before moving to the tail, receiving trips again at rates $\{\eta_i^{(1)}\}_{i\in\mathcal{L}}$. Deviating from $\beta^*(q, Q^*) = 0$ is therefore not a useful strategy either. This implies that σ^* is a best response for drivers in the first bin, and completes the proof of the base case with $k = 1$.

Step 2: induction step for $1 < k < m$. Assume that when the length of the queue is Q^*, and when every other driver adopts strategy σ^*, it is a best response for a driver any position $q \in [0, \bar{b}^{(k-1)}]$

to adopt strategy σ^*. We now prove in this induction step, that it is also a best response for any driver at positions $q \in (\bar{b}^{(k-1)}, \bar{b}^{(k)}]$ to adopt strategy σ^*.

We take the following steps in proving this result:

Step 2.1 Under strategy σ^*, the equilibrium continuation payoff $\pi(q, Q^*, \sigma^*, \sigma^*)$ is linearly decreasing in q when $q \in [\underline{b}^{(k-1)}, \underline{b}^{(k)}]$, and constant for $q \in [\underline{b}^{(k)}, \bar{b}^{(k)}]$:

$$\pi(q, Q^*, \sigma^*, \sigma^*) = \begin{cases} \min\limits_{i \in \mathcal{L}^{(k-1)}} \{w_i\} - c\left(q - \bar{b}^{(k-1)}\right) \Big/ \sum\limits_{i \in \cup_{k'=1}^{k-1} \mathcal{L}^{(k')}} \mu_i, & \text{if } q \in [\bar{b}^{(k-1)}, \underline{b}^{(k)}], \\ \min\limits_{i \in \mathcal{L}^{(k)}} \{w_i\}, & \text{if } q \in [\underline{b}^{(k)}, \bar{b}^{(k)}]. \end{cases} \tag{39}$$

Step 2.2 Under any feasible strategy $\sigma = (\alpha, \beta, \gamma)$ s.t. the driver does not leave the queue or move to the tail of the queue (i.e. if $\beta(q, Q^*) = \gamma(q, Q^*) = 0$ for all $q \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$), the continuation payoff cannot exceed that under σ^*:

$$\pi(q, Q^*, \sigma, \sigma^*) \leq \pi(q, Q^*, \sigma^*, \sigma^*), \ \forall q \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}].$$

Step 2.3 σ^* is a best response for drivers at positions $q \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$ in the queue.

Step 2.1 implies that for any driver at some $q \in [\underline{b}^{(k)}, \bar{b}^{(k)}]$, it cannot be a useful deviation from σ^* to accept any trip in later bins $\cup_{k'>k} \mathcal{L}^{(k')}$ since the driver gets $\pi(q, Q^*, \sigma^*, \sigma^*) \geq \min_{i \in \mathcal{L}^{(k)}} w_i > \max_{i \in \cup_{k'>k} \mathcal{L}^{(k')}} w_i$ from following strategy σ^*. Moreover, all best-response strategies must have $\gamma(q, Q^*) = 0$ for all $q \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$, because $\min_{i \in \mathcal{L}^{(k)}} \{w_i\} > 0$ thus leaving the queue and getting zero cannot be a useful deviation.

With Step 2.2, we know that across all feasible strategies where the driver does not move to the tail of the queue, σ^* is a best strategy for drivers at any $q \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$. With 2.1 and 2.2, we prove the final step, that even when we consider all feasible strategies where people may move to the tail of the queue, there is still no strategy that results in a higher payoff than σ^*.

We start from Step 2.1.

Step 2.1. We prove (39) in this step. First, we show that under σ^*, the continuation payoff of drivers satisfy $\pi(\bar{b}^{(k-1)}, Q^*, \sigma^*, \sigma^*) = \min_{i \in \mathcal{L}^{(k-1)}} w_i$, and $\pi(\underline{b}^{(k)}, Q^*, \sigma^*, \sigma^*) = \pi(\bar{b}^{(k)}, Q^*, \sigma^*, \sigma^*) = \min_{i \in \mathcal{L}^{(k)}} w_i$. For simplicity of notation, denote the continuation payoff under strategy σ^* given the equilibrium queue length Q^* as:

$$\pi^*(q) \triangleq \pi(q, Q^*, \sigma^*, \sigma^*). \tag{40}$$

Moreover, denote the total trip volume and average net earnings for a given subset of partitions as:

$$s_{k_1:k_2} \triangleq \sum_{i \in \cup_{k'=k_1}^{k_2} \mathcal{L}^{(k')}} \mu_i, \tag{41}$$

$$\bar{w}_{k_1:k_2} \triangleq \sum_{i \in \cup_{k'=k_1}^{k_2} \mathcal{L}^{(k')}} w_i \mu_i \Big/ \sum_{i \in \cup_{k'=k_1}^{k_2} \mathcal{L}^{(k')}} \mu_i = \sum_{i \in \cup_{k'=k_1}^{k_2} \mathcal{L}^{(k')}} w_i \mu_i \Big/ s_{k_1:k_2}. \tag{42}$$

Consider now a driver who had just reached the position position $\bar{b}^{(k-1)}$ in the queue. Under σ^*, a driver at $[0, \bar{b}^{(k-1)}]$ in the queue only accept trips in the top $k-1$ partitions $\cup_{k' \leq k-1} \mathcal{L}^{(k')}$.

When all drivers adopt the same strategy σ^*, *a priori* there is no difference in the waiting times or earnings from trips of any driver who are at $\bar{b}^{(k-1)}$ in the queue. The average net earnings a driver at $q = \bar{b}^{(k-1)}$ will get from the trip she will accept in the future is therefore $\bar{w}_{1:k-1}$. By Little's Law, the average amount of time the driver spends waiting in the queue is $\bar{b}^{(k-1)}/s_{1:k-1}$. Thus the average continuation payoff for the driver at $q = \underline{b}^{(k)}$ is:

$$\pi^*(\bar{b}^{(k-1)}) = \bar{w}_{1:k-1} - c\bar{b}^{(k-1)}/s_{1:k-1}.$$

Given $\bar{b}^{(k)}$ as defined in (14), we know:

$$\pi^*(\bar{b}^{(k-1)}) = \bar{w}_{1:k-1} - \left(\sum_{i \in \cup_{k'<k} \mathcal{L}^{(k')}} \mu_i \left(w_i - \min_{i' \in \mathcal{L}^{(k-1)}} \{w_{i'}\} \right) \right) \Big/ s_{1:k-1} = \min_{i \in \mathcal{L}^{(k-1)}} \{w_i\}. \tag{43}$$

Similarly, by reasoning about the net earnings an average driver gets from an average trip, and the average waiting cost a driver incurs, we can show that $\pi(\underline{b}^{(k)}, Q^*, \sigma^*, \sigma^*) = \pi(\bar{b}^{(k)}, Q^*, \sigma^*, \sigma^*) = \min_{i \in \mathcal{L}^{(k)}} w_i$. Under σ^*, drivers at some position $q \in (\bar{b}^{(k-1)}, \underline{b}^{(k)})$ will wait for $(q - \underline{b}^{(k-1)})/s_{1:k-1}$ units of time before reaching $\bar{b}^{(k-1)}$ in the queue, therefore her continuation payoff is of the form:

$$\pi^*(q) = \min_{i \in \mathcal{L}^{(k-1)}} \{w_i\} - c \left(q - \bar{b}^{(k-1)} \right) \Big/ s_{1:k-1}, \text{ if } q \in [\underline{b}^{(k-1)}, \underline{b}^{(k)}). \tag{44}$$

It is straightforward to verify that π^* is left continuous at $\underline{b}^{(k)}$:

$$\lim_{q \to \underline{b}^{(k)}-} \pi^*(q) = \min_{i \in \mathcal{L}^{(k-1)}} \{w_i\} - c \left(\underline{b}^{(k)} - \bar{b}^{(k-1)} \right) \Big/ s_{1:k-1} = \min_{i \in \mathcal{L}^{(k)}} \{w_i\}.$$

What is left to prove for Step 2.1 is that $\pi^*(q)$ remains constant where $q \in [\underline{b}^{(k)}, \bar{b}^{(k)}]$. This is trivial when $|\mathcal{L}^{(k)}| = 1$, in which case $\bar{b}^{(k)} = \underline{b}^{(k)}$. Therefore, we now consider the case where $|\mathcal{L}^{(k)}| > 1$ s.t. $\bar{b}^{(k)} > \underline{b}^{(k)}$. When all drivers adopt strategy σ^*, all trips in the first $k-1$ partitions $\cup_{k'=1}^{k-1} \mathcal{L}^{(k')}$ are accepted before reaching the k^{th} bin. For a driver in the k^{th} bin, the rate at which she receives trip dispatches to each location $i \in \mathcal{L}$ is therefore:

$$\eta_i^{(k)} = \begin{cases} 0, & \text{if } i \in \cup_{k'=1}^{k-1} \mathcal{L}^{(k')} \\ \mu_i/(\bar{b}^{(k)} - \underline{b}^{(k)}), & \text{if } i \in \cup_{k' \geq k} \mathcal{L}^{(k')} \end{cases} \tag{45}$$

Note that the rates $\{\eta_i^{(k)}\}_{i \in \mathcal{L}}$ are independent to the strategies taken by drivers in later bins of the queue. With a slight abuse of notation, let

$$\eta^{(k)} \triangleq \sum_{i \in \mathcal{L}^{(k)}} \eta_i^{(k)}$$

be the total rate at which drivers in the k^{th} bin receives trips in $\mathcal{L}^{(k)}$.

Fix an arbitrary point in time and call it time $t = 0$, and consider a driver who is at position $\bar{b}^{(k)}$ at time $t = 0$. Let $g(t)$ be the position of the driver in the queue, if the driver has not yet accepted a trip and leave the queue. We know $g(t) = \bar{b}^{(k)}$. Before the driver reaches position $\underline{b}^{(k)}$ in the queue, we know that in the next $\mathrm{d}t$ units of time, when every other driver adopts strategy σ^*, there are $s_{1:k-1}\mathrm{d}t$ drivers who are dispatched from queue positions earlier than $\underline{b}^{(k)}$, and there

are $\mathrm{d}t s_{k:k}(g(t)-\underline{b}^{(k)})/(\bar{b}^{(k)}-\underline{b}^{(k)})$ drivers who are dispatched from the k^{th} bin, ahead of this driver. As a result, the time derivative of the driver's queue position satisfies

$$\frac{\mathrm{d}g(t)}{\mathrm{d}t} = -s_{1:k-1} - s_{k:k}\frac{g(t)-\underline{b}^{(k)}}{\bar{b}^{(k)}-\underline{b}^{(k)}}, \tag{46}$$

i.e. the driver moves forward in the queue at a rate of $s_{1:k-1}+s_{k:k}(g(t)-\underline{b}^{(k)})/(\bar{b}^{(k)}-\underline{b}^{(k)})$ positions per unit of time. Since $s_{1:k-1}>0$, we know that the driver will reach $\underline{b}^{(k)}$ within finite time.

$\pi^*(g(t))$ denotes continuation payoff of this driver as a function of time. For a driver at some position $g(t)\in(\underline{b}^{(k)},\ \bar{b}^{(k)}]$ at time t, the probability for the driver to be dispatched a trip she will accept under σ^* in the next $\mathrm{d}t$ units of time is $\eta^{(k)}\mathrm{d}t$. If the driver is not dispatched, she moves forward in the queue to position $g(t+\mathrm{d}t)$ after incurring a cost of $c\mathrm{d}t$. If the driver is dispatched, she takes a trip with an average net earnings of $\bar{w}_{k:k}$ after incurring a waiting cost in the order of $cO(\mathrm{d}t)$. Therefore, the driver's continuation payoff as a function of time t can be written as:

$$\pi^*(g(t)) = (1-\eta^{(k)}\mathrm{d}t)(\pi^*(g(t+\mathrm{d}t))-c\mathrm{d}t)+\eta^{(k)}\mathrm{d}t(\bar{w}_{k:k}-cO(\mathrm{d}t)) \tag{47}$$

Reorganizing (47), and taking the limit as $\mathrm{d}t\to 0$, we have

$$\frac{\mathrm{d}\pi^*(g(t))}{\mathrm{d}t} = c+\eta^{(k)}\left(\pi^*(g(t))-\bar{w}_{k:k}\right) = \eta^{(k)}\left(\pi^*(g(t))-\left(\bar{w}_{k:k}-c/\eta^{(k)}\right)\right), \tag{48}$$

and this implies

$$\pi^*(g(t)) = \bar{w}_{k:k}-c/\eta^{(k)}+Ce^{\eta^{(k)}t}, \tag{49}$$

where C is some constant. Given that $g(0)=\bar{b}^{(k)}$ and $\pi(\bar{b}^{(k)})=\min_{i\in\mathcal{L}^{(k)}}\{w_i\}$, we have:

$$\pi^*(g(0)) = \min_{i\in\mathcal{L}^{(k)}}\{w_i\} = \bar{w}_{k:k}-c/\eta^{(k)}+C.$$

Given (13) and (14), the size of the k^{th} bin is:

$$\bar{b}^{(k)}-\underline{b}^{(k)} = \frac{1}{c}\sum_{i\in\mathcal{L}^{(k)}}(w_i-\min_{i'\in\mathcal{L}^{(k)}}\{w_{i'}\})\mu_i. \tag{50}$$

$\bar{w}_{k:k}-c/\eta^{(k)}$ therefore satisfies

$$\bar{w}_{k:k}-c/\eta^{(k)} = \left(\sum_{i\in\mathcal{L}^{(k)}}w_i\mu_i - c(\bar{b}^{(k)}-\underline{b}^{(k)})\right)\Big/ s_{k:k} = \min_{i\in\mathcal{L}^{(k)}}\{w_i\}.$$

As a result, $C=0$ must hold, meaning that $\pi^*(q)$ remains constant w.r.t. t for t such that $g(t)\le\bar{b}^{(k)}$ and $g(t)\ge\underline{b}^{(k)}$. This completes the proof that $\pi^*(q)=\bar{w}_{k:k}-c/\eta^{(k)}=\min_{i\in\mathcal{L}^{(k)}}\{w_i\}$ for all $q\in(\underline{b}^{(k)},\bar{b}^{(k)}]$.

This completes the proof of Step 2.1. What we know from this step and Lemma 2 is that for any driver at some $q\in[\underline{b}^{(k)},\bar{b}^{(k)}]$, it cannot be a useful deviation from σ^* to accept any trip in later bins $\cup_{k'>k}\mathcal{L}^{(k')}$ since the driver gets $\pi(q,Q^*,\sigma^*,\sigma^*)\ge\min_{i\in\mathcal{L}^{(k)}}w_i>\max_{i\in\cup_{k'>k}\mathcal{L}^{(k')}}w_i$ from following strategy σ^*. Moreover, all best-response strategies must have $\gamma(q,Q^*)=0$ for all $q\in[\bar{b}^{(k-1)},\bar{b}^{(k)}]$, because $\min_{i\in\mathcal{L}^{(k)}}\{w_i\}\ge 0$ thus leaving the queue and getting zero cannot be a useful deviation.

Step 2.2. We now prove that under any feasible strategy $\sigma = (\alpha, \beta, \gamma)$ s.t. the driver does not leave the queue or move to the tail of the queue (i.e. if $\beta(q, Q^*) = \gamma(q, Q^*) = 0$ for all $q \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$), for any position $q \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$,

$$\pi(q, Q^*, \sigma, \sigma^*) \leq \pi(q, Q^*, \sigma^*, \sigma^*). \tag{51}$$

First, (51) is straightforward to establish for $q \in [\bar{b}^{(k-1)}, \underline{b}^{(k)})$, since if the driver does not leave or move to the tail of the queue, the driver will wait in line until she reaches position $\bar{b}^{(k-1)}$ in the queue, and this is aligned with σ^*. Once a driver is at $\bar{b}^{(k-1)}$, the best strategy moving forward is σ^* (by induction assumption). As a result, it is impossible to achieve a better continuation payoff than that under σ^*.

Now consider $q \in [\underline{b}^{(k)}, \bar{b}^{(k)}]$, and there are two cases depending on whether $|\mathcal{L}^{(k)}| = 1$. When $|\mathcal{L}^{(k)}| = 1$, Lemma 4 implies that $\bar{b}^{(k)} = \underline{b}^{(k)}$, thus all trips in $\cup_{k' \geq k} \mathcal{L}^{(k')}$ are dispatched to the driver at position $\underline{b}^{(k)}$ in the queue. Under σ^*, a driver at $q = \underline{b}^{(k)}$ gets a continuation payoff of w_i where $i \in \mathcal{L}^{(k)}$ is the only trip in the k^{th} partition, regardless of whether the driver accepts a trip or moved forward in the queue. An argument very similar to the proof of the induction step of Lemma 1 shows no alternative strategy may achieve a higher continuation payoff.

What is left to study in the case where $|\mathcal{L}^{(k)}| > 1$ and $\bar{b}^{(k)} - \underline{b}^{(k)} > 0$. Assume towards a contradiction that there exists some $q \in (\underline{b}^{(k)}, \bar{b}^{(k)}]$ s.t. $\pi(q, Q^*, \sigma, \sigma^*) > \pi(q, Q^*, \sigma^*, \sigma^*) = \min_{i \in \mathcal{L}^{(k)}} w_i$. We introduce the following notation:

- If the driver did not accept any trip dispatches under σ before reaching position $\underline{b}^{(k)}$ in the queue, denote the time it takes for the driver to move from q to $\underline{b}^{(k)}$ as $\kappa(q)$.
- Denote the probability for the driver to be dispatched a trip she is willing to accept under σ before the driver reaches $\underline{b}^{(k)}$ (i.e. within the next $\kappa(q)$ units of time, given strategy σ) as $\xi(\kappa(q))$.
- If the driver did receive a trip within the $\kappa(q)$ units of time while following strategy σ, let $\omega(\kappa(q))$ be the driver's expected payoff, which includes both the net earnings from the trip the driver accepts and the waiting cost the driver incurs.

The driver's continuation payoff at position q under strategy σ can therefore be written as:

$$\pi(q, Q^*, \sigma, \sigma^*) = \xi(\kappa(q))\omega(\kappa(q)) + (1 - \xi(\kappa(q))) \left(\pi(\underline{b}^{(k)}, Q^*, \sigma, \sigma^*) - c\kappa(q) \right). \tag{52}$$

$\pi(\underline{b}^{(k)}, Q^*, \sigma, \sigma^*)$ shows up in the second term because if a driver did not accept a dispatch before time $\kappa(q)$ had passed (starting from the time of her being at position q), the driver has now reached $\underline{b}^{(k)}$. We have just argued that this continuation payoff is bounded by $\pi(\underline{b}^{(k)}, Q^*, \sigma, \sigma^*) \leq \pi^*(\underline{b}^{(k)}) = \min_{i \in \mathcal{L}^{(k)}} w_i$. When $\xi(\kappa(q)) = 0$, $\pi(q, Q^*, \sigma, \sigma^*) \leq \min_{i \in \mathcal{L}^{(k)}} w_i$ trivially holds. When $\xi(\kappa(q)) > 0$, combining (52) and the assumption that $\pi(q, Q^*, \sigma, \sigma^*) > \min_{i \in \mathcal{L}^{(k)}} w_i$, we get

$$\omega(\kappa(q)) > \min_{i \in \mathcal{L}^{(k)}} \{w_i\} + (1 - \xi(\kappa(q))) \cdot c\kappa(q)/\xi(\kappa(q)). \tag{53}$$

Observe that in the first $\kappa(q)$ units of time, the driver receives trip dispatches at rates $\{\eta_i^{(k)}\}_{i \in \mathcal{L}}$. Now consider a stationary setting that we analyzed in Lemma 3, where a driver always receives trip dispatches at rates $\{\eta_i^{(k)}\}_{i \in \mathcal{L}}$.[21] If the driver employs strategy σ (restricted to the first $\kappa(q)$

[21] In other words, we allow the driver to remain in the k^{th} bin forever, instead of forcing her to move past $\underline{b}^{(k)}$.

units of time starting from position q in the queue) in this stationary setting, the driver's expected utility can be written as:

$$\hat{\pi}(\sigma) = \xi(\kappa(q))\omega(\kappa(q)) + (1 - \xi(\kappa(q)))(\hat{\pi}(\sigma) - c\kappa(q)).$$

Intuitively, if the driver gets dispatched in the first $\kappa(q)$ units of time given strategy σ, she gets an expected payoff of $\omega(\kappa(q))$, and this happens with probability $\xi(\kappa(q))$. If the driver is not dispatched in the first $\kappa(q)$ units of time, the driver's continuation payoff starting from that point of time onward is again $\hat{\pi}(\sigma)$. Reorganizing this expression, and applying (53), we get:

$$\begin{aligned}\xi(\kappa(q))\hat{\pi}(\sigma) =& \xi(\kappa(q))\omega(\kappa(q)) - (1 - \xi(\kappa(q)))c\kappa(q) \\ >& \xi(\kappa(q)) \min_{i\in\mathcal{L}^{(k)}}\{w_i\} + (1 - \xi(\kappa(q))) \cdot c\kappa(q) - (1 - \xi(\kappa(q)))c\kappa(q) \\ =& \xi(\kappa(q)) \min_{i\in\mathcal{L}^{(k)}}\{w_i\}.\end{aligned}$$

This implies $\hat{\pi}(\sigma) > \min_{i\in\mathcal{L}^{(k)}}\{w_i\}$, meaning that there exists a strategy for a driver to get a continuation payoff strictly above $\min_{i\in\mathcal{L}^{(k)}}\{w_i\}$ in the stationary setting where the driver always receives trip dispatches at rate $\{\eta_i^{(k)}\}_{i\in\mathcal{L}}$ given by (45). We now prove that this is not possible, and as a result we have a contradiction.

Given (45) and (50), for a driver who receives trip dispatches at rates $\{\eta_i^{(k)}\}_{i\in\mathcal{L}}$, the expected utility from accepting top j trips (as defined in (25)) for some $j \in \mathcal{L}$ can be rewritten as:

$$\rho_j^{(k)} = \left(\sum_{i\le j} w_i\eta_i^{(k)} - c\right) \Big/ \sum_{i\le j}\eta_i^{(k)} = \left(\sum_{i=\underline{j}^{(k)}}^{j} w_i\mu_i - c(\bar{b}^{(k)} - \underline{b}^{(k)})\right) \Big/ \sum_{i=\underline{j}^{(k)}}^{j}\mu_i \tag{54}$$

$$= \left(\sum_{i=\underline{j}^{(k)}}^{j} w_i\mu_i - \left(\sum_{i\in\mathcal{L}^{(k)}}(w_i - \min_{i'\in\mathcal{L}^{(k)}}\{w_{i'}\})\mu_i\right)\right) \Big/ \sum_{i=\underline{j}^{(k)}}^{j}\mu_i. \tag{55}$$

Recall that $\underline{j}^{(k)}$ and $\bar{j}^{(k)}$ are the lower-index and highest-index trips in $\mathcal{L}^{(k)}$, respectively. When $j \le \bar{j}^{(k)}$,

$$\rho_j^{(k)} = \min_{i'\in\mathcal{L}^{(k)}}\{w_{i'}\} - \left(\sum_{i=j+1}^{\bar{j}^{(k)}}(w_i - \min_{i'\in\mathcal{L}^{(k)}}\{w_{i'}\})\mu_i\right) \Big/ \sum_{i=\underline{j}^{(k)}}^{j}\mu_i \le \min_{i'\in\mathcal{L}^{(k)}}\{w_{i'}\}.$$

When $j \ge \bar{j}^{(k)}$, we also have:

$$\rho_j^{(k)} = \left(\sum_{i=\underline{j}^{(k+1)}}^{j} w_i\mu_i + \min_{i'\in\mathcal{L}^{(k)}}\{w_{i'}\}\sum_{i\in\mathcal{L}^{(k)}}\mu_i\right) \Big/ \sum_{i=\underline{j}^{(k)}}^{j}\mu_i \le \min_{i'\in\mathcal{L}^{(k)}}\{w_{i'}\}.$$

As a result, $\rho_j^{(k)}$ is optimized at $j = \bar{j}^{(k)}$, taking value $\min_{i\in\mathcal{L}^{(k)}}\{w_i\}$. Lemma 3 then implies that a driver in such a stationary setting cannot achieve a utility strictly higher than $\min_{i\in\mathcal{L}^{(k)}}\{w_i\}$. This completes the proof of Step 2.2.

Step 2.3. We now complete the induction step by proving that σ^* is a best response for drivers at any position $q \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$ in the queue.

Assume towards a contradiction, that there exists a strategy σ s.t. $\pi(q, Q^*, \sigma, \sigma^*) > \pi^*(q)$ for some $\hat{q} \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$. We know from Step 2.1 that it is without loss of generality to restrict our analysis to strategies σ where the driver does not leave the queue (i.e. $\gamma(q, Q^*) = 0$, $\forall q \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$). This is because if a driver leaves with a non-zero probability at some $q \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$, we may construct an alternative strategy where instead of leaving, the driver stays in the queue, follows σ^* from then on and gets $\pi^*(q) \geq \min_{i \in \mathcal{L}^{(k)}}\{w_i\} > 0$. This increases the driver's continuation payoff, thus we get an alternative strategy that improves over σ^*, and also satisfies $\gamma(q, Q^*) = 0$ for all $q \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$.

Step 2.2 then implies that there must exist some $\tilde{q} \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$ such that $\beta(\tilde{q}, Q^*) > 0$ because otherwise, σ becomes a strategy s.t. $\gamma(q, Q^*) = \beta(q, Q^*) = 0$ for all $q \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$, and we have proved in Step 2.2 that among all such strategies, achieving a higher continuation payoff than that under σ^* is not possible.

We claim that we can find a $\tilde{q} \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$ where $\beta(\tilde{q}, Q^*) > 0$ and the continuation payoff satisfies $\pi(\tilde{q}, Q^*, \sigma, \sigma^*) \leq \pi(Q^*, Q^*, \sigma, \sigma^*)$. Assume otherwise, that $\pi(q, Q^*, \sigma, \sigma^*) > \pi(Q^*, Q^*, \sigma, \sigma^*)$ for all $q \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$ s.t. $\beta(q, Q) > 0$. We can improve the continuation payoff under σ by reducing the probability of moving to the tail to zero everywhere, and this gives us a strategy with $\gamma(q, Q^*) = \beta(q, Q^*) = 0$ for all $q \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$, but achieves a higher payoff than that under σ^* at some $\hat{q} \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$. This, again violates Step 2.2.

Now consider an alternative strategy σ' which is the same as σ except that the driver always moves back to the tail of the queue at the position $\tilde{q}$, for which $\pi(\tilde{q}, Q^*, \sigma, \sigma^*) \leq \pi(Q^*, Q^*, \sigma, \sigma^*)$ holds. This must be a weak improvement over σ, therefore we still have $\pi(q, Q^*, \sigma', \sigma^*) > \pi^*(q)$ for some $\hat{q} \in [\bar{b}^{(k-1)}, \bar{b}^{(k)}]$. We now argue that $\pi(Q^*, Q^*, \sigma', \sigma^*) > \min_{i \in \mathcal{L}^{(k)}}\{w_i\}$ must hold, because if otherwise, we may improve the continuation payoff under σ' by replacing "moving to the tail of the queue" with "remain in the queue and adopt σ^* from then on". This gives us a strategy where the driver does not move to the tail of the queue, but still achieves a higher continuation payoff than σ^* at $\hat{q}$, and this violates Step 2.2.

With $\pi(Q^*, Q^*, \sigma', \sigma^*) > \min_{i \in \mathcal{L}^{(k)}}\{w_i\}$, we claim that

$$\pi(\bar{b}^{(k)}, Q^*, \sigma', \sigma^*) > \pi(Q^*, Q^*, \sigma', \sigma^*) > \min_{i \in \mathcal{L}^{(k)}}\{w_i\}. \tag{56}$$

This is because a driver at the tail of the queue $q = Q^*$ will not receive any trip with net earnings weakly above $\min_{i \in \mathcal{L}^{(k)}}\{w_i\}$ until the driver reaches position $\bar{b}^{(k)}$ in the queue, thus in the scenarios where the driver is dispatched under σ' before reaching $\bar{b}^{(k)}$, the driver's payoff is strictly below $\min_{i \in \mathcal{L}^{(k)}}\{w_i\}$. $\pi(Q^*, Q^*, \sigma', \sigma^*)$ is a weighted average of this payoff and $\pi(\bar{b}^{(k)}, Q^*, \sigma', \sigma^*)$ minus the waiting cost a driver incurs before reaching $\bar{b}^{(k)}$, we must have $\pi(\bar{b}^{(k)}, Q^*, \sigma', \sigma^*) > \pi(Q^*, Q^*, \sigma', \sigma^*)$ in order for $\pi(Q^*, Q^*, \sigma', \sigma^*) > \min_{i \in \mathcal{L}^{(k)}}\{w_i\}$ to hold.

Now consider an alternative setting, where the driver always moves back to $\bar{b}^{(k)}$ instead of the tail of the queue, whenever the driver is prescribed to move to the tail of the queue under σ'. (56) implies that this will be an improvement, s.t. the continuation payoff under this new setting must also satisfy $\hat{\pi}(\bar{b}^{(k)}, Q^*, \sigma, \sigma^*) > \min_{i \in \mathcal{L}^{(k)}}\{w_i\}$.

This is, however, not possible. Observe that in this alternative setting, under σ', the driver at $\bar{b}^{(k)}$ will either accept a trip and leave the queue before reaching $\tilde{q} \in [\bar{b}^{(k-1)}, \underline{b}^{(k)}]$, or move back to $\bar{b}^{(k)}$ before reaching $\tilde{q}$ or at $\tilde{q}$. As a result, the driver will either be receiving trip dispatches at rates $\{\eta_i^{(k)}\}_{i \in \mathcal{L}}$ defined in (45) (when the driver is at some position in $[\underline{b}^{(k)}, \bar{b}^{(k)}]$), or not receive any trip dispatches at all (when the driver is in $[\bar{b}^{(k-1)}, \underline{b}^{(k)})$). The driver's payoff is therefore upper bounded by the scenario where she is in a stationary setting, always receiving trip dispatches at rates $\{\eta_i^{(k)}\}_{i \in \mathcal{L}}$, but we have proved in Step 2.2 that the highest achievable expected payoff in this setting is $\min_{i \in \mathcal{L}^{(k)}}\{w_i\}$.

This is a contradiction, and concludes the proof of the induction step, that σ^* is a best response for a driver at any position $[\bar{b}^{(k-1)}, \bar{b}^{(k)}]$ in the queue.

Step 3: The last bin $k = m$ and beyond. Given Steps 1 and 2, we know that σ^* is a best response for a driver at any position $q \leq \bar{b}^{(m-1)}$ in the queue. What is left to prove that σ^* is also a best response for any driver at $q \in (\bar{b}^{(m-1)}, Q^*]$ in the queue.

First, with the same arguments as in Step 2.1, we can show that

$$\pi(q, Q^*, \sigma^*, \sigma^*) = \min_{i \in \mathcal{L}^{(m-1)}} \{w_i\} - c(q - \bar{b}^{(m-1)})/s_{1:m-1}, \text{ if } q \in [\bar{b}^{(m-1)}, \underline{b}^{(m)}], \tag{57}$$

and that

$$\pi(\underline{b}^{(m)}, Q^*, \sigma^*, \sigma^*) = \pi(\bar{b}^{(m)}, Q^*, \sigma^*, \sigma^*) = \min_{i \in \mathcal{L}^{(m)}} \{w_i\} = w_{i^*}. \tag{58}$$

In the case where $|\mathcal{L}^{(m)}| = 1$, i.e. when $\mathcal{L}^{(m)} = \{i^*\}$, Lemma 4 implies that $\bar{b}^{(m)} = \underline{b}^{(m)} = n_{i^*}$. Under σ^*, all trips to locations $j < i^*$ are accepted by drivers in the top $m-1$ bins, and all trips to locations $j \geq i^*$ are dispatched (for the last time) to drivers at position n_{i^*} in the queue, where drivers accept only trips to location i^*. The equilibrium queue length is $Q^* = n_{i^*}$ when $\lambda \leq \sum_{i \in \mathcal{L}} \mu_i$. When $\lambda > \sum_{i \in \mathcal{L}} \mu_i$, the equilibrium queue length is $Q^* = \bar{Q}$, and $\pi^*(q)$ decreases linearly in q when $q \geq n_{i^*} = n_\ell$, with $\pi^*(\bar{Q}) = 0$ at the tail of the queue. Using the same arguments as those in the proof of Lemma 1, we can show that when the length of the queue is Q^* and when the rest of the drivers adopt σ^*, it is also a best response for a driver at any $q \in [\bar{b}^{(m-1)}, Q^*]$ to adopt σ^*. We do not repeat the same reasoning here.

What is left to analyze is the case where $|\mathcal{L}^{(m)}| > 1$, in which case $\bar{b}^{(m)} - \underline{b}^{(m)} > 0$. When all drivers adopt strategy σ^*, for any driver in the last bin $[\underline{b}^{(m)}, \bar{b}^{(m)}]$, the rates at which the driver receives trip dispatches to each location are given by

$$\eta_i^{(m)} = \begin{cases} 0, & \text{if } i \in \cup_{k'=1}^{m-1} \mathcal{L}^{(k')}, \\ \mu_i/(\bar{b}^{(m)} - \underline{b}^{(m)}), & \text{if } i \in \mathcal{L}^{(m)}. \end{cases} \tag{59}$$

Applying Lemma 3 in the same way as we did in Step 2.2 above, we can show that for a driver in a stationary setting, where the driver always receives trips to all locations at rates $\{\eta_i^{(m)}\}_{i \in \mathcal{L}}$, the highest expected payoff a driver may get is $\min_{i \in \mathcal{L}^{(m)}} \{w_i\} = w_{i^*}$.

We prove that under σ^*, $\pi^*(q) = w_{i^*}$ holds for all $q \in [\underline{b}^{(m)}, \bar{b}^{(m)}]$. Denote $\tilde{\mu}_{i^*} \triangleq \min\{\mu_{i^*},\ \lambda - \sum_{j<i^*} \mu_j\}$. Under σ^*, a driver in the last bin accepts trip dispatches to location i^* with probability $\tilde{\mu}_{i^*}/\mu_{i^*}$ (see (35)). As a result, for a driver at position $q \in [\underline{b}^{(m)}, \bar{b}^{(m)}]$ in the queue, under σ^*, the total rate at which the driver receives and accepts trip dispatches is

$$\tilde{\eta}^{(m)} \triangleq \sum_{i \in \mathcal{L}^{(m)},\ i<i^*} \eta_i^{(m)} + \tilde{\mu}_{i^*}/(\bar{b}^{(m)} - \underline{b}^{(m)}).$$

Consider now a driver who is at position $\bar{b}^{(m)}$ at time $t = 0$, and denote the driver's position as a function of time t as $g(t)$. The same argument as in the proof of Step 2.1 implies that for all t such that $g(t) \in [\underline{b}^{(m)}, \bar{b}^{(m)}]$ the derivative of the continuation payoff $\pi^*(g(t))$ w.r.t. t is of the form:

$$\frac{\mathrm{d}\pi^*(g(t))}{\mathrm{d}t} = \tilde{\eta}^{(m)} \left(\pi^*(g(t)) - \left(\left(\sum_{i \in \mathcal{L}^{(m)}, i<i^*} w_i \mu_i + w_{i^*} \tilde{\mu}_{i^*} \right) \Big/ \left(\sum_{i \in \mathcal{L}^{(m)}, i<i^*} \mu_i + \tilde{\mu}_{i^*} \right) - c/\tilde{\eta}^{(k)} \right) \right).$$

It is straightforward to verify that

$$\left(\left(\sum_{i\in\mathcal{L}^{(m)},i<i^*} w_i\mu_i + w_{i^*}\tilde{\mu}_{i^*}\right)\Big/\left(\sum_{i\in\mathcal{L}^{(m)},i<i^*} \mu_i + \tilde{\mu}_{i^*}\right) - c/\tilde{\eta}^{(m)}\right) = w_{i^*}, \tag{60}$$

thus solving $\frac{\mathrm{d}\pi^*(g(t))}{\mathrm{d}q} = \tilde{\eta}^{(m)}\left(\pi^*(g(t)) - w_{i^*}\right)$ with boundary condition $\pi^*(g(t)) = w_{i^*}$, we know that $\pi^*(g(t)) = w_{i^*}$ holds for all t such that $g(t) \in [\underline{b}^{(m)}, \bar{b}^{(m)}]$.

Regarding the tail of the queue: in the under-supplied scenario where $\lambda \le \sum_{i\in\mathcal{L}}\mu_i$, the equilibrium queue length is $Q^* = \bar{b}^{(m)} = n_{i^*}$, thus there is no more driver in the queue beyond $\bar{b}^{(m)}$. In the over-supplied scenario with $\lambda > \sum_{i\in\mathcal{L}}\mu_i$, $i^* = \ell$ and $\bar{b}^{(m)} = n_\ell$, and we have:

$$\pi^*(q) = w_\ell - c(q - n_\ell)\Big/\sum_{i\in\mathcal{L}}\mu_i,\ \forall q \in [n_\ell, \bar{Q}]. \tag{61}$$

Assume that the queue length is Q^* and the rest of the drivers adopt strategy σ^*. To prove that σ^* is a best-response for drivers in $[\bar{b}^{(m-1)}, \bar{b}^{(m)}]$ i.e. $\pi^*(q) \ge \pi(q, Q^*, \sigma, \sigma^*)$ for all $q \in [\bar{b}^{(m-1)}, \bar{b}^{(m)}]$ and any feasible strategies σ, we use arguments very similar to those in Steps 2.2 and 2.3, and therefore do not repeat the details here. Intuitively, if σ^* is not a best response, we are able to construct a strategy under which a driver gets a payoff strictly higher than $\min_{i\in\mathcal{L}^{(m)}}\{w_i\} = w_{i^*}$ in the stationary setting where the river always receives trips to all locations at rates $\{\eta_i^{(m)}\}_{i\in\mathcal{L}}$. This is not possible, as we have discussed above.

This establishes that the highest continuation payoff a driver at $q = \bar{b}^{(m)}$ may get under any strategy is $\pi^*(\bar{b}^{(m)})$. To show that σ^* is also a best response for any driver at $q \in [\bar{b}^{(m)}, \bar{Q}]$ in the setting with $\lambda > \sum_{i\in\mathcal{L}}\mu_i$, the same arguments used in the proof of Lemma 1 applies, thus we again refer the readers to Appendix A.1.

This completes the proof for Case 2 $m > 1$, and establishes that when the length of the queue is Q^*, strategy σ^* forms a Nash equilibrium in steady state among the drivers. As we have discussed earlier, this equilibrium outcome is optimal for trip throughput and the platform's net revenue, since it has the same queue length and completes the same set of trips as the equilibrium outcome under direct FIFO (which is optimal - see Theorem 2). Combining the three steps of Case 2, we also see that the continuation payoff $\pi^*(q)$ is non-negative and monotonically non-increasing in q. As a result, the equilibrium outcome of Case 2 is also individually rational and envy-free.

This completes the proof of this theorem. □

B Equilibrium Outcome Under Various Mechanisms

In this section, we derive the steady state equilibrium outcome under various benchmarks and mechanisms that we discussed in this paper. For each mechanism, (the first best, strict FIFO, direct FIFO, random dispatching, and randomized FIFO), we compute the equilibrium trip throughput, net revenue, average driver payoff, length of the queue, the minimum and maximum waiting times in the queue, and the variance in drivers' total payoffs.

Recall that i^* as defined in (1) is the lowest earning trip that is (partially) completed under the first best outcome, and that $\tilde{\mu}_{i^*} \triangleq \min\{\mu_{i^*},\ \lambda - \sum_{i=1}^{i^*-1}\mu_j\}$ denotes the amount of type i^* jobs fulfilled per unit of time in steady state. Moreover, for any $i \in \mathcal{L}$, $\tau_{i,i+1}$ denotes the amount of time a driver is willing to wait for a trip to location i, assuming that the driver has the option to immediately accept a trip to location $i+1$: $\tau_{i,i+1} = (w_i - w_{i+1})/c$.

B.1 The First Best

We first derive the steady state first best outcome as discussed in Section 2, which dispatches available drivers (upon arrival) to destinations in decreasing order of w_i, until either all drivers are dispatched, or all riders are picked up.

- Trip throughput: $T_{\text{FB}} = \min\{\lambda, \ \sum_{i\in\mathcal{L}} \mu_i\}$.
- Net revenue: $R_{\text{FB}} = \sum_{i=1}^{i^*-1} w_i\mu_i + w_{i^*}\tilde{\mu}_{i^*}$.
- The average payoff of each driver who arrived at the queue: $u_{\text{FB}} = R_{\text{FB}}/\lambda$, since no driver incurs any waiting cost, and R_{FB} is equal to the total payoff achieved by all drivers per unit of time.
- The first best outcome maintains an empty driver queue, therefore the queue length, the minimum and maximum waiting time of any driver, and the average driver waiting time are all zero.
- Since drivers are either dispatched a random trip or asked to leave the queue without waiting any time in the queue, the variance in drivers' payoffs can be computed as follows:
  - When the platform is under-supplied, i.e. $\lambda \le \sum_{i\in\mathcal{L}} \mu_i$, every driver gets dispatched a trip to some location $i \le i^*$. The variance in drivers' payoffs is therefore the variance of the net earnings of the completed trips:

$$\text{Var}\left(U_{\text{FB}}\right) = \frac{1}{\lambda}\left(\sum_{i=1}^{i^*-1} \mu_i(w_i - u_{\text{FB}})^2 + \tilde{\mu}_{i^*}(w_{i^*} - u_{\text{FB}})^2\right).$$

  - When the platform is over-supplied, there are μ_j drivers per unit of time each getting a payoff of w_j, and the rest of the drivers all get zero since they leave the airport a rider trip. Therefore,

$$\text{Var}\left(U_{\text{FB}}\right) = \frac{1}{\lambda}\left(\sum_{i\in\mathcal{L}} \mu_i(w_i - u_{\text{FB}})^2 + (\lambda - \sum_{i\in\mathcal{L}} \mu_i)(0 - u_{\text{FB}})^2\right).$$

B.2 The Direct FIFO Mechanism

As we have proved in the paper, the direct FIFO mechanism has the same trip throughput as the first best throughput, and zero variance in drivers' earnings. Moreover:

- As we have shown in Theorem 2, the steady state equilibrium queue length is $Q^*_{\text{direct}} = \bar{Q}$ if $\lambda > \sum_{i\in\mathcal{L}} \mu_i$, and $Q^*_{\text{direct}} = n_{i^*}$, otherwise.
- Net revenue:
  - When $\lambda > \sum_{i\in\mathcal{L}} \mu_i$, all trips are completed, thus $R_{\text{direct}} = \sum_{i\in\mathcal{L}} \mu_i w_i - Q^*_{\text{direct}} c_p$.
  - When $\lambda \le \sum_{i\in\mathcal{L}} \mu_i$, we have $R_{\text{direct}} = \sum_{i=1}^{i^*-1} \mu_i w_i + \tilde{\mu}_{i^*} w_{i^*} - Q^*_{\text{direct}} c_p$.
- The average driver payoff in equilibrium is $u^*_{\text{direct}} = 0$ if $\lambda > \sum_{i\in\mathcal{L}} \mu_i$, and $u^*_{\text{direct}} = w_{i^*}$ otherwise.
- Regarding the maximum, minimum, and average waiting times in queue:
  - When $\lambda \le \sum_{i\in\mathcal{L}} \mu_i$, the minimum waiting time for a trip (in this case, a trip to location i^*) would be zero. The maximum waiting time (which would be for a trip to location 1) is $\sum_{i=1}^{i^*-1} \tau_{i,i+1} = (w_1 - w - i^*)/c$. The average waiting time for a driver who arrived at the queue is $Q^*_{\text{direct}}/\lambda$, and the average waiting time for a driver who joined the queue is the same.

- When the queue is $\lambda > \sum_{i\in\mathcal{L}} \mu_i$, the minimum amount of time the driver needs to wait in the queue for a trip (which would be for a trip to location ℓ) is w_ℓ/c. The maximum waiting time is w_1/c. The average waiting time for a driver who arrived at the airport is $Q^*_{\text{direct}}/\lambda$, and the average waiting time for a driver who joined the queue $Q^*_{\text{direct}}/\sum_{i\in\mathcal{L}} \mu_i$.

B.3 Strict FIFO Dispatching

Under strict FIFO, when $P \geq n_{i^*}$, all trips that are completed under direct FIFO will be able to reach a driver who is willing to accept them under strict FIFO. As a result, the equilibrium outcome will be identical to that under direct FIFO.

When $P < n_{i^*}$, some trips that are completed under direct FIFO will not be completed under strict FIFO, thus there exists excess drivers who need to leave the queue without a rider trip. Let $i^*(P)$ be the lowest earning trip with $n_i \leq P$, the equilibrium outcome is as follows:

- Trip throughput: $T_{\text{strict}} = \sum_{i=1}^{i^*(P)} \mu_i$.
- The average payoff of drivers is thereby also $u^*_{\text{strict}} = 0$, since drivers will join the queue until when they are indifferent between joining the queue and leaving without a rider.
- Drivers will be willing to wait $w_{i^*(P)}/c$ units of time for a trip to location $i^*(P)$, thus the total length of the queue would be $Q^*_{\text{strict}} = n_{i^*(P)} + T_{\text{strict}} w_{i^*(P)}/c$, which is equal to $\sum_{i=1}^{i^*(P)} \mu_i w_i/c$.
- Net revenue: $R_{\text{strict}} = \sum_{i=1}^{i^*(P)} \mu_i w_i - Q^*_{\text{strict}} c_p$.
- $w_{i^*(P)}/c$ is the minimum amount of time a driver has to wait for any trip, and the maximum waiting time (which would be for a trip to location 1) is w_1/c.
- On average, the total amount of time spent by all drivers on waiting is Q^* units of time, per unit of time. Therefore, the average waiting time for a driver who joined the virtual queue is Q^*/T_{strict}, and the average waiting time for a driver who arrived at the origin is Q^*/λ.
- Every driver gets a zero net payoff, thus the variance in drivers' earnings is also zero.

B.4 Random Dispatching

As we have proved in Proposition 2, random dispatching achieves the same equilibrium trip throughput, net revenue, and queue length as those under the direct FIFO mechanism. As a result, drivers also have the same average payoff and average waiting time. Given the fact that dispatching is random, theoretically drivers might not have to wait any time for a trip dispatch, and there is also no upper bound on a driver's waiting time in the queue.

What is left to compute is the variance in drivers' total payoff. We discuss the over-supplied and the under-supplied settings separately.

Over-supplied. With $\lambda > \sum_{i\in\mathcal{L}} \mu_i$, the average net earnings from a completed trip is $\bar{w} \triangleq \sum_{i\in\mathcal{L}} w_i\mu_i \Big/ \sum_{i\in\mathcal{L}} \mu_i$. The average waiting time of a driver who joined the queue is $\bar{w}/c$, since in equilibrium drivers are indifferent towards whether to join the queue.

Note that (i) a driver's waiting time in the queue is independent to the driver's net earnings from the trip she accepts, and (ii) whether a driver gets dispatched in memoryless, thus a driver's

waiting time is exponentially distributed, with mean $\bar{w}/c$. The variance in drivers' waiting times is therefore $(\bar{w}/c)^2$, thus the variance in drivers waiting costs is $\bar{w}^2$.

In steady state, $\sum_{i\in\mathcal{L}}\mu_i$ drivers join the queue per unit of time. The rest of the drivers do not join the queue thus get 0, which is equal to the average payoff of all drivers. The total variance in the payoff of a driver who arrived at the airport is therefore:

$$\left(\bar{w}^2+\sum_{i\in\mathcal{L}}(w_i-\bar{w})^2\mu_i\Big/\sum_{i\in\mathcal{L}}\mu_i\right)\sum_{i\in\mathcal{L}}\mu_i/\lambda.$$

Under-supplied. Consider now the case when the queue is not over-supplied, and the lowest-earning trips that's completed is i^*. In equilibrium, $\tilde{\mu}_{i^*}$ units of trips to location i^* are completed in each unit of time. The average net earnings of the trip completed by each driver

$$\bar{w}=\left(\sum_{i=1}^{i^*}w_i\mu_i+\tilde{\mu}_{i^*}w_{i^*}\right)\Big/\lambda,$$

and the variance in drivers' net earnings from trips is

$$\left(\sum_{i=1}^{i^*-1}(w_i-\bar{w})^2\mu_i+(w_{i^*}-w_{i^*})^2\tilde{\mu}_{i^*}\right)\Big/\lambda.$$

The average waiting time for a driver is $Q^*/T=(\bar{w}-w_{i^*})/c$, and the distribution of waiting times is exponential. Therefore, the variance in drivers waiting costs is $(cQ^*/T)^2=(\bar{w}-w_{i^*})^2$, and the total variance in drivers' payoff is:

$$(\bar{w}-w_{i^*})^2+\sum_{i=1}^{i^*-1}(w_i-\bar{w})^2\mu_i\Big/\lambda.$$

B.5 Randomized FIFO

As we have proved in Theorem 3, the randomized FIFO mechanisms achieve the same equilibrium trip throughput, net revenue, and queue length as those under the direct FIFO mechanism. As a result, drivers also have the same average payoff and average waiting time in the queue. With randomization in dispatching, it is generally possible for drivers to wait zero or infinite units of time for a dispatch, although there exist special cases where the minimum and maximum waiting times in the queue are non-zero or finite.

We now derive the minimum and maximum waiting times, and the variance in drivers' payoffs. We discuss the same set of cases as analyzed in the proof of Theorem 3 in Appendix A.4.

Case 1.1: $m=i^*=1$. As we have proved in Appendix A.4, the outcome under randomized FIFO in this case is identical to that under the direct FIFO mechanism.

Case 1.2: $i^*>1$, $m=P=1$. As we've discussed in Appendix A.4, the outcome under randomized FIFO in this case will be identical to that under random dispatching, when the queue is not over supplied, i.e. when $\lambda\le\sum_{i\in\mathcal{L}}\mu_i$.

What is left to discuss in the over-supplied case with $\lambda>\sum_{i\in\mathcal{L}}\mu_i$. In this case, every trip is randomly dispatched to drivers at positions $[0,n_\ell]$ in the queue, and no driver declines any dispatches

in equilibrium. With this randomization, the maximum waiting time for a driver in the queue can be infinite, but the minimum time a driver has to wait for a trip would be $(\bar{Q}-n_\ell)/\sum_{i\in\mathcal{L}}\mu_i = w_\ell/c$, since a driver does not receive any dispatch until she has moved from the tail of the queue $Q^* = \bar{Q}$ to position n_ℓ in the queue.

For a driver who joined the queue, the variance in her earnings from the trip she completes is $\sum_{i\in\mathcal{L}}(w_i-\bar{w})^2\mu_i \Big/ \sum_{i\in\mathcal{L}}\mu_i$, where $\bar{w} = \sum_{i\in\mathcal{L}} w_i\mu_i \Big/ \sum_{i\in\mathcal{L}}\mu_i$ is the average net earnings from a rider trip. Once a driver had reached n_ℓ in the queue, the additional time she has to wait for a trip is exponentially distributed with mean $(\bar{w}-w_\ell)/c$. As a result, for a driver who joined the queue, the variance in her waiting cost is $(\bar{w}-w_\ell)^2$, and the overall variance of all drivers who arrived at the queue is:

$$\left((\bar{w}-w_\ell)^2 + \sum_{i\in\mathcal{L}}(w_i-\bar{w})^2\mu_i \Big/ \sum_{i\in\mathcal{L}}\mu_i\right)\sum_{i\in\mathcal{L}}\mu_i/\lambda.$$

Case 2: $m > 1$**.** We now consider the case where there are multiple bins under randomized FIFO.

No driver receives any dispatch until the driver reaches $\bar{b}^{(m)}$. Since $\bar{b}^{(m)} = n_{i^*}$, the minimum amount of time any driver has to wait for a trip is equal to $(Q^*_{\text{rand}} - n_{i^*})/T_{\text{rand}}$, which is equal to 0 if the queue is not over-supplied, and is equal to $(\bar{Q}-n_\ell)/\sum_{i\in\mathcal{L}}\mu_i$ otherwise. What is left to compute is drivers' maximum waiting time in queue, and the variance in drivers' total payoffs.

Recall that $U^* = U(Q^*, Q^*, \sigma^*, \sigma^*)$ is the random variable representing the payoff of all drivers who arrived at the queue, and that $u^* \triangleq \mathbb{E}[U^*] = \pi^*(Q^*)$. From Theorem 3, the average payoff of all drivers who arrived at the queue is the same as that under direct FIFO, i.e. $u^* = w_{i^*}$ when the queue is not oversupplied, and $u^* = 0$, otherwise.

Let $U^{(k)}$ represent the (equilibrium, steady state) total payoff of a driver who is dispatched from the k^{th} bin, and let $U^{(0)} = 0$ denote the payoff of drivers who did not join the queue (if any). We know that

- U^* takes value $U^{(0)}$ with probably $\psi^{(0)} \triangleq \max\{\lambda - \sum_{i\in\mathcal{L}}\mu_i,\ 0\}/\lambda$.
- U^* takes value $U^{(k)}$ with probability $\psi^{(k)} \triangleq s_{k:k}/\lambda = \sum_{i\in\mathcal{L}^{(k)}}\mu_i/\lambda$ for each $k = 1,\ldots,m-1$, and
- U^* takes value $U^{(m)}$ with probability $\psi^{(m)} \triangleq \min\{\lambda - s_{1:m-1},\ s_{m:m}\}/\lambda$.

The overall variance in drivers' total payoff $\text{Var}(U^*)$ can be written as:

$$\mathbb{E}\left[(U^*-u^*)^2\right] = \sum_{k=0}^{m}\psi^{(k)}\mathbb{E}\left[\left(U^{(k)}-u^*\right)^2\right] = \sum_{k=0}^{m}\psi^{(k)}\left(\text{Var}\left(U^{(k)}\right) + \left(\mathbb{E}\left[U^{(k)}\right]-u^*\right)^2\right).$$

$\mathbb{E}\left[U^{(0)}\right] = \text{Var}\left(U^{(0)}\right) = 0$. We now compute $\mathbb{E}\left[U^{(k)}\right]$ and $\text{Var}\left(U^{(k)}\right)$ for each $k \geq 1$.

The last bin: $k = m$. When $|\mathcal{L}^{(m)}| = 1$, $\mathcal{L}^{(m)} = \{i^*\}$ and $\underline{b}^{(m)} = \bar{b}^{(m)} = n_{i^*}$. In this case, there is no variance in the payoffs of all drivers who are dispatched from the last bin: $\text{Var}\left(U^{(m)}\right) = 0$. Moreover, drivers dispatched from the last bin gets the average payoff of all drivers: $\mathbb{E}\left[U^{(m)}\right] = u^*$.

Now consider the setting where $|\mathcal{L}^{(m)}| > 1$. Recall that $\tilde{\mu}_{i^*} \triangleq \min\{\mu_{i^*},\ \lambda - \sum_{j<i^*}\mu_j\}$ is the unit of location i^* trip completed per unit of time in steady state. The rate at which drivers are dispatched from the last bin is:

$$\tilde{s}_{m:m} \triangleq \sum_{i\in\mathcal{L}^{(m)},\ i<i^*}\mu_i + \tilde{\mu}_{i^*}$$

and the average net earnings from a trip accepted by a driver dispatched from the last bin is:

$$\bar{w}^{(m)} \triangleq \left(\sum_{i \in \mathcal{L}^{(m)},\ i<i^*} w_i \mu_i + w_{i^*} \tilde{\mu}_{i^*} \right) \Big/ \tilde{s}_{m:m}.$$

$U^{(m)}$ is equal to the earning from the trip the driver (who is dispatched from the last bin) completes, minus the total waiting costs the driver incurred. The expected net earnings from trip is $\bar{w}^{(m)}$. For the expected waiting cost, note that once a driver reached $\bar{b}^{(m)}$, the driver's additional waiting time for a dispatch should be exponentially distributed, truncated at the time the driver reaches $\underline{b}^{(m)}$. The parameter of this exponential distribution is $\tilde{\eta}^{(m)} \triangleq \tilde{s}_{m:m}/(\bar{b}^{(m)} - \underline{b}^{(m)})$, and by the time the driver reaches $\underline{b}^{(m)}$, $\tilde{s}_{m:m}$ out of the $T_{\text{rand}} = \min\{\lambda, \sum_{i \in \mathcal{L}} \mu_i\}$ drivers who reached position $\bar{b}^{(m)}$ in the queue are dispatched. Denote

$$\zeta^{(m)} \triangleq \tilde{s}_{m:m}/T_{\text{rand}}$$

as the fraction of drivers who are dispatched in the last bin (out of all of the drivers who joined the queue), and denote $\Delta^{(m)}$ as the time it takes for a driver to move from $\bar{b}^{(m)}$ to $\underline{b}^{(m)}$ in the queue if the driver is not dispatched before reaching $\underline{b}^{(m)}$, we know:

$$1 - e^{-\Delta^{(m)} \tilde{\eta}^{(m)}} = \zeta^{(m)} \Rightarrow \Delta^{(m)} = -\log(1 - \zeta^{(m)})/\tilde{\eta}^{(m)}. \tag{62}$$

Denote $\nu^{(m)} \triangleq \Delta^{(m)} + (Q^* - \bar{b}^{(m)})/T_{\text{rand}}$, we know that $\nu^{(m)}$ is the time it takes for a driver to move from the tail of the queue to the lower bound of the last bin $\underline{b}^{(m)}$, if the driver is not dispatched before reaching $\underline{b}^{(m)}$.

For a trip for a driver who is dispatched from the last bin, the average waiting time the driver spends in the last bin is therefore:

$$\int_0^{\Delta^{(m)}} t \cdot \tilde{\eta}^{(m)} e^{-\tilde{\eta}^{(m)} t} \mathrm{d}t = \frac{1}{\tilde{\eta}^{(m)}} (1 + (1/\zeta^{(m)} - 1) \log((1 - \zeta^{(m)})). \tag{63}$$

Let $\kappa^{(m)}$ be the random variable representing the total waiting time of a driver who is dispatched from the last bin, we know that

$$\mathbb{E}\left[\kappa^{(m)}\right] = (Q^* - \bar{b}^{(m)})/T_{\text{rand}} + \frac{1}{\tilde{\eta}^{(m)}} (1 + (1/\zeta^{(m)} - 1) \log((1 - \zeta^{(m)})),$$

and with this we can compute $\mathbb{E}\left[U^{(m)}\right] = \bar{w}^{(m)} - c\mathbb{E}\left[\kappa^{(m)}\right]$.

What is left to compute is $\mathrm{Var}\left(U^{(m)}\right)$. For a driver who is dispatched from the last bin, the earning from the trip the driver receives is independent to the time at which the driver receives the trip. As a result, the variance $\mathrm{Var}\left(U^{(m)}\right)$ should be the sum of the variance in trip earnings and the variance in the waiting costs. The former is equal to:

$$\left(\sum_{i \in \mathcal{L}^{(m)},\ i<i^*} (w_i - \bar{w}^{(m)})^2 \mu_i + (w_{i^*} - \bar{w}^{(m)}) \tilde{\mu}_{i^*} \right) \Big/ \tilde{s}_{m:m},$$

and the latter is of the form:

$$\begin{aligned} & c^2 \int_0^{\Delta^{(m)}} \left(t - \frac{1}{\tilde{\eta}^{(m)}} (1 + (1/\zeta^{(m)} - 1) \log((1 - \zeta^{(m)})) \right)^2 \tilde{\eta}^{(m)} e^{-\tilde{\eta}^{(m)} t} \mathrm{d}t \\ = & \left(\frac{c}{\tilde{\eta}^{(m)} \zeta^{(m)}} \right)^2 \left((\zeta^{(m)})^2 + (-1 + \zeta^{(m)}) \left(\log(1 - \zeta^{(m)}) \right)^2 \right). \end{aligned}$$

A middle bin. Now consider each $k = m-1, m-2, \ldots, 2$. When $|\mathcal{L}^{(k)}| = 1$, there is no variance in the payoffs of drivers who are dispatched from the k^{th} bin: $\text{Var}\left(U^{(k)}\right) = 0$. It takes the driver a total of $\nu^{(k+1)} + (\underline{b}^{(k+1)} - \bar{b}^{(k)})/s_{1:k}$ units of time to move from the tail of the queue to $\bar{b}^{(k)}$, and once the driver gets to $\bar{b}^{(k)}$, the driver receives $\bar{w}_{k:k}$, which is equal to the net earnings from the only trip in $\mathcal{L}^{(k)}$. Therefore, $\mathbb{E}\left[U^{(k)}\right] = \bar{w}_{k:k} - c\left(\nu^{(k+1)} + (\underline{b}^{(k+1)} - \bar{b}^{(k)})/s_{1:k}\right)$.

For the case where $|\mathcal{L}^{(k)}| > 1$, recall that $\eta^{(k)} \triangleq s_{k:k}/(\bar{b}^{(k)} - \underline{b}^{(k)})$, and denote:

$$\zeta^{(k)} \triangleq s_{k:k}/s_{1:k}.$$

With the same argument as those for the last bin, the time it takes a driver to move from $\bar{b}^{(k)}$ to $\underline{b}^{(k)}$ (if the driver is not dispatched before reaching $\underline{b}^{(k)}$) is

$$\Delta^{(k)} \triangleq -\log(1-\zeta^{(k)})/\eta^{(k)}.$$

This implies that the total waiting time for a driver to reach $\underline{b}^{(k)}$ (if the driver is not dispatched before then) is $\nu^{(k)} \triangleq \nu^{(k+1)} + (\underline{b}^{(k+1)} - \bar{b}^{(k)})/s_{1:k} + \Delta^{(k)}$.

The expected time a driver waits in the k^{th} bin, if the driver is dispatched from the k^{th} bin, is of the form:

$$\int_0^{\Delta^{(k)}} t \cdot \eta^{(k)} e^{-\eta^{(k)} t} \mathrm{d}t = \frac{1}{\eta^{(k)}}(1 + (1/\zeta^{(k)} - 1)\log((1-\zeta^{(k)})).$$

thus the expected payoff of a driver who is dispatched in the k^{th} bin is:

$$\mathbb{E}\left[U^{(k)}\right] = \bar{w}_{k:k} - c\left(\nu^{(k+1)} + (\underline{b}^{(k+1)} - \bar{b}^{(k)})/s_{1:k} + \frac{1}{\eta^{(k)}}(1 + (1/\zeta^{(k)} - 1)\log((1-\zeta^{(k)}))\right).$$

The variance $\text{Var}\left(U^{(k)}\right)$ is similarly consisted of two parts. The variance from the net earnings from a trip a driver accepts in the k^{th} bin is

$$\sum_{i \in \mathcal{L}^{(k)}} (w_i - \bar{w}^{(k)})^2 \mu_i / s_{k:k},$$

and the variance of drivers' waiting costs is:

$$\begin{aligned} &c^2 \int_0^{\Delta^{(k)}} \left(t - \frac{1}{\eta^{(k)}}(1 + (1/\zeta^{(k)} - 1)\log((1-\zeta^{(k)}))\right)^2 \eta^{(k)} e^{-\eta^{(k)} t} \mathrm{d}t \\ =& \left(\frac{c}{\eta^{(k)}\zeta^{(k)}}\right)^2 \left((\zeta^{(k)})^2 + (-1+\zeta^{(k)})\left(\log(1-\zeta^{(k)})\right)^2\right). \end{aligned}$$

The first bin. When $|\mathcal{L}^{(1)}| = 1$, $\text{Var}\left(U^{(1)}\right) = 0$. The total waiting time for a driver to reach $\bar{b}^{(1)} = 0$ is $\nu^{(2)} + \underline{b}^{(2)}/\mu_1$, thus the expected payoff of a driver dispatched from the first bin is $\mathbb{E}\left[U^{(1)}\right] = w_1 - c\left(\nu^{(2)} + \underline{b}^{(2)}/\mu_1\right)$. Moreover, in this case, the maximum waiting time for any driver who have joined the queue is $\nu^{(2)} + \underline{b}^{(2)}/\mu_1$.

When $|\mathcal{L}^{(1)}| > 1$, the waiting time a driver may spend waiting in the queue is unbounded. Once a driver reached $\bar{b}^{(1)}$, the driver's waiting time for a trip is exponentially distributed with parameter $\eta^{(1)} \triangleq s_{1:1}/(\bar{b}^{(1)} - \underline{b}^{(1)})$. The expected waiting time in the first bin is $1/\eta^{(1)}$, thus the expected payoff of a driver who is dispatched from the first bin is

$$\mathbb{E}\left[U^{(1)}\right] = \bar{w}_{1:1} - c\left(1/\eta^{(1)} + \nu^{(2)} + (\underline{b}^{(2)} - \bar{b}^{(1)})/s_{1:1}\right),$$

and $\text{Var}\left(U^{(1)}\right)$, the variance of the earnings of a driver who is dispatched from the first bin is:

$$\text{Var}\left(U^{(1)}\right) = \sum_{i \in \mathcal{L}^{(1)}} (w_i - \bar{w}_{1:1})^2 \mu_i / s_{1:1} + (c/\eta^{(1)})^2.$$

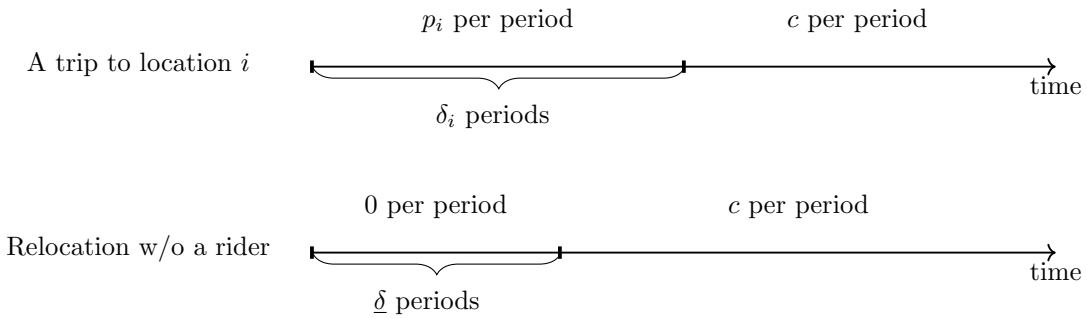


Figure 10: The timeline of a trip to location i (above) and relocation without a rider (below).

C Additional Discussion and Examples

C.1 Net Earnings from Prices and Distances

For each location $i \in \mathcal{L}$, the net earnings w_i from a trip to location i represents the total payoff of a driver who completed a trip to location i, minus the total payoff to a driver who left without a rider. This incorporates the driver's payment from the trip, the potential difference in drivers' continuation earnings after arriving at different destinations (which are affected by market conditions at the destinations), as well as the potential relocation costs of those drivers who left the queue without a rider. In this way, the payoff of a driver who left without a rider is normalized to be zero.

Assuming that drivers get the same continuation earnings from every destination, we discuss how net earnings of trips can be derived from the prices and distances of trips to different destinations. For each destination $i \in \mathcal{L}$, let $\delta_i > 0$ denote the amount of time (e.g. minutes) it takes for a driver to complete a trip to destination i, which includes time it takes for a driver to pick up the rider. $p_i > 0$ denotes the effective earnings rate from a trip to location $i \in \mathcal{L}$, meaning that the total payment to a driver for a trip to location i is $p_i\delta_i$. The earnings rates are induced by the time and distance rates the platform pays the drivers, and may vary across destinations due to differences in trip lengths and traffic conditions, etc. For drivers who decides to relocate without a rider and drive elsewhere, $\underline{\delta} > 0$ is the minimum relocation distance, i.e. the amount of time a driver needs to spend driving from e.g. the airport in order to start making an earnings rate of c.

A driver who accepts a trip from the virtual queue to some location $i \in \mathcal{L}$ will make p_i per unit of time for δ_i periods, followed by making c per period after arriving at location i. A driver who relocates back to the city without a rider makes 0 for the first $\underline{\delta}$ periods, and then starts to earn c per period. See Figure 10. We refer to the additional earnings from a trip to location i, relative to that from relocating without a rider (and then driving in the city), as the *net earnings* from this trip. For each location $i \in \mathcal{L}$, the net earnings w_i is of the form:

$$w_i = \delta_i p_i - (\underline{\delta} \cdot 0 + (\delta_i - \underline{\delta})c) = \delta_i(p_i - c) + \underline{\delta}c. \tag{64}$$

C.2 Additional Examples

Example 3 shows that a randomized FIFO mechanism may not achieve the second best outcome, when some trips with lower earnings than w_{i^*} is included in the last partition.

Example 3 (Last bin w/ trips to location $i \geq i^*$)**.** Consider an economy with three destinations, where $\mu_1 = 1$, $w_1 = 100$, $\mu_2 = 2$, $w_2 = 40$, and $\mu_3 = 5$, $w_3 = 10$. Assume that the arrival rate of

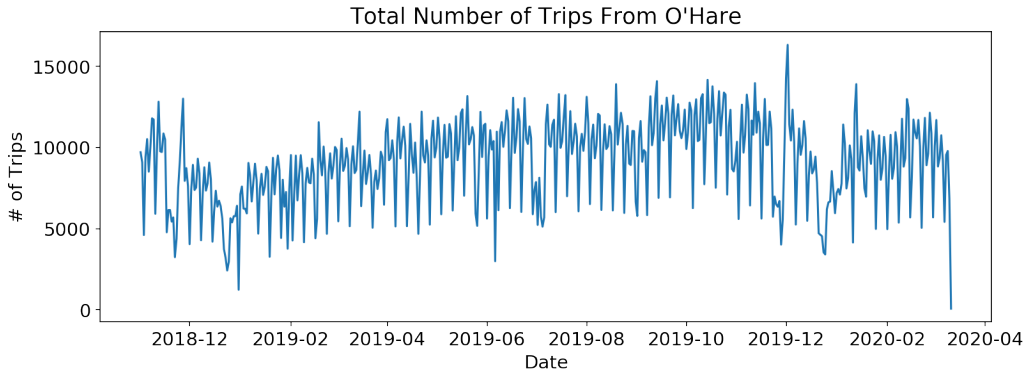


Figure 11: Total number of trips per day from the Chicago O'Hare International Airport.

drivers is $\lambda = 2$, and that the opportunity cost for drivers' time is $c = 1$. Under the second best outcome, one unit of trips to location 1 and one unit of trips to location 2 are completed per unit of time. With $i^* = 2$, the average net payoff of drivers would be $w_{i^*} = w_2 = 40$, and the equilibrium queue length is $Q^* = n_{i^*} = \mu_1(w_1 - w_2)/c = 60$.

Now consider a randomized FIFO mechanism associated with the ordered partition $\mathcal{L}^{(1)} = \{1\}$ and $\mathcal{L}^{(2)} = \{2,\ 3\}$. Constructing the bins according to (13) and (14), we have $\underline{b}^{(1)} = \bar{b}^{(1)} = 0$ and

$$\underline{b}^{(2)} = \frac{1}{c}\mu_1(w_1 - w_3) = 90.$$

Note that $\underline{b}^{(2)}$ is higher than the equilibrium queue length under the second best outcome, and we show that the randomized FIFO mechanism will not achieve the second best in this case. For a driver in the first bin, i.e. at the head of the queue, the driver will only be willing to accept a trip to location 1. When the queue is shorter than $\underline{b}^{(2)}$, trips to location 2 or 3 will not be dispatched by the randomized FIFO mechanism, thus all but location 1 trips become unfulfilled. But when the queue is longer than $\underline{b}^{(2)}$, we will not be achieving our second best outcome either, since the total waiting costs incurred by the drivers will be higher than that under the second best outcome, even when we are completing the same set of trips. □

D Additional Simulations

We include in this section simulation results for the Midway airport, as well as additional simulation results for O'Hare that are omitted from Section 5 of the paper.

D.1 Chicago O'Hare International Airport

D.1.1 Trip Volume by Day and Hour-of-Week

We include in this section the volume of trips originating from the Chicago O'Hare International Airport, and the average duration and earning rates by destination. Figure 11 shows the number of trips that originate from the O'Hare airport on each day, from November 1, 2018 to mid March, 2020. We can see strong seasonality patterns (e.g. low numbers of trips during Christmas through New Year), and also the sharp decline in trip volume after the onset of the COVID-19 pandemic.

The average number of trips originating from O'Hare during each hour-of-week is as shown in Figure 12. Here, the 0^{th} hour-of-week corresponds to midnight - 1am on Mondays, and the 1^{st} hour-of-week corresponds to 1am - 2am on Mondays, and so on. Gray stripes indicate the morning rush

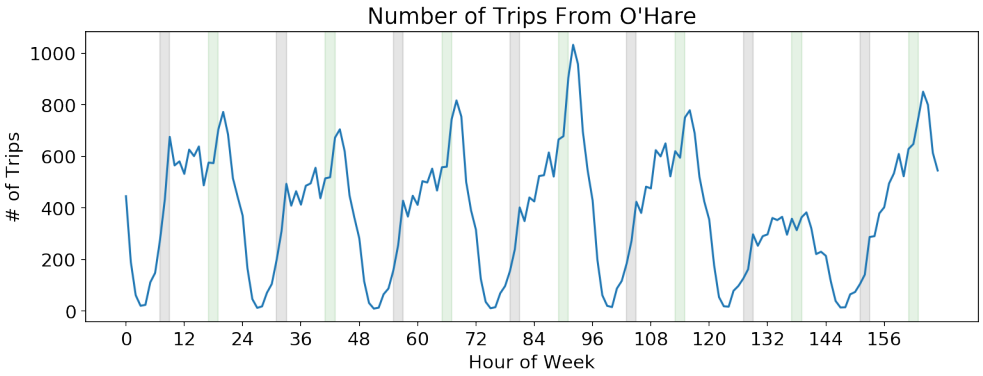
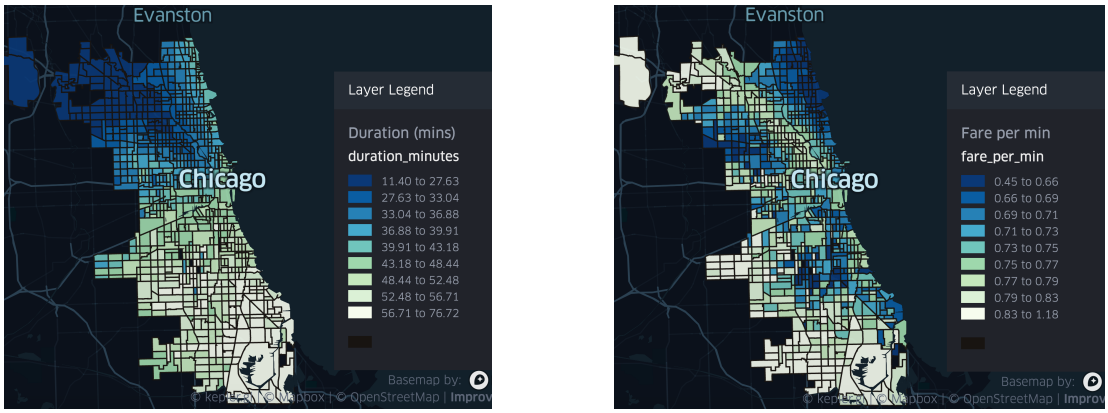


Figure 12: Average number of trips from the O'Hare International Airport, by hour-of-week.



(a) Average trip duration.

(b) Average fare per minute.

Figure 13: Average trip duration (in minutes) and the average fare per minute by destination Census Tract in Chicago, for trips originating from the Chicago O'Hare International Airport.

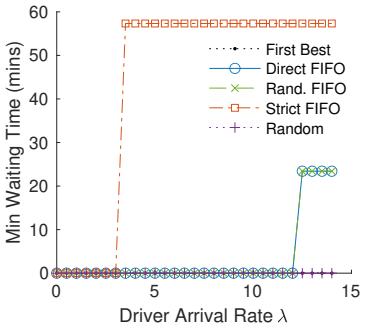
hours (7am - 9am) and the green stripes indicate the evening rush hours (5pm - 7pm). We can see that the number of trips originating from the airport peaks during early evenings, averaging around 12 trips per minute during the weekdays, and peaking at over 15 trips per minute on Thursday. Note that these are completed trips, thus the rider request rates should be strictly higher.

In Figure 13, we illustrate the average duration of trips, and the average earning rates (trip fare divided by trip duration) for each destination census tract. We can see that longer trips take more time on average, and trips ending closer to the major highways have better earnings rates.

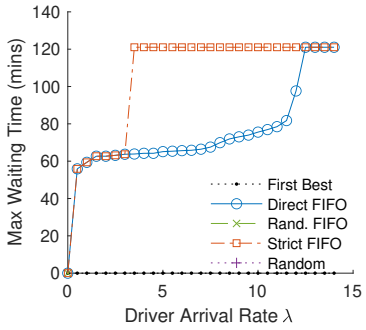
D.1.2 Counterfactual Simulations

We now provide additional results for O'Hare that are omitted from Section 5 of the paper.

Varying Driver Supply As the arrival rate of driver varies, Figure 14 presents the minimum and maximum waiting times for drivers who joined the queue in equilibrium in steady state. For strict FIFO, and direct FIFO, the minimum waiting time is the time a driver need to wait in the queue for the lowest earning trip that is completed in equilibrium. Under randomized FIFO, the minimum waiting time is the time it takes for a driver to move from the tail of the queue to the

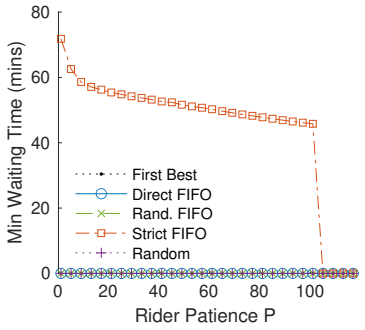


(a) Minimum waiting time.

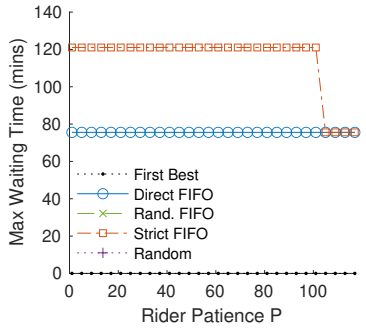


(b) Maximum waiting time.

Figure 14: The minimum and maximum waiting time for drivers who join the queue, in equilibrium in steady state, as the arrival rate of drivers varies. Chicago O'Hare.



(a) Minimum waiting time.



(b) Maximum waiting time.

Figure 15: The minimum and maximum waiting time for drivers who join the queue, in equilibrium in steady state, as the patience level of the riders varies. Chicago O'Hare.

upper bound of the last bin $\underline{b}^{(m)}$. When the queue is not over-supplied, the minimum waiting times under direct FIFO and randomized FIFO are zero.

Under strict FIFO and direct FIFO, the maximum waiting times are the time a driver needs to wait in the queue for a trip to location 1, the highest earning trip. Under random dispatching or randomized FIFO, there is no upper bound on how long a driver may need to wait in the queue.

Varying Rider Patience Figure 15 presents the minimum and maximum waiting times for drivers who joined the queue, as we vary the patience level of riders. We can see that under strict FIFO, the minimum waiting time for drivers decreases very slowly as riders' patience level increases, despite the fact that no driver has to wait at all under every other mechanism if they would like to leave the queue as quickly as possible.

D.2 Chicago Midway International Airport

In this section, we present simulation results for trips originating from the Chicago Midway International airport. First, we plot in Figures 16 and 17 the daily number of trips originating from Midway, and the average number of of trips by hour-of-week. Figure 18, shows the total trip count by destination census tract, and the estimated net earnings by destination assuming that drivers' opportunity costs is $c = 1/3$ per minute.

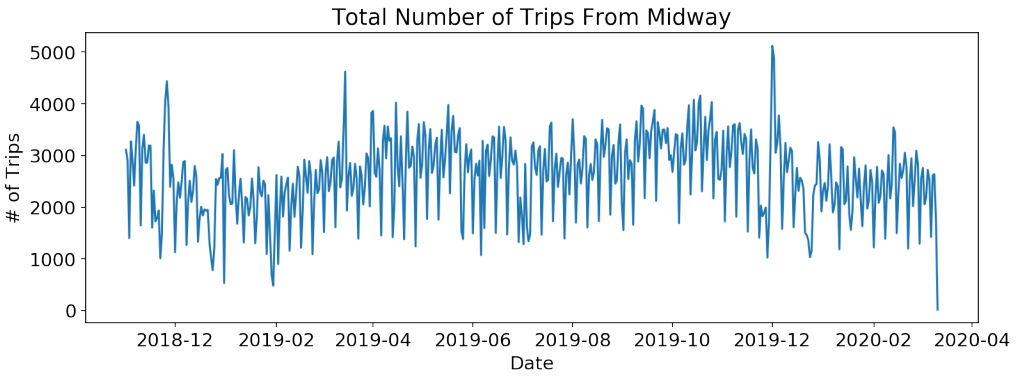


Figure 16: Total number of trips per day from the Chicago Midway International Airport.

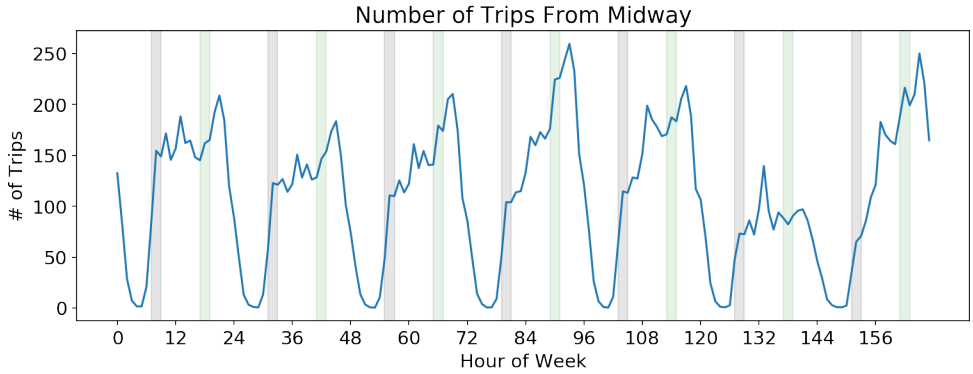
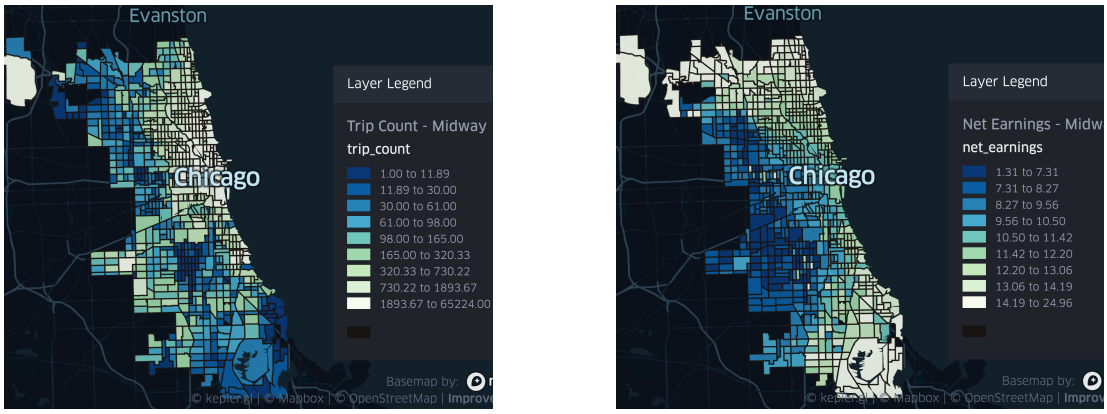


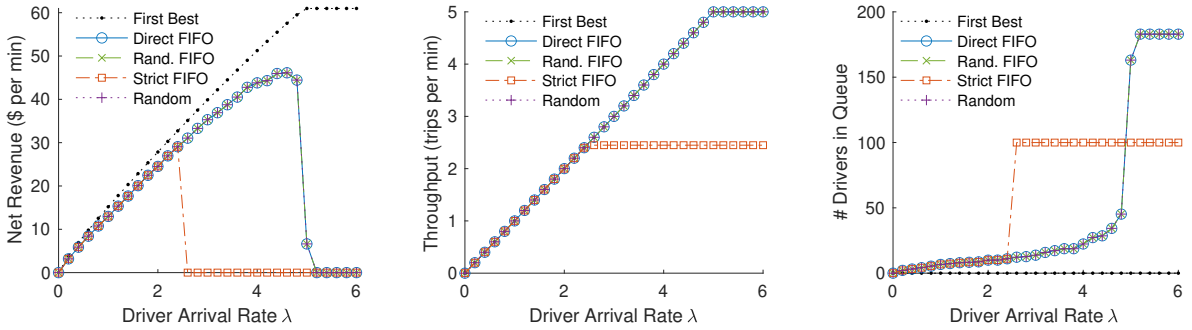
Figure 17: Average number of trips from the Midway International Airport, by hour-of-week.



(a) Trip count by destination.

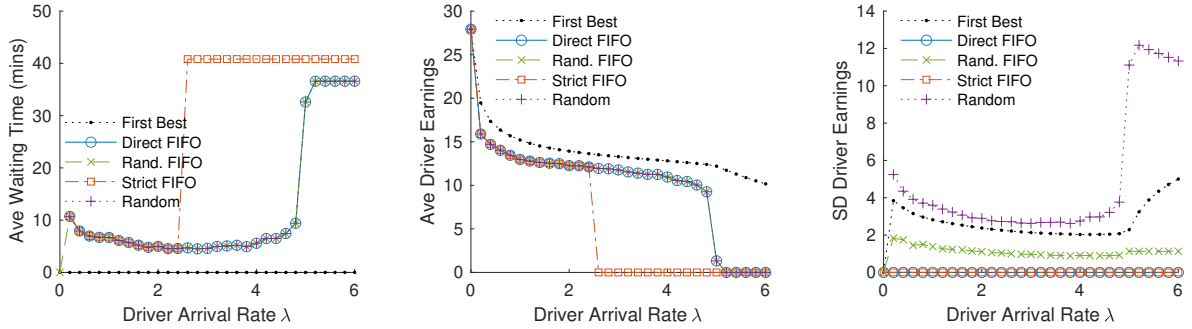
(b) Net earnings by destination.

Figure 18: Trip volume and net earnings (assuming $c = 1/3$) by destination Census Tract in Chicago, for trips originating from the Chicago Midway International Airport.



(a) Net revenue. (b) Trip throughput. (c) Equilibrium queue length.

Figure 19: Equilibrium net revenue, trip throughput, and length of the queue in steady state, as the arrival rate of drivers varies. Chicago Midway.



(a) Average waiting time. (b) Average driver payoff. (c) SD of driver payoffs.

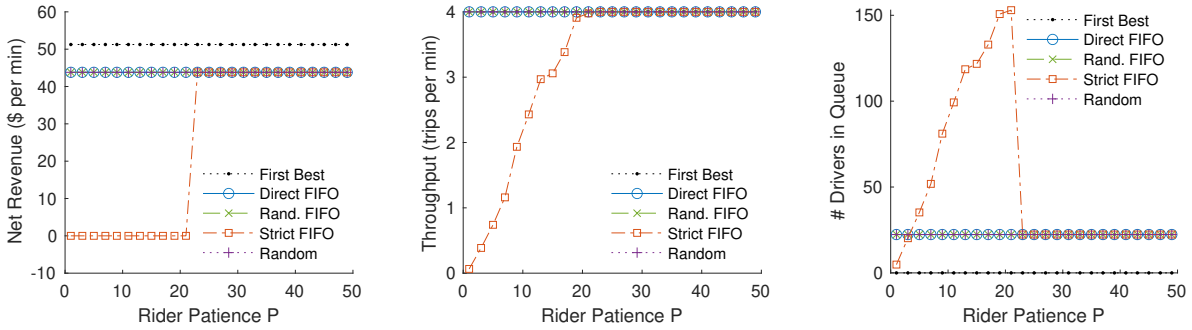
Figure 20: Drivers' average waiting times, total payoff, and the standard deviation (SD) in drivers' total payoff in equilibrium in steady state, as the arrival rate of drivers varies. Chicago Midway.



(a) Minimum waiting time. (b) Maximum waiting time.

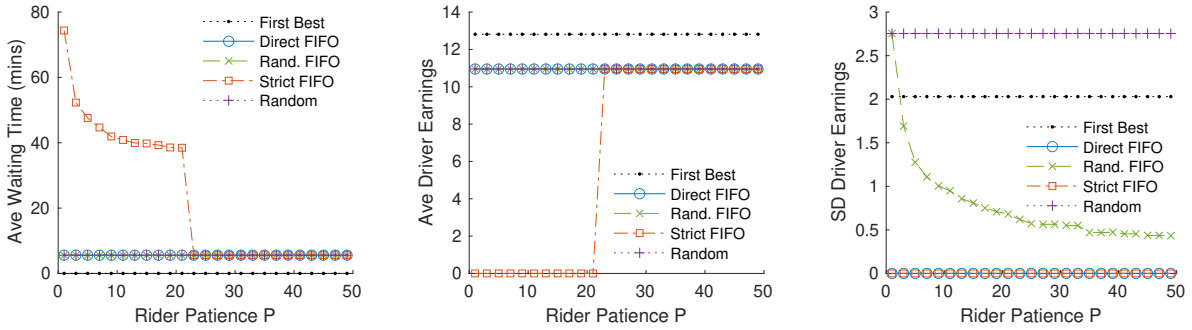
Figure 21: The minimum and maximum waiting time for drivers who join the queue, in equilibrium in steady state, as the arrival rate of drivers varies. Chicago Midway.

Varying Driver Supply We now compare the equilibrium, steady state outcome under various mechanisms and benchmarks, as we vary the arrival rate of drivers from 0 to 6 drivers per minute. We fix the total arrival rate of riders at $\sum_{i\in\mathcal{L}} \mu_i = 5$, and the rider patience level at $P = 12$. See Figures 19, 20 and 21. The observations here are fully aligned with those for the O'Hare airport presented in Section 5.



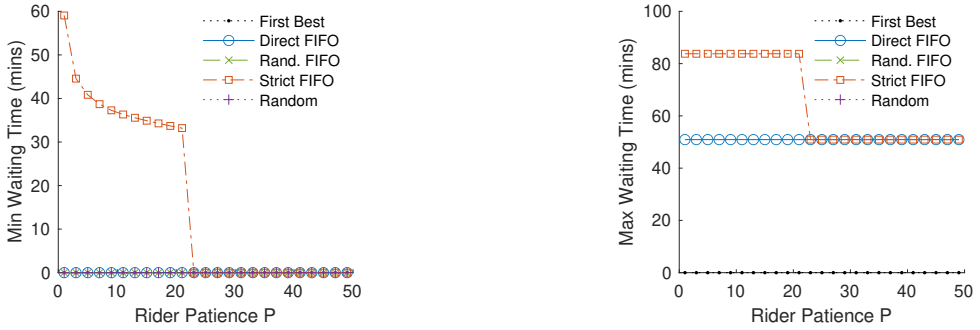
(a) Net revenue. (b) Trip throughput. (c) Equilibrium queue length.

Figure 22: Equilibrium net revenue, trip throughput, and length of the queue in steady state, as the rider patience level varies. Chicago Midway.



(a) Average waiting time. (b) Average driver payoff. (c) SD of driver payoffs.

Figure 23: Drivers' average waiting times, total payoff, and the standard deviation (SD) in drivers' total payoff in equilibrium in steady state, as the rider patience level varies. Chicago Midway.



(a) Minimum waiting time. (b) Maximum waiting time.

Figure 24: The minimum and maximum waiting time for drivers who join the queue, in equilibrium in steady state, as the patience level of the riders varies. Chicago Midway.

Varying Rider Patience Fixing the arrival rate of drivers at $\lambda = 4$, Figures 22, 23 and 24 compare the equilibrium outcome under various mechanisms and benchmarks as we vary the patience level of the riders. The observations are, again, fully aligned with those for the O'Hare airport, presented in Section 5 of the paper.