6-2018

Coordinating supply and demand on an on-demand service platform with impatient customers

Jiaru BAI
Binghamton University--SUNY

Kut C. SO
University of California, Irvine

Christopher S. TANG
University of California, Los Angeles

Xiqun CHEN
Zhejiang University

Hai WANG
Singapore Management University, haiwang@smu.edu.sg

DOI: https://doi.org/10.1287/msom.2018.0707

Coordinating Supply and Demand on an On-demand Service Platform with Impatient Customers

Jiaru Bai
School of Management, Binghamton University, Binghamton, NY 13902, USA
jbai@binghamton.edu

Kut C. So
The Paul Merage School of Business, University of California, Irvine, CA 92697, USA
rick.so@uci.edu

Christopher S. Tang
UCLA Anderson School, 110 Westwood Plaza, Los Angeles, CA 90095, USA
chris.tang@anderson.ucla.edu

Xiqun (Michael) Chen
College of Civil Engineering and Architecture, Zhejiang University, Hangzhou 310058, China
chenxiqun@zju.edu.cn

Hai Wang
School of Information Systems, Singapore Management University, Singapore
haiwang@smu.edu.sg

December 20, 2017

Abstract

We consider an on-demand service platform using earning sensitive independent providers with heterogeneous reservation price (for work participation) to serve its time and price sensitive customers with heterogeneous valuation of the service. As such, both the supply and demand are "endogenously" dependent on the price the platform charges its customers and the wage the platform pays its independent providers. We present an analytical model with endogenous supply (number of participating agents) and endogenous demand (customer request rate) to study this on-demand service platform. To coordinate endogenous demand with endogenous supply, we include the steady-state waiting time performance based on a queueing model in the customer utility function to characterize the optimal price and wage rates that maximize the profit of the platform (as well as the total welfare). We first analyze a base model that uses a fixed payout ratio (i.e., the ratio of wage over price), and then extend our model to allow the platform to adopt a time-based payout ratio. We find that it is optimal for the platform to charge a higher price when demand increases; however, the optimal price is not necessarily monotonic when the provider capacity or the waiting cost increases. Furthermore, the platform should offer a higher payout ratio as demand increases, capacity decreases or customers become

more sensitive to waiting time. We also find that the platform should lower its payout ratio as it grows with the number of providers and customer demand increasing at about the same rate. We use a set of actual data from a large on-demand ride-hailing platform to calibrate our model parameters in numerical experiments to illustrate some of our main insights.

Keywords: On-Demand Services, Endogenous Supply and Demand, Queueing Models.

1 Introduction

Recent advances in internet/mobile technologies have enabled the creation of various innovative on-demand service platforms for providing *on-demand* services anytime/anywhere. Examples include grocery delivery services (e.g., Instacart, Google Express), meal delivery services (e.g., Sprig, Blue Apron), and food delivery services directly from restaurants (e.g., DoorDash, Deliveroo (U.K.), UberEats, Yelp Eat24), consumer goods delivery services (e.g., UberRush), dog-walking services (e.g., Wag), and ride-hailing services (e.g., Uber, Lyft, Didi). Furthermore, the adoption of mobile applications as well as the availability of on-demand service platforms increase the expectations and demands of impatient customers for quick services.

To meet dynamic customer demand anytime/anywhere, it is economical for on-demand service firms to use independent providers (or agents) to fulfill customer requests quickly. However, using independent agents to deliver on-demand services can be challenging, as work participation of independent providers is primarily driven by earnings. As independent agents do not get compensated for idle times, earnings depends on wage rate and utilization, whereas utilization depends on customer demand. At the same time, the demand associated with time and price sensitive customers depends on two key factors: price and waiting time. Since customer's waiting time is highly dependent on the number of participating agents (which is a function of agent's wage and customer demand), the "supply" of participating agents and the "demand" of customer requests are endogenously dependent on the wage and the price specified by the firm.

An on-demand service firm needs to analyze the underlying interactions between supply and demand so as to select the optimal wage and price. The firm must carefully coordinate endogenous supply and demand in different time periods by: (1) setting the right wage (i.e., compensation) to get the right supply (i.e., the right number of earning sensitive participating agents); and (2) charging the right price to control the right demand (i.e., the right number of time and price sensitive customers). To elaborate, consider the simple case when the demand is fixed. If the firm offers a higher wage, more agents will participate and customer satisfaction will increase due to a quicker service. However, each participating agent will earn less due to low utilization. On the other hand, if the firm offers a lower wage, fewer agents will participate and customer satisfaction will decrease due to longer waiting times.

In view of the intricate relationship between endogenous supply and demand through wage and price selections, we develop an analytical framework to examine how an on-demand service firm should set its price rate for the customers and its wage rate for the providers. In our framework, we use a queueing model to capture the underlying waiting time where both supply (i.e., number of providers) and demand (i.e., customer arrival rate) are "endogenously" dependent on wage, price and other operating factors. Our model captures an operating environment where (1) time and price sensitive customers are "heterogeneous" in their *valuation* of the service; and (2) earning sensitive independent providers are "heterogeneous" in their *reservation earning rate* (i.e., the minimum wage for work participation).

We only consider time-based pricing (instead of dynamic pricing) in our analytical framework, i.e., the price rate can change across different time periods, but is known in advance to customers. Also, we only consider time-based wages so that the wage schedule is known in advance to service providers. Besides the fact that time-based pricing (and wages) is practical, it is considered to be fairer than the dynamic pricing/wages that is not known to customers/providers in advance. For instance, the behavior experiments conducted by Haws and Bearden (2006) reveal that consumers viewed price changes within very short time periods as being more "unfair" than price changes over a more extended period of time. Therefore, for practical reasons and for tractability, we focus on time-based strategy in this paper.

We first use the analytical framework to construct a base model for a common situation in which an on-demand service platform adopts a *fixed payout ratio* of wage over price to pay its service providers. (Throughout this paper, we refer to "payout ratio" as the wage offered to the providers as a percentage of the price paid by the customers.) By including waiting time performance based on an $M/M/k$ queueing model in the customer utility function, we analyze the optimal price and wage rates that maximize the expected profit of the service platform. We conduct extensive numerical experiments to generate managerial insights on how to select the optimal price and wage rates for the platform. We further develop a good approximation of the steady-state waiting time function to provide analytical results that support the insights derived from our numerical experiments.

For the base model, we find that the platform should increase the price rate (and the wage rate accordingly) when customer demand increases. This result thus supports an on-demand ride-hailing service platform that uses a fixed payout ratio (such as Uber) of charging a higher price (and

offering a higher wage) during rush hours when the customer demand is high. Interestingly, we find that the optimal price (and wage) rate is not necessarily monotonic in the maximum number of available service providers.

We then extend our base model to analyze the general situation in which the on-demand service platform can use a *time-based payout ratio* to pay the providers in order to maximize its profit. We analyze the optimal price and wage rates and evaluate the potential benefit of using a time-based payout ratio over a fixed payout ratio. Similar to the base model, we use an approximation of the waiting time function to provide analytical results that support the insights derived from our numerical experiments. We further extend our analysis to a more general setting under which the objective is to maximize the platform's profit plus the welfare of the consumers and providers.

For the general model (based on time-based payout ratio), we find that the optimal price and wage rates increase as customer demand increases. Interestingly, the impact of service capacity on the optimal price rate is more subtle. We find that the optimal price is not necessarily monotonic as the maximum number of available service providers increases. Similarly, the optimal price is not necessarily monotonic as waiting cost increases. This non-monotonic property can be explained by the queueung effects captured in the customer utility function. We also find that, when the customer's valuation of the service and the provider's earning reservation are uniformly distributed, the optimal payout ratio increases when demand increases, service capacity decreases, or customers become more sensitive to waiting time. In other words, the platform should increase its payout ratio at time periods with high demand, but reduce its payout ratio when the number of registered independent providers increases. For urgent on-demand services with highly time sensitive customers, the firm needs to increase its payout ratio to attract more service providers to handle the increasingly impatient customers. We also find that the platform should lower its payout ratio as it grows with the number of providers and customer demand increasing at about the same rate.

Our results also show that the profit can be greatly reduced if the platform uses a fixed payout ratio that is far from the optimal time-based payout ratio, and that the optimal time-based payout ratio can vary widely depending on specific operating characteristics. This implies that, while it is simple for the platform to share a fixed percentage of its revenue with its independent providers, the platform should adopt a time-based payout ratio to maximize profitability across different time periods when the underlying operating characteristics can change significantly. We hope our results

might motivate on-demand service firms adopting a fixed payout scheme to carefully re-evaluate the effectiveness of such a fixed payout scheme.

This paper is organized as follows. We provide a brief review of related literature in Section 2. Section 3 presents our modeling framework of endogenous supply and demand along with heterogeneous providers and customers. In Section 4, we develop our base model for analyzing a common situation in which the on-demand platform adopts a *fixed payout ratio*. We analyze the optimal price and wage rates that maximize the platform's profit using extensive numerical experiments. We further develop a good approximation scheme to provide analytical support of the insights derived from the numerical experiments. In Section 5 we extend our base model to analyze the general situation in which the platform can use a *time-based payout ratio* in order to maximize its profit. We further extend our analysis to the case when the objective of the platform is to maximize its own profit plus the welfare of the consumers and providers. In Section 6, we construct some illustrative numerical examples based on actual data provided by Didi: the leading on-demand ride-hailing service in China. We conclude the paper in Section 7. For ease of exposition, all mathematical proofs for the results in the main text are provided in an Appendix A.

2 Literature Review

Our paper relates to pricing strategies in two-sided markets in the industrial organization literature. Our framework is akin to the models developed by Rochet and Tirole (2003, 2006) and Armstrong (2006) in the following sense. Our framework studies a service platform that maximizes its profit by charging prices (wage can be viewed as a negative price) to both sides of the market, which captures some positive "cross-group" externalities, i.e., the utility of an agent in one side increases with the number of agents in the other side. However, our framework differs from their setting in two important aspects. First, our framework incorporates a queueing model, which is a salient feature of a ride-sharing platform. As such, our framework also captures the within-group effects in which an increase in customer demand would reduce customer utility due to an increase in waiting time in the demand side, and an increase in service providers would reduce provider earnings due to lower utilization in the supply side. We shall further discuss how the non-linear queueing effect can affect the structural results in Section 5.2. Second, our framework considers a different objective function. Rochet and Tirole (2003, 2006) use the product of the difference in price and wage rate, supply and

demand in the objective function. Hu and Zhou (2017) use the product of the difference in price and wage rate and the minimum of supply and demand in the objective function. In contrast, our framework uses the product of the difference in price and wage the "throughput rate". Notice that the throughput rate is a non-separable function of the arrival rate (or throughput) and the number of servers in an equilibrium, and these two factors depend on the underlying price and wage rates.

Our paper belongs to an emerging stream of research that examines operations and pricing issues arising from the *sharing economy*; see e.g., Benjaafar et al. (2015), Fraiberger and Sundararajan (2015), and Jiang and Tian (2015) examined a customer's decision to purchase or to rent assets in the presence of "product sharing platforms" such as Airbnb. By crawling data from Airbnb, Li et al. (2015) showed empirically that "professional" owners earned more. For many of such sharing platforms, the owners set the price, the platforms set the payout amounts to the owners, and customers often reserve the service in advance. In contrast, our paper studies on-demand service platforms which provide time-sensitive service in an on-demand manner and addresses different decision issues in managing the underlying service request mechanisms.

Recent developments of various on-demand service platforms such as Uber and DoorDash (see Kokalitcheva (2015), Wirtz and Tang (2016), and Shoot (2015)) have motivated researchers to explore various operational issues. First, there is an on-going debate regarding the definition of independent contractors for various on-demand service platforms (e.g., see Roose (2014)). At the same time, it is of interest to examine how dynamic wage affects supply, especially when independent providers can freely choose whether and when to work. Chen and Sheldon (2015) examined transactional data associated with 25 million trips obtained from Uber and showed empirically that dynamic wage (due to surge pricing) could entice independent drivers to work for longer hours. Sheldon (2016) analyzed data from a peer-to-peer ride-sharing firm to examine the supply elasticity of individual contractors in the ride-sharing market. Moreno and Terwiesch (2014) also examined empirically the independent contractor's bidding behavior on freelancing platforms. Allon et al. (2012) explored the process for matching providers to consumers when capacities were exogenous.

A number of researchers have recently studied the impact of wage and price on supply and for on-demand services and examined whether it would be beneficial for an on-demand service firm to adjust its prices and wages dynamically based on real-time system information including the current number of customers requesting service and the number of providers in the system. Riquelme et al.

(2015) and Cachon et al. (2015) compared the impact of static versus dynamic prices and wages. When customers were heterogeneous in terms of valuation and the payout ratio was exogenously given, Riquelme et al. (2015) found that static pricing performed well. On the other hand, Cachon et al. (2015) found that surge pricing performed well when customers were homogeneous and the payout ratio was endogenously determined. When the profit function of the platform is the minimum of demand (a linear function of price) and supply (a linear function of wage), Hu and Zhou (2017) showed that it is optimal for the platform to offer a constant payout ratio, which depends on the price and wage sensitivity coefficients of the linear demand and supply functions. Moreover, their main focus is to provide performance bounds for an endogenized fixed payout ratio. Gurvich et al. (2015) developed a newsvendor-style model to examine the optimal price and wage decisions. This stream of research has assumed that customer demand is independent of waiting time and service capacity is independent of system utilization over time. In contrast, our model captures the rational behavior of customers who are sensitive to waiting time (and price) and independent providers who are sensitive to earnings which depend on the system utilization.

One research stream in the queueing literature has studied pricing decisions for services where customers can incur waiting or delay costs. In particular, a number of research papers have examined an operating environment that uses a static uniform (non-discriminatory) pricing strategy for heterogeneous customers. Afeche and Mendelson (2004) analyzed the revenue-maximizing and socially optimal equilibria under uniform pricing for heterogeneous customers and found that the classical result that the revenue-maximizing admission price was higher than the socially-optimal price (e.g., see Naor (1969)) could be reversed under a more generalized delay cost structure. Zhou et al. (2014) analyzed the structure of the optimal uniform pricing strategies for two classes of customers with different service valuations and waiting time sensitivities. Armony and Haviv (2003) and Afanasyev and Mendelson (2010) studied the competition between two firms under uniform pricing for two classes of heterogeneous customers. All the above research papers were based on the assumption that capacity was exogenously given. In contrast, our paper considers the situation where service capacity is endogenously dependent on wage and system utilization.

Finally, our model is closely related to some recent work by Taylor (2016). To our knowledge, Taylor (2016) is the first to examine pre-committed price and wage based on customer demand and other operating factors in the context of on-demand services. He compared the optimal prices

when the providers were independent contractors or regular employees, and examined the impact of waiting time sensitivity on the optimal price and wage using a two-point distribution for both the customer valuation of the service and the provider's reservation earning rate. Our model allows these two distributions to be continuous, and complements Taylor's work in two important ways. First, our focus is to examine the impact of demand rate, waiting time sensitivity, service rate, and the size of available providers (who are on-reserve) on the optimal price, wage and payout ratio. Second, in addition to maximizing its profit, we also consider the case when the firm maximizes the sum of its own profit and the total consumer and provider surplus.

3 A Modeling Framework with Price and Time Sensitive Customers and Earning Sensitive Service Providers

We consider an on-demand service platform that coordinates randomly arriving (price and time sensitive) customers with (earning sensitive) independent service providers. To simplify our exposition, we shall use on-demand ride-hailing service platforms (such as Uber) to illustrate our model formulation and results throughout this paper. However, our model can also be used to study other on-demand service applications.

Customers arrive randomly at the platform to request for service, and each service request consists of an (random) amount of service units to be processed by a service provider (e.g., travel distance in km). Throughout this paper, we assume that the requested service by any customer can be met by any of the available service providers. The platform charges each customer a *fixed price rate* p per service unit (e.g., dollar per km), and offers a *fixed wage rate* w per service unit to each participating service provider. Here, we use "wage rate" per service unit so that the payout ratio $\frac{w}{p}$ is well defined. We shall compute "earning rate" per unit time later for providers who decide whether to participate or not.

In the same spirit as in Taylor (2016), the price rate p and wage rate w are pre-committed, but their values can vary across different time periods depending on the specific market characteristics such as the average customer demand rate and the expected number of available providers. In other words, we focus on time-based pricing/wage instead of real-time dynamic pricing/wage that depends on real-time system status such as the number of customers requesting service and number

of available providers in real time.[1]

Each customer decides whether to use the platform to request for service, and each independent provider decides whether to participate. We assume that the price rate p and wage rate w are known to the customers and the providers in advance so that they can make their informed decisions. For each service request, the platform will assign one of the available participating providers to serve the customer.[2] The primary objective of the service platform is to select the optimal price rate and wage rate, denoted by p^* and w^*, so as to maximize its average profit.

3.1 Realized customer request rate λ and price rate p

Consider a certain time period (e.g., peak hours from 8am to 10am). The maximum potential customer demand rate for the service during this time period is given by $\bar{\lambda}$, each of which has a valuation of the service that is based on a value rate v per service unit, where v varies across customers. To model heterogeneous customers, we assume that there is a continuum of customer types so that the value rate v spreads over the range $[0, 1]$ according to a cumulative distribution function $F(.)$, where $F(.)$ is a strictly increasing function with $F(0) = 0$ and $F(1) = 1$.

For a customer with valuation v and a service request of D units, the customer's service surplus is equal to $(v - p)D$.[3] To simplify exposition, we assume that the service units requested D is independent of the customer type v. (If D and v are dependent, we can still apply our analysis by treating the random variable vD as the new "valuation".) As our focus is on the steady state analysis, it suffices to use the average service units requested by customers in our analysis. Let $d = E(D)$ denote the average service units requested by customers. To capture the notion of waiting time sensitivity, we assume that the expected utility function of a customer of value rate type v is given by

$$U(v) = (v - p)d - cW_q, \tag{1}$$

[1] As articulated in MacMillan (2015) and Taylor (2016), many customers resist real time dynamic pricing due to fairness concerns and most on-demand service providers, other than Uber and Lyft, tend to adopt this form of time-based pricing.

[2] Our model does not consider any specific assignment mechanism. For instance, the service platform can assign an available participating provider based on certain specific criteria (e.g, Uber assigns an available driver closest to the pickup location), or can announce a service request to all available participating service providers and assign the request to the first respondent.

[3] By leveraging internet and mobile technologies, customer requests (e.g., pick up and drop off locations) and the service operations (e.g., route) can be monitored or controlled by the on-demand platform. As such, we assume that the number of service units (e.g., travel distance) in each requests is dictated by the customers, and the service providers cannot manipulate or maximize their earnings by deliberately increasing the service units (e.g., travel distance) due to information transparency and real-time location tracking capabilities.

where c denotes the cost of waiting per unit time and W_q represents the expected waiting time for the service. (For instance, Uber and Lyft provide estimated pick-up time to customers.)

Using (1) and assuming that a rational customer with valuation v will request for service only if $U(v) \geq 0$,[4] the platform can use p and w to indirectly control the effective demand (i.e., the realized customer request rate) λ so that

$$\lambda = Prob\{U(v) \geq 0\} \cdot \bar{\lambda} = Prob\{v \geq p + \frac{c}{d} W_q\} \cdot \bar{\lambda}.$$

Define the "target" service level $s = Prob\{v \geq p + \frac{c}{d} W_q\}$. Then, the realized customer request rate λ is given by:

$$\lambda = s\bar{\lambda}. \tag{2}$$

Since $v \sim F(.)$, it follows from (1) that the price rate p satisfies the following equation:

$$p = F^{-1}(1 - \frac{\lambda}{\bar{\lambda}}) - \frac{c}{d} W_q. \tag{3}$$

Note that the price rate p decreases in the expected waiting time W_q and the unit waiting cost c.

3.2 Realized number of participating providers k and wage rate w

Let K be the (maximum) number of potential earning sensitive providers who may decide to participate over the same time period, i.e., K represents the number of registered providers who are eligible to participate. For any given (p, w), let k be the realized number of providers participating in the platform, where $k \leq K$. Also, let μ denote the average service speed (number of service units processed per unit time; e.g., travel speed measured in terms of km per hour) of the service providers so that μ/d represents the service rate of the providers (i.e., average number of customers served per hour).[5] Given the realized customer request rate λ and the realized number of participating providers k, the utilization of these k participating providers is equal to $\frac{\lambda}{k \cdot (\mu/d)}$, where $\lambda d < k\mu$ to ensure system stability. The average wage per unit time of a participating provider (when working) is equal to the wage per service unit w multiplied by the average service speed μ. Accounting for the utilization, the average "earning rate" per unit time of a participating provider is equal to $w\mu \cdot \frac{\lambda d}{k\mu} = w\frac{\lambda d}{k}$.[6]

[4] In other words, in equilibrium, only customers with value rate $v \geq p + \frac{c}{d} W_q$ will use the platform to request for service, and customer requests with value rate $v < p + \frac{c}{d} W_q$ will not use the platform to meet their service need.

[5] If the service units d are already measured in terms of time units, we can simply set $\mu = 1$ in this case.

[6] For independent service providers, utilization and wage rate are the two key factors for their participation. For example, DePillis (2016) reported that Uber drivers obtain higher earnings primarily because their utilization rate

To model the notion of earning-sensitivity, we assume that each potential provider has a reservation earning rate r per unit time (i.e., corresponding to his outside option), where r varies across different providers. To model the heterogeneity among providers, we assume that there is a continuum of provider types so that the reservation rate r spreads over the range $[0, 1]$ according to a cumulative distribution function $G(.)$, where $G(.)$ is a strictly increasing function with $G(0) = 0$ and $G(1) = 1$. For a (potential) provider with reservation rate r, he will participate to offer service only if his average earning rate $w\frac{\lambda d}{k}$ is at least equal to r.

Let β denote the proportion of providers who participate in the platform to offer service during this time period. Then, $\beta = Prob\{r \leq w\frac{\lambda d}{k}\} = G(w\frac{\lambda d}{k})$, and the realized number of participating providers k (i.e., supply) is given by

$$k = \beta K. \tag{4}$$

Also, in equilibrium, $\beta = G(w\frac{\lambda d}{k})$ so that:

$$G^{-1}(\beta) = w\frac{\lambda d}{k}. \tag{5}$$

From (4) and (5), we can express the wage rate w as a function of the number of participating providers k:

$$w = G^{-1}(\beta)\frac{k}{\lambda d} = G^{-1}(\frac{k}{K})\frac{k}{\lambda d}. \tag{6}$$

3.3 Problem Formulation

Since the platform earns an average profit of $(p - w)d$ for each customer request, the platform's average total profit is then equal to $\pi = \lambda(p - w)d$. By substituting (3) and (6) into the profit function, we can express the profit function π as a function of (k, λ) below:

$$\pi(k, \lambda) = \lambda d\left[F^{-1}(1 - \frac{\lambda}{\bar{\lambda}}) - \frac{c}{d}W_q - G^{-1}(\frac{k}{K})\frac{k}{\lambda d}\right]. \tag{7}$$

Considering the system stability condition $\lambda d < k\mu$, the optimization problem of the platform can be formulated as

$$\max_{k,\lambda}\ \pi(k, \lambda) \equiv \lambda d\left[F^{-1}(1 - \frac{\lambda}{\bar{\lambda}}) - \frac{c}{d}W_q - G^{-1}(\frac{k}{K})\frac{k}{\lambda d}\right],\ \text{subject to}\ \frac{\lambda d}{k\mu} < 1, \tag{8}$$

(measured in terms of percentage of miles driven with a passenger) is much higher than that for taxi drivers. For instance, Uber driver's utilization is 64.2%, while taxi driver's utilization is only 40.7% in Los Angeles.

from which we can determine the optimal supply (i.e., the number of participating providers k^*) and the optimal demand (i.e., the realized customer request rate λ^*). Then, we can use (3) and (6) to retrieve the corresponding optimal price rate p^* and optimal wage rate w^* from k^* and λ^*.

3.4 Notation

For ease of reference, we list below the basic notation used in the paper.

- K : Maximum number of potential service providers who may opt to participate;
- k : Realized number of participating service providers ($k \leq K$);
- $\bar{\lambda}$: Customer demand rate who may opt to use the platform to request for service;
- λ : Realized customer request rate ($\lambda \leq \bar{\lambda}$);
- s : Target service level;
- D : Random amount of service units per service request;
- d : Average amount of service units per service request, i.e., $d = E(D)$;
- μ : Average service speed of the service providers;
- v : Value rate per service unit of a customer;
- $F(.)$: Cumulative distribution of value rate of customers v;
- r : Reservation earning rate of service providers;
- $G(.)$: Cumulative distribution of reservation rate of service providers r;
- c : Unit waiting cost of customers;
- p : Price rate (price per service unit) charged to customers;
- w : Wage rate (wage per service unit) paid to service providers.

4 The Base Model with A Fixed Payout Ratio

A common practice for many on-demand service platforms is to set the wage rate as a fixed proportion of the price rate, i.e., $w = \alpha p$ for some fixed α, $0 < \alpha < 1$. For example, Uber set $\alpha = 0.8$ for its first cohort of drivers in San Francisco (Huet (2014)). We can use our modeling framework to analyze this common practice by imposing an additional constraint of $w = \alpha p$ in the optimization problem as given in (8). We refer to this model as the base model with a fixed payout ratio, or simply the "base model", in our subsequent discussions.

We model the expected waiting time W_q used in the customer's utility function (1) based on an $M/M/k$ queue. For an $M/M/k$ queue with arrival rate λ and service rate $\frac{\mu}{d}$, it is well-known (see e.g., Gross et al. (2008)) that the expected waiting time is given by

$$W_q = \frac{1}{1 + \frac{k!(1-\rho)}{k^k \rho^k} \sum_{i=0}^{k-1} \frac{k^i \rho^i}{i!}} \left[\frac{\rho}{\lambda(1-\rho)} \right], \tag{9}$$

where $\rho = \lambda d / k\mu$ represents the system utilization with $\rho < 1$.

To simplify our analysis here, we shall assume that the distributions of value rate v and reservation earning rate r are uniformly distributed over the range [0,1] so that $F(v) = v$ and $G(r) = r$ in our models for the rest of this paper. However, all our analytical and numerical results can be directly extended to the more general case where the positive support of the uniform distribution of $F(.)$ or $G(.)$ is within the range of $[a, b]$ rather than $[0, 1]$, as used in our illustrative numerical examples in Section 6.

With the above assumptions, the respective price, wage and profit functions given in (3), (6) and (7) can be expressed as follows:

$$p = \left(1 - \frac{\lambda}{\overline{\lambda}}\right) - \frac{c}{1 + \frac{k!(1-\rho)}{k^k \rho^k} \sum_{i=0}^{k-1} \frac{k^i \rho^i}{i!}} \left[\frac{\rho}{\lambda d(1-\rho)} \right] \tag{10}$$

$$w = \frac{k^2}{K\lambda d} \tag{11}$$

$$\pi(k, \lambda) = \lambda d(p - w) = \lambda d \left\{ (1 - \frac{\lambda}{\overline{\lambda}}) - \frac{c}{1 + \frac{k!(1-\rho)}{k^k \rho^k} \sum_{i=0}^{k-1} \frac{k^i \rho^i}{i!}} \left[\frac{\rho}{\lambda d(1-\rho)} \right] - \frac{k^2}{K\lambda d} \right\}, \tag{12}$$

where the system utilization $\rho = \frac{\lambda d}{k\mu} < 1$. Using (10) and (11), the fixed payout ratio constraint, $w = \alpha p$, can be written as

$$\frac{k^2}{K\lambda d} = \alpha \left\{ \left(1 - \frac{\lambda}{\overline{\lambda}}\right) - \frac{c}{1 + \frac{k!(1-\rho)}{k^k \rho^k} \sum_{i=0}^{k-1} \frac{k^i \rho^i}{i!}} \left[\frac{\rho}{\lambda d(1-\rho)} \right] \right\},$$

or equivalently,

$$k^2 = K\alpha \left\{ \lambda d \left(1 - \frac{\lambda}{\overline{\lambda}}\right) - \frac{c}{1 + \frac{k!(1-\rho)}{k^k \rho^k} \sum_{i=0}^{k-1} \frac{k^i \rho^i}{i!}} \left(\frac{\rho}{1-\rho} \right) \right\}. \tag{13}$$

Also, as $w = \alpha p$, we can use (11) to rewrite the profit function (12) as

$$\pi(k, \lambda) = \lambda d(p - w) = \lambda d(\frac{w}{\alpha} - w) = \frac{k^2(1-\alpha)}{K\alpha}. \tag{14}$$

Then, the optimization problem is to maximize the profit function (14) subject to the constraints (13) and $k \leq K$. It is easy to see that the optimal k^* is given by the largest value of k, with $k \leq K$, that possesses a feasible λ to (13).

While it is difficult to derive tractable results using (13), it is straightforward to numerically search for any feasible λ satisfying (13) for each fixed value of k. For each fixed value of k, with $k = 1, 2, ..., K$, we search through all possible values of λ, with $\lambda d/k\mu < 1$, that would satisfy (13). The optimal solution k^* is given by the largest value of k with a feasible λ to (13), and the optimal p^* and w^* are given by (10) and (11) accordingly.

The left panel of Table 1 provides a sample set of results in our numerical experiments. For this set of numerical experiments, we set $\alpha = 0.5$, $c = 1$, $K = 50$, $\mu = 1$, and $d = 1$ with values of $\bar{\lambda}$ ranging from 10 to 100. Table 1 shows that the optimal value of k^* (and the optimal profit π^*) is non-decreasing in $\bar{\lambda}$, i.e., the optimal number of participating providers and the optimal expected profit of the platform would increase (or remain the same) as the customer demand rate who may opt to use the service increases. However, the optimal values of λ^* and p^* are not necessarily monotone in $\bar{\lambda}$.[7] In particular, the optimal price could possibly decrease when the customer demand increases. We shall provide an explanation of why this seemingly counter-intuitive result could occur in our numerical results later.

Insert Table 1 about here

4.1 An Approximation Scheme

To obtain some analytical results that can enable us to understand why λ^* and p^* are not necessarily monotonic in $\bar{\lambda}$, we next develop an approximation scheme by using a simpler function for the expected waiting time function W_q. The approximation scheme serves two purposes. It gives a more efficient way of finding a near-optimal solution numerically and provides analytical results for supporting the insights obtained from our numerical experiments.

Our approximation scheme is motivated by the following well-studied approximation for the

[7] With the integer constraint on k, there generally exists two feasible values of λ^* corresponding to the optimal integer solution k^*. For consistent comparisons, we always present the smaller value of λ^*. The larger value of λ^* also shows similar non-monotonic property as well.

expected waiting time of an $M/M/k$ queue with arrival rate λ and service rate $\frac{\mu}{d}$:

$$W_q = \frac{\rho\sqrt{2(k+1)}}{\lambda(1-\rho)}, \tag{15}$$

where $\rho = \frac{\lambda d}{k\mu}$ represents the system utilization. The approximation formula (15) is exact for an $M/M/1$ queue, i.e., (9) reduces to (15) when $k = 1$, and it has been shown (see Sakasegawa (1977)) to provide a very good estimate of (9) when $k > 1$.

However, using (15) for W_q is still too complex for developing tractable results for the base model. The decision variable k appears in both $\rho = \frac{\lambda d}{k\mu}$ and the exponent of the expression given in (15), which makes the first-order conditions of the optimization problem difficult to analyze. Therefore, we use a simpler approximation for W_q by assuming that:

$$W_q = \frac{\rho\sqrt{2(n+1)}}{\lambda(1-\rho)}, \tag{16}$$

where $\rho = \frac{\lambda d}{k\mu} < 1$ and n is some fixed positive number. By using (16) for W_q, the price and profit functions given in (10) and (12) now become:

$$p = \left(1 - \frac{\lambda}{\bar{\lambda}}\right) - \frac{c\rho\sqrt{2(n+1)}}{\lambda d(1-\rho)} \tag{17}$$

$$\pi(k,\lambda) = \lambda d\left[(1 - \frac{\lambda}{\bar{\lambda}}) - \frac{c\rho\sqrt{2(n+1)}}{\lambda d(1-\rho)} - \frac{k^2}{K\lambda d}\right], \tag{18}$$

and the fixed payout ratio constraint given by (13) becomes

$$k^2 = K\alpha\left\{\lambda d\left(1 - \frac{\lambda}{\bar{\lambda}}\right) - \frac{c\rho\sqrt{2(n+1)}}{1-\rho}\right\}. \tag{19}$$

Then, for each fixed value of k, we now use (19) instead of (13) to find a feasible λ numerically, and the optimal solution k^* is given by the largest value of k, with $k \leq K$, that possesses a feasible λ to (19).

The only difference between (15) and (16) is that the exponent in (15) is based on the decision variable k, whereas the exponent in (16) is based on a fixed parameter n. Thus, (16) would be very close to (15) when n is close to k. We next provide an iterative procedure to determine the parameter n in (16) such that the resulting optimal value of k^* is equal to n itself. By setting $n = k^*$, (16) can be approximated by (15) for $k \approx k^*$, and so the approximation (16) would be close to the exact formula (9) when $k \approx k^*$.

An iterative procedure for determining n:

1. Initialize $n = 0$.

2. Solve for the optimal $k^*(n)$, and set $n = k^*(n)$.

3. Repeat Step 2 until the values of n and $k^*(n)$ converge.

Note that we allow both n and $k^*(n)$ to be positive numbers rather than positive integers in the iterative procedure. The next proposition shows that the above iterative procedure will always converge to a (unique) fixed point, i.e., $n^* = k^*(n^*)$.

Proposition 1 *There exists a unique fixed point $n^* = k^*(n^*)$, and the iterative procedure always converges to n^*.*

We performed a comprehensive set of numerical experiments to examine the performance of our approximation scheme using (16) for W_q with $n = n^*$. For each numerical example, we find the fixed point n^* using the iterative procedure. As the decision variable k is required to be a positive integer under the exact formula (9), we round n^* down to the integer below and set $k^* = n = \lfloor n^* \rfloor$, for comparison purposes. (Rounding up n^* would give an infeasible solution in our approximation scheme.)

In our numerical experiments, we set $\alpha \in [0.4, 0.5, ..., 0.9]$, $c \in [0.5, 0.75, 1]$, $K \in [50, 60, ..., 150]$, $\mu \in [1, 3, 5]$, $\bar{\lambda} \in [10, 20, ..., 100]$ and $d = 1$ for a total of 5940 cases. The right panel of Table 1 provides the corresponding results for the same set of numerical experiments using our approximation scheme. Note that the optimal solutions given by the approximation scheme exhibit similar patterns as those using the exact formula; e.g., Table 1 shows that both k^* and π^* are non-decreasing in $\bar{\lambda}$, whereas p^* and λ^* are not necessarily monotone in $\bar{\lambda}$.

Table 2 summarizes the performance of our approximation scheme, as compared with the results using the exact formula (9). The numbers in Table 2 represent the mean absolute percent difference for the optimal values between using the approximation scheme and the exact formula. For comparison purposes with the exact formula, n^* is rounded down to an integer due to the integral constraint on k. Overall, our numerical experiments suggest that the approximation scheme provides very good approximation results.

Insert Table 2 about here

We next derive some analytical results for the base model using the approximation formula (16) for W_q and allowing the decision variable k to take on positive numbers rather than positive integers only. We can establish the following analytical results under the formal assumptions as stated below:

Assumption 1: *$F(.) \sim U[0,1]$, $G(.) \sim U[0,1]$, and W_q is given by (16) where n is a fixed positive number. Also, the decision variable k is not restricted to positive integers only.*

Proposition 2 *Suppose that Assumption 1 holds and $\frac{w}{p} = \alpha$, $0 < \alpha < 1$. Then,*
(i) p^* (and the corresponding $w^* = \alpha p^*$), k^*, λ^* and ρ^* increase in $\bar{\lambda}$; and
(ii) p^* (and the corresponding $w^* = \alpha p^*$), k^* and ρ^* increase in d, and λ^* decreases in d.

We note that the monotonicity results given in Proposition 2 are established for any fixed positive number n. In our approximation scheme, n is chosen such that $n = k^*(n)$ using the iterative procedure, which changes as the values of the model parameters change. However, the monotonicity properties stated in Proposition 2 remain valid for all our numerical results using the approximation scheme. For instance, our numerical results using the approximation scheme (when k can take on any positive number) have confirmed that both the optimal price p^* and realized customer demand rate λ^* increase in $\bar{\lambda}$, as given in Proposition 2(i).

With the integer constraint on k, the results in Table 1 show that p^* and λ^* are not necessarily monotone in $\bar{\lambda}$. We can now explain this non-monotonic behavior of p^* and λ^* observed in Table 1 as follows. When k is restricted to be (positive) integers, it is not possible to increase k^* by any amount less than one. Consequently, with a small increase in $\bar{\lambda}$, k^* might stay the same, and the optimal p^* and λ^* would then need to be reduced. Without the integer constraint on k, this behavior will no longer occur. Any increase in $\bar{\lambda}$ will cause k^* to increase, and the resulting p^* and λ^* will always increase, as shown in Proposition 2(i).

4.2 Main Insights

We performed an extensive set of numerical experiments using our approximation scheme (with k being a continuous variable). Based on these numerical results, together with analytical support of Proposition 2, we summarize below the main insights for the base model.

First, the optimal price rate p^* increases when the customer demand rate $\bar{\lambda}$ is higher (or when the average service unit d is higher). Note that the profit of the platform is equal to the product of the realized customer request rate λ and the profit margin $p(1-\alpha)$. When $\bar{\lambda}$ increases, the platform can increase its price p^* while sustaining a higher demand request rate λ^*, resulting in a higher profit. A higher price rate p^* also corresponds to a higher wage rate (as $w^* = \alpha p^*$), which attracts more participating providers k^* to handle the higher demand request rate λ^*. Thus, our results suggest that an on-demand ride-hailing service platform using a fixed payout ratio should charge a higher price to increase profitability during rush hours when the customer demand is high.

Second, while a higher customer demand rate $\bar{\lambda}$ (or a higher d) would increase the optimal price rate p^* and wage rate w^*, the optimal price and wage rates are not necessarily monotone as service capacity increases (with a higher K or μ). We can explain this contrast as follows. When the number of available providers K (or service rate μ) increases, the platform can decrease its wage rate w^* while still attracting more participating providers k^*. Also, the corresponding decrease in price rate p^* would increase the realized demand request rate λ^* as its capacity increases. However, this does not necessarily increase the profit as the profit margin $p^*(1-\alpha)$ would reduce. Overall, the optimal price p^* is not monotonic in K, but depends on the relative changes in demand request rate λ^* and profit margin $p^*(1-\alpha)$.

Similarly, the optimal price and wage rates are not necessarily monotonic in the unit waiting cost c. As c increases, a direct effect is a decrease in demand request rate, and the platform needs to adjust its price rate (and the corresponding wage rate) to reduce the adverse effect of a higher waiting cost. If the platform increases its wage rate w to attract more participating providers to reduce waiting time, the corresponding price increase p^* would further reduce demand request rate λ and possibly lead to a lower profit. On the other hand, if the platform reduces its price rate p to stimulate demand request rate λ, the corresponding reduction in wage rate w would reduce supply capacity k and profit margin $p^*(1-\alpha)$. Therefore, the impact of c on the optimal price and wage rates are not necessarily monotonic, but depends on specific values of the model parameters.

5 The General Model with A Time-based Payout Ratio

Our base model is based on the situation where the platform uses a fixed payout ratio for its service providers. While a fixed ratio payout scheme is easy to implement and widely adopted in practice,

it raises an interesting question of whether a time-based payout ratio that depends on specific time-based market characteristics could significantly improve the profitability of an on-demand service platform. To answer this question, we now analyze the general situation where the optimal price and wage rates are determined without imposing the constraint of $w = \alpha p$ in solving the decision problem of the platform. We refer to this model as the general model with a time-based payout ratio, or simply the "general model", in our subsequent discussions.

For the general model, the decision problem is to find the optimal values (k, λ) that maximize the profit function $\pi(k, \lambda)$ given in (12) subject to the utilization constraint $\rho = \frac{\lambda d}{k\mu} < 1$. As for the base model, the profit function (12) is too complex for conducting tractable analysis, but we can solve the problem numerically. Specifically, we can perform an exhaustive numerical search for the optimal λ that maximizes (12) for each fixed value of k, $k = 1, 2, ..., K$, and we then compare the optimal profit for each value of k to select the optimal k^* and the corresponding optimal λ^*.

The left panel of Table 3 provides the results for the same set of numerical experiments as given in Table 1. Table 3 shows that k^* and π^* for the general model are also non-decreasing in $\bar{\lambda}$, i.e., both the optimal number of participating providers and the optimal expected profit of the platform increase (or remain the same) as the customer demand rate increases. Furthermore, p^* and w^* generally (but not always) increase in $\bar{\lambda}$, i.e., the platform would most likely increase price and wage rates when the customer demand rate increases.

Insert Table 3 about here

We next illustrate in Table 4 how the optimal expected profit would be affected if the platform uses a fixed payout ratio instead of the optimal time-based payout ratio. By using the parameters associated with the numerical experiments discussed above, we conduct the following analysis. For a given $\bar{\lambda}$, we compute the ratio (in percentage) between the expected profit under a fixed payout ratio α (value is given in the first row) and the expected profit under the optimal time-based ratio α^* (value is given in the second column). The results in Table 4 show that the expected profit can be greatly reduced if α is substantially different from α^*. For example, when $\bar{\lambda} = 10$, the platform can only obtain 31% of the expected profit under the optimal time-based payout ratio $\alpha^* = .35$ if a fixed payout ratio $\alpha = .8$ is used.

Insert Table 4 about here

Table 5 provides the values of the optimal time-based payout ratio α^* for the above set of numerical experiments with $\bar{\lambda}$ ranging from 10 to 100 and K ranging from 10 to 100. Observe that the optimal dynamic payout ratio α^* can vary widely depending on the specific values of the model parameters. Thus, the combined results in Tables 4 and 5 suggest that, when the operating characteristics (such as $\bar{\lambda}$ or K) can change significantly at different time periods, it is not possible to choose one single fixed payout ratio that would be close to the optimal payout ratios across all time periods. Consequently, the platform using a fixed payout ratio scheme can achieve near-optimal results for only certain time periods. Instead, the platform needs to adopt a time-based payout ratio scheme to maximize profitability across different time periods.

Insert Table 5 about here

5.1 An Approximation Scheme

We can use (16) for W_q to develop a similar approximation scheme for finding near-optimal solutions for the general model in which the price and profit functions are given by (17) and (18), respectively. In this case, the optimal (k^*, λ^*) can be obtained from the following two first-order conditions:

$$\frac{\partial \pi}{\partial k} = c\mu \frac{\rho^{\sqrt{2(n+1)}}}{k\mu(1-\rho)} \left(\sqrt{2(n+1)} + \frac{\rho}{1-\rho} \right) - \frac{2k}{K} = 0 \tag{20}$$

$$\frac{\partial \pi}{\partial \lambda} = d \left\{ \left(1 - 2\frac{\lambda}{\bar{\lambda}} \right) - c \frac{\rho^{\sqrt{2(n+1)}-1}}{k\mu(1-\rho)} \left(\sqrt{2(n+1)} + \frac{\rho}{1-\rho} \right) \right\} = 0. \tag{21}$$

For each fixed value of n, we can use the two first-order conditions (20) and (21) to find the optimal values of (k, λ), as denoted by $(k^*(n), \lambda^*(n))$. We can also use an iterative procedure to select the parameter n given in (16) such that the resulting optimal value of $k^*(n)$ is equal to n itself.

We can establish the following result under the approximation scheme for the general model:

Proposition 3 *There exists a unique fixed point $n^* = k^*(n^*)$ for the general model.*

We can use a simple bisection search to find the unique fixed point n^* as follows:

1. Initialize $l = 0$ and $u = K$.
2. Set $n = \frac{l+u}{2}$ and solve for the optimal $k^*(n)$. If $k^*(n) > n$, set $l = n$; otherwise set $u = n$.
3. Repeat Step 2 until the values of n and $k^*(n)$ converge.

We also performed numerical experiments to evaluate the performance of the approximation scheme by comparing the optimal solutions with those using the exact formula (9) for W_q. For consistent comparisons, we also restricted k to positive integers under the approximation scheme by rounding up the value of n^* obtained from the iterative procedure, i.e., $n = \lceil n^* \rceil$. We then set the optimal $k^* = n$ and use the first-order condition (21) with $k = k^*$ to find the corresponding optimal λ^*.

In our numerical experiments, we set $c \in [0.5, 0.6, ..., 1]$, $K \in [50, 60, ..., 150]$, $\mu \in [1, 2, ..., 10]$, $\bar{\lambda} \in [10, 20, ..., 100]$ and $d = 1$ for a total of 6600 cases. The right panel of Table 3 provides the corresponding results for the same set of numerical experiments discussed earlier for the base model. Observe that while k^* under the exact formula (9) and approximation (16) are slightly different in some cases (e.g., $\bar{\lambda} = 40$), λ^* is adjusted accordingly to achieve near-optimal profit. Also, p^* and w^* are mostly (though still not always) increasing in $\bar{\lambda}$.

Table 6 summarizes the performance of our approximation scheme for the general model, as compared with the results using the exact formula (9). The numbers in Table 6 represent the mean absolute percent difference for the optimal values between using the approximation scheme and the exact formula. Overall, our numerical experiments suggest that the approximation scheme provides very good approximation results, even when k^* is relatively small (i.e., $k^* > 10$). Thus, our numerical results show that our approximation scheme could provide near-optimal results for most service platforms in practice.

Insert Table 6 about here

We can also establish the following monotonicity results for the general model using approximation (16) for W_q.

Proposition 4 *Under Assumption 1, the optimal solution for the general model exhibits the following characteristics:*

(i) When K or μ increases, w^* decreases, π^* increases, but p^* is not necessarily monotonic.

(ii) When c increases, w^* increases, π^* decreases, but p^* is not necessarily monotonic.

(iii) When $\bar{\lambda}$ or d increases, w^*, p^* and π^* increase.

(iv) The optimal payout ratio $\alpha^* = \frac{w^*}{p^*}$ decreases in K and μ, and increases in c, $\bar{\lambda}$ and d.

It is important to observe from Proposition 4 that, even though the optimal price rate is not necessarily monotonic in K, μ and c, the optimal time-based payout ratio α^* is monotone in all model parameters. In particular, the optimal time-based payout ratio α^* decreases when the service capacity increases (with a higher K or μ), but increases when the waiting cost c is higher or when customer demand increases (with a higher $\bar{\lambda}$ or d).

As shown in the proof of Proposition 4, we also obtain monotonicity properties for other system performance measures as summarized in Table 7. The monotonicity properties given in Proposition 4 and Table 7 are established for a fixed value of n, whereas the value of n used in (16) in our approximation scheme changes as the values of the model parameters change. We also performed numerical experiments to validate these properties for the optimal solutions using our approximation scheme. Results from all our numerical experiments are consistent with the analytical results given in Proposition 4. For example, Table 8 provides the optimal values of α^* for the numerical experiments discussed earlier, which is consistent with Proposition 4(iv) that α^* generally decreases in K and increases in $\bar{\lambda}$. Consequently, Proposition 4(iv) provides analytical support that the optimal time-based payout ratio α^* generally decreases in K and increases in $\bar{\lambda}$, as observed in Table 5 using the exact formula (9) and in Table 8 using the approximation scheme.

Insert Tables 7 and 8 about here

Proposition 4 shows the impact on the optimal wage, time-based payout ratio and the profit of the platform when either the number of providers K or the customer demand rate $\bar{\lambda}$ increases. As a platform grows, it is common that both K and $\bar{\lambda}$ would increase at the same time. Therefore, it would be useful to understand how these optimal results would change as both K and $\bar{\lambda}$ increase. It is clear from Proposition 4 that the optimal profit of the platform would increase when both K and $\bar{\lambda}$ increase. However, it is unclear as how the platform would adjust its wage and price rates as well as its payout ratio as the platform grows, as both the optimal wage rate w^* and the optimal payout ratio α^* would change in opposite directions with respect to the changes in K and $\bar{\lambda}$.

It is intuitive that the changes in the optimal wage rate, price rate and payout ratio would generally depend on the relative growth rates of the number of providers K and customer demand rate $\bar{\lambda}$. However, we can derive the following results for the special case when K and $\bar{\lambda}$ increase at the same rate. Specifically, suppose that the initial number of providers and customer demand

rate are given by $\hat{K}$ and $\hat{\lambda}$, respectively. Let $\epsilon > 1$ represent the (same) growth rate of number of providers and customer demand rate, i.e., $K = \epsilon\hat{K}$ and $\bar{\lambda} = \epsilon\hat{\lambda}$. The following proposition shows the effect of ϵ on the optimal wage and price rates, and the optimal payout ratio.

Proposition 5 *Under Assumption 1, both w^* and $\alpha^* = \frac{w^*}{p^*}$ decrease as ϵ increases. However, p^* is not necessarily monotonic in ϵ.*

Proposition 5 shows that, under Assumption 1, a platform should lower its wage rate and payout ratio as both the number of providers and customer demand rate grow at the same rate. We further performed some numerical experiments to confirm these monotonicity results using our approximation scheme $n = k^*(n)$. Table 9 provides some sample results of our numerical experiments that also illustrate the monotonicity results. For this set of numerical examples, we set $\hat{K} = \hat{\lambda} = 10$, $c = 1$, $\mu = 1$ and $d = 1$, with ϵ increasing from 1 to 5. Observe that both w^* and α^* decreases as ϵ increases, as supported by the analytical results of Proposition 5.

Insert Table 9 about here

5.2 Main Insights

Based on our numerical experiments, together with analytical support from Propositions 4 and 5 and Table 7, we summarize the main insights for the general model below.

First, the platform should reduce the wage rate w^* as the number of available providers K (or average service speed μ) increases. In addition, the optimal profit π^* increases as K or μ increases, which implies that it is beneficial for the platform to recruit more providers to join the platform and to help providers increase their average service speed. However, the optimal price p^* is not necessarily monotonic in K.[8] Our numerical results suggest that the optimal price could first increase and then decrease in K, and we can explain this behavior using the well-known "queueing effect" that the expected waiting time increases convexly in the system utilization as follows.

When K is small (relative to the customer demand rate $\bar{\lambda}$), the constraint is on the supply side, and the platform needs to operate in high utilization. In this case, an increase in supply capacity from a higher K can significantly reduce the waiting time W_q (due to the non-linear queueing

[8]For a numerical example using the exact formula (9), set $c = 5$, $\mu = 1$, $\bar{\lambda} = 500$ and $d = 1$. The optimal price p^* increases as K increases from 50 to 70, but then decreases as K increases further from 70 to 150.

effect), so the platform can afford to increase the optimal price p^* to maintain a higher realized customer request rate λ^* and achieve a higher profit π^*. This explains why the optimal price p^* could initially increase in K when K is small. On the other hand, when K is large, the constraint is now on the demand side, and the system can operate in lower utilization. In this case, an increase in K would only reduce the waiting time W_q slightly (due to the non-linear queueing effect), and the platform now chooses to reduce the optimal price p^* in order to stimulate a higher customer request rate λ^* and achieve a higher profit π^*. This explains why the optimal price p^* would decrease in K when K is large. Overall, we show that the queueing effect has caused the optimal price p^* to be non-monotonic in K.

Our results show that the optimal price and the optimal wage may move in the same direction or opposite direction when the maximum number of service providers increases. This non-monotonic property of the optimal price in our model is apparently due to the fact that our model captures the nonlinear effect of utilization on waiting time. When the queueing effect on customer demand is not captured in our model (i.e., $c = 0$), it is straightforward to show that both p^* and w^* decrease in K, which provides a further justification that the non-monotonic property in the optimal price rate is due to the queueing effect captured in the customer utility function (1) of our model.[9]

Second, we find that the platform should offer a higher wage rate w^* as the waiting cost c increases. This helps to attract more providers k^* to participate, but will reduce the optimal profit of the platform π^*. However, the optimal price p^* is not necessarily monotonic in c.[10] Our numerical results suggest that the optimal price p^* could first increase in c when c is small, but then decrease in c as c increases. This non-monotonic behavior can be again explained by how the queueing effect captured in our model.

When c is small (relative to the price p), the platform can operate in high utilization (with few providers) since customers are less sensitive to waiting time than price. In this case, an increase in c would reduce demand and decrease waiting time significantly at high utilization (due to the non-linear queueing effect). Consequently, the platform can take advantage of the significant waiting time reduction by increasing the optimal price p^* to maximize its profit. On the other hand, when

[9]When the profit function is not a multiplicative form of demand and supply, Hu and Zhou (2017) show that the optimal price has a U-shape relationship with the exogenous wage when the profit function of the platform is the minimum of demand and supply.

[10]For a numerical example using the exact formula (9), set $K = 50$, $\mu = 1$, $\bar{\lambda} = 50$, $d = 1$. The optimal price p^* first increases as c increases from 10 to 80, but then decreases as c increases further from 80 to 100.

c is large, customers are now more sensitive to waiting time than price, and the platform now needs to operate at lower utilization. In this case, an increase in c would reduce demand, but would provide only marginal waiting time reduction (due to the non-linear queueing effect). As a result, the platform would now choose to reduce the optimal price p^* in order to stimulate the customer request rate λ^* to maximize its profit. This explains why the optimal price p^* would decrease in c when c is large. Overall, we explain that the non-linear queueing effect has caused the optimal price p^* to be non-monotonic in c

Third, the platform should increase its price rate p^* as customer demand rate $\bar{\lambda}$ (or average service units d) increases. At the same time, the platform should also increase its wage rate w^*in order to attract more participating providers k^* to handle the higher customer request rate λ^*. Overall, the profit of the platform π^* increases as $\bar{\lambda}$ (or d) increases.

Finally, the platform should reduce its payout ratio α^* when the service capacity (i.e., a higher K or μ) increases. This implies that the platform can lower its payout ratio as it attracts more providers to the platform. Also, the platform should increase the payout ratio when the customer waiting cost c is higher or when customer demand increases (i.e., a higher $\bar{\lambda}$ or d). One interesting implication of this result is that an on-demand ride-hailing service platform should increase the payout ratio to its participating drivers during rush hours when the customer demand rate $\bar{\lambda}$ is higher and/or the travel speed μ is lower. More interestingly, the platform should also reduce its payout ratio as it expands with the number of providers and customer demand growing at about the same rate. This result could provide an economic justification for Uber's strategy as reported by Huet (2014) of offering a payout ratio of 0.8 for its first cohorts of drivers in San Francisco initially, but lowering its payout ratio to 0.75 for its second cohorts of drivers in 2014, as both the number of registered drivers and customer demand rate had increased.

5.3 Extension to include consumer and provider surplus

Besides profit, the platform may have an interest in managing the welfare of its customers and providers carefully, especially when the practices of some on-demand service platforms could be potentially controversial. For example, Uber has been challenged by consumer rights group due to concerns about public safety including sexual assaults, physical attacks, by independent drivers due to their concerns about being treated as regular employees without benefits, by the government due

to concerns over regulations, and by other taxi drivers due to their concerns over unfair competition; see Rogers (2015) for a comprehensive list of social costs of Uber including public safety, privacy, discrimination, and labor law violations. Our general model can be extended to incorporate the welfare of the consumers and providers to help address these concerns.

In the same spirit as Cachon et al. (2015), we can extend the objective of our general model to maximize the firm's profit plus the total consumer and provider surplus. For a customer who requests for service with a value rate of $v \geq F^{-1}(1-\frac{\lambda}{\bar{\lambda}})$, her surplus is given by $\{(v-p)d - cW_q\}$. Therefore, the total customer surplus is equal to

$$C_s = \bar{\lambda} \int_{F^{-1}(1-\frac{\lambda}{\bar{\lambda}})}^{1} \left[(v-p)d - cW_q\right] dF(v) = \bar{\lambda} \left[\left(\int_{F^{-1}(1-\frac{\lambda}{\bar{\lambda}})}^{1} v \; dF(v) - \frac{\lambda}{\bar{\lambda}} p \right) d - \frac{\lambda}{\bar{\lambda}} cW_q \right]. \tag{22}$$

For a participating provider with a wage reservation rate of $r \leq w\frac{\lambda d}{k}$, his surplus is given by $w\frac{\lambda d}{k} - r$. Therefore, the total provider surplus is equal to

$$P_s = K \int_0^{G^{-1}(\frac{k}{K})} \left(w\frac{\lambda d}{k} - r \right) dG(r) = w\lambda d - G^{-1}(\frac{k}{K})k + K \int_0^{G^{-1}(\frac{k}{K})} G(r)dr. \tag{23}$$

Thus, the objective function of the platform for the general model can be expressed as

$$\begin{aligned}
\Pi(k,\lambda) =& (1-\gamma)\pi(k,\lambda) + \gamma(C_s + P_s) \\
=& (1-\gamma)\pi(k,\lambda) + \gamma \left\{ \bar{\lambda} \left[\left(\int_{F^{-1}(1-\frac{\lambda}{\bar{\lambda}})}^{1} v \; dF(v) - \frac{\lambda}{\bar{\lambda}} p \right) d - \frac{\lambda}{\bar{\lambda}} cW_q \right] + w\lambda d - G^{-1}(\frac{k}{K})k \right. \\
& \left. + K \int_0^{G^{-1}(\frac{k}{K})} G(r)dr \right\} \\
=& (1-\gamma)\pi(k,\lambda) + \gamma \left\{ \bar{\lambda} d \left[\int_{F^{-1}(1-\frac{\lambda}{\bar{\lambda}})}^{1} v \; dF(v) - \frac{\lambda}{\bar{\lambda}} F^{-1}(1-\frac{\lambda}{\bar{\lambda}}) \right] + K \int_0^{G^{-1}(\frac{k}{K})} G(r)dr \right\},
\end{aligned} \tag{24}$$

where $\gamma \in [0,1]$ is the "equitable payoff" parameter which represents the willingness of the platform to give up some of its profit for a more equitable (or fairer) outcome for its customers and providers in setting price and wage rates; see Cui et al. (2007). For example, when $\gamma = \frac{1}{2}$, the platform assigns equal weights on its profit and the total consumer and provider surplus. When $\gamma = 0$, the platform completely ignores the consumer and provider surplus, and $\Pi(k,\lambda)$ simply reduces to the profit function $\pi(k,\lambda)$ as given in (7).

For this extension, the decision problem is to determine the optimal values of (k,λ) that maximize the total welfare function $\Pi(k,\lambda)$ subject to the system stability constraint of $\rho = \frac{\lambda d}{k\mu} < 1$.

We can follow the same approach to analyze the general model for this extension. In particular, we can establish the following results:

Proposition 6 *Suppose that Assumption 1 holds and* $\gamma \leq \frac{2}{3}$. *When the platform maximizes the total welfare function (24), the optimal solution exhibits the following characteristics:*
(i) All results as stated in Propositions 4 and 5 continue to hold.
(ii) When γ *increases, the optimal wage rate* w^* *increases (and both* k^* *and* λ^* *increase), but the optimal price rate* p^* *is not necessarily monotonic.*

Proposition 6(i) shows that our results for the general model are robust even when we include the total customer and provider surplus in the objective function as long as $\gamma \leq \frac{2}{3}$. Proposition 6(ii) further shows that when a higher weight γ is placed on the total consumer and provider surplus, the platform would increase the wage rate w^* to attract more participating providers k^* and serve more customers λ^*. However, the optimal price p^* is not necessarily monotonic as γ increases.

We also conducted numerical experiments to illustrate the results for the extension of the general model using the exact formula (9) for W_q. Table 10 shows the results for the case with the same set of experiments given in Table 3 (with $\bar{\lambda} = 100$) for different values of γ. Observe that as γ increases, the optimal wage rate w^* increases (as supported by Proposition 6), and the optimal wage rate p^* decreases (although Proposition 6 suggests that it is not necessarily monotonic in general). Also, as γ increases, the optimal payout ratio α^* increases, the platform's profit decreases and the total consumer and provider surplus increases. When the platform puts a larger weight on the total consumer and provider surplus than its profit (i.e., $\gamma \geq 0.5$), the optimal payout ratio α^* exceeds one, which implies that the platform is willing to increase the total consumer and provider surplus at the expense of a profit loss. This observation suggests that the platform needs to select a low "equitable payoff" γ to be financially viable in the long run, while an emerging service platform might adopt the strategy of placing a higher weight on the welfare of the consumers and providers initially in order to increase market share; e.g., this strategy was used by Didi during its early stage of competition with taxis and other ride-hailing firms.

Insert Table 10 about here

6 Numerical Illustrations Based on Didi Data

6.1 Background information

To calibrate our model parameters, we collected real data from Didi, the largest on-demand ride-hailing service platform in China that was founded in June 2012.[11] Our data was based on rides that took place in Hangzhou, the capital city of Zhejiang province with an urban population of over 7 millions, during the time periods between September 7-13 and November 1-30 in 2015.

In Hangzhou city, Didi offers different types of services including Taxi (traditional taxi service), Express/Private (equivalent to UBER X/Black with on-demand drivers), and Hitch (equivalent to UBER Pool)[12]. For our numerical illustrations here, we focus on the data associated with the Express/Private service, which accounts for 60% of all rides provided by Didi in Hangzhou. Didi had approximately 13,000 registered drivers for all services in Hangzhou, but the exact number of Express/Private drivers was not known to us. So, we simply assume that 60% of Didi drivers were Express/Private drivers so that the number of registered Express/Private drivers in Hangzhou was assumed to be around $7,800$.

6.2 Number of rides and drivers across different hours

Figure 1 depicts the average number of Express/Private rides and drivers across different hours on any given day. (Here, Hour 8 represents one-hour interval 8am-9am, Hour 19 for 7pm- 8pm, and so on. Data for Hours 1-7 were omitted due to incomplete data in the database.) We observe from the Didi data that the pattern depicted in Figure 1 is consistent throughout the weekdays (even though the average number of rides and drivers were slightly lower on Saturdays and Sunday) and that the peak hours are being Hours 9 and 19, and the slowest hours are being Hours 23 and 24. For instance, during the peak Hour 19, there were an average of 1,211 drivers and an average of 2,006 Express/Private rides in a weekday. However, there were only an average of 597 drivers and an average of 1,029 rides during the late night Hour 23. (The mean and standard deviation of the number of drivers and number of rides over the weekdays are provided in the Appendix B.)

[11] http://www.xiaojukeji.com/en/company.html. Didi merged with Kuaidi (a major competitor) in February 2015 as a way to defend its market share when Uber officially launched its service in China in July 2014. In August 2016, Uber decided to retreat from China and its China operations merged with Didi.

[12] Unlike Uber's business model that aims to displace the traditional taxi services, Didi integrates taxi services into its business model by providing its mobile hailing service to taxi drivers free of charge. Chen et al. (2017) have recently used the data provided by Didi to analyze ridesplitting behavior of passengers using on-demand ride-hailing services.

6.3 Travel distance and travel speed

While the average number of rides and drivers vary substantially across different hours of the day, Figure 2 shows that the average travel distance for each Express/Private ride was rather stable across different hours. For example, the average travel distance d during the peak Hour 19 and during the late night Hour 23 were 6.3 km and 6.6 km, respectively. The Didi database also provided the average travel times μ across different hours from which we can estimate the average travel speed across different hours. For example, we estimated that the average travel speeds were about 19 km/hour for Hour 19 and 26 km/hour for Hour 23. These numbers are consistent with the actual expected traffic conditions, where traffic is much less congested during late night hours. (The travel distance and travel time distributions across different hours are provided in the Appendix B.)

6.4 Price and wage rates

Didi's price p for its service consists of two components so that $p = p_1 + p_2$, where p_1 represents the fare that is primarily based on the travel distance, and p_2 represents surcharges (e.g., tolls). Accordingly, Didi paid its drivers based on the following scheme. When a passenger pays a total fee of p, the driver receives $(p_1 * 80\% - 0.5) * (100\% - 1.77\%) + p_2 * (100\% - 1.77\%)$, but the driver needs to cover the surcharges p_2. Thus, the actual wage that Didi paid its drivers was approximately 80% of the total price; i.e., $w \approx 0.8p$.

Figure 2 also shows that the average price per km charged by Didi (excluding the surcharges) was relatively stable across different hours of the day. Overall, the price per km had a mean of 3.07 RMB and a standard deviation of 1.45 RMB. In particular, the average prices per km charged were RMB 3.13 for Hour 19 (peak hour) and RMB 2.76 for Hour 23 (non-peak hour). We also observe from the Didi data that the average price per km p was highly correlated with the number of rides λ over the peak (non-peak) hours, with a correlation coefficient of 0.81. In other words, the price per km was usually higher during peak hours when the customer request rate is high, and was lower during non-peak hours when the customer request rate is low. This pricing pattern is consistent with the results obtained from our base model (see Proposition 2) that p^* increases as $\bar{\lambda}$ increases. (The mean and standard deviation of the average price per km across different hours are provided in the Appendix B.)

6.5 Strategic factors and their implications

It is important to note that the observed price that Didi charged its passengers was heavily discounted during the data collection periods for two strategic reasons: (a) Didi wanted to attract more passengers by pricing its service below the traditional taxi services;[13] and (b) Didi was engaged in a price war with Uber by offering discount coupons to compete for market share. In addition to offering heavily discounted price to attract passengers, Didi also provided extra "side payments" to its drivers to entice drivers to join its platform due to the intense market competition. For instance, Didi had offered an extra bonus if the number of rides provided by a driver exceeds a certain quota within a 7-day period. BBC (2016) had reported that the extra payment can be as high as 110% of the fare paid by the passengers. With such generous payments, more drivers reported to work and Didi did not need to use surge pricing during peak hours, which explains why Didi was able to offer relatively stable pricing in Hangzhou as depicted in Figure 2. Furthermore, the waiting time for Didi's service was reasonably short with an adequate supply of drivers. Specifically, the average waiting time of all Express/Private rides over the aforementioned time periods was about 6 minutes, of which the waiting time for accepting a ride request was approximately 1 minute and the waiting time for picking up a passenger was approximately 5 minutes.

In view of the heavily discounted price due to the above strategic reasons, the average price per km p as reported in Figure 2 was biased and did not accurately represent the regular prices p that the firm should quote and the actual wages w should offer in equilibrium. Nevertheless, we use the data given in the Didi database to calibrate our model parameters for constructing realistic numerical examples to illustrate some of our model results.

6.6 Numerical examples for illustrative purposes

We next provide some numerical results using parameter values calibrated from the Didi data. As Hangzhou is a large urban area of over $5,000\ km^2$, it is not possible to assign any available driver to serve a call request due to a long pickup time. Instead, only nearby drivers can be used to serve a local request. For simplicity, we assume that the city is divided into 20 zones with equal passenger and driver distributions such that only drivers and riders within the same zone would be matched.

[13] In Hangzhou, taxi charges RMB 11 initially and then RMB 2.6 per km. As a way to entice passengers to choose Didi over taxi service, Didi had priced its service below taxi rates to increase market share. Based on our discussions with passengers in China, there was an expectation that Didi's price rate was lower than the taxi rate.

As such, we simply re-scale the demand and number of available drivers by a factor of 20 and set the maximum number of drivers $K = 7,800/20 = 390$.

We examine the average income for taxi drivers in Hangzhou and the average major out-of-pocket expenses borne by the Didi drivers (including car insurance, license, fuel cost, etc.). We estimate that a minimum hourly wage of RMB 30 is required for a Didi driver to offer service. Thus, the hourly wage reservation r is assumed to be distributed uniformly between RMB 30 to RMB 40.

As discussed earlier, the data were collected during the time when Didi was offering large fare discounts to attract riders such that riders expected that Didi price wound be around or even less than the taxi rate of RMB 2.6 per km in Hangzhou. Thus, we use the taxi rate as a benchmark and assume that the customer value per km v is distributed uniformly between RMB 2 to RMB 4.

As shown in Figure 2, the average travel distances did not vary significantly across hours, so we simply set the average travel distance $d = 6$ km across all hours. It is difficult to provide an accurate estimate of the waiting cost per hour c. Gomez-Ibanez et al. (1999) reported that the waiting cost for a working class passenger in San Francisco is approximately 195% of the passenger's after-tax wages. Using this estimate and the fact that the average hourly wage of workers in Hangzhou is approximately RMB 40 per hour (China Daily, 2016), one can argue that the waiting cost for an average passenger in Hangzhou is approximately RMB 80 per hour. Accounting for the income inequality and the impatient characteristics of most city dwellers in China (Li (2016)), we simply choose the range of c from RMB 0 to RMB 1,000.

We use data from two specific time periods to illustrate our model results. In particular, we use Hour 19 to represent peak-hour characteristics with high demand and travel congestion levels, and Hour 23 to represent non-peak hour characteristics with lower demand and congestion levels. For Hour 19, we set the average customer demand rate $\bar{\lambda} = 200$ with an average service speed $\mu =$ 19 km/hour so that the average demand request rate is equal to 100 ($\approx 1969/20$) when the price rate is equal to RMB 3 to match the Didi data. For Hour 23, we set $\bar{\lambda} = 100$ and $\mu = 26$ km/hour such that the average request rate is equal to 50 ($\approx 1033/20$) when the price rate is equal to RMB 3. We summarize the parameter values used in our illustrative examples in Table 11.

Insert Table 11 about here

In each numerical experiment, we solve for the optimal price and wage rates numerically for the general model using the exact formula (9) for W_q. Figures 3 and 4 show the optimal number of participating drivers k^* (in each zone), price rate p^* and wage rate w^* for the peak hour and non-peak hour scenarios, respectively, as the waiting cost c increases from 0 to $1,000$. Observe that w^* increase as c increases in both Figures 3 and 4, and that k^* (scale on the left), p^* and w^* (scale on the right) are all higher during the peak hour (Figure 3) than those during the non-peak hour (Figure 4), which are intuitive as the peak hour period has a higher customer demand rate $\bar{\lambda}$ and a slower service speed μ than that during non-peak hour period. Also, k^* increases and p^* slightly increases as c increases.

Figure 5 shows that the optimal payout ratio α^* increases from 0.57 to 0.78 for the peak hour scenario and increases from 0.45 to 0.70 for the non-peak hour scenario, respectively, as c increases from 0 to $1,000$. Observe that the optimal payout ratio is always higher during the peak hour than that during the non-peak hour. As the optimal payout ratio α^* increases significantly when c increases, this suggests that a fixed payout ratio would not perform well across different time periods. To illustrate, Figure 6 shows the result for the peak hour scenario (Hour 19) that using the optimal time-based payout ratio α^* can substantially increase the profit of from using a fixed payout ratio of 0.8, especially when c is small in which α^* is much lower than 0.8. In particular, when $c = 0$ (i.e., ignoring waiting cost), the optimal profit is equal to 843 with optimal payout ratio $\alpha^* = 0.57$, as compared with an optimal profit of 479 with a fixed payout ratio of 0.8. Thus, our numerical results suggest that the platform should deploy a time-based payout ratio scheme to achieve a much higher profit across all time periods, especially when the waiting cost c is small.

7 Conclusion

Motivated by the increasing popularity of on-demand service platforms with independent service providers and time sensitive customers, we develop an analytical framework to understand how such platforms should set their optimal price and wage to match the needs of providers and customers taking into account the underlying supply and demand characteristics. Our framework incorporates waiting time performance based on a queueing model in customer utility and captures some important market characteristics including time sensitive customers and earning sensitive service providers. We conduct extensive numerical experiments to illustrate the behavior of the optimal

price and wage rates as predicted by our modeling framework. We further derive analytical results to support the main insights observed in our numerical experiments. Our findings provide some interesting implications in managing prices and wages for on-demand service platforms.

Using some actual data collected from a major on-demand ride-hailing company in China, we calibrate our model parameters to construct realistic numerical examples to illustrate some implications on the optimal price and wage with respect to the underlying operating characteristics. Although our framework does not capture certain important practical issues due to intense competition existed in China when the data were collected (and thus cannot be used to accurately predict the actual behavior of the players in the market), our model results can help to illustrate and explain some observations that are consistent with the actual data provided by the company. More importantly, our model results can serve as a guideline for potentially increasing profitability when the underlying market conditions were to evolve to be consistent with the operating environment captured in our modeling framework. We also illustrate the potential benefits if the company were to adopt a time-based payout ratio versus their current practice of using a fixed payout ratio.

Our results are obtained under the assumption that the customer's valuation of the service and the provider's earning reservation are uniformly distributed. We also conducted some numerical experiments using exponential distributions for both the customer's valuation of the service and the provider's earning reservation, and the results are consistent with those under uniform distributions. However, a comprehensive numerical study is needed to confirm the robustness of our results under more general distributions.

Our model considers price and wage rates that are pre-committed, and we analyze the equilibrium behavior of the system. One future research direction is to study dynamic pricing strategies in which the platform can offer dynamic prices and wages to customers and providers based on the real-time status of the system. Specifically, one can develop a modeling framework that considers the real-time interactions among the customers, providers and the platform where the customers and providers need to make real-time decisions on whether to accept the dynamic prices and wages offered by the service platform. Another possible future research direction is to study platform competition so as to characterize the optimal demand-contingent price and wage strategies in a competitive setting.

References

[1] Afanasyev, M., H. Mendelson. (2010). Service provider competition: Delay cost structure, segmentation and cost advantage. *Manufacturing & Service Operation Management* 12(2): 213-235.

[2] Afeche, P., H. Mendelson. (2004). Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Management Science* 50(7): 869-882.

[3] Allon, G., A. Bassamboo, E.B. Cil. (2012). Large-scale service marketplaces: The role of the moderating firm. *Management Science* 58(10): 1854-1872.

[4] Armstrong, M. (2006). Competition in two-sided markets. *The RAND Journal of Economics* 37(3): 668-691.

[5] Armony, M., and M. Haviv (2003). Price and delay competition between two service providers. *European Journal of Operational Research* 147(1): 32-50.

[6] Benjaafar, S., G. Kong, X. Li, and C. Courcoubetis. (2015). Peer-to-peer product sharing. Working paper, University of Minnesota.

[7] Cachon G.P., K.M. Daniels, R. Lobel. (2015) The role of surge pricing on a service platform with self-scheduling capacity. Available at SSRN.

[8] Chen, M.K., M. Sheldon. (2015) Dynamic Pricing in a Labor Market: Surge Pricing and Flexible Work on the UBER Platform. Working paper, UCLA Anderson School.

[9] Chen, X.M., M. Zahiri, S. Zhang. (2017). Understanding ridesplitting behavior of on-demand ride services: An ensemble learning approach. *Transportation Research Part C* 76: 51-70.

[10] China Daily. 2016. Average salary in major Chinese cities is US$900 and growing. January 21, 2016. http://www.chinadaily.com.cn/china/2016-01/21/content_23183484.htm

[11] Cui, T.H., J.S. Raju, J. Zhang. (2007). Fairness and channel coordination. *Management Science* 53(8): 1303-1314.

http://www.forbes.com/sites/aswathdamodaran/2014/06/10/a-disruptive-cab-ride-to-riches-the-uber-payoff/.

[12] DePillis, L. (2016). One Reason You Might be Better Off Driving for Uber than in a Taxi. *The Washington Post* (March 15).

[13] Fraiberger, S.P., A. Sundararajan. (2015). Peer-to-peer rental markets in the sharing economy. Working paper, New York University.

[14] Gomez-Ibanez, J., W. Tye, C. Winston. (1999). *Essays in Transportation Economics and Policy*, 42. Brookings Institution Press, Washington D.C.

[15] Gross, D., J.F. Shortle, J.M. Thompson, C.M. Harris. (2008). *Fundamentals of Queueing Theory.* Fourth Edition. John Wiley & Sons, Inc., New Jersey.

[16] Gurvich, I., M. Lariviere, A. Moreno-Garcia. (2015). Operations in the on-demand economy: Staring services with self-scheduling capacity. Technical report, Northwestern University.

[17] Haws, K.L., X.O. Bearden. (2006). Dynamic pricing and consumer fairness perceptions. *Journal of Consumer Research* 33: 304-311.

[18] Hu, M., Y. Zhou. (2017) Price, wage and fixed commision in on-demand matching. Working paper. Rotman School of Management, University of Toronto.

[19] Huet, E. (2014) Uber Now Taking its Biggest UberX Commission Ever – 25 Percent. *Forbes*, September 22, 2014.

[20] Jiang, B., L. Tian. (2015). Collaborative consumption: Strategic and economic implications of product sharing. Working paper, Washington University.

[21] Kokalitcheva, K. (2015). Uber and Lyft face a new challenger in Boston. *Fortune.com* (October 5).

[22] Lee, H.L., M.A. Cohen. (1983). A Note on the Convexity of Performance Measures of $M/M/c$ Queueing Systems. *Journal of Applied Probability* 20: 920-923.

[23] Li, A., (2016). Why are Chinese tourists so rude? A few insights. *South China Morning Post*, August 10, 2016.

[24] Li, J., A. Moreno, D.J. Zhang. (2015). Agent behavior in the sharing economy: Evidence from Airbnb. Working paper, University of Michigan .

[25] MacMillan, D. (2015). The $50 billion question: Can Uber deliver? *Wall Street Journal* (June 16) A1-A12.

[26] Moreno, A., C. Terwiesch. (2014). Doing business with strangers: Reputation in online service marketplaces. *Information Systems Research* 25(4): 865-886.

[27] Naor, P. (1969). The regulation of queue size by levying tolls. *Econometrica* 37(1): 15-24.

[28] Riquelme, C., S. Banerjee, R. Johari. (2015). Pricing in ride-share platforms: A queueing-theoretic approach. Working paper, Stanford University.

[29] Rochet, J.C., J. Tirole. (2003). Platform competition in two-sided markets. *Journal of the European Economic Association* 1(4): 990-1029.

[30] Rochet, J.C., J. Tirole. (2006). Two-sided markets: A progress report. *The RAND Journal of Economics* 37(3): 645-667

[31] Rogers, B. (2015). The Social Costs of Uber. *The University of Chicago Law Review*, 82: 85-104.

[32] Roose, K. (2014). Does Silicon Valley have a contract-worker problem? *NYMag.com* (September 18).

[33] Sakasegawa, H. (1977). An approximation formula $L_q \simeq \alpha\rho^{\beta}/(1-\rho)$. *Annals of the Institute of Statistical Mathematics* 29(1): 67-75.

[34] Sheldon, M. (2016). Income targeting and the ride-sharing market. Working paper, University of Chicago. Available at http://www.michaelsheldon.org/working-papers-section/.

[35] Shoot, B. (2015). Hot food, fast. *Entrepreneur* 68 (Aug.).

[36] Taylor T. (2016). On-Demand Service Platforms. Working paper, University of California, Berkeley. Available at SSRN 2722308.

[37] Wirtz, J., C.S. Tang. (2016). UBER: Competiting as Market Leader in the US versus Being a Distant Second in China. Case Study published in Wirtz and Lovelock (2016), Service Marketing: People, Technology and Strategy, 8th edition. World Scientific.

[38] Zhou, W., X. Chao, X. Gong. (2014). Optimal uniform pricing strategy of a service firm when facing two classes of customers. *Production and Operations Management* 23(4): 676-688.

Appendix

A Mathematical Proofs:

A.1 Proof of Proposition 1

When $k^* < K$, it is clear that the right hand of (19) is increasing in n for any given fixed parameter values. Since the optimal k is obtained by maximizing k subject to (19), we can conclude that k^* increases in n, i.e., $\frac{\partial k^*}{\partial n} > 0$. We next show that k^* is concave in n, i.e., $\frac{\partial^2 k^*}{\partial n^2} < 0$.

Let $y = \frac{c\rho\sqrt{2(n+1)}}{1-\rho}$. Since $s = \frac{\lambda}{\bar{\lambda}}$, we can rewrite (19) as

$$k^2 = K\alpha\{s\bar{\lambda}d(1-s) - y\}.$$

Taking the derivative with respect to n on both sides, we obtain

$$2k\frac{\partial k}{\partial n} = K\alpha\left[\bar{\lambda}d(1-2s)\frac{\partial s}{\partial n} - \frac{\partial y}{\partial s}\frac{\partial s}{\partial n} - \frac{\partial y}{\partial k}\frac{\partial k}{\partial n} - \frac{\partial y}{\partial n}\right] \tag{25}$$

Also,

$$\frac{\partial y}{\partial s} = \bar{\lambda}dc\frac{\rho^{\sqrt{2(n+1)}-1}}{k\mu(1-\rho)}\left(\sqrt{2(n+1)} + \frac{\rho}{1-\rho}\right),$$

from which we can use (30) to deduce that

$$\frac{\partial y}{\partial s} = \bar{\lambda}d(1-2s). \tag{26}$$

We substitute (26) into (25) to obtain

$$(2k + K\alpha\frac{\partial y}{\partial k})\frac{\partial k}{\partial n} = -K\alpha\frac{\partial y}{\partial n}.$$

It is straightforward to show that $\frac{\partial y}{\partial n} < 0$. Since $\frac{\partial k}{\partial n} > 0$, we have

$$2k + K\alpha\frac{\partial y}{\partial k} > 0. \tag{27}$$

We next take the derivative of (25) with respect to n and obtain

$$2(\frac{\partial k}{\partial n})^2 + 2k\frac{\partial^2 k}{\partial n^2} = K\alpha\left[-2\bar{\lambda}d(\frac{\partial s}{\partial n})^2 + \bar{\lambda}d(1-2s)\frac{\partial^2 s}{\partial n^2} - \frac{\partial y}{\partial s}\frac{\partial^2 s}{\partial n^2} - \frac{\partial^2 y}{\partial s^2}(\frac{\partial s}{\partial n})^2 - \frac{\partial^2 y}{\partial k^2}(\frac{\partial k}{\partial n})^2 - \frac{\partial y}{\partial k}\frac{\partial^2 k}{\partial n^2} - \frac{\partial^2 y}{\partial n^2}\right].$$

We substitute (26) into the above equation to obtain

$$(2k + K\alpha\frac{\partial y}{\partial k})\frac{\partial^2 k}{\partial n^2} = K\alpha\left[-2\bar{\lambda}d(\frac{\partial s}{\partial n})^2 - \frac{\partial^2 y}{\partial s^2}(\frac{\partial s}{\partial n})^2 - \frac{\partial^2 y}{\partial k^2}(\frac{\partial k}{\partial n})^2 - \frac{\partial^2 y}{\partial n^2}\right] - 2(\frac{\partial k}{\partial n})^2. \tag{28}$$

It is straighforward to verify that $\frac{\partial^2 y}{\partial s^2} > 0$, $\frac{\partial^2 y}{\partial k^2} > 0$ and $\frac{\partial^2 y}{\partial n^2} > 0$, which implies that the right side of (28) is less than zero. Using (27), we can conclude that $\frac{\partial^2 k}{\partial n^2} < 0$.

Clearly, we have $k^*(0) > 0$ when $n = 0$. Also, $k^*(n) \leq K$ for all n, as $k^*(n)$ is given by the largest value of k, with $k \leq K$, that possesses a feasible λ to (19). Thus, $k^*(K) \leq K$. Since we have shown that k^* is concave and increasing in n, we can conclude that there exists one unique fixed point such that $k^*(n^*) = n^*$, and the iterative procedure will always converge to n^*.

A.2 Proof of Proposition 2

With the constraint that $\frac{w}{p} = \alpha$, the objective function can be expressed as $\pi = \lambda d(p - w) = \lambda d(\frac{1}{\alpha} - 1)w$. We can solve the constrained problem as an unconstrained Lagrange optimization problem with the Lagrange function of $L(p, w, z) = \lambda d(\frac{1}{\alpha} - 1)w + z\lambda d(\alpha p - w)$, where z is the nonzero Lagrange multiplier.[14] We substitute the values of p and w given by (17) and (11) and the fact that $\lambda = \bar{\lambda}s$ into the Lagrange function $L(p, w, z)$, and can obtain the following optimality conditions from the three first-order conditions, $\frac{\delta L}{\delta k} = 0$, $\frac{\delta L}{\delta s} = 0$, and $\frac{\delta L}{\delta z} = 0$, respectively:

$$(\frac{1}{\alpha} - 1)\frac{2k}{K} + z\left[c\mu\alpha\frac{\rho\sqrt{2(n+1)}}{k\mu(1-\rho)}\left(\sqrt{2(n+1)} + \frac{\rho}{1-\rho}\right) - \frac{2k}{K}\right] = 0, \tag{29}$$

$$(1 - 2s) - c\frac{\rho\sqrt{2(n+1)}-1}{k\mu(1-\rho)}\left(\sqrt{2(n+1)} + \frac{\rho}{1-\rho}\right) = 0, \tag{30}$$

$$\alpha\left[1 - s - \frac{c}{k\mu}\left(\frac{\rho\sqrt{2(n+1)}-1}{1-\rho}\right)\right] - \frac{k^2}{K\lambda d} = 0. \tag{31}$$

We next use the optimality conditions (30) and (31) to establish the following properties:

(i) ρ^* and k^* change in the same direction for any fixed n, α, K, c and μ;

(ii) λ^* and k^* change in the same direction for any fixed n, α, K, c, d and μ;

(iii) ρ^* and w^* change in the same direction for any fixed n, α, K, c and μ;

(iv) ρ^* and $\frac{W_q^*}{d}$ change in the same direction for any fixed n, α, K, c and μ;

(v) ρ^* and s^* change in the opposite direction for any fixed n, α, K and c.

First, we can substitute (30) into (31) and use $\rho = \frac{\lambda d}{k\mu} < 1$ to obtain

$$\frac{\alpha}{2}\left[1 + c\frac{\rho\sqrt{2(n+1)}-1}{k\mu(1-\rho)}\left(\sqrt{2(n+1)} + \frac{\rho}{1-\rho} - 2\right)\right] - \frac{k}{\rho K\mu} = 0. \tag{32}$$

[14] We ignore the constraints that $0 \leq s \leq 1$ and $0 \leq k \leq K$ to simplify our exposition in the proof, but the analysis can be easily adapted to include these constraints as well.

We can show that the left side of (32) increases in ρ but decreases in k for any fixed n, α, K, c and μ. Thus, ρ^* and k^* must change in the same direction, which prove (i).

Since $\lambda = \frac{\rho k \mu}{d}$, and ρ^* and k^* must change in the same direction as proved in (i), ρ^*, k^* and λ^* must all change in the same direction for any fixed n, α, K, c, d and μ. This prove (ii).

We can use (11) to rewrite (32) as

$$\frac{\alpha}{2}\left\{1+\frac{c}{wK\mu^2(1-\rho)}\left[\rho^{\sqrt{2(n+1)}-2}\left(\sqrt{2(n+1)}-2\right)+\frac{\rho^{\sqrt{2(n+1)}-1}}{1-\rho}\right]\right\}-w=0. \tag{33}$$

The left side of (33) increases in ρ but decreases in w for any fixed n, α, K, c and μ. Thus, the values of ρ and w at optimality must change in the same direction, which proves (iii).

We can use (16) to rewrite (32) as

$$\rho\left(1-\frac{c}{d}W_q\right)-\frac{c}{d}W_q(1-\rho)\left[(\sqrt{2(n+1)}-2)\right]+\frac{2d\rho^{\sqrt{2(n+1)}-2}}{K\alpha\mu^2W_q}-1=0. \tag{34}$$

The left side of (34) decreases in $\frac{W_q}{d}$ but increases in ρ for any fixed n, α, K, c and μ. This shows that the values of $\frac{W_q}{d}$ and ρ at optimality must change in the same direction for any fixed n, α, K, c and μ, which proves (iv).

Finally, we can again use (16) to rewrite (30) as

$$(1-2s)-\frac{c}{d}W_q\left(\sqrt{2(n+1)}+\frac{\rho}{1-\rho}\right)=0,$$

or equivalently,

$$s=\frac{1}{2}\left[1-\frac{c}{d}W_q\left(\sqrt{2(n+1)}+\frac{\rho}{1-\rho}\right)\right].$$

We show in (iv) that the values of ρ and $\frac{W_q}{d}$ at optimality change in the same direction for any fixed n, α, μ, K and c. It then follows that the value of ρ and s at optimality must change in the opposite direction, which proves (v).

We can now rewrite (30) as

$$(1-2\frac{\lambda}{\bar{\lambda}})-\frac{c}{d}W_q\left(\sqrt{2(n+1)}+\frac{\rho}{1-\rho}\right)=0. \tag{35}$$

We have shown in (i), (ii) and (iv) that the values of λ, $\frac{W_q}{d}$ and ρ at optimality change in the same direction for any fixed n, α, K, c, d and μ. We can deduce from (35) that the values of λ, $\frac{W_q}{d}$ and ρ at optimality must all increase when $\bar{\lambda}$ increases. It then follow from (iii) that w^* (and thus $p^* = \frac{w^*}{\alpha}$) increases when $\bar{\lambda}$ increases.

We can also use $\rho = \frac{\bar{\lambda} s^* d}{k\mu}$ to express d as $d = \frac{\rho k \mu}{\bar{\lambda} s}$. It follows from (i) and (v) that the value of ρ at optimality must change in the same direction as k, but in opposite direction of s for any fixed n, α, K, c and μ. Therefore, we can deduce that, at optimality, the values of ρ and k must increase while the value of s must decrease when d increases. From (iii), we can also conclude that w^* (and thus $p^* = \frac{w^*}{\alpha}$) increases when d increases. ■

A.3 Proof of Proposition 3

To prove that there is a unique fixed point, $n^* = k^*(n^*)$, we shall show that there exists only one solution (λ^*, k^*) that can satisfy the two first-order conditions, (20) and (21), at the same time if we set $n = k$ in (20) and (21). As $s = \frac{\lambda}{\bar{\lambda}}$, we use s instead of λ in our analysis here.

First, observe that the left side of (20) decreases in k but increases in s (or λ). Therefore, for any fixed s, there exists only one k, denoted by $k(s)$, such that $(s, k(s))$ satisfies (20), and that $k(s)$ increases in s. Next, let

$$h(\rho, k) = c\frac{\rho^{\sqrt{2(k+1)}-1}}{k\mu(1-\rho)}\left(\sqrt{2(k+1)} + \frac{\rho}{1-\rho}\right),$$

so that we can espress the first-order condition (21) as

$$1 - 2s - h(\rho, k) = 0 \tag{36}$$

Taking the second derivative of (36) with respect to s, we obtain

$$-\frac{\partial h^2(\rho,k)}{\partial \rho^2}\left(\frac{\partial \rho}{\partial s}\right)^2 - \frac{\partial h(\rho,k)}{\partial \rho}\frac{\partial^2 \rho}{\partial s^2} - \frac{\partial h^2(\rho,k)}{\partial k^2}\left(\frac{\partial k}{\partial s}\right)^2 - \frac{\partial h(\rho,k)}{\partial k}\frac{\partial^2 k}{\partial s^2} = 0. \tag{37}$$

We can further expand $\frac{\partial^2 \rho}{\partial s^2} = \frac{\partial^2 \rho}{\partial k^2}\left(\frac{\partial k}{\partial s}\right)^2 + \frac{\partial \rho}{\partial k}\left(\frac{\partial^2 k}{\partial s^2}\right)$ and rewrite (37) as

$$-\frac{\partial h^2(\rho,k)}{\partial \rho^2}\left(\frac{\partial \rho}{\partial s}\right)^2 - \frac{\partial h(\rho,k)}{\partial \rho}\frac{\partial^2 \rho}{\partial k^2}\left(\frac{\partial k}{\partial s}\right)^2 - \frac{\partial h^2(\rho,k)}{\partial k^2}\left(\frac{\partial k}{\partial s}\right)^2 - \left[\frac{\partial h(\rho,k)}{\partial k} + \frac{\partial \rho}{\partial k}\right]\frac{\partial^2 k}{\partial s^2} = 0. \tag{38}$$

Since both $\frac{\rho^{\sqrt{2(k+1)}-1}}{1-\rho}$ and $\frac{\rho}{1-\rho}$ are convexly increasing in ρ, we have $\frac{\partial h^2(\rho,k)}{\partial \rho^2} > 0$. Also, it is straightforward to show that $\frac{\partial h(\rho,k)}{\partial \rho} > 0$ and $\frac{\partial^2 \rho}{\partial k^2} = \frac{2\lambda d}{k^3 \mu} > 0$. Furthermore, we can prove that both $\frac{\rho^{\sqrt{2(k+1)}-1}}{1-\rho}$ and $\frac{\sqrt{2(k+1)}+\frac{\rho}{1-\rho}}{k}$ are convexly decreasing k, so that $h(\rho, k)$ is also convexly decreasing in k. Therefore, $\frac{\partial h^2(\rho,k)}{\partial k^2} > 0$ and $\frac{\partial h(\rho,k)}{\partial k} < 0$. Finally, $\frac{\partial \rho}{\partial k} = -\frac{\lambda d}{k^2 \mu} < 0$. We can then infer from (38) that $\frac{\partial^2 k}{\partial s^2} > 0$. Thus, we have established that $k(s)$ is convexly increasing in s, with $k(0) = 0$.

For any (s^*, k^*) that satisfies both (20) and (21), we have

$$\frac{2k^2}{K} = \bar{\lambda} d(1-2s)s. \tag{39}$$

Thus, for any fixed s, there exists only one $k(s)$ satisfying (39). Furthermore, $k(s)$ first increases in s, and then decreases in s with $k(0) = k(\frac{1}{2}) = 0$ and $\frac{\partial k}{\partial s}|_{s=0} \to \infty$.

Overall, we use (36) to show that $k(s)$ is convexly increasing in s with $k(0) = 0$, and use (39) to show that $k(s)$ first increases in s, and then decreases in s with $k(0) = k(\frac{1}{2}) = 0$. Therefore, there exists only one solution (s^*, k^*) that can satisfy both (36) and (39) when $n = k$. In other words, there exists a unique fixed point, $k^*(n^*) = n^*$.

A.4 Proof of Proposition 4

To establish the analytical results for the general model with a dynamic payout ratio, we first provide some preliminary result for a special case for the general model by imposing a fixed target service level s. In particular, assume that the model parameter s (or equivalently, the customer request rate λ because $\lambda = s\bar{\lambda}$) is exogenously given and the optimization problem (8) is reduced to:

$$\max_k \ \pi(k) \equiv \lambda d \left[(1-s) - c\frac{\rho\sqrt{2(n+1)}-1}{k\mu(1-\rho)} - \frac{k^2}{K\lambda d} \right], \text{ subject to } \frac{\lambda d}{k\mu} < 1.$$

The results for this special case are based on the following more general settings than the simplifying assumptions as stated in Assumption 1. Specifically, we assume that W_q, $F(.)$ and $G(.)$ satisfy the following assumptions:

Assumption 2: *The expected waiting time function W_q is convex and increasing in λ, and is convex and decreasing in both k and μ. Furthermore, $\frac{\partial}{\partial\lambda}(\frac{\partial W_q}{\partial k}) < 0$, $\frac{\partial}{\partial d}(\frac{\partial W_q}{\partial k}) < 0$ and $\frac{\partial}{\partial\mu}(\frac{\partial W_q}{\partial k}) > 0$.*

Observe that the convexity of the waiting time function W_q is valid for an $M/M/k$ queueing model with arrival rate λ and service rate $\frac{\mu}{d}$; e.g., see Lee and Cohen (1983). The three conditions, $\frac{\partial}{\partial\lambda}(\frac{\partial W_q}{\partial k}) < 0$, $\frac{\partial}{\partial d}(\frac{\partial W_q}{\partial k}) < 0$ and $\frac{\partial}{\partial\mu}(\frac{\partial W_q}{\partial k}) > 0$, basically require that the marginal decrease in waiting time due to an additional service provider is larger at a higher system utilization level. This assumption is reasonable, and is also supported by the waiting time function of an $M/M/k$ queuing system. However, we do not require any specific functional form of W_q.

Assumption 3: *The cumulative value distribution $F(.)$ is strictly increasing. The cumulative wage distribution $G(.)$ is concave and strictly increasing.*

Assumption 3 stipulates that the density of the reservation wage rate r is decreasing, which implies that there are more service providers who would be willing to participate and offer service at a lower minimum earning rate. Clearly, Assumption 1 implies Assumptions 2 and 3.

Under Assumptions 2 and 3 and with a fixed value of s, we can obtain the following result:

Lemma 1 *The profit function $\pi(k)$ given in (7) is concave in k. Also, the optimal number of participating providers k^* satisfies the following first-order condition:*

$$-c\lambda\frac{\partial W_q}{\partial k} = G^{-1}(\frac{k}{K}) + G'^{-1}(\frac{k}{K})\frac{k}{K} = G^{-1}(\beta) + \beta G'^{-1}(\beta) = \frac{\partial(\beta G^{-1}(\beta))}{\partial\beta}. \tag{40}$$

Proof of Lemma 1: Differentiate the profit function given in (7) with respect to k and obtain

$$\pi'(k) = -c\lambda\frac{\partial W_q}{\partial k} - \left[G^{-1}(\frac{k}{K}) + G'^{-1}(\frac{k}{K})\frac{k}{K}\right] \tag{41}$$

and

$$\pi''(k) = -\lambda c\frac{\partial^2 W_q}{\partial k^2} - \left[2G'^{-1}(\frac{k}{K})\frac{1}{K} + G''^{-1}(\frac{k}{K})\frac{k}{K^2}\right]. \tag{42}$$

Assumption 3 implies that $G^{-1}(.)$ is convex and increasing. Together with Assumption 2, it follows that $\pi''(k) < 0$, which shows that $\pi(k)$ is concave in k. Therefore, the optimal value of k is given by the first-order condition $\pi'(k) = 0$, which is given in (40). This completes our proof. ■

The first-order condition given in (40) can be interpreted as follows. The left side of (40) measures the marginal reduction in waiting cost for each additional service provider joining the platform. In view of (5), the term $G^{-1}(\beta) = w \cdot \frac{\lambda d}{k}$ represents the average earning rate of a provider. Hence, by noting that $\beta = k/K$, the right side of (40) can be interpreted as the marginal benefit associated with the increase in the average earning rate for each additional service provider participating in the platform in terms of β. Therefore, the first-order condition (40) shows that the optimal value of k is achieved when marginal cost equals marginal benefit.

By using the implicit function theorem to analyze the first-order condition (40), we can establish the following proposition:

Proposition A1: Suppose that Assumptions 2 and 3 hold. Then,

(a) When K increases, both k^* and p^* increase, but the ratio $\beta = \frac{k^*}{K}$ decreases.

(b) When μ increases, both k^* and w^* decrease.

(c) When c increases, both k^* and w^* increase.

(d) When $\bar{\lambda}$ (or s) increases, k^* increases.

(e) When d increases, k^* increases.

Proof of Proposition A1: (a) Suppose that k_0 denotes the optimal value of k for $K = K_0$. Using the first-order condition (40) and expressing the profit π as a function of k, we have

$$\pi'(k_0) = -\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K_0}) + G'^{-1}(\frac{k_0}{K_0})\frac{k_0}{K_0} \right] = 0 \tag{43}$$

Note that $G'^{-1}(.)$ is an increasing function since $G^{-1}(.)$ is convex as $G(.)$ is concave by Assumption 3, which implies that $\{G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K}\}$ is decreasing in K. Therefore, for any $K_1 > K_0$,

$$-\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K_1}) + G'^{-1}(\frac{k_0}{K_1})\frac{k_0}{K_1} \right] > 0.$$

Since the profit function is concave in k, the optimal k^* must be greater than k_0 for any fixed $K = K_1 > K_0$, which shows that the optimal k^* is increasing in K. Since the waiting time W_q is decreasing in k, the optimal p^* given in (3) is also increasing in K.

Let $\beta_0 = \frac{k_0}{K_0}$, and rewrite the derivative of the profit function (43) as

$$\pi'(k_0) = -\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - \left\{ G^{-1}(\beta_0) + G'^{-1}(\beta_0)\beta_0 \right\} = 0, \tag{44}$$

Let k_1 be the optimal value of k when $K = K_1 > K_0$, and define $\beta_1 = \frac{k_1}{K_1}$. Then,

$$\pi'(k_1) = -\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_1} - \left\{ G^{-1}(\beta_1) + G'^{-1}(\beta_1)\beta_1 \right\} = 0. \tag{45}$$

Since k^* is increasing in K, we have $k_1 > k_0$. Thus, $-\lambda c \frac{\partial W_q(\lambda,k,\mu,d)}{\partial k}|_{k=k_1} < -\lambda c \frac{\partial W_q(\lambda,k,\mu,d)}{\partial k}|_{k=k_0}$. From (44) and (45), we can obtain

$$G^{-1}(\beta_1) + G'^{-1}(\beta_1)\beta_1 < G^{-1}(\beta_0) + G'^{-1}(\beta_0)\beta_0.$$

Since $G^{-1}(\beta) + G'^{-1}(\beta)\beta$ is an increasing function in β, we can conclude that $\beta_1 < \beta_0$. Therefore, β^* is decreasing in K.

(b) Let k_0 denote the optimal k when $\mu = \mu_0$. Then,

$$\pi'(k_0) = -\lambda c \frac{\partial W_q(\lambda, k, \mu_0, d)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K} \right] = 0$$

We have $\frac{\partial}{\partial \mu}(\frac{\partial W_q}{\partial k}) > 0$ from Assumption 2. Then, for any $\mu_1 > \mu_0$,

$$-\lambda c \frac{\partial W_q(\lambda, k, \mu_1, d)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K}\right] < 0,$$

Therefore, the optimal k^* must be smaller than k_0 for any fixed $\mu = \mu_1 > \mu_0$, which shows that the optimal k^* is decreasing in μ. From (6), the wage rate is increasing in k^*, therefore, w^* is decreasing in μ.

(c) Suppose that k_0 denotes the optimal value of k for $c = c_0$. Then,

$$\pi'(k_0) = -\lambda c_0 \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K}\right] = 0$$

It is clear that for any $c_1 > c_0$,

$$-\lambda c_1 \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K}\right] > 0.$$

Therefore, the optimal k^* must be greater than k_0 for any fixed $c = c_1 > c_0$, which shows that the optimal k^* is increasing in c. It is clear from (6) that the wage rate w is increasing in k. Therefore, the optimal w^* is also increasing in c.

(d) Since $\lambda = \bar{\lambda}s$, it suffices to show that k^* is increasing in λ. Let k_0 denote the optimal k when $\lambda = \lambda_0$. Then,

$$\pi'(k_0) = -\lambda_0 c \frac{\partial W_q(\lambda_0, k, \mu, d)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K}\right] = 0.$$

We have $\frac{\partial}{\partial \lambda}(\frac{\partial W_q}{\partial k}) < 0$ from Assumption 2, which implies that $-\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0}$ is increasing in λ. Therefore, for any $\lambda_1 > \lambda_0$, we have

$$-\lambda_1 c \frac{\partial W_q(\lambda_1, k, \mu, d)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K}\right] > 0.$$

Therefore, the optimal k^* must be greater than k_0 for any fixed $\lambda = \lambda_1 > \lambda_0$, which shows that the optimal k^* is increasing in λ.

(e) When $d = d_0$, let k_0 denote the optimal k. The first-order condition shows,

$$\pi'(k_0) = -\lambda c \frac{\partial W_q(\lambda, k, \mu, d_0)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K}\right] = 0$$

Since from assumption 1, we know $\frac{\partial}{\partial d}(\frac{\partial W_q}{\partial k}) < 0$. Therefore, for any $d_1 > d_0$, we must have,

$$-\lambda c \frac{\partial W_q(\lambda, k, \mu, d_1)}{\partial k}|_{k=k_0} - \left[G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K}\right] > 0,$$

The optimal k^* must be greater than k_0 for any $d = d_1 > d_0$, which shows that the optimal k^* is increasing in d. ■

We can now proceed to prove the results of Proposition 4. First. we shall show that the optimal (λ^*, k^*) is given by the two first-order conditions (20) and (21). As $s = \frac{\lambda}{\lambda}$, we use s instead of λ in our analysis here. For easier reference, we provide the two conditions below:

$$\frac{\partial \pi}{\partial k} = c\mu \frac{\rho\sqrt{2(n+1)}}{k\mu(1-\rho)} \left(\sqrt{2(n+1)} + \frac{\rho}{1-\rho} \right) - \frac{2k}{K} = 0 \tag{46}$$

$$\frac{\partial \pi}{\partial s} = \bar{\lambda} d \left\{ (1-2s) - c \frac{\rho^{\sqrt{2(n+1)}-1}}{k\mu(1-\rho)} \left(\sqrt{2(n+1)} + \frac{\rho}{1-\rho} \right) \right\} = 0 \tag{47}$$

We can use the same argument in the proof of Proposition 3 to establish the result that for any given n, there exists a unique solution (s^*, k^*) to (46) and (47). Furthermore, we shall show that this unique stationary point is a global maximum. Suppose that this stationary point is not a global maximum. Then, the maximum point must be at the boundaries, i.e., 1) $s = 0$; 2) $k = 0$; 3) $\rho = 1$; 4) $s = 1$; or 5) $k = \infty$. For the first three cases, the profit function (18) takes a value of $-\infty$. For $s = 1$, it is easy to verify that $\frac{\partial \pi}{\partial s} < 0$, which implies that the platform can increase profit by reducing s, and so $s = 1$ cannot be optimal. Finally, when $k = \infty$, $\frac{\partial \pi}{\partial k} < 0$, which shows that the platform can increase profit by reducing k, and so $k = \infty$ cannot be optimal. Therefore, the unique stationary point must be a global maximum.

In the following, we shall use (47) to study the behavior of s^*, and use (46) to characterize the behavior of k^* as a function of s^*.

(a) Let (s_0, k_0) and (s_1, k_1) be the optimal values of (s, k) when $K = K_0$ and $K = K_1$, respectively. Suppose that $K_1 > K_0$. We shall show that $s_1 \geq s_0$ and $k_1 \geq k_0$, which implies that both s^* and k^* increase when K increases.

We use the notation $k^*(K, s)$ to denote the optimal value of k for the general model with parameter K and fixed service level s. In particular, $k^*(K_0, s_0) = k_0$ and $k^*(K_1, s_1) = k_1$. Since $K_1 > K_0$, it follows from Proposition A1(a) that $k^*(K_1, s_0) \geq k^*(K_0, s_0) = k_0$. It is clear that the derivative $\frac{\partial \pi}{\partial s}$ given in (47) is increasing in k. Since (s_0, k_0) satisfies the first-order condition $\frac{\partial \pi}{\partial s} = 0$ and $k^*(K_1, s_0) \geq k_0$, we have

$$(1-2s_0) - c \frac{\left[\frac{\bar{\lambda} s_0 d}{k^*(K_1, s_0)\mu} \right]^{\sqrt{2(n+1)}-1}}{k^*(K_1, s_0)\mu(1 - \frac{\bar{\lambda} s_0 d}{k^*(K_1, s_0)\mu})} \left(\sqrt{2(n+1)} + \frac{\frac{\bar{\lambda} s_0 d}{k^*(K_1, s_0)\mu}}{1 - \frac{\bar{\lambda} s_0 d}{k^*(K_1, s_0)\mu}} \right) \geq 0.$$

Therefore, the optimal value of s must be greater than s_0 when $K = K_1$, as $\pi(s, k^*(s))$ is concave in s. Since (s_1, k_1) is optimal at $K = K_1$, this proves that $s_1 \geq s_0$. Also, it follows from Proposition A1(d) that $k_1 = k^*(K_1, s_1) \geq k^*(K_1, s_0) \geq k_0$. Thus, we prove that both s^* and k^* increase in K.

Using (11), (46) and (47), we have

$$1 - 2s = c\frac{\rho^{\sqrt{2(n+1)}-1}}{k\mu(1-\rho)}\left(\sqrt{2(n+1)} + \frac{\rho}{1-\rho}\right) = 2w. \tag{48}$$

This proves that w^* decreases in K since s^* increases in K.

We next show that W_q^* decreases in K. First, we can rewrite (47) as

$$\bar{\lambda}sd(1-2s) = \lambda d\left[c\frac{\rho^{\sqrt{2(n+1)}-1}}{k\mu(1-\rho)}\left(\sqrt{2(n+1)} + \frac{\rho}{1-\rho}\right)\right] = c\frac{\rho^{\sqrt{2(n+1)}}}{1-\rho}\left(\sqrt{2(n+1)} + \frac{\rho}{1-\rho}\right). \tag{49}$$

Clearly, the right side of (49) is increasing in ρ. Also, the left side of (49) implies that $0 < s^* \leq \frac{1}{2}$. Suppose that $0 < s^* < \frac{1}{4}$. In this case, the left side of (49) increases in s. Since s^* increases in K as proved earlier, we can conclude that ρ^* must also increase in K in this case. We can use the first-order condition (47) and ((16) to deduce that

$$(1 - 2s^*) = c\frac{W_q}{d}\left(\sqrt{2(n+1)} + \frac{\rho}{1-\rho}\right). \tag{50}$$

Since s^* increases in K, the left side of (50) must be decreasing as K increases. On the other hand, we have shown that ρ^* increases in K in this case, which implies that $\frac{\rho^*}{1-\rho^*}$ must be increasing in K in this case. We can conclude from (50) that W_q^* must be decreasing as K increases in this case.

Now suppose that $\frac{1}{4} \leq s^* \leq \frac{1}{2}$. In this case, the left side of (49) decreases in s. Since s^* increases in K, we can conclude from (49) that ρ^* must be decreasing in K in this case. Also, we can deduce from (16) that

$$\frac{W_q^*}{d} = \frac{\rho^{*\sqrt{2(n+1)}-1}}{k^*\mu(1-\rho^*)}. \tag{51}$$

As K increases, k^* increases and ρ^* decreases. Since the right side of (51) decreases in k^* but increases in ρ^*, we can conclude from (51) that $\frac{W_q^*}{d}$ (or equivalently W_q^*) must also be decreasing in K in this case.

Furthermore,

$$\pi = \lambda d(p - w) = \lambda d(1 - s - \frac{c}{d}W_q - w) = \lambda d(\frac{1}{2} - \frac{c}{d}W_q), \tag{52}$$

where the last equality follows from (48). Since λ^* increases in K and W_q^* decreases in K, we can conclude that π^* increases in K.

(b) Similarly, let (s_0, k_0) and (s_1, k_1) be the optimal values of (s, k) when $\mu = \mu_0$ and $\mu = \mu_1$, respectively. Suppose that $\mu_1 > \mu_0$. Again, we use the notation $k^*(\mu, s)$ to denote the optimal value of k for the general model with parameter μ and fixed service level s such that $k^*(\mu_0, s_0) = k_0$ and $k^*(\mu_1, s_1) = k_1$. We also use the notation $\rho^*(\mu, s)$ and $W_q^*(\mu, s)$ to denote the corresponding optimal values of ρ and W_q for the general model with fixed μ and s.

Proposition A1(b) shows that $k^*(\mu, s)$ decreases in μ. We next show that $\rho^*(\mu, s)$ and $W_q^*(\mu, s)$ also decrease in μ. We can rewrite the first-order condition (46) as

$$c\frac{[\rho^*(\mu,s)]^{\sqrt{2(n+1)}}}{1-\rho^*(\mu,s)}\left(\sqrt{2(n+1)}+\frac{\rho^*(\mu,s)}{1-\rho^*(\mu,s)}\right)=\frac{2[k^*(\mu,s)]^2}{K},$$

or equivalently,

$$k^*(\mu,s)=\sqrt{cK\frac{[\rho^*(\mu,s)]^{\sqrt{2(n+1)}}}{2[1-\rho^*(\mu,s)]}\left(\sqrt{2(n+1)}+\frac{\rho^*(\mu,s)}{1-\rho^*(\mu,s)}\right)}.$$

It is clear from the above equation that $\rho^*(\mu, s)$ increases as $k^*(\mu, s)$ increases. Since $k^*(\mu, s)$ decreases in μ, $\rho^*(\mu, s)$ must also decrease in μ.

Using (16), we can also rewrite the first-order condition (46) as

$$c\left[\frac{\mu W_q^*(\mu,s)^*}{d}\right]^2\frac{1}{\rho^*(\mu,s)^{\sqrt{2(n+1)}-2}}\left[(\sqrt{2(n+1)}-1)(1-\rho^*(\mu,s))+1\right]=\frac{2}{K}.$$

This implies that $\frac{\mu}{d}W_q^*(\mu, s)$ increases as $\rho^*(\mu, s)$ increases. Since $\rho^*(\mu, s)$ decreases in μ, $W_q^*(\mu, s)$ must also decrease in μ. Then, the function

$$H(\mu)=c\frac{\rho^*(\mu,s)^{\sqrt{2(n+1)}-1}}{k^*(\mu,s)\mu(1-\rho^*(\mu,s))}\left(\sqrt{2(n+1)}+\frac{\rho^*(\mu,s)}{1-\rho^*(\mu,s)}\right)=c\frac{W_q^*(\mu,s)}{d}\left(\sqrt{2(n+1)}+\frac{\rho^*(\mu,s)}{1-\rho^*(\mu,s)}\right)$$

decreases in μ.

Since (s_0, k_0) is the optimal solution when $\mu = \mu_0$, they satisfy the first-order condition (47):

$$\bar{\lambda}d\left\{(1-2s_0)-c\frac{W_q^*(\mu_0,s_0)}{d}\left(\sqrt{2(n+1)}+\frac{\rho^*(\mu_0,s_0)}{1-\rho^*(\mu_0,s_0)}\right)\right\}=0. \tag{53}$$

Since $H(\mu)$ decreases in μ and $\mu_1 > \mu_0$, it follows from (53) that

$$\bar{\lambda}d\left\{(1-2s_0)-c\frac{W_q^*(\mu_1,s_0)}{d}\left(\sqrt{2(n+1)}+\frac{\rho^*(\mu_1,s_0)}{1-\rho^*(\mu_1,s_0)}\right)\right\}\geq 0.$$

Therefore, the optimal value of s must be greater than s_0 when $\mu = \mu_1$, i.e., $s_1 \geq s_0$. This proves that s^* increases in μ. It then follows immediately from (48) that w^* decreases in μ.

Similarly, we then show that W_q^* decreases in μ. Suppose that $0 < s^* < \frac{1}{4}$. In (49), the left side increases in s and the right side is increasing in ρ. Since s^* increases in μ, therefore, ρ^* must also increase in μ. In (50), the left side decreases in μ, as s^* is increasing in μ. Since ρ^* increases in μ in this case, $\frac{\rho^*}{1-\rho^*}$ must also increase in μ. We can conclude from (50) that W_q^* must be decreasing as μ.

On the other hand, suppose that $\frac{1}{4} \leq s^* \leq \frac{1}{2}$. In this case, the left side of (49) decreases in s. As we have proved that s^* increases in μ, it then follows from (49) that ρ^* must be decreasing in μ in this case.

Using (46) and (16), we obtain

$$c(\frac{\mu {W_q}^*}{d})^2 \frac{1}{\rho^* \sqrt{2(n+1)}-2} \left[(\sqrt{2(n+1)} - 1)(1-\rho^*) + 1 \right] = \frac{2}{K}. \tag{54}$$

The left side of (54) decreases in ρ^* but increases in W_q^*. Since ρ^* increases in μ in this case, we can conclude that W_q^* must be decreasing in μ in this case. Since $\lambda^* = \bar{\lambda} s^*$ increases and W_q^* decreases in μ, it follows from (52) that π^* increases in μ. This proves part (i).

(c) Let (s_0, k_0) and (s_1, k_1) be the optimal values of (s, k) when $c = c_0$ and $c = c_1$, respectively. Suppose that $c_1 > c_0$. Here, we use the notation $k^*(c, s)$ to denote the optimal value of k for the general model with parameter c and fixed service level s such that $k^*(c_0, s_0) = k_0$ and $k^*(c_1, s_1) = k_1$.

Since $c_1 > c_0$, it follows from Proposition A1(c) that $k^*(c_1, s_0) \geq k^*(c_0, s_0) = k_0$. Therefore, $\rho^*(c_1, s_0) = \frac{\bar{\lambda} d}{k^*(c_1, s_0)\mu} \leq \rho^*(c_0, s_0) = \rho_0$. Then,

$$\begin{aligned} c_0 \frac{\rho_0^{\sqrt{2(n+1)}}}{1-\rho_0} \left(\sqrt{2(n+1)} + \frac{\rho_0}{1-\rho_0} \right) &= \frac{2k_0^2}{K} \leq \frac{2k^*(c_1, s_0)^2}{K} \\ &= c_1 \frac{\rho^*(c_1, s_0)^{\sqrt{2(n+1)}}}{1-\rho^*(c_1, s_0)} \left(\sqrt{2(n+1)} + \frac{\rho^*(c_1, s_0)}{1-\rho^*(c_1, s_0)} \right), \end{aligned} \tag{55}$$

where the two equalities come from the first-order condition (46) and the fact that k_0 and $k^*(c_1, s_0)$ are the optimal values of k for the general model with $s = s_0$ when $c = c_0$ and $c = c_1$, respectively.

Since (s_0, k_0) is the optimal solution when $c = c_0$, they satisfy the first-order condition (47):

$$(1 - 2s_0) - c_0 \frac{\rho_0^{\sqrt{2(n+1)}-1}}{k_0 \mu (1-\rho_0)} \left(\sqrt{2(n+1)} + \frac{\rho_0}{1-\rho_0} \right) = 0. \tag{56}$$

Combining (55) and (56), we obtain

$$(1-2s_0) - c_1 \frac{\rho^*(c_1,s_0)^{\sqrt{2(n+1)}-1}}{k^*(c_1,s_0)\mu[1-\rho^*(c_1,s_0)]}\left(\sqrt{2(n+1)} + \frac{\rho^*(c_1,s_0)}{1-\rho^*(c_1,s_0)}\right) \leq 0.$$

Therefore, the optimal value of s must be smaller than s_0 when $c = c_1$, as $\pi(s, k^*(s))$ is concave in s. This proves that s^* is decreasing in c. It then follows from (48) that w^* is increasing in c.

Also, we can use (6) to express $\rho^* = \sqrt{\frac{\bar{\lambda}s^*d}{K\mu^2 w^*}}$, where $\rho^* = \frac{\lambda^* d}{k^*\mu}$ and $\lambda^* = \bar{\lambda}s^*$. Since s^* is decreasing in c and w^* is increasing in c, ρ^* is decreasing in c. From (54) we know that W_q^* is decreasing in c since ρ^* is decreasing in c.

Since both s^* and ρ^* decrease in c, it follows from (50) that cW_q increases in c. Since $\lambda^* = \bar{\lambda}s^*$ decreases in c, it follows from (52) that π^* is decreasing in c. This proves part (ii).

(d) Now let (s_0, k_0) and (s_1, k_1) be the optimal values of (s, k) when $\bar{\lambda} = \bar{\lambda}_0$ and $\bar{\lambda} = \bar{\lambda}_1$, respectively. Suppose that $\bar{\lambda}_1 > \bar{\lambda}_0$. Again, we use the notation $k^*(\bar{\lambda}, s)$ to denote the optimal value of k for the general model with parameter $\bar{\lambda}$ and fixed service level s. In particular, $k^*(\bar{\lambda}_0, s_0) = k_0$ and $k^*(\bar{\lambda}_1, s_1) = k_1$. We also use the notation $\rho^*(\bar{\lambda}, s)$ and $W_q^*(\bar{\lambda}, s)$ denote the corresponding optimal values of ρ and W_q for the general model with fixed $\bar{\lambda}$ and s.

Proposition A1(d) shows that $k^*(\bar{\lambda}, s)$ increases in $\bar{\lambda}$. We can use the same argument as given in part (b) to show that $\rho^*(\bar{\lambda}, s)$ and $W_q^*(\bar{\lambda}, s)$ also increase in $\bar{\lambda}$. This implies that the function

$$H(\bar{\lambda}) = \frac{\rho^*(\bar{\lambda},s)^{\sqrt{2(n+1)}-1}}{k^*(\bar{\lambda},s)\mu(1-\rho^*(\bar{\lambda},s))}\left(\sqrt{2(n+1)} + \frac{\rho^*(\bar{\lambda},s)}{1-\rho^*(\bar{\lambda},s)}\right) = \frac{W_q^*(\bar{\lambda},s)}{d}\left(\sqrt{2(n+1)} + \frac{\rho^*(\bar{\lambda},s)}{1-\rho^*(\bar{\lambda},s)}\right)$$

increases in $\bar{\lambda}$ since $\frac{\rho}{1-\rho}$ is an increasing function in ρ.

Since (s_0, k_0) is the optimal solution when $\bar{\lambda} = \lambda_0$, they satisfy the first-order condition (47):

$$\bar{\lambda}_0 d\left\{(1-2s_0) - c\frac{W_q^*(\bar{\lambda}_0,s_0)}{d}\left(\sqrt{2(n+1)} + \frac{\rho^*(\bar{\lambda}_0,s_0)}{1-\rho^*(\bar{\lambda}_0,s_0)}\right)\right\} = 0. \tag{57}$$

Since $H(\bar{\lambda})$ increases in $\bar{\lambda}$ and $\bar{\lambda}_1 > \bar{\lambda}_0$, it follows from (57) that

$$\bar{\lambda}_1 d\left\{(1-2s_0) - c\frac{W_q^*(\bar{\lambda}_1,s_0)}{d}\left(\sqrt{2(n+1)} + \frac{\rho^*(\bar{\lambda}_1,s_0)}{1-\rho^*(\bar{\lambda}_1,s_0)}\right)\right\} \leq 0.$$

Therefore, the optimal value of s must be smaller than s_0 when $\bar{\lambda} = \bar{\lambda}_1$, i.e., $s_1 \leq s_0$. This proves that s^* decreases in $\bar{\lambda}$. Then, it follows immediately from (48) that w^* increases in $\bar{\lambda}$.

Using (54), ρ^* and W_q^* must change in the same direction as $\bar{\lambda}$ increases. Also we can rewrite (46) as

$$c\frac{\rho\sqrt{2(n+1)}}{(1-\rho)}\left(\sqrt{2(n+1)}+\frac{\rho}{1-\rho}\right)=\frac{2k^2}{K}, \tag{58}$$

Since the left side is increasing in ρ and the right side is increasing in k, ρ^* and k^* must change in the same direction as $\bar{\lambda}$ increases. Thus, we can conclude that k^*, ρ^* and W_q^* must all change in the same direction when $\bar{\lambda}$ increases.

Since s^* decreases in $\bar{\lambda}$, the left side of (50) increases in $\bar{\lambda}$, which implies that the right side of (50) also increases in $\bar{\lambda}$. Since W_q^* and ρ^* must change in the same direction as $\bar{\lambda}$ increases, we can conclude that both W_q^* and ρ^* increases in $\bar{\lambda}$. As k^*, ρ^* and W_q^* all change in the same direction when $\bar{\lambda}$ increases, we must have k^* increases in $\bar{\lambda}$ and that $\lambda^*=\frac{\rho^* k^* \mu}{d}$ increases in $\bar{\lambda}$.

Using (47) and (16), we obtain

$$s=\frac{1}{2}\left\{1-c\frac{\rho\sqrt{2(n+1)}-1}{k\mu(1-\rho)}\left(\sqrt{2(n+1)}+\frac{\rho}{1-\rho}\right)\right\}=\frac{1}{2}\left\{1-\frac{c}{d}W_q\left(\sqrt{2(n+1)}+\frac{\rho}{1-\rho}\right)\right\}.$$

We substitute the above equation into (3) to obtain

$$p=1-\frac{1}{2}\left\{1-\frac{c}{d}W_q\left(\sqrt{2(n+1)}+\frac{\rho}{1-\rho}\right)\right\}-\frac{c}{d}W_q=\frac{1}{2}\left\{1+\frac{c}{d}W_q\left(\sqrt{2(n+1)}+\frac{\rho}{1-\rho}-1\right)\right\}. \tag{59}$$

Since both W_q^* and ρ^* increases in $\bar{\lambda}$, it follows from (59) that p^* increases in $\bar{\lambda}$.

Let π_1^* and π_0^* denote the optimal profit when $\bar{\lambda}=\bar{\lambda}_1$ and $\bar{\lambda}=\bar{\lambda}_0$, respectively. Also, let $\pi^*(\bar{\lambda},s)$ denote the optimal profit for the general model with fixed values of $\bar{\lambda}$ and s. For any $(\bar{\lambda},s)$ with a fixed value of $\lambda=\bar{\lambda}s$, observe from (46) that the optimal values of k remain the same. Furthermore, it follows from (11) and (16) that the corresponding values of W_q^* and w^* are also the same. Using (52), this implies that

$$\pi^*(\bar{\lambda}_1,\frac{\bar{\lambda}_0 s_0}{\bar{\lambda}_1})=\bar{\lambda}_0 s_0 d(1-\frac{\bar{\lambda}_0 s_0}{\bar{\lambda}_1}-\frac{c}{d}W_q^*-w^*)\geq\bar{\lambda}_0 s_0 d(1-s_0-\frac{c}{d}W_q^*-w^*)=\pi^*(\bar{\lambda}_0,s_0),$$

when $\bar{\lambda}_1>\bar{\lambda}_0$. Then,

$$\pi_1^*=\pi^*(\bar{\lambda}_1,s_1)\geq\pi^*(\bar{\lambda}_1,\frac{\bar{\lambda}_0 s_0}{\bar{\lambda}_1})\geq\pi^*(\bar{\lambda}_0,s_0)=\pi_0^*,$$

where the first inequality is due to the fact that (k_1,s_1) is the optimal solution when $\bar{\lambda}=\bar{\lambda}_1$. This proves that π^* increases in $\bar{\lambda}$.

(e) Let (s_0, k_0) and (s_1, k_1) be the optimal values of (s, k) when $d = d_0$ and $d = d_1 > d_0$, respectively. We use the notation $k^*(d, s)$ to denote the optimal value of k for the general model with parameter d and fixed service level s such that $k^*(d_0, s_0) = k_0$ and $k^*(d_1, s_1) = k_1$. We also use the notation $\rho^*(d, s)$ and $W_q^*(d, s)$ to denote the corresponding optimal values of ρ and W_q for the general model with fixed d and s.

Proposition A1(e) shows that $k^*(d, s)$ increases when d increases. We can use the same argument as given in part (b) to show that $\rho^*(d, s)$ and $\frac{W_q^*(d,s)}{d}$ also increase when d increases. Then,

$$H(d) = \frac{\rho^*(d,s)\sqrt{2(n+1)}-1}{k^*(d,s)\mu(1-\rho^*(d,s))}\left(\sqrt{2(n+1)} + \frac{\rho^*(d,s)}{1-\rho^*(d,s)}\right) = \frac{W_q^*(d,s)}{d}\left(\sqrt{2(n+1)} + \frac{\rho^*(d,s)}{1-\rho^*(d,s)}\right)$$

increases in d.

Since (s_0, k_0) is the optimal solution when $d = d_0$, they satisfy the first-order condition (47):

$$\bar{\lambda} d_0 \left\{(1 - 2s_0) - c\frac{W_q^*(d_0, s_0)}{d_0}\left(\sqrt{2(n+1)} + \frac{\rho^*(d_0, s_0)}{1-\rho^*(d_0, s_0)}\right)\right\} = 0. \tag{60}$$

Since $H(d)$ increases in d and $d_1 > d_0$, it follows from (60) that

$$\bar{\lambda} d_1 \left\{(1 - 2s_0) - c\frac{W_q^*(d_1, s_0)}{d_1}\left(\sqrt{2(n+1)} + \frac{\rho^*(d_1, s_0)}{1-\rho^*(d_1, s_0)}\right)\right\} \leq 0.$$

Therefore, the optimal value of s must be smaller than s_0 when $d = d_1$, i.e., $s_1 \leq s_0$. This proves that s^* decreases in d. It then follows immediately from (48) that w^* increases in d.

We can use (54) to deduce that ρ^* and $\frac{W_q^*}{d}$ must change in the same direction when d increases. We can also use (58) to deduce that ρ^* and k^* must change in the same direction when d increases. Thus, we can conclude that k^*, ρ^* and $\frac{W_q^*}{d}$ must all change in the same direction when d increases. Since s^* decreases in d, we can use (50) and the fact that both ρ^* and $\frac{W_q^*}{d}$ must change in the same direction to conclude that both ρ^* and $\frac{W_q^*}{d}$ increase in d, which implies that k^* and W_q^* increase in d. Also, it follows from (59) that p^* increases in d.

Let π_1^* and π_0^* denote the optimal profit when $d = d_1$ and $d = d_0$, respectively. Also, let $\pi^*(d, s)$ denote the optimal profit for the general model with any fixed values of d and s. For any (d, s) with a fixed ratio of ds in the general model, it is easy to check from (46) that the optimal values of k remain the same, and from (11) and (16) that the corresponding values of $\tilde{W}_q^* = \frac{W_q^*}{d}$ and w^* are also the same. Then,

$$\pi^*(d_1, \frac{d_0 s_0}{d_1}) = \bar{\lambda} s_0 d_0 (1 - \frac{d_0 s_0}{d_1} - c\tilde{W}_q^* - w^*) \geq \bar{\lambda} s_0 d_0 (1 - s_0 - c\tilde{W}_q^* - w^*) = \pi^*(d_0, s_0),$$

when $d_1 > d_0$. Then,

$$\pi_1^* = \pi^*(d_1, s_1) \geq \pi^*(d_1, \frac{d_0 s_0}{d_1}) \geq \pi^*(d_0, s_0) = \pi_0^*,$$

where the first inequality is due to the fact that (k_1, s_1) is the optimal solution when $d = d_1$. Therefore, π^* increases in d. This proves part (iii).

Let $\alpha^* = \frac{w^*}{p^*}$. As shown in (52), we can express

$$p^* - w^* = (\frac{1}{\alpha^*} - 1)w^* = \frac{1}{2} - \frac{c}{d}W_q^*. \tag{61}$$

We have shown in the proof of Proposition 4(c) that w^* and cW_q^* increase in c and in Proposition 4(d) and (e) that w^* and $\frac{W_q^*}{d}$ increase in $\bar{\lambda}$ and d. We can then conclude from (61) that α^* is increasing in c, $\bar{\lambda}$ and d. On the other hand, we have shown in the proof of Proposition 4(a) that w^* and W_q^* decrease in K and Proposition 4(b) that w^* and W_q^* decrease in μ. Again, we can conclude from (61) that α^* is decreasing in K and μ. ■

A.5 Proof of Proposition 5

Suppose $\epsilon_1 > \epsilon_0 \geq 1$. Let $K_i = \epsilon_i \hat{K}$ and $\bar{\lambda}_i = \epsilon_i \hat{\lambda}$, where $i = 0, 1$. Let (s_0, k_0) and (s_1, k_1) be the optimal values of (s, k) when $\epsilon = \epsilon_0$ and $\epsilon = \epsilon_1$, respectively. We shall first show that $s_1 \geq s_0$ and $k_1 \geq k_0$, which implies that both s^* and k^* increase in ϵ.

We use the notation $k^*(\epsilon, s)$ and $\rho^*(\epsilon, s)$ to denote the optimal value of k and ρ for the general model with parameter ϵ and a fixed service level s. In particular, $k^*(\epsilon_0, s_0) = k_0$ and $k^*(\epsilon_1, s_1) = k_1$. Since $K_1 > K_0$ and $\bar{\lambda}_1 > \bar{\lambda}_0$, it follows from Proposition A1(a)(d) that $k^*(\epsilon_1, s_0) \geq k^*(\epsilon_0, s_0) = k_0$. We can rewrite (46) as

$$c\frac{\rho^{\sqrt{2(n+1)}-1}}{k\mu(1-\rho)}\left(\sqrt{2(n+1)} + \frac{\rho}{1-\rho}\right) - \frac{2s\bar{\lambda}d}{K\mu^2\rho^2} = 0. \tag{62}$$

It is clear that the left side of (62) increases in ρ but decreases in k. Note that $\frac{\bar{\lambda}_1}{K_1} = \frac{\bar{\lambda}_0}{K_0} = \frac{\bar{\lambda}}{K}$. Since $k^*(\epsilon_1, s_0) \geq k^*(\epsilon_0, s_0)$, we must have $\rho^*(\epsilon_1, s_0) \geq \rho^*(\epsilon_0, s_0)$.

We can use (62) to rewrite (47) as

$$\frac{\partial \pi}{\partial s} = \bar{\lambda}d\left\{(1-2s) - \frac{2s\bar{\lambda}d}{K\mu^2\rho^2}\right\} = 0.$$

It is then clear that $\frac{\partial \pi}{\partial s}$ increases in ρ. Since (s_0, k_0) satisfies the first-order condition $\frac{\partial \pi}{\partial s} = 0$ and $\rho^*(\epsilon_1, s_0) \geq \rho^*(\epsilon_0, s_0)$, we must have

$$\bar{\lambda}d\left\{(1-2s_0) - \frac{2s_0\bar{\lambda}d}{K\mu^2[\rho^*(\epsilon_1, s_0))]^2}\right\} \geq 0.$$

Therefore, the optimal value of s must be greater than s_0 when $\epsilon = \epsilon_1$, which proves that s^* increases in ϵ. Then, it follows immediately from (48) that w^* decreases in ϵ.

We next prove that ρ^* increases in ϵ by contradiction. Suppose that ρ^* decreases in ϵ. Since $w^* = \frac{(k^*)^2}{\epsilon \hat{K} \lambda^* d} = \frac{k^*}{\epsilon \hat{K} \mu \rho^*}$ decreases in ϵ, $\frac{k^*}{\epsilon}$ must be decreasing in ϵ. On the other hand, as $\rho^* = \frac{\epsilon \hat{\lambda} d s^*}{k^* \mu}$ decreases in ϵ and s^* increases in ϵ, $\frac{\epsilon}{k^*}$ must be decreasing in ϵ. This contradicts with the fact that $\frac{k^*}{\epsilon}$ must be decreasing in ϵ. Therefore, ρ^* increases in ϵ.

Since s^* increases in ϵ, the left side of (50) decreases in ϵ. On the other hand, we have shown that ρ^* increases in ϵ, which implies that $\frac{\rho^*}{1-\rho^*}$ increases in ϵ. Therefore, we can conclude from (50) that W_q^* must be decreasing as ϵ increases. Since both w^* and W_q^* decrease in ϵ, we can then conclude from (61) that $\alpha^* = \frac{w^*}{p^*}$ decreases in ϵ. ■

A.6 Proof of Proposition 6

(i) Let us first adapt the proof of Proposition A1 to establish the same results to this extension. To establish the result of Proposition A1(a), let k_0 denotes the optimal value of k for $K = K_0$. Under Assumption 1, the first-order condition for $\Pi(k)$ now becomes

$$\Pi'(k_0) = -(1-\gamma)\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - (2-3\gamma)\frac{k_0}{K_0} = 0, \tag{63}$$

and we can show that, for any $K_1 > K_0$,

$$-(1-\gamma)\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - (2-3\gamma)\frac{k_0}{K_1} > 0.$$

Therefore, the optimal value of k must be greater than k_0 for any fixed $K_1 > K_0$, which implies that k^* is increasing in K. Using the same argument as before, we can show that p^* is increasing in K.

Let $\beta_0 = \frac{k_0}{K_0}$, we can rewrite the first-order condition (63) as

$$\pi'(k_0) = -(1-\gamma)\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_0} - (2-3\gamma)\beta_0 = 0, \tag{64}$$

Let k_1 denote the optimal value of k when $K = K_1 > K_0$, and define $\beta_1 = \frac{k_1}{K_1}$. Then,

$$\pi'(k_1) = -(1-\gamma)\lambda c \frac{\partial W_q(\lambda, k, \mu, d)}{\partial k}|_{k=k_1} - (2-3\gamma)\beta_1 = 0. \tag{65}$$

As k^* is increasing in K, we have $k_1 > k_0$. Therefore, $-\lambda c \frac{\partial W_q(\lambda,k,\mu,d)}{\partial k}|_{k=k_1} < -\lambda c \frac{\partial W_q(\lambda,k,\mu,d)}{\partial k}|_{k=k_0}$. Since $(2-3\gamma) > 0$, we can conclude that $\beta_1 < \beta_0$. Thus, β^* is decreasing in K. This proves Proposition A1(a).

The results for Proposition A1(b),(c),(d) and (e) can be proved using the same arguments from the proof of Proposition A1.

We next adapt the proofs of Propositions 4 and 5 to establish all the corresponding results to this extension. To illustrate the adaptation, we next outline the proof of the results of Proposition 4(a). We can use similar arguments to prove the rest of the results, but omit the details here.

The two first-order conditions (46) and (47) given in the proof of Proposition 4 now become

$$\frac{\partial \pi}{\partial k} = (1-\gamma)c\mu \frac{\rho^{\sqrt{2(n+1)}}}{k\mu(1-\rho)}\left(\sqrt{2(n+1)}+\frac{\rho}{1-\rho}\right) - \frac{(2-3\gamma)k}{K} = 0 \quad (66)$$

$$\frac{\partial \pi}{\partial s} = \bar{\lambda} d\left\{[(1-\gamma)-(2-3\gamma)s] - (1-\gamma)c\frac{\rho^{\sqrt{2(n+1)}-1}}{k\mu(1-\rho)}\left(\sqrt{2(n+1)}+\frac{\rho}{1-\rho}\right)\right\} = 0. \quad (67)$$

Let (s_0, k_0) and (s_1, k_1) be the optimal values of (s, k) when $K = K_0$ and $K = K_1$, respectively. Suppose that $K_1 > K_0$. We use the notation $k^*(K, s)$ to denote the optimal value of k for the general model with parameter K and fixed service level s. Then, $k^*(K_1, s_0) > k^*(K_0, s_0)$ in view of Proposition A1(a). Furthermore, we can use the first-order condition to establish that

$$[(1-\gamma)-(2-3\gamma)s_0] - (1-\gamma)c\frac{\left[\frac{\bar{\lambda}s_0 d}{k^*(K_1,s_0)\mu}\right]^{\sqrt{2(n+1)}}}{k^*(K_1,s_0)\mu\left[1-\frac{\bar{\lambda}s_0 d}{k^*(K_1,s_0)\mu}\right]}\left(\sqrt{2(n+1)}+\frac{\frac{\bar{\lambda}s_0 d}{k^*(K_1,s_0)\mu}}{1-\frac{\bar{\lambda}s_0 d}{k^*(K_1,s_0)\mu}}\right) \geq 0.$$

Therefore, the optimal value of s must be greater than s_0 when $K = K_1$, as $\pi(s, k^*(s))$ is concave in s. Since (s_1, k_1) is optimal at $K = K_1$, this proves that $s_1 \geq s_0$. Also, it follows from Proposition A1(d) that $k_1 = k^*(K_1, s_1) \geq k^*(K_1, s_0) \geq k_0$. Therefore, we prove that both s^* and k^* increase in K.

Using (66) and (67), we can obtain

$$s + w = \frac{1-\gamma}{2-3\gamma}.$$

It then follows that w^* decreases in K, as s^* increases in K.

We next show that W_q^* decreases in K. We can rewrite (67) as

$$\bar{\lambda} s d[(1-\gamma)-(2-3\gamma)s] = (1-\gamma)\left[c\frac{\rho^{\sqrt{2(n+1)}}}{1-\rho}\left(\sqrt{2(n+1)}+\frac{\rho}{1-\rho}\right)\right]. \quad (68)$$

Clearly, the right side of (68) is increasing in ρ. Also, the left side of (68) implies that $0 < s^* \leq \frac{1-\gamma}{2-3\gamma}$. Suppose that $0 < s^* < \frac{1-\gamma}{2(2-3\gamma)}$. In this case, the left side of (68) increases in s. Since s^* increases in

K as proved earlier, we can conclude that ρ^* must also increase in K in this case. The first-order condition (67) implies that

$$[(1-\gamma)-(2-3\gamma)s^*] = (1-\gamma)\frac{c}{d}W_q^*\left(\sqrt{2(n+1)}+\frac{\rho^*}{1-\rho^*}\right). \tag{69}$$

Since s^* increases in K, the left side of (69) must be decreasing as K increases. On the other hand, we have shown that ρ^* increases in K in this case, which implies that $\frac{\rho^*}{1-\rho^*}$ must be increasing in K in this case. We can conclude from (69) that W_q^* must be decreasing as K increases in this case.

Now suppose that $\frac{1-\gamma}{2(2-3\gamma)} \le s^* \le \frac{1-\gamma}{2-3\gamma}$. In this case, the left side of (68) decreases in s. Since s^* increases in K, we can conclude from (68) that ρ^* must be decreasing in K in this case. Also, it follows from (16) that

$$\frac{W_q^*}{d} = \frac{\rho^*\sqrt{2(n+1)}-1}{k^*\mu(1-\rho^*)}. \tag{70}$$

As K increases, k^* increases and ρ^* decreases. Since the right side of (70) decreases in k^* but increases in ρ^*, we can conclude from (70) that $\frac{W_q^*}{d}$ (or equivalently W_q^*) must also be decreasing in K in this case.

Furthermore, we can show that

$$\pi^* = \lambda^* d(p^*-w^*) = \lambda^* d(\frac{1-2\gamma}{2-3\gamma} - \frac{c}{d}W_q^*). \tag{71}$$

Since λ^* increases in K and W_q^* decreases in K, we can conclude that π^* increases in K. Also,

$$p^*-w^* = (\frac{1}{\alpha^*}-1)w^* = \frac{1-2\gamma}{2-3\gamma} - \frac{c}{d}W_q^*. \tag{72}$$

Since w^* and W_q^* decrease in K, α^* is also decreasing in K.

(ii) We first consider the case where s is fixed and show that the optimal k^* is increasing in γ. Let k_0 denotes the optimal value of k when $\gamma = \gamma_0 \ge 0$. Then,

$$\Pi'(k_0) = (1-\gamma_0)\left[-\lambda c\frac{\partial W_q(\lambda,k,\mu,d)}{\partial k}|_{k=k_0} - \left\{G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K}\right\} + \frac{\gamma_0}{1-\gamma_0}\frac{k_0}{K}G'^{-1}(\frac{k_0}{K})\right] = 0.$$

For any $\gamma_1 > \gamma_0$, we must have $\frac{\gamma_1}{1-\gamma_1}\frac{k_0}{K}G'^{-1}(\frac{k_0}{K}) > \frac{\gamma_0}{1-\gamma_0}\frac{k_0}{K}G'^{-1}(\frac{k_0}{K})$. Therefore,

$$(1-\gamma_1)\left[-\lambda c\frac{\partial W_q(\lambda,k,\mu,d)}{\partial k}|_{k=k_0} - \left\{G^{-1}(\frac{k_0}{K}) + G'^{-1}(\frac{k_0}{K})\frac{k_0}{K}\right\} + \frac{\gamma_1}{1-\gamma_1}\frac{k_0}{K}G'^{-1}(\frac{k_0}{K})\right] > 0.$$

Since $\Pi(k)$ is concave in k, the optimal k^* must be greater than k_0 when $\gamma = \gamma_1 > \gamma_0$.

Now consider the joint optimization problem of (s,k). Let (s_0,k_0) and (s_1,k_1) be the optimal values of (s,k) when $\gamma=\gamma_0$ and $\gamma=\gamma_1$, respectively. Suppose that $\gamma_1>\gamma_0$. The two first-order conditions are given by

$$\frac{\partial\pi}{\partial k}|_{s=s_0,k=k_0}=-(1-\gamma_0)c\lambda d^2\frac{\partial W_q(\lambda,k,\mu,d)}{\partial k}|_{\lambda=s_0\bar{\lambda},k=k_0}-\left[(1-\gamma_0)G^{-1}(\frac{k_0}{K_0})+(1-2\gamma_0)G'^{-1}(\frac{k_0}{K_0})\frac{k_0}{K_0}\right]=0$$

$$\frac{\partial\pi}{\partial s}|_{s=s_0,k=k_0}=\bar{\lambda}\Big\{d\left[(1-\gamma_0)F^{-1}(1-s_0)-(1-2\gamma_0)s_0F'^{-1}(1-s_0)\right]-(1-\gamma_0)cW_q(s_0\bar{\lambda},k_0,\mu,d)$$
$$-(1-\gamma_0)cs_0\bar{\lambda}\frac{\partial W_q(\lambda,k,\mu,d)}{\partial\lambda}|_{\lambda=\bar{\lambda}s_0,k=k_0}\Big\}=0$$

Let $k^*(\gamma,s)$ be the optimal value of k with fixed values of γ and s. As we have shown that the optimal k^* is increasing in γ for fixed s, we have $k^*(\gamma_1,s_0)\geq k^*(\gamma_0,s_0)$. As both $\frac{\partial W_q(s\bar{\lambda},k,\mu,d)}{\partial\lambda}$ and $W_q(s_0\bar{\lambda},k,\mu,d)$ decrease in k, we have

$$d\left[(1-\gamma_1)F^{-1}(1-s_0)-(1-2\gamma_1)s_0F'^{-1}(1-s_0)\right]-(1-\gamma_1)cW_q(s_0\bar{\lambda},k^*(\gamma_1,s_0),\mu,d)$$
$$-(1-\gamma_1)cs_0\bar{\lambda}\frac{\partial W_q(\lambda,k,\mu,d)}{\partial\lambda}|_{\lambda=\bar{\lambda}s_0,k=k^*(\gamma_1,s_0)}\geq 0.$$

Therefore, the optimal value of s must be greater than s_0 when $\gamma=\gamma_1$, i.e., $s_1\geq s_0$. Also, $k_1=k^*(\gamma_1,s_1)\geq k^*(\gamma_1,s_0)\geq k^*(\gamma_0,s_0)=k_0$. Therefore, both s^* and k^* are increasing in γ.

We can use (66) and (67) to cancel out γ and obtain

$$k=c\left(1+\rho^2\frac{K\mu^2}{\bar{\lambda}d}\right)\frac{\rho^{\sqrt{2(n+1)}-1}}{\mu(1-\rho)}\left(\sqrt{2(n+1)}+\frac{\rho}{1-\rho}\right). \tag{73}$$

Since k^* is increasing in γ and the right side of (73) is increasing in ρ, we can conclude that ρ^* is increasing in γ. Also, we can rewrite (73) as

$$c\left(1+\rho^2\frac{K\mu^2}{\bar{\lambda}d}\right)\left(\sqrt{2(n+1)}+\frac{\rho}{1-\rho}\right)\frac{W_q}{d}=1.$$

Since ρ^* is increasing in γ, we can conclude that W_q^* is decreasing in γ.

Finally, we can use (11) and (66) to obtain

$$c\frac{\rho^{\sqrt{2(n+1)}-2}}{\mu^2(1-\rho)}\left(\sqrt{2(n+1)}+\frac{\rho}{1-\rho}\right)=\frac{2-3\gamma}{1-\gamma}Kw^2. \tag{74}$$

Since ρ^* is increasing in γ, the left side of (74) is increasing in γ. Since $\frac{2-3\gamma}{1-\gamma}$ is decreasing in γ, we can conclude that w^* is increasing in γ. ■

B Some Detailed Statistics from Didi Data

Table 12: Mean and standard deviation of number of drivers, number of rides, and price per km across differrent hours over weekdays.

	Number of drivers		Number of rides		Price per km	
Hour	Mean	Std.Dev.	Mean	Std. Dev.	Mean	Std. Dev.
8	879	26	1273	44	2.919	1.196
9	1196	24	2229	37	3.056	1.264
10	1099	31	1725	64	3.028	1.231
11	771	34	1170	69	2.953	1.191
12	707	40	1208	75	2.959	1.229
13	676	33	1210	88	2.915	1.221
14	699	20	1251	44	2.938	1.232
15	691	21	1173	19	2.994	1.240
16	738	79	1177	108	3.025	1.250
17	904	80	1426	138	3.066	1.255
18	1080	42	1713	93	3.134	1.275
19	1211	59	2006	60	3.214	1.319
20	992	56	1550	91	3.039	1.250
21	811	64	1452	97	2.855	1.183
22	752	67	1406	127	2.817	1.197
23	597	64	1029	99	2.799	1.213
24	348	53	552	83	2.727	1.124

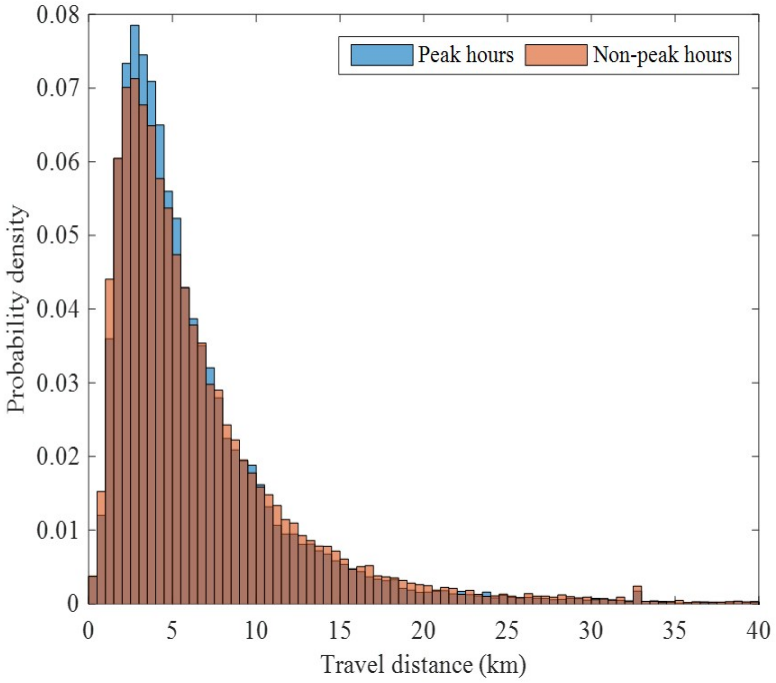

Figure 7: Travel distance distribution across all hours.

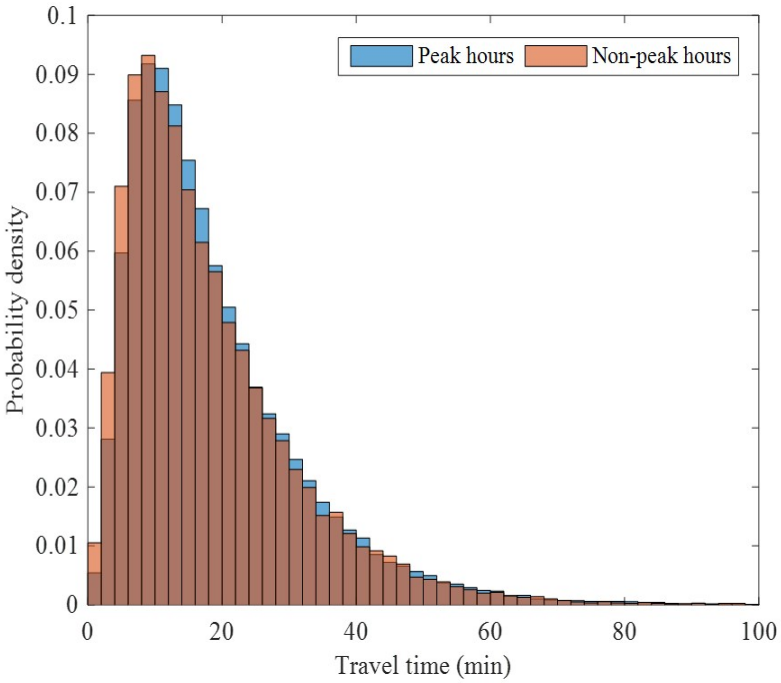

Figure 8: Travel time distribution across all hours.

Table 1: Comparisons of results for the base model with exact formula (9) and approximation (16).

	W_q is given by exact formula (9)				W_q is given by (16) with $n = n^*$				
$\bar{\lambda}$	k^*	λ^*	p^*	π^*	n^*	k^*	λ^*	p^*	π^*
10	7	2.71	0.72	0.98	7.48	7	2.76	0.73	0.98
20	10	5.79	0.69	2.00	10.02	10	6.22	0.64	2.00
30	11	6.20	0.78	2.42	11.57	11	6.34	0.76	2.42
40	12	7.14	0.81	2.88	12.64	12	7.29	0.79	2.88
50	13	8.32	0.81	3.38	13.41	13	8.53	0.79	3.38
60	14	9.80	0.80	3.92	13.99	13	8.05	0.84	3.38
70	14	9.29	0.84	3.92	14.45	14	9.50	0.83	3.92
80	15	11.16	0.81	4.50	14.82	14	9.18	0.85	3.92
90	15	10.62	0.85	4.50	15.13	15	10.97	0.82	4.50
100	15	10.36	0.87	4.50	15.39	15	10.60	0.85	4.50

Table 2: Performance of the approximation scheme for the base model.

	k^*	λ^*	p^*	w^*	π^*
k^* (5940 cases)	2%	6%	6%	6%	2%

Table 3: Comparisons of results for the general model with exact formula (9) and approximation (16).

	W_q is given by exact formula (9)					W_q is given by (16) with $n = n^*$					
$\bar{\lambda}$	k^*	λ^*	p^*	w^*	π^*	n^*	k^*	λ^*	p^*	w^*	π^*
10	6	3.32	0.613	0.217	1.32	5.48	6	3.28	0.603	0.220	1.25
20	8	5.14	0.677	0.249	2.20	7.90	8	5.11	0.663	0.259	2.11
30	10	6.87	0.706	0.291	2.85	9.61	10	6.86	0.692	0.292	2.75
40	12	8.61	0.723	0.335	3.34	10.87	11	7.82	0.722	0.310	3.22
50	13	9.55	0.745	0.354	3.73	11.90	12	8.74	0.742	0.330	3.60
60	14	10.47	0.761	0.375	4.04	12.70	13	9.65	0.756	0.350	3.92
70	14	10.55	0.780	0.372	4.31	13.38	14	10.55	0.767	0.371	4.18
80	15	11.44	0.789	0.393	4.53	13.91	14	10.61	0.782	0.369	4.38
90	15	11.49	0.802	0.392	4.71	14.43	15	11.50	0.789	0.391	4.58
100	16	12.39	0.807	0.413	4.88	14.84	15	11.55	0.799	0.390	4.73

Table 4: Ratio of expected profits between using a fixed payout ratio and using the optimal time-based payout ratio.

		$\alpha =$							
$\bar{\lambda}$	α^*	.2	.3	.4	.5	.6	.7	.8	.9
10	.35	.55	.89	.82	.74	.65	.53	.31	.17
20	.37	.58	.76	.87	.91	.73	.56	.38	.20
30	.41	.45	.80	.85	.85	.79	.68	.45	.23
40	.46	.38	.68	.90	.86	.78	.66	.48	.27
50	.48	.34	.80	.97	.91	.80	.74	.54	.26
60	.49	.49	.74	.90	.97	.84	.77	.55	.29
70	.48	.46	.69	.84	.91	.89	.72	.56	.30
80	.50	.44	.66	.80	.99	.85	.76	.58	.31
90	.49	.42	.63	.92	.95	.92	.80	.56	.32
100	.51	.41	.61	.89	.92	.89	.78	.59	.31

Table 5: Values of the optimal time-based payout ratio α^*.

	$K =$									
$\bar{\lambda}$	10	20	30	40	50	60	70	80	90	100
10	.68	.56	.47	.35	.35	.29	.31	.28	.24	.22
20	.78	.57	.45	.46	.37	.35	.35	.30	.31	.28
30	.75	.62	.54	.46	.41	.38	.37	.36	.32	.31
40	.74	.59	.51	.48	.46	.42	.40	.38	.36	.33
50	.73	.58	.55	.50	.48	.43	.40	.40	.39	.35
60	.72	.57	.53	.52	.49	.44	.44	.41	.39	.37
70	.72	.63	.57	.51	.48	.46	.45	.41	.41	.39
80	.72	.63	.56	.54	.50	.47	.46	.42	.42	.40
90	.71	.62	.56	.53	.49	.49	.47	.43	.43	.40
100	.71	.62	.55	.52	.51	.48	.48	.45	.44	.41

Table 6: Performance of the approximation scheme for the general model.

	k^*	λ^*	p^*	w^*	π^*
$k^* \leq 10$ (4495 cases)	11%	4%	2%	20%	3%
$k^* > 10$ (2105 cases)	2%	1%	1%	3%	1%

Table 7: Impact of model parameters on s^*, k^*, W_q^*, λ^* and ρ^*.

	s^*	k^*	W_q^*	λ^*	ρ^*
K	$\uparrow$	$\uparrow$	$\downarrow$	$\uparrow$	$\times$
μ	$\uparrow$	$\times$	$\downarrow$	$\uparrow$	$\times$
c	$\downarrow$	$\times$	$\downarrow$	$\downarrow$	$\downarrow$
$\bar{\lambda}$	$\downarrow$	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$
d	$\downarrow$	$\uparrow$	$\uparrow$	$\downarrow$	$\uparrow$

$\uparrow$(increasing); $\downarrow$(decreasing); $\times$(non-monotonic)

Table 8: Optimal values of α^* using the approximation formula (16) with $n = n^*$.

	$K =$									
$\bar{\lambda}$	10	20	30	40	50	60	70	80	90	100
10	.67	.49	.42	.36	.33	.30	.27	.25	.24	.22
20	.67	.52	.46	.41	.37	.34	.32	.30	.28	.27
30	.67	.54	.48	.44	.40	.38	.35	.33	.31	.30
40	.67	.55	.49	.46	.42	.40	.38	.36	.34	.32
50	.67	.56	.51	.47	.44	.42	.39	.38	.36	.34
60	.68	.56	.51	.48	.45	.43	.41	.39	.38	.36
70	.68	.57	.52	.49	.46	.44	.42	.40	.39	.37
80	.68	.57	.53	.49	.47	.45	.43	.41	.40	.39
90	.68	.57	.53	.50	.48	.46	.44	.42	.41	.40
100	.68	.57	.53	.50	.48	.46	.45	.43	.42	.41

Table 9: Impact of growth rate ϵ on p^*, w^*, α^* and π^*.

ϵ	p^*	w^*	α^*	π^*
1	0.64	0.43	0.67	0.15
2	0.71	0.37	0.52	0.86
3	0.73	0.35	0.48	1.71
4	0.74	0.34	0.46	2.63
5	0.74	0.33	0.44	3.60

Table 10: Numerical results for the general model to include the total consumer and provider surplus.

γ	p^*	w^*	α^*	π^*	$C_s^* + P_s^*$	$\Pi^* = (1-\gamma)\pi^* + \gamma(C_s^* + P_s^*)$
0.0	0.81	0.41	0.51	4.88	3.33	4.88
0.1	0.80	0.43	0.54	4.84	3.77	4.73
0.2	0.79	0.46	0.58	4.75	4.25	4.65
0.3	0.78	0.50	0.64	4.42	5.28	4.68
0.4	0.75	0.59	0.79	3.11	7.67	4.93
0.5	0.68	0.76	1.12	-2.14	13.84	5.85
0.6	0.52	1.15	2.19	-27.17	34.52	9.84

Table 11: Summary of parameter values for our illustrative examples.

Parameters	Peak hour	Non-peak hour	Data source
K	390	390	Didi data with assumption of 20 equal zones
$\bar{\lambda}$	200 /hour	100 /hour	Didi data with assumption of 20 equal zones
d	6 km	6 km	Didi data
μ	19 km/hour	26 km/hour	Didi data
v	U[2,4] RMB/km	U[2,4] RMB/km	Benchmarked against taxi rate
r	U[30,40] RMB/hour	U[30,40] RMB/hour	Estimated from taxi driver wages
c	0 to 1,000 RMB/hour	0 to 1,000 RMB/hour	Assumption for sensitivity analysis

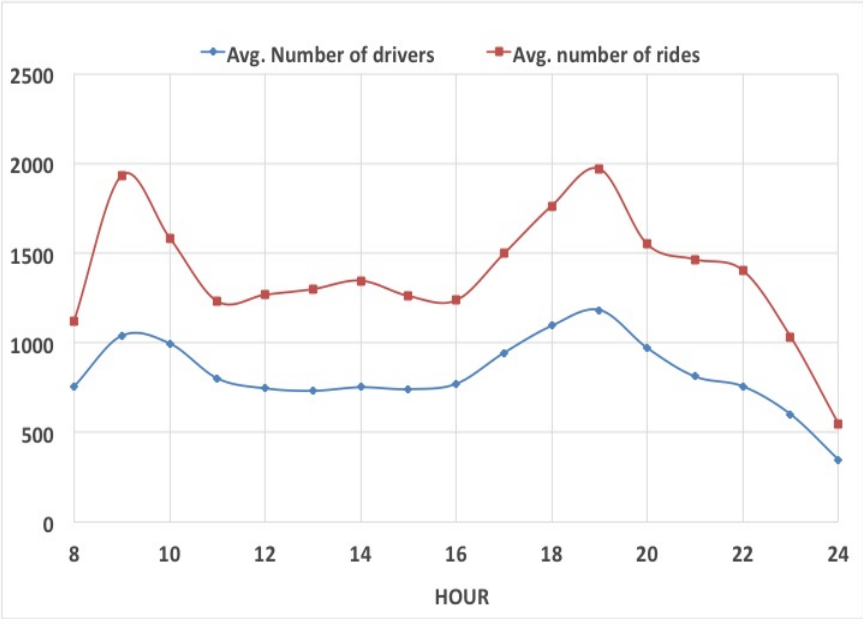

Figure 1: Number of rides and drivers across different hours.

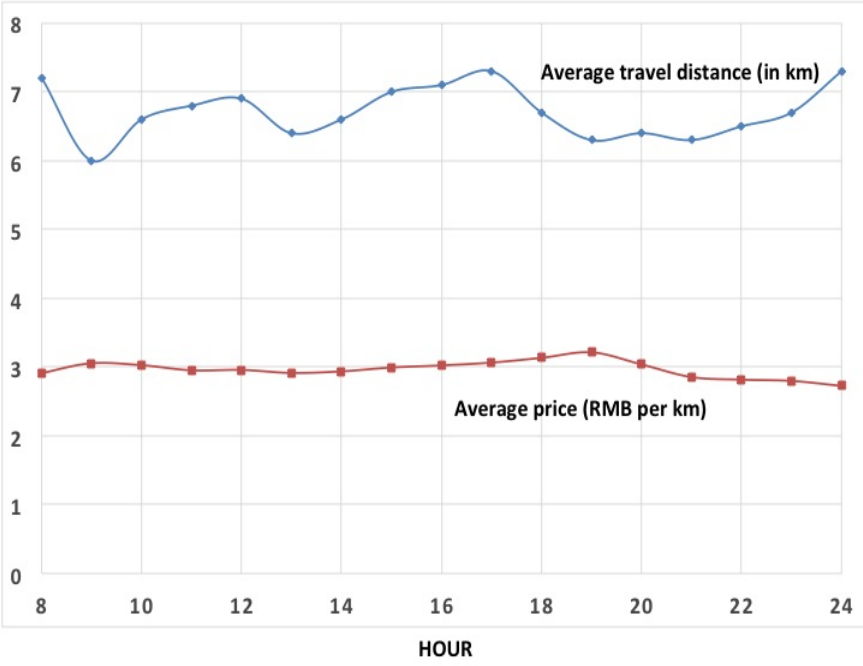

Figure 2: Average travel distance and average price per kilometer across different hours.

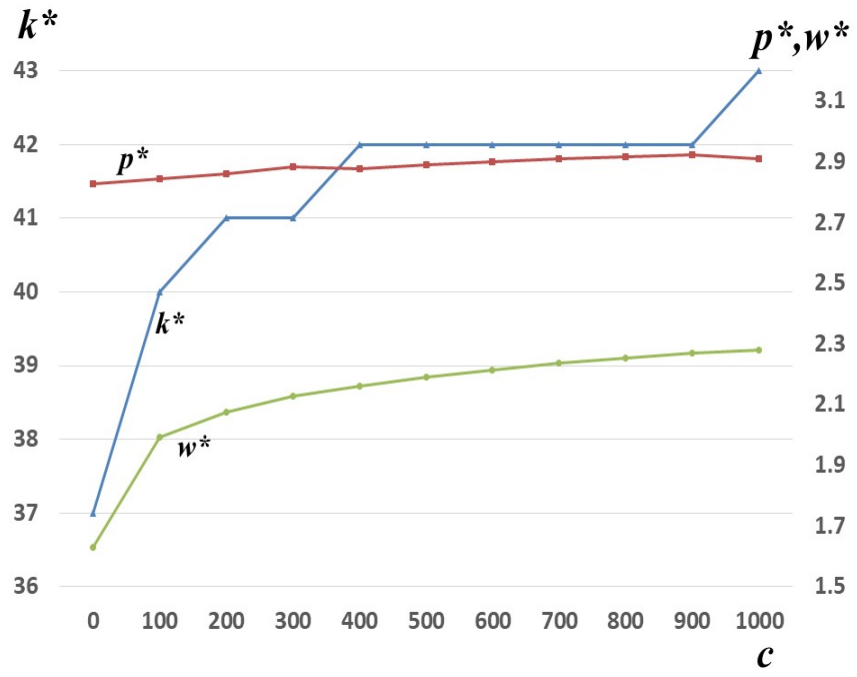

Figure 3: Optimal number of participating drivers, optimal price and wage rates during peak hours ($\bar{\lambda} = 200$ and $\mu = 19$ km/hour).

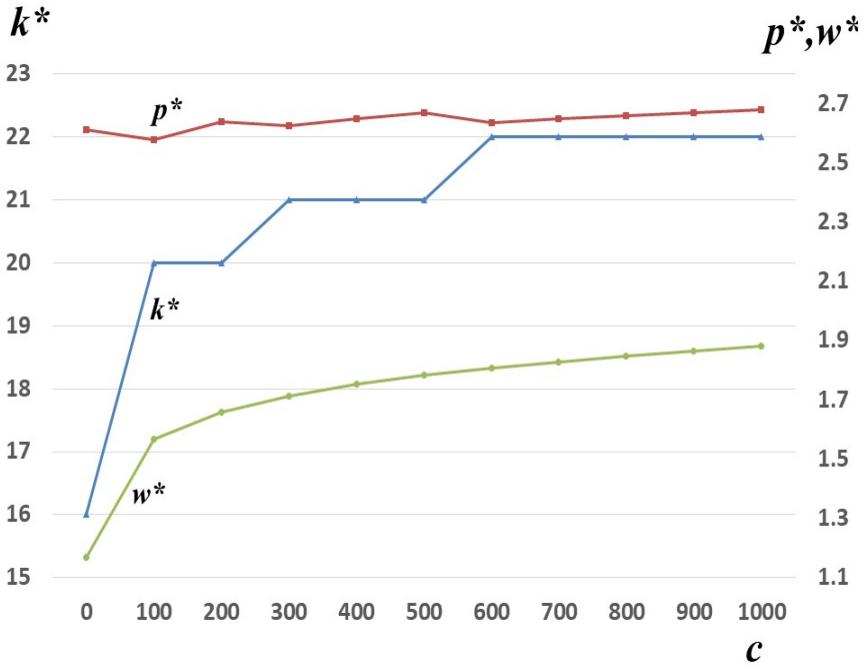

Figure 4: Optimal number of participating drivers, optimal price and wage rates during non-peak hours ($\bar{\lambda} = 100$ and $\mu = 26$ km/hour).

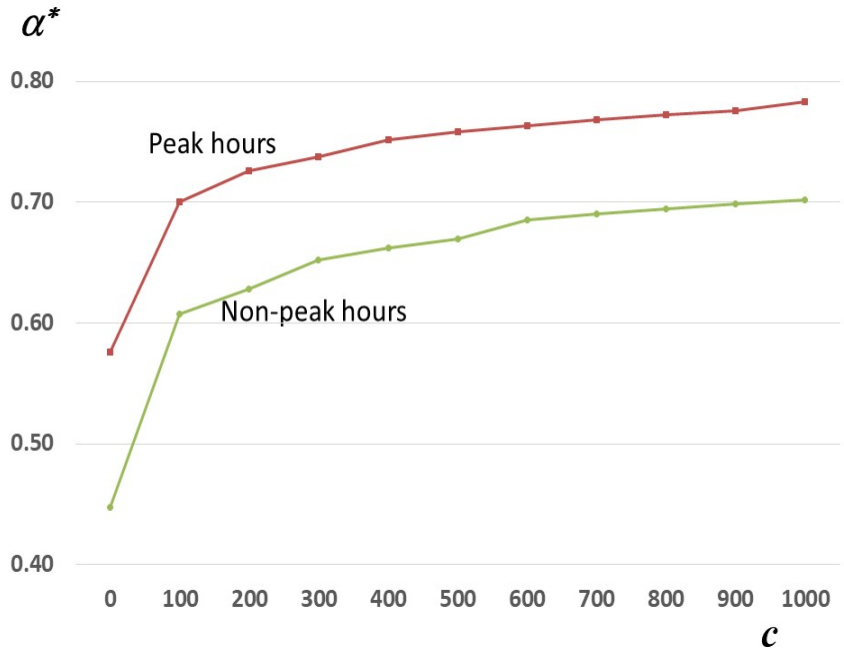

Figure 5: Comparisons of the optimal time-based payout ratio between peak and non-peak hours.

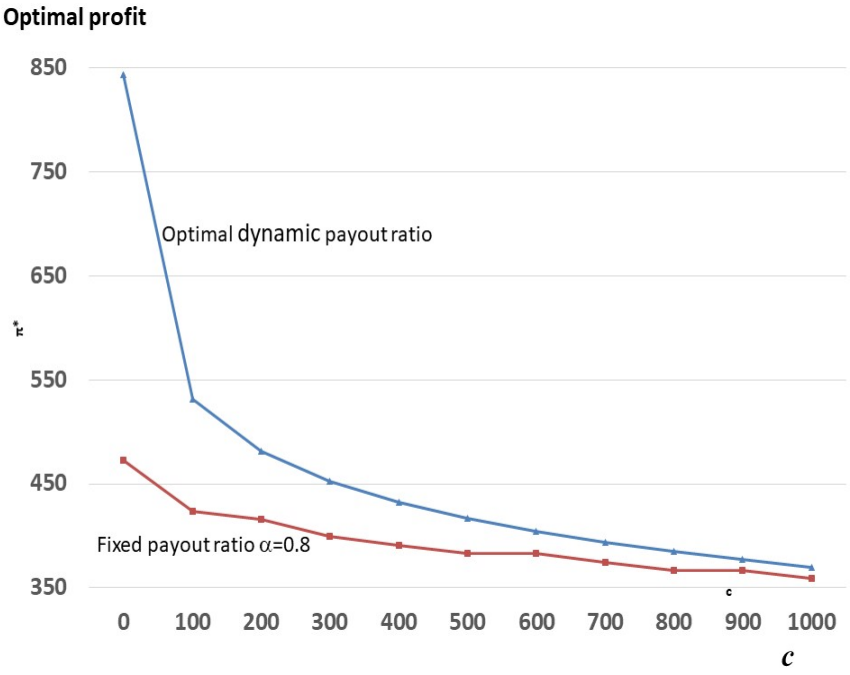

Figure 6: Comparisons of optimal profit between the optimal time-based payout ratio and a fixed payout ratio for the peak hour scenario.