https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 70, No. 6, June 2024, pp. 3789–3807
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

# Courier Dispatch in On-Demand Delivery

**Mingliu Chen,**[a,*] **Ming Hu**[b]

[a] Naveen Jindal School of Management, The University of Texas at Dallas, Richardson, Texas 75074; [b] Rotman School of Management, University of Toronto, Toronto, Ontario M5S 1A1, Canada

*Corresponding author

**Contact:** mingliu.chen@utdallas.edu, https://orcid.org/0000-0001-7218-9269 (MC); ming.hu@rotman.utoronto.ca, https://orcid.org/0000-0003-0900-7631 (MH)

**Received:** January 9, 2021
**Revised:** October 14, 2021; August 1, 2022
**Accepted:** October 5, 2022
**Published Online in Articles in Advance:** July 21, 2023

**https://doi.org/10.1287/mnsc.2023.4858**



**Abstract.** We study a courier dispatching problem in an on-demand delivery system in which customers are sensitive to delay. Specifically, we evaluate the effect of temporal pooling by comparing systems using the dedicated strategy, with which only one order is delivered per trip, versus the pooling strategy, with which a batch of consecutive orders is delivered on each trip. We capture the courier delivery system's spatial dimension by assuming that, following a Poisson process, demand arises at a uniformly generated point within a service region. With the same objective of revenue maximization, we find that the dispatching strategy depends critically on customers' patience level, the size of the service region, and whether the firm can endogenize the demand. We obtain concise but informative results with a single courier and assuming that customers' underlying arrival rate is large enough, meaning a crowded market, such as rush hour delivery. In particular, when the firm has a growth target and needs to achieve an exogenously given demand rate, using the pooling strategy is optimal if the service area is large enough to fully exploit the pooling efficiency in delivery. Otherwise, using the dedicated strategy is optimal. In contrast, if the firm can endogenize the demand rate by varying the delivery fee, using the dedicated strategy is optimal for a large service area. The reason is that it is optimal for the firm to sustain a relatively low demand rate by charging a high fee for a large service radius: within this large area, the pooling strategy leads to a long wait because it takes a long time for multiple orders to accumulate. Moreover, with an exogenous demand rate to meet, customers' patience level has no impact on the dispatch strategy. However, when the demand rate can be endogenized, the dedicated strategy is preferable if customers are impatient. Furthermore, we extend our model to account for social welfare maximization, a hybrid contingent delivery policy, a general arrival rate that does not have to be large, a nonuniform distribution of orders in the service region, and multiple couriers. We also conduct numerical analysis and simulations to complement our main results and find that most insights in our base model still hold in these extensions and numerical studies.

**History:** Accepted by Jeannette Song, operations management.
**Funding:** This work was supported by the Natural Sciences and Engineering Research Council of Canada [Grants RGPIN-2015-06757 and RGPIN-2021-04295].
**Supplemental Material:** The online appendix and data are available at https://doi.org/10.1287/mnsc.2023.4858.

**Keywords:** on-demand economy • smart city • food delivery • spatial queueing • pricing • temporal pooling • sharing economy

## 1. Introduction

On-demand delivery of food and groceries has gained traction nowadays. Given the prevalence of smart devices and a flexible labor force of independent contractors, many food and grocery stores have started on-demand delivery for relatively small orders. For example, Starbucks plans to expand its coffee delivery services across the United States and has already established delivery services in China in 30 cities and more than 2,000 stores (Jargon 2018). Unlike traditional package delivery services, coffee delivery involves spontaneous orders for small quantities. Typically, customers who order consumables such as coffee do not order in advance and expect the coffee to still be hot on arrival. A customer may choose not to order if the expected delivery time is too long.

In hyper-fast (or so-called instant) delivery, companies offer a wait time expectation coupled with a price tag, for example, 10-minute grocery delivery for $2 by Gorillas and 30-minute grocery and food delivery for $1.95 by Gopuff with additional markups on product prices. Companies such as Gorillas and Gopuff employ and staff couriers dedicated to multihour shifts, fulfilling orders from "dark" warehouses or microfulfillment centers to meet the promise of rapid delivery. The Covid-19 pandemic has solidified this trend. Many more

vendors are hiring dedicated couriers for delivery.[1] According to Rana and Haddon (2021b), about half of the 150 registered restaurants on Spread, a start-up delivery platform, hire dedicated drivers for their deliveries. Furthermore, they set much lower delivery prices to carve out a market share to compete with large platforms. Haddon (2021) reports that Domino's has established market penetration by using dedicated drivers and offering cheaper than market price pies. During the pandemic, Domino's market share increased by 31%.

Because on-demand deliveries are sensitive to delay, many delivery systems dispatch a courier whenever an order arrives. Thus, the couriers can serve only one order per trip in the hope of reducing delivery time for each customer. The empirical analysis of Mao et al. (2022) shows that delivery delay significantly reduces future orders. However, there are still many occasions when a firm can utilize batch delivery if multiple orders are placed around the same time in the same area. A courier may deliver multiple orders per trip; we refer to this as the temporal pooling strategy. In this strategy, a courier is not necessarily dispatched as soon as an order arrives; orders are allowed to accumulate over time, and then a batch of sequential orders is delivered in one trip. We show that this strategy achieves delivery efficiency in the form of a shorter expected travel distance per order and lower variability in traveling distance per trip. However, whereas this pooling strategy benefits the supply side, it undoubtedly affects customers' experiences on the demand side, which may deter them from using the service or require monetary compensation for the long wait, reducing the strategy's attractiveness. Therefore, each delivery strategy has its advantages: the dedicated delivery may mean a shorter wait for each customer, whereas batch delivery appears more efficient from the firm's perspective.

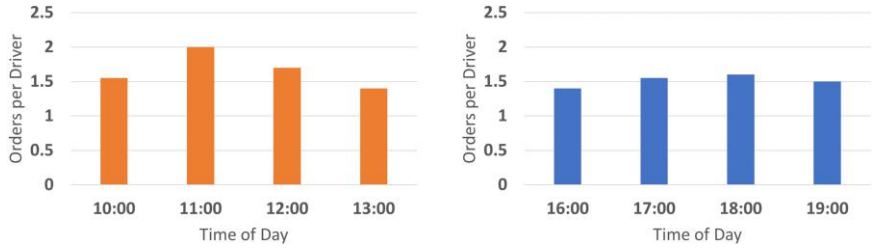
The on-demand courier dispatch problem differs from traditional delivery problems (such as the celebrated traveling salesman problem (TSP)), in which there are many stops per trip. Orders containing on-demand supplies (such as coffee, food, and medicine) typically have short delivery windows. According to Rana and Kang (2021), food delivery platforms such as DoorDash and Uber are researching bundling orders together. Still, unlike traditional delivery services, they also plan to deliver all orders in an hour. Thus, on-demand delivery services cannot deliver with large batch sizes consistently. In particular, according to an internal study conducted by one of the largest delivery platforms in China, for food delivery, their couriers carry fewer than two orders on average per trip even during peak lunch and dinner hours (see Figure 1).

Another critical factor in the operations of delivery systems is whether the demand is exogenous or can be endogenized through pricing. On the one hand, a new delivery platform needs to maintain growth and carve out its market share by sustaining a fixed demand, also known as market penetration (Rana and Haddon 2021a).[2] Studies on market penetration can be traced back to Buzzell et al. (1975), followed by empirical evidence (see, e.g., Szymanski et al. 1993), stating there is a positive correlation between the market share and (long-term) profitability. Thus, the demand can be exogenously determined for a vendor in its early stage of operations to achieve a particular market share. On the other hand, a vendor that has already established a stable market base can endogenize the demand by varying delivery fees to further optimize its revenue.

In this paper, we take the perspective of a vendor providing delivery service and address the following research questions: when is temporal pooling beneficial, and when should a courier be dedicated to one order per trip? More specifically, we consider scenarios in which the delivery system with dedicated couriers has exogenous and endogenous demand, respectively, and identify the key factors affecting its operating strategy. We use the vendor's revenue as the performance measure in either scenario. For simplicity, we refer to the delivery strategy with temporal pooling as the batch or pooling strategy and the one serving a single order per trip as the dedicated strategy. In the exogenous demand case, depending on the expected wait time associated with each strategy, the vendor sets the price to achieve the targeted demand rate. In the endogenous demand case, the vendor has complete freedom at varying the price to moderate the demand rate.

We build a stylized model capturing the spatial aspect of delivery systems under different dispatch strategies. Following a Poisson process, demand arises at a uniformly distributed point in a service region. We obtain

**Figure 1.** (Color online) The Distribution of Orders per Courier During Peak (Left) Lunch and (Right) Dinner Hours




*Note.* By courtesy of Hongyan Dai.

concise but informative analytical results by using a disk-shaped service area and recognizing the similarities between delivery and (spatial) queueing systems. Whether the demand is endogenized critically affects the vendor's optimal dispatch strategy. In our base model, we assume there is a single courier for dispatch (which we relax in an extension). We first analyze a large market in which customers' potential arrival rate is large (relaxed in another extension). We show that, in such a crowded market, if the demand rate is exogenously given as under market penetration (e.g., in the "scale-up" stage of a start-up), there is a threshold size for the service area below which it is optimal to use the dedicated strategy and above which it is optimal to use the pooling strategy. We find that whichever strategy produces a shorter expected wait time under exogenous demand is optimal for the vendor. Thus, customers' patience level does not directly impact the decision on the delivery strategy because it does not affect the length of wait time itself.

The situation is very different if the firm can endogenize the demand rate (e.g., as in the "sustainment" stage of a start-up where profit maximization is focused on). With endogenized demand, there is a threshold size for the service area below which it is optimal for the firm to deliver in batches and above which it is optimal to adopt dedicated delivery. This result is in stark contrast to the one for exogenous demand. It runs counter to popular belief that serving in batches leads to higher delivery efficiency in a large service area than dedicated delivery (which is likely gained under the assumption that the demand rate is exogenously given). The intuition of our finding is that, in a relatively large service area, both strategies involve substantial travel distances, leading to long wait times. By maintaining a high demand rate, the firm needs to sacrifice a lot of profit margin to ensure customers join the service. As a result, the firm favors a relatively low endogenized demand rate for both strategies. The pooling strategy loses its efficiency edge in this case because it takes a long time to accumulate multiple orders with a low demand rate. The dedicated strategy is more efficient because its optimal demand rate is lower than the one under the pooling strategy. Furthermore, we also find that there is a threshold on customers' patience level below which the pooling strategy is optimal and above which the dedicated strategy is optimal. We summarize these results in Table 1.

**Table 1.** Optimal Delivery Strategy According to Nature of Demand

| | Exogenous demand | Endogenous demand |
|---|---|---|
| Small area | Dedicated | Batch |
| Large area | Batch | Dedicated |
| Patient customers | — | Batch |
| Impatient customers | — | Dedicated |

We then examine a variety of extensions of the base model, including social welfare maximization, hybrid policies that use dedicated or batch delivery contingently, general arrival rates that do not rely on the large market assumption, larger batch sizes, a nonuniform demand distribution inside the service area, and finally multiple couriers. Our main insights carry through in these extensions.

## 2. Literature Review

Cao and Qi (2022) and Yildiz and Savelsbergh (2019) are most closely related to ours. Cao and Qi (2022) study the optimal deployment strategy for vendors with high mobility, often referred to as the stall economy. Although their primary focus is on using the analytical model and machine learning algorithms to explain the scalability of the stall economy, the authors also empirically evaluate the benefit of demand pooling. They divide the service area into several subregions. They consider demand pooling that serves orders arriving within the same time window in the same subregion together before moving to the next subregion. Their empirical study finds that such demand pooling is more beneficial when customers are patient, which is consistent with our analytical results under the endogenous demand rate. Yildiz and Savelsbergh (2019) also consider a disk-shaped delivery area similar to that in our model, in which a single restaurant at the center of the disk serves the entire area. They only consider the dedicated strategy. Their focus is on the optimal service radius and compensation for crowdsourced couriers, whereas ours is evaluating the benefit of temporal demand pooling.

Our paper belongs to the stream of research on spatial queueing models. This literature typically considers a logistical setting in which vehicles are modeled as servers, and their traveling time to serve customers equals the service time. Berman et al. (1985, 1987) focus on finding one or multiple service hubs in a network to minimize the expected response time to random demand. They model the service system using queueing models incorporating the spatial features of the network. Bertsimas and van Ryzin (1990, 1992) consider stochastic and dynamic routing of vehicles to serve service requests that are randomly generated over a service region. The authors evaluate the performances of various policies and identify optimal and near-optimal policies under light and heavy traffic. Recently, spatial queueing models are also utilized in smart city design (see, e.g., He et al. 2017, Mak 2022) and warehouse operations (see, e.g., Besbes and Cachon 2021). Besbes and Cachon (2021) compare temporal pooling of robots versus human pickers, in which robots move a pod with pooled items to the drop-off station and human workers pick multiple items in a trip similar to our pooling strategy. In contrast, we focus on comparing the dedicated and pooling strategies

and also incorporate demand moderation to examine the interaction between the demand side's pricing decision and the supply side's dispatch decision.

Our paper is also related to papers using queueing models to study the on-demand economy. Taylor (2018) and Bai et al. (2019) treat freelancers in the on-demand economy as servers in queueing models. They approximate the customers' wait time with M/M/$k$ queues. Daniels and Turcic (2021) capture the competition between taxis and Uber for wait-sensitive riders using queueing models. Feldman et al. (2022) examine different contracts between a delivery platform and a single restaurant and compare their performance to that in a centralized setting in which the restaurant controls prices. Chen et al. (2022) study a similar problem by examining a setting with two streams of customers: tech-savvy and traditional. Both papers model the food-serving restaurant as a stylized M/M/1 queue. Cui et al. (2020, 2021) model line-sitting and queue-scalping, respectively, based on M/M/1 queues. They treat line-sitting and queue-scalping as innovative service models as opposed to traditional first come, first served and compare their performances in equilibria.

Similar to our paper, a stream of literature in operations management also uses couriers' travel distances to quantify the delivery cost. These papers typically deal with a large number of orders per delivery trip and resort to the asymptotic analysis of variants of the TSP to quantify the expected travel distance (see, e.g., Cachon 2014, Carlsson and Song 2017, Qi et al. 2018, Cao et al. 2020). In contrast, we assume that a courier delivers no more than a few orders per trip, supported by empirical evidence (see Figure 1). Furthermore, with a spatial queueing formulation, our analysis is anchored by the expected travel distance and the variability in traveling during delivery trips. More recently, He et al. (2021) also recognize that using TSP may not accurately depict the trip length in food delivery as couriers and the platform may not share the same information. They propose prediction models on travel time using machine learning.

Many papers also discuss the impact of dispatch policies on operational efficiency and profitability. Klapp et al. (2018a, b) consider the dynamic dispatch wave problem. In their setting, dispatch decisions are made at predetermined times of a day, and the decision maker decides on which orders to be delivered in each wave. The major trade-off in whether to deliver an order is between reducing the number of outstanding orders so they can be delivered by the end of the day versus waiting for nearby orders to show up so the delivery efficiency can be improved. Voccia et al. (2019) also consider a multivehicle dynamic pickup and delivery problem with same-day delivery as the time constraint. Other papers such as Azi et al. (2012) and Ulmer et al. (2019) also study the optimal order assignment and the optimal timing for vehicle departure in a single-depot setup. In addition, at the operational level, Farahani et al. (2022) optimize the dispatch to minimize the costs of earliness and tardiness benchmarked with a common quote time. At the tactic level, He and Goh (2022) study the optimal order allocation between in-house employees and freelancers with a thicker market for the latter attracting more customers over the long term. Unlike these papers, we consider the pricing decision besides the short-run dispatching policies.

Finally, our spatial modeling approach relates to Hotelling's circular city model in economics. That model has suppliers and consumers evenly dispersed on a circle, and consumers have preferences over suppliers based on their relative locations. We extend the original circular city model (see, e.g., Salop 1979) to have the supplier sitting at the center of the circle and customers located inside the circle, forming a disk-shaped service area. In an extension, we also investigate the extreme case in which customers only reside on the edge of the disk. Some recent papers in operations management also use spatial models based on a circular city. Chen et al. (2021) consider a matching problem in ride-sharing in which drivers and riders depart from the center of a circle going to different locations on the edge of the circle. Feng et al. (2021) also use a circular city to study ride-hailing in which drivers travel clockwise or counterclockwise, picking up riders on the circle. Unlike our spatial model, none of these papers considers areas inside the circle as part of the service region.

## 3. Model

Consider a vendor with a facility located at the center of a disk-shaped region with a radius $r > 0$ and a single courier serving customers in the area. We relax the single-courier assumption and consider multiple couriers in Section 6.7. The structure of our service area is a generalization of the circular city model (see, e.g., Salop 1979) in the sense that customers also occupy areas inside the disk. In contrast, the original model only considers the edge of the disk. The centrally located facility can be a store, urban warehouse, restaurant, or ghost kitchen. We assume the arrival process of customers is Poisson with rate $\Lambda r^2$, which scales with the area $\pi r^2$ of the service region.[3] Upon arrival, each customer's location is independent and uniformly distributed on the disk. Each customer is also subjected to a wait cost with a rate $c$ per unit of time. Furthermore, we assume that each customer has a valuation $v$ for the delivery service, which follows a general distribution with the cumulative density function (CDF) $F$. Without loss of generality, we normalize the support of $F$ to $[0, 1]$. The vendor can decide the charge for each delivery service at a price $p$. We assume that customers are sensitive toward the wait time from their order time to the time of seeing

"your order is on its way" (a treatment consistent with those papers that ignore the delivery time such as Chen et al. 2022 and Feldman et al. 2022), which represents the time between when an order is placed and the delivery courier starts to be en route. (This assumption is relaxed in Chen et al. 2023.) That is, a customer is satisfied once a courier is on the way to make an exclusive delivery of the order: in dedicated delivery, every trip is exclusive; in batch delivery with size two, a customer anticipates being either the first or second stop in a delivery trip with an equal probability and, if the customer happens to be the second, expects that the courier needs to deliver the first order to a random location. We assume that the vendor commits and announces its delivery strategy. Thus, the resulting demand process is a homogeneous spatial Poisson process, which not only makes our model more elegant, but also has practical relevance. For example, Ulmer et al. (2021) study a stochastic dynamic pickup and delivery problem and conduct experiments on a dynamic food delivery problem in the Iowa City metropolitan area. They use a homogeneous spatial Poisson process to model the demand request pattern that is confirmed by the local providers. In Section 6.6 and Online Appendix D, we further investigate the case in which customers are also sensitive to en route delays, which leads to distance-dependent wait time. As a result, we would have a nonuniformly distributed demand over the space.

Thus, in our base model, given the expected wait time $w$ from the order time to the expected starting time of making an exclusive delivery, a customer's utility from using the delivery service is simply $v - p - cw$. Note that this utility expression assumes that the wait cost is linear in time. This is indeed a simplification of reality for model tractability. In practice, most instant delivery services offer a "soft promise" in wait time (e.g., 10 minutes for Gorillas and 30 minutes for Gopuff). This implies that the wait cost by customers may be negligible if the wait time is below a cutoff, whereas above it, the wait cost can be convexly increasing in the wait time because the wait beyond the promise can be increasingly painful as each additional minute passes by; see Online Appendix H for such an extension.

Only customers with nonnegative utilities use the delivery service from the vendor. Customers with negative utilities may choose to pick up the orders themselves or not order at all. Denote by $\lambda \in [0, \Lambda r^2]$ the effective demand rate of the delivery service. Because each customer's valuation $v$ follows a distribution with the CDF $F$, the demand rate $\lambda$ satisfies $\lambda/\Lambda r^2 = 1 - F(p + cw)$, which implies that, for all $\lambda \in (0, \Lambda r^2]$, we have

$$p = F^{-1}\left(1 - \frac{\lambda}{\Lambda r^2}\right) - cw, \tag{1}$$

where function $F^{-1}$ is the inverse function of the CDF $F$. Thus, a one-to-one mapping exists between price $p$ and positive demand rate $\lambda$. Note that, if the demand rate $\lambda$ is exogenous, then it is possible to have $p < 0$ as the vendor needs to subsidize customers for the service, which may happen when the vendor wants to grow a market. This would not happen when the demand rate is endogenized. Using the expression in (1) for positive demand rate, the vendor's revenue function can be written as

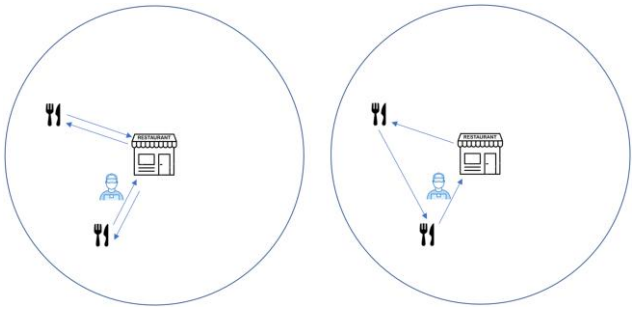
$$V(\lambda, w) = \lambda p = \lambda\left(F^{-1}\left(1 - \frac{\lambda}{\Lambda r^2}\right) - cw\right), \quad \lambda > 0. \tag{2}$$

The vendor makes operational decisions based on the revenue it generates according to (2).

We emphasize that wait time for each customer, $w$, in a steady state also depends on the effective demand rate $\lambda$. In later sections, when comparing the vendor's revenue functions under different delivery modes, we replace $w$ by the expected wait time for each customer, which is a function of the demand rate $\lambda$. The underlying assumption is that customers anticipate a wait time and use it to decide whether to adopt the service. In equilibrium, their expected wait time is consistent with their experiences over repeated interactions.

We consider and compare two delivery strategies: the dedicated and pooling strategies. On the one hand, with the dedicated strategy, the courier serves orders one by one in the first come, first served fashion (referred to as dedicated delivery). On the other hand, with the pooling strategy, the courier is not en route for delivery until exactly two[4] orders are accumulated, which can be interpreted as serving orders in batches of two (referred to as batch delivery). Figure 2 illustrates the differences between the two strategies. When serving dedicated delivery, a courier leaves the restaurant immediately when an order arrives. After delivering the food, the courier returns to the restaurant to pick up or wait for the following order. When serving batch delivery, the courier does not leave the restaurant until two orders have arrived. Then, the courier delivers both orders in a single delivery trip before returning to the restaurant for the next batch. We do not specify the fulfillment sequence within a batch as long as the resulting order is random; for example, the sequence can follow the time

**Figure 2.** (Color online) Serving Dedicated vs. Serving Batch



or spatial order of arrivals, such as always traveling clockwise. The fulfillment sequence within a batch does not affect the total travel distance of a courier but may affect the wait time of a specific order. If the resulting fulfillment order is random, customers still have the same expected wait time over repeated interactions with the system. We assume that there is no delivery delay at each drop-off location, which is relaxed in Online Appendix G.

We recognize the similarity between our delivery system and a single-server queue in which a courier acts as the server and customers' orders queue up. Because potential customer arrivals follow a Poisson process and a fraction of the customers choose the delivery service based on the expected wait, the arrival process of orders is also Poisson with the rate equal to the effective demand rate $\lambda$. As for the service process, we assume that the courier instantly picks up the delivery goods at the centrally located facility and spends no time at each customer's location. Thus, the service time only consists of the courier's traveling time between the facility and the customers' location(s). We define a delivery trip as the process starting when the courier picks up the delivery goods at the facility and ending when the courier returns. We utilize the queueing literature results to derive customers' expected wait time under each delivery strategy in the following two sections.

### 3.1. Dedicated Delivery

Suppose the courier uses dedicated delivery to serve customers. As mentioned, orders arrive following a Poisson process with a rate $\lambda$ in equilibrium. The service time is the time the courier spends delivering each order. When dedicated delivery is adopted, each delivery trip is the round trip between the facility and a random customer's location. Assuming a constant travel speed and normalizing it to one, the service time equals the travel distance per delivery trip.

Denote by a random variable $X_D$ the shortest Euclidean distance of a delivery trip when serving orders under dedicated delivery. So $X_D$ is two times the distance between the disk's center with radius $r$ and a uniformly distributed point on the disk. According to the disk point picking literature (see, e.g., Solomon 1978), we have

$$\mathbb{E}[X_D] = \frac{1}{2\pi r^2}\int_0^{r^2}\int_0^{2\pi} 2\sqrt{x}\ \mathrm{d}\theta\ \mathrm{d}x = \frac{4}{3}r, \text{ and}$$
$$\mathbb{E}[X_D^2] = \frac{1}{2\pi r^2}\int_0^{r^2}\int_0^{2\pi} 4x\ \mathrm{d}\theta\ \mathrm{d}x = 2r^2. \tag{3}$$

Note that the first moment of random variable $X_D$ represents the expected distance of the delivery trip, which is also the expected service time under our normalization of the travel speed. Then, we can treat this delivery system as an M/G/1 queue with the service rate and load factor equal to

$$\mu_D = \frac{1}{\mathbb{E}[X_D]} = \frac{3}{4r}, \text{ and } \rho_D = \frac{\lambda}{\mu_D} = \frac{4}{3}\lambda r, \text{ respectively.} \tag{4}$$

We define the expected wait time by $W_D$ when using dedicated delivery as a function of demand rate, service rate, and the coefficient of variation of the arrival and service processes:

$$W_D(\lambda, \mu, C) := \frac{\lambda}{\mu(\mu-\lambda)}\frac{C}{2}, \quad \forall \lambda, C \geq 0, \mu > 0, \tag{5}$$

where the term represents in-line delay of an M/G/1 queue (see, e.g., Gross et al. 2008). The summation of coefficients of variation of our M/G/1 queue's arrival and service processes is

$$C_D = 1 + \frac{\mathbb{E}[X_D^2] - (\mathbb{E}[X_D])^2}{(\mathbb{E}[X_D])^2} = \frac{9}{8}. \tag{6}$$

Thus, according to (5), $W_D(\lambda, \mu_D, C_D)$ represents the expected wait time for each customer when the courier uses dedicated delivery. Therefore, we can rewrite the revenue function in (2) as

$$V_D(\lambda, W_D(\lambda, \mu_D, C_D)) = \lambda\left[F^{-1}\left(1 - \frac{\lambda}{\Lambda r^2}\right) - cW_D(\lambda, \mu_D, C_D)\right], \tag{7}$$

representing the revenue rate of the delivery service when the vendor adopts dedicated delivery.

### 3.2. Batch or Pooling Strategy

Instead of serving orders with dedicated delivery, the courier can also deliver orders using batch delivery. In this paper, we assume that each batch consists of two orders and, inside each batch, orders are delivered following a predetermined rule. The courier does not leave the facility until two orders have arrived. Thus, when comparing our delivery system to a queueing system, we consider orders entering the queue in pairs of two. An arriving order does not technically enter the queue if all outstanding orders in the system are already in pairs of two. Instead, it waits and joins the queue together with the following order that arrives. Therefore, when the demand rate is $\lambda$, we can effectively treat the interarrival time as being Erlang distributed with order two and having a mean of $2/\lambda$ (with the arrival rate being $\lambda/2$).

Next, we analyze the service process of the delivery system using batch. A delivery trip needs to include three parts: travel between the facility and the first order's location, between the first and second orders' locations, and finally back to the facility from the second order's location. Denote by random variable $X_B$ the shortest distance a courier needs to travel per trip.

According to the disk line picking literature (see, e.g., Solomon 1978), we have[5]

$$\mathbb{E}[X_B] = \frac{1}{\pi r^4}\int_0^{r^2}\int_0^{r^2}\int_0^{\pi}\left(\sqrt{x+y-2\sqrt{xy}\cos(\theta)} + \sqrt{x}+\sqrt{y}\right)\mathrm{d}\theta\,\mathrm{d}x\,\mathrm{d}y = \left(\frac{128}{45\pi}+\frac{4}{3}\right)r,$$
$$\mathbb{E}[X_B^2] = \frac{1}{\pi r^4}\int_0^{r^2}\int_0^{r^2}\int_0^{\pi}\left(\sqrt{x+y-2\sqrt{xy}\cos(\theta)} + \sqrt{x}+\sqrt{y}\right)^2\mathrm{d}\theta\,\mathrm{d}x\,\mathrm{d}y \approx 5.428r^2. \quad (8)$$

Because the travel speed is normalized to one, the travel distance in each delivery trip is the service time for the courier. Using the first moment of $X_B$, we can derive the service rate and load factor of this service queue as

$$\mu_B = \frac{1}{\mathbb{E}[X_B]} = \frac{45\pi}{4r(32+15\pi)}, \text{ and}$$
$$\rho_B = \frac{\lambda}{2\mu_B} = \frac{2\lambda r(32+15\pi)}{45\pi}, \quad (9)$$

respectively. With both arrival and service processes characterized, we recognize that our batch service can be analyzed through an $E_2/G/1$ queue.

Because the interarrival time follows an Erlang-2 distribution, using the first and second moments of $X_B$, the summation of the coefficients of variation for arrival and service processes is

$$C_B = \frac{1}{2} + \frac{\mathbb{E}[X_B^2]-(\mathbb{E}[X_B])^2}{(\mathbb{E}[X_B])^2} \approx 0.583. \quad (10)$$

Unfortunately, we do not have a closed-form expression for the expected in-line delay of the $E_2/G/1$ queues. Seeking analytical results, we use Kingman's formula (see, e.g., Gross et al. 2008) to approximate the in-line delay of this $E_2/G/1$ queue as a $G/G/1$ queue. That is, we have

$$W_q \approx \frac{1}{2\mu_B}\frac{\rho_B}{1-\rho_B}C_B = \frac{C_B}{2}\frac{\lambda}{\mu_B(2\mu_B-\lambda)}, \quad (11)$$

where $C_B$ is defined in (10). The Kingman's formula we adopt serves as an upper bound (see, e.g., Kingman 1962) on the in-line delay and is asymptotically exact in the heavy traffic regime. All our results in favor of the pooling strategy can be refined analytically exact as we use the upper bound of the in-line delay under batch delivery compared with the dedicated strategy. Our results still hold for a numerical verification in which the expected in-line delay is computed from a simulated system of the $E_2/G/1$ queue. In Online Appendix F, we provide simulation results on the accuracy of all the approximations in this paper. In summary, this paper's closed-form approximations are reasonably accurate.

Note that the batch delivery has a shorter in-line delay compared with a hypothetical $M/G/1$ dedicated delivery system in which the arrival rate is $\lambda/2$. The reason is that the batch system has a lower coefficient of variation, that is, $C_B \le C_D$, which means there is less variability in both the arrival and service processes of the batch system. More specifically, the variability in the arrival process is reduced from 1 in the dedicated system to 1/2 in the batch system because of temporal pooling of orders. The variability in the service process is reduced from 1/8 in the dedicated system to about 0.083 in the batch system because of spatial pooling of two delivery trips into one.

Recall that, when using an $E_2/G/1$ queue to analyze our batch system, a single order does not enter the queue until a second order arrives. In other words, the in-line delay does not include the time to form a batch of two orders, which is on average $1/\lambda$. We assume that the customer does not know the exact state of the system as is the case in practice. That is, the customer has no information on the customer's position in the queue. Thus, from a customer's perspective, the expected wait time consists of three parts: the expected wait time for a second order to arrive if the customer's order does not enter the queue immediately, the average in-line delay once the customer's batch enters the queue, and if the customer is the second in the batch to be served, the time it takes to serve the first. Define the expected wait time $W_B$ as a function of the demand rate, service rate, and the coefficient of variation. That is, we have

$$W_B(\lambda,\mu,C) := \frac{1}{2\lambda} + \frac{\lambda}{\mu(2\mu-\lambda)}\frac{C}{2} + \frac{1}{2}\frac{\mathbb{E}[X_D]}{2} = \frac{1}{2\lambda} + \frac{\lambda}{\mu(2\mu-\lambda)}\frac{C}{2} + \frac{r}{3}, \quad \forall \lambda, C \ge 0, \mu > 0, \quad (12)$$

where the components correspond to the three parts in the customer's expected wait time, respectively. In particular, the last term $\mathbb{E}[X_D]/4$ represents the expected extra delay if the courier serves the customer's order in the second. So, half of the time, the customer must wait for the courier to deliver the other order first (taking $\mathbb{E}[X_D]/2$ time in expectation) before being en route with the customer's order. Thus, $W_B(\lambda,\mu_B,C_B)$ represents a customer's expected wait time when the courier is serving batch. Note that $W_B(\lambda,\mu_B,C_B)$ approaches infinity as $\lambda$ goes to zero. The reason is that the courier never leaves the facility with a single order, so a customer may need to wait for a long time when a second order takes some time to arrive. Thus, the revenue function in (2) becomes

$$V_B(\lambda, W_B(\lambda,\mu_B,C_B)) = \lambda\left[F^{-1}\left(1-\frac{\lambda}{\Lambda r^2}\right) - cW_B(\lambda,\mu_B,C_B)\right], \lambda \in (0, 2\mu_B). \quad (13)$$

It is worth pointing out that $\lim_{\lambda\to 0} V_B(\lambda, W_B(\lambda,\mu_B,C_B)) = -\frac{c}{2} < 0$ as the expected wait time $W_B(\lambda,\mu_B,C_B)$ approaches infinity when $\lambda$ approaches zero. Thus, in

batch serving, if the vendor needs to maintain a low demand rate close to zero, the vendor has a negative revenue rate. In other words, maintaining a low demand rate in batch serving is unprofitable for the vendor because it requires a significant subsidy to customers. However, we only use this limit case to provide intuitions on a disadvantage of batch serving because, to gain profitability, the vendor can serve dedicated, generating nonnegative revenue when the demand is very low.

## 4. Exogenous Demand Rate

In this section, we evaluate the performance of adopting the dedicated and batch deliveries when the demand is exogenous. The base model uses the vendor's revenue as the performance measure. Although the demand rate is exogenous, the vendor can still decide on which delivery mode to operate, coupled with the corresponding price, to achieve the targeted demand rate and attain a higher revenue. This is the case when the firm has an exogenously given demand segment to cover because of the needs of growing or penetrating a market or other goals that are not directly related to revenue creation from delivery services, for example, the need to match the delivery capacity with the kitchen capacity. We observe that serving batch can sustain a higher demand rate than serving dedicated delivery because, when comparing the load factors in (4) and (9), we have $\rho_B < \rho_D$ if $\lambda > 0$ is fixed. Furthermore, because both $\rho_D$ and $\rho_B$ are linearly increasing in $r$, we also observe that serving batch allows the delivery service to handle a larger service region than serving dedicated delivery.

Comparing the revenue functions in (7) and (13), if the demand rate $\lambda$ is exogenous, the delivery strategy that has the shorter expected wait time leads to higher revenue. Thus, the operating strategy with exogenous demand is efficiency-driven. We compare the revenues generated via the two delivery strategies and their corresponding expected wait times in the following two propositions.

**Proposition 1.** *If the demand rate is exogenously given, there exists a threshold on the demand rate below which serving dedicated leads to a shorter expected wait time and, thus, higher revenue and above which serving batch leads to a shorter expected wait time and, thus, higher revenue.*

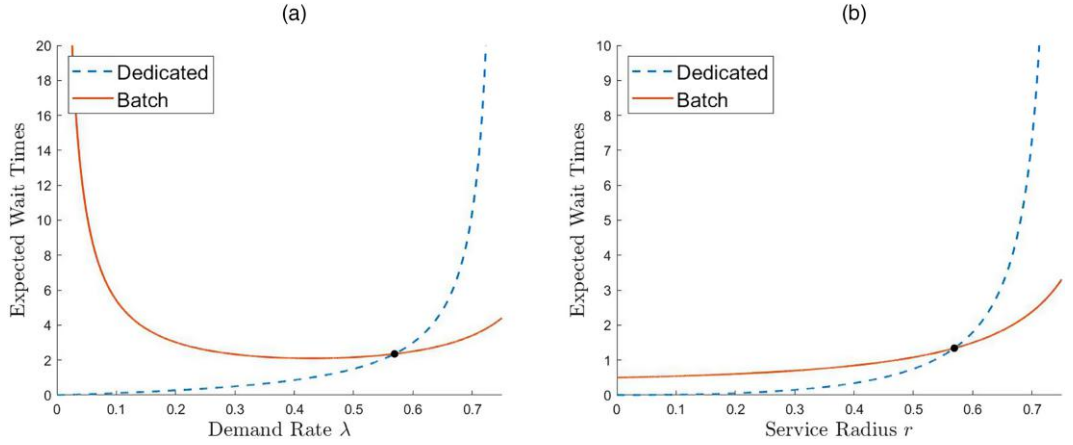
Proposition 1 states that operating dedicated delivery is better than batch when the exogenous demand rate is low. The intuition is that, when the demand rate is low, it takes a very long time to accumulate two orders so that the courier can make a batch delivery trip. Figure 3(a) provides a visual representation of the wait times. As an extreme case, when the demand rate goes to zero, the expected wait time for each customer approaches infinity under batch. However, adopting dedicated delivery leads to a much shorter expected wait time.

As the average time to accumulate two orders drastically decreases when the demand rate increases, the overall expected wait time under batch also decreases. When the demand rate becomes very high, the in-line delay of customers dominates the average wait time for a pair of two orders to accumulate. Thus, the expected wait time increases with a sufficiently high demand rate. As mentioned, serving batch can handle a higher demand rate than serving dedicated because the average travel distance associated with delivering an order is shorter. In Figure 3(a), we observe that the expected wait time under dedicated delivery approaches infinity faster when $\lambda$ becomes sufficiently large than that under batch delivery does.

Not only is there a threshold on the demand rate that changes the vendor's delivery strategy, but the next proposition also states that there is such a threshold on the size of the service region.

**Proposition 2.** *If the demand rate is exogenously given, there exists a threshold on the service radius below which serving dedicated leads to a shorter expected wait time and,*

**Figure 3.** (Color online) Expected Wait Time When Serving Dedicated or Batch




*Notes.* (a) $r = 1$. (b) $\lambda = 1$.

*thus, higher revenue and above which serving batch leads to a shorter expected wait time and, thus, higher revenue.*

Proposition 2 states that operating dedicated delivery is better if the service radius is small and serving batch is better otherwise. This result appears to be intuitive as one may think that when the service radius is large, serving batch can reduce the total travel distance of the courier. However, the first moments of the lengths of delivery trips under both dedicated delivery and batch scale with $r$ when other parameters are fixed in (3) and (8), respectively. Thus, one can verify that, for any service radius, compared with dedicated delivery, serving batch leads to a longer average total travel distance but a shorter distance per order, that is, $\mathbb{E}[X_B]/2 \le \mathbb{E}[X_D] \le \mathbb{E}[X_B]$. The main reason behind Proposition 2 is that, when the service radius is small, the time to accumulate two orders when serving batch is much longer than the actual travel time. On the other hand, if the service radius is large, the travel time becomes longer than the time to accumulate two orders, independent of the service radius when the demand rate is exogenous. Thus, serving batch is more beneficial when the service radius is large. Figure 3(b) provides a visual illustration of the expected wait time of a customer when the courier serves dedicated delivery and batch, respectively.

**Corollary 1.** *Suppose the demand rate is exogenously given.*

i. *There exist thresholds in demand rate and service radius (same as those in Propositions 1 and 2, respectively) such that below which the price is higher when using dedicated delivery and above which serving batch leads to a higher price.*

ii. *There exist thresholds in demand rate and service radius (same as those in Propositions 1 and 2, respectively) such that below which the expected wait time per order is shorter when serving dedicated and above which serving batch leads to a shorter expected wait time per order.*

Corollary 1 extends the results in Propositions 1 and 2 to price and delivery efficiency. When the demand rate is exogenous, the price is nonincreasing with the wait time. Furthermore, as we use the expected wait time per order as the measure of delivery efficiency, serving batch is more efficient when either the demand rate is high enough or the service radius is large enough. Otherwise, dedicated delivery is more efficient as it bypasses the order accumulation time.

As mentioned, the case with an exogenous demand rate can describe the market penetration stage experienced by many start-up companies or applications in public or other business settings with rigid demand requirements. For example, consider a newly formed ghost kitchen in a mega city, which hires a given number of the kitchen staff (so the maximum kitchen throughput is given) at the operational level or aims to carve a targeted market share in the local takeaway food market at the tactic level. Thus, the kitchen needs to maintain a targeted demand rate by offering delivery promotions, which greatly limits its pricing decision. If the service area is fixed, dedicated delivery outperforms batch delivery if and only if the targeted demand rate is relatively low. Serving batch is only beneficial if a relatively high demand rate needs to be maintained, so temporal pooling can add efficiency en route without losing too much time accumulating orders. Further, dedicated delivery leads to a shorter expected wait time for customers and higher revenue if the service area is relatively small. However, with a predetermined larger service area, it is better to serve batch, taking advantage of the efficiency en route.

We conclude this section by pointing out that, if the demand rate is exogenously determined, only the effective demand rate $\lambda$ and the service radius $r$ impact the vendor's delivery decision because we only need to compare the expected wait times for customers under the two strategies. That is, the underlying arrival rate of customers $\Lambda$, wait cost parameter $c$, and the distribution function $F$ of customer valuations do not affect the delivery strategy once the targeted demand rate is determined. In the next section, we compare and contrast the results of this section to the case in which the demand rate $\lambda$ can be optimized.

## 5. Endogenous Demand Rate

The previous section covers the scenario with an exogenous demand rate that needs to be sustained. In this section, the vendor aims at maximizing its revenue with an endogenized demand rate. That is, there is no exogenous constraint on the demand rate, and the vendor maximizes its revenue by designing the optimal demand rate. Therefore, unlike Section 4, in which the vendor can only choose in which delivery mode to operate with a given demand rate, in this section, the vendor also chooses the optimal demand rate in each mode that can be achieved via the freedom in varying the price.

Seeking tractable analytical results, we first take advantage of a crowded market setting in which the underlying arrival rate of customers is high enough. Suppose the arrival rate scales with a density factor $n \in \mathbb{N}$. As $n$ increases, the arrival rate $n\Lambda$ increases, meaning that the market gets more and more crowded. Thus, with customer valuations drawn from the CDF $F$ (with its support normalized to $[0,1]$), the revenue function in (2) can be modified to

$$V_n(\lambda, w) = \lambda\left(F^{-1}\left(1 - \frac{\lambda}{n\Lambda r^2}\right) - cw\right), \quad \lambda \ge 0. \tag{14}$$

As in this section, the vendor maximizes the revenue rate by choosing the demand rate; $\lambda = 0$ is not the optimal choice.

Define function

$$V_\infty(\lambda, w) := \lim_{n\to\infty} V_n(\lambda, w) = \lambda(1 - cw), \quad \lambda \geq 0, \tag{15}$$

where the equality follows that the upper bound on customer valuations has been normalized to one. The expression in (15) represents the limiting revenue when the density factor $n$ goes to infinity. According to (15), when the underlying arrival of customers goes to infinity, the vendor only serves those with a valuation almost equal to one, the upper bound. Thus, at the limit, the vendor's revenue is independent of customer valuation distribution. In general, under a given delivery strategy, when the targeted demand rate $\lambda$ increases, two terms in (14) change: (i) the base price $F^{-1}\left(1 - \frac{\lambda}{n\Lambda r^2}\right)$ needs to be adjusted downward to incentivize more adoption, and (ii) the expected wait time $w$ increases as a result of a higher joining rate, and thus, more discount $cw$ needs to be paid to compensate customers for the longer wait. The crowded market assumption assumes away the first effect, which is verified by Lemma 1. We relax this assumption in Section 6.3.

First, we present a lemma on utilizing the expression in (15), which greatly simplifies our analysis for a crowded market.

**Lemma 1.** *Consider $n \in \mathbb{N}$ and a CDF $F$ such that $F^{-1}$ is Lipschitz continuous. We have*

$$\lim_{n\to\infty} \max_{\lambda\in[0,\mu_D)} V_n(\lambda, W_D(\lambda, \mu_D, C_D)) = \max_{\lambda\in[0,\mu_D)} V_\infty(\lambda, W_D(\lambda, \mu_D, C_D)), \tag{16}$$

*and*

$$\lim_{n\to\infty} \max_{\lambda\in[0,2\mu_B)} V_n(\lambda, W_B(\lambda, \mu_B, C_B)) = \max_{\lambda\in[0,2\mu_B)} V_\infty(\lambda, W_B(\lambda, \mu_B, C_B)). \tag{17}$$

Lemma 1 implies that we can simply optimize the demand rates for serving dedicated and batch using the limiting revenue function in (15) when $n$ approaches infinity. Therefore, the vendor's demand-rate decision is independent of the customer valuation distribution. Because function $V_\infty$ has a much more concise expression than the nonlimiting revenue function, it is much easier to analyze and use for comparing optimal solutions under different delivery strategies. In particular, the next two propositions summarize the results for a crowded market when the vendor can endogenize the demand rate.

**Proposition 3.** *Assume a large market and suppose the demand rate can be endogenized.*

i. *There exists a threshold $c_\infty$ on customers' wait cost parameter $c$, below which serving batch leads to higher revenue and above which serving dedicated leads to higher revenue.*

ii. *As $c$ crosses the threshold $c_\infty$ such that the optimal strategy switches from serving batch to serving dedicated delivery, the optimal demand rate has a discontinuous drop, that is, $\lim_{c\to c_\infty^-} \lambda^*(c) > \lim_{c\to c_\infty^+} \lambda^*(c)$, where $\lambda^*(c)$ is the optimal demand rate as a function of the wait cost coefficient $c$, and the corresponding optimal price has a discontinuous surge.*

Proposition 3(i) states that, if the vendor can optimize the revenue rate by endogenizing the demand rate, serving dedicated is better if customers are impatient (i.e., $c$ is sufficiently high). With patient customers, it is optimal to serve batch (i.e., $c$ is sufficiently low). This is in contrast to the result in Section 4: when the demand rate is fixed, the wait cost parameter $c$ has no impact on the vendor's delivery decision because it does not affect the expected wait time. Proposition 3(ii) says that there is a sudden drop in the optimal demand rate and a surge in the optimal price when the cost of waiting crosses the threshold such that the optimal delivery strategy changes from batch to dedicated. When customers are impatient, the vendor should have a less crowded system with a relatively low demand rate, which gives an edge to dedicated fulfillment. If customers are patient, it is better to sustain a higher demand rate when implementing batch strategy. This shortens the time needed to accumulate two orders and, hence, the overall expected wait time. This intuition is consistent with Proposition 1 that a low (respectively, high) demand rate favors dedicated (batch).

**Proposition 4.** *Assume a large market and suppose the demand rate can be endogenized.*

i. *There exists a threshold $r_\infty$ on the service radius $r$ below which serving batch leads to higher revenue and above which serving dedicated leads to higher revenue.*

ii. *As $r$ crosses the threshold $r_\infty$ such that the optimal strategy switches from serving batch to dedicated, the optimal demand rate has a discontinuous drop, that is, $\lim_{r\to r_\infty^-} \lambda(r) > \lim_{r\to r_\infty^+} \lambda(r)$, where $\lambda(r)$ is the optimal demand rate as a function of the service radius $r$, and the corresponding optimal price has a discontinuous surge.*

Proposition 4 states that the vendor should serve dedicated when the market is crowded if the service radius $r$ is large enough. Instead, serving batch is optimal if the service radius is sufficiently small. This result contrasts with Proposition 2, in which the demand rate is exogenous. With a large service radius, the courier's travel time is long under either dedicated or batch, which leads to a relatively long expected wait time for customers. Thus, the vendor should sustain a relatively low demand rate. Otherwise, the compensation for the long wait would be significant. Again, there is a sudden drop in the optimal demand rate and a surge in the optimal price when the service radius crosses the threshold at which the optimal delivery strategy changes from batch to dedicated. Recall that serving batch is less profitable

than serving dedicated when the demand rate is low because serving batch has a much longer expected wait time. That is, an order may have to wait for a long time for another order to arrive and form a batch before it is en route for delivery. When the service radius is small, it is beneficial to operate under a relatively high demand rate as the average travel distance is shorter under either delivery strategy than with a large service radius. As mentioned, serving batch is more profitable for a relatively high demand rate.

Next, we discuss the practical implications of our results by discussing a few examples. During rush hour for a delivery system, the vendor may have far more potential customers than it can serve. Customers ordering a cup of coffee may be impatient because hot coffee will be cold if not delivered in time. In contrast, a grocery vendor or restaurant that only serves cold dishes such as sushi may have more patient customers. Thus, as implied by Proposition 3, even though the two businesses have the same service area, the coffee shop may prefer the dedicated strategy, and the grocery vendor or sushi restaurant, the pooling strategy. As an implication of Proposition 4, even if their customers have the same patience level, a restaurant serving only a 10-block radius in Midtown Manhattan may prefer the batch strategy, but a restaurant with similar characteristics delivering throughout Midtown Manhattan may want to use the dedicated strategy because the latter has a much bigger service area. This implication may seem counterintuitive at first glance as a larger service area may require more emphasis on delivery efficiency that the pooling strategy may achieve (as conveyed in Proposition 2). The key to understanding this seemingly counterintuitive insight is that the dedicated strategy is coupled with a high delivery price for a large service area. In contrast, the pooling strategy needs to keep the delivery price relatively low to compensate customers for the wait. With the profit margin being considered as the demand rate is endogenized, the dedicated strategy becomes optimal for a large service area.

We conclude this section by summarizing the results and contrasting them with those when the demand rate is exogenously given. First, we observe that with an endogenous demand rate, it is optimal to serve dedicated if the service area is large. This result directly contrasts with the one for an exogenous demand rate, where it is optimal to serve batch for a large service area. Second, customers' patience level, which has no impact if the demand rate is exogenous, greatly affects the vendor's delivery strategy for the endogenized demand rate. With the demand rate endogenously determined, the vendor should serve batch if customers are patient. However, if customers are impatient, serving dedicated generates higher revenue. Finally, for a crowded market, we can identify the optimal delivery strategy analytically for the entire spectrum of customers' patience level and the service area's size, respectively.

# 6. Extensions

This section considers a set of extensions of our base model. We investigate each one and examine the robustness of our results and intuitions obtained from Sections 4 and 5.

## 6.1. Social Welfare

Another objective of interest is the social welfare generated by the delivery system. We define the social welfare generated per order as the summation of the vendor's revenue and the customer's profit, that is, $v - cw$, where $w$ is the expected wait time because the price is an internal transfer between the vendor and a customer. Thus, the social welfare generated per order is

$$\begin{aligned} SW(\lambda, w) &= \Lambda r^2 \mathbb{P}(v \geq p + cw)\mathbb{E}[v - cw \mid v \geq p + cw] \\ &= \Lambda r^2 \int_{F^{-1}\left(1-\frac{\lambda}{\Lambda r^2}\right)}^{1} (v - cw) dF(v). \end{aligned} \tag{18}$$

The next proposition characterizes the impacts on the social welfare when the vendor focuses on market penetration or maximizing revenue, respectively.

**Proposition 5.**

i. *Suppose the demand rate is exogenous. There exist thresholds on the demand rate and service radius below which serving dedicated leads to higher social welfare and above which serving batch leads to higher social welfare.*

ii. *Suppose the demand rate is endogenous and the market is crowded. There exist thresholds on the service radius and customers' patience level below which serving batch leads to higher social welfare and above which serving dedicated leads to higher social welfare.*

Essentially, we recover the results in Sections 4 and 5 in Proposition 5. Thus, our major insights in the previous sections still hold even when the performance measure changes from the vendor's revenue to social welfare. When the demand rate is exogenous, the key factor in operations is delivery efficiency. On the other hand, when the demand can be endogenized, the vendor needs to consider the optimal demand rate to sustain, which tremendously impacts the system efficiency.

## 6.2. Contingent Policy

Another natural extension to our base model is to consider a contingent policy alternating between serving dedicated and batch depending on the queue size.[6] Suppose the courier serves the orders in batch if and only if there is more than one outstanding order in the queue and serves dedicated otherwise (i.e., when there is a single unfilled order). At first glance, it seems this contingent policy takes advantage of both delivery methods considered in this paper. In the next proposition, we show its relationship with dedicated and batch delivery.

**Proposition 6.** *For any demand rate $\lambda > 0$, the contingent policy leads to a shorter expected wait time for customers*

*than dedicated delivery. Furthermore, if the demand rate is large enough, batch delivery leads to a shorter expected wait time than the contingent policy, whereas if the demand rate is low enough, the contingent policy leads to a shorter expected wait time than batch delivery.*

Proposition 6 states that the contingent policy always dominates dedicated delivery in terms of the expected wait time. Thus, we can conclude that the contingent policy indeed outperforms dedicated delivery. However, the major trade-off between dedicated and batch delivery persists between this contingent policy and batch delivery. As batch serving always waits to accumulate two orders before dispatch, it can take advantage of a large demand rate setting in which the expected wait time to accumulate another order is shorter than a delivery trip with a single order. On the other hand, the contingent policy is better suited when the demand rate is relatively low, providing the flexibility to avoid long wait times for order accumulation.

To better analyze the performance of the contingent policy considered here or any other state-dependent delivery policy, we believe a dynamic program model is needed, and this is beyond the scope of this paper. We hope our discussion can stimulate future research in this direction.

### 6.3. General Arrival Rate

In this section, we investigate whether observations such as Propositions 3 and 4 still hold without the arrival rate being at the limit. To keep our results concise and informative, we assume that customers' valuations are uniformly distributed on $[0,1]$. That is, $F(v)=v$ for $v\in[0,1]$ and $F(v)=0$ otherwise. Note that our result does not anchor on the uniform distribution assumption. Statements in this section can also be generalized to more general valuation distributions. We leave the detailed discussion to Online Appendix C.2.

When the courier serves dedicated, the revenue maximization problem for the vendor is

$$\max_{\lambda\in[0,\mu_D)} V_D(\lambda, W_D(\lambda,\mu_D,C_D)), \tag{19}$$

where the constraint on the demand rate $\lambda$ reflects the load factor $\rho_D<1$ so that the system is stable. Similarly, when the courier serves batch, the maximization problem is

$$\max_{\lambda\in[0,2\mu_B)} V_B(\lambda, W_B(\lambda,\mu_B,C_B)), \tag{20}$$

where functions $V_D$ and $V_B$ are defined in (7) and (13), respectively; the constraint on $\lambda$ reflects $\rho_B<1$. Note that we do not include constraint $\lambda\le\Lambda r^2$ in either (19) or (20). The reason is that, for any demand rate greater than $\Lambda r^2$ (which is still mathematically possible), the corresponding revenue function has a negative value, so it cannot be optimal. The next two propositions summarize the results when the vendor optimizes its revenue according to (19) and (20).

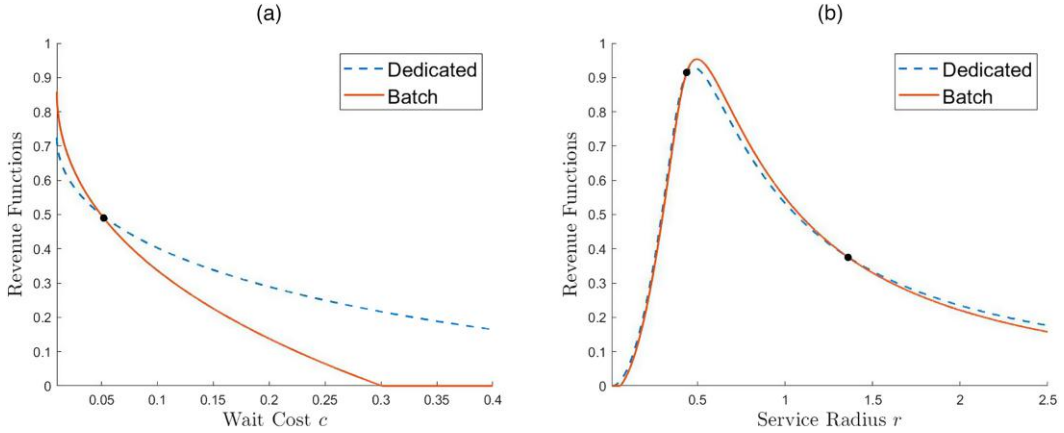
**Proposition 7.** *Fix $r,\Lambda>0$. Consider $F(v)=v$ for $v\in[0,1]$ and $F(v)=0$ otherwise. With the demand rate endogenized, there exists a threshold $c_{en}$ on the customers' wait cost parameter $c$ such that, for all $c\ge c_{en}$, it is optimal to serve dedicated.*

Proposition 7 complements the results in Proposition 3, assuming that each customer's valuation follows an independent standard uniform distribution. Even with the general arrival rates of customers, it is still optimal to serve dedicated when customers are impatient (i.e., $c$ is large enough). Unfortunately, it is challenging to demonstrate analytically that it is optimal with general arrival rates to serve batch when customers are very patient unlike the case in the limiting regime. With general arrival rates, both the distribution of customers' valuations and the expected wait time affect the overall revenue as mentioned in Section 5. The distribution of valuations determines the optimal base price, which, unlike the crowded market, is no longer independent of the demand. Furthermore, finite arrival rates may prevent the delivery system from achieving the optimal demand rate when customers are patient. This hurts serving batch specifically because the pooling strategy shines under a high demand rate, and its efficiency may not be fully exploited in this case. Moreover, the price compensation has to be significant to sustain a large demand rate with finite arrivals. However, we can still numerically verify that there exists a threshold on wait cost parameter $c$ below which is optimal to serve batch. Figure 4(a) provides a visual illustration: the optimal revenue functions of serving dedicated and batch only cross once.

**Proposition 8.** *Fix $c,\Lambda>0$ and constant $L$ such that $(\Lambda/c^3)>L$ (with the exact expression of constant $L$ provided in the online appendix). Consider $F(v)=v$ for $v\in[0,1]$ and $F(v)=0$ otherwise. With the demand rate endogenized, there exists a threshold $r_{en}$ on the service radius $r$ such that, for all $r\ge r_{en}$, it is optimal to serve dedicated.*

Proposition 8 extends the result in Proposition 4 when each customer's valuation follows an independent standard uniform distribution. We show that, with general customer arrival rates, it is still optimal to serve dedicated when the service radius is large enough. We only require an extra minor condition that either the arrival rate of customers is high enough or their wait cost parameter is low enough. Similar to Proposition 7, it is very difficult to establish optimal conditions for serving batch. In fact, in our numerical experiments, we find counterexamples in which it may not be optimal to serve batch when the radius is small. Instead, as in the counterexample shown in Figure 4(b), it is only optimal to serve batch when the service radius is medium. For sufficiently small or large service radii, it is always better

**Figure 4.** (Color online) Revenue Functions Under Dedicated and Batch Delivery



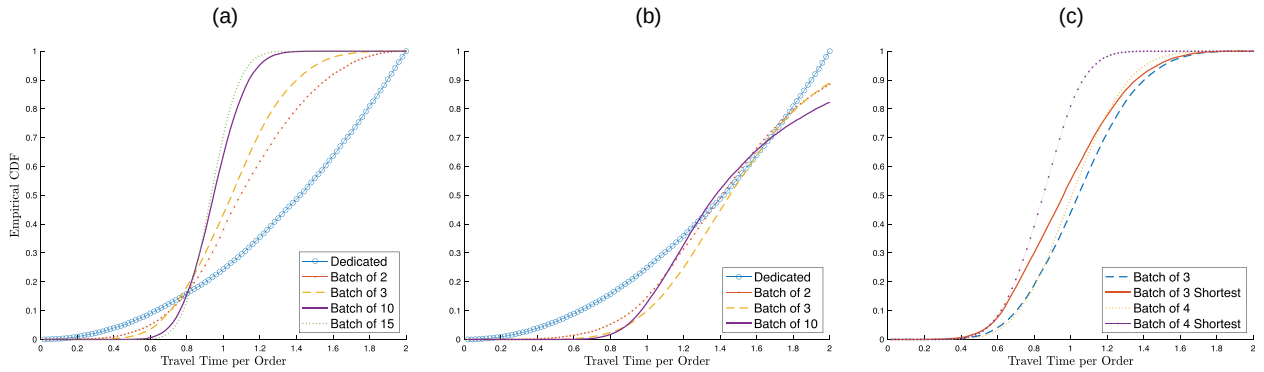

*Notes.* (a) $\bar{r} = 0.6$, $\underline{r} = 0.4$, $\Lambda = 25$. (b) $c = 0.2$, $\Lambda = 25$.

to serve dedicated. As mentioned, serving batch has the edge over dedicated when the demand rate is relatively high. When the service radius is sufficiently small, sustaining a high demand rate for both dedicated and batch is beneficial. However, because of the finite arrival rate of customers, the demand rate cannot reach the magnitude at which serving batch outperforms serving dedicated; otherwise, the price discount to sustain a high demand rate for batch delivery would be too great. This also explains why we only observe a single threshold on the service radius in Proposition 4 in the large market limiting regime.

### 6.4. Batch Size Greater Than Two

In our base model, we consider batches with the size of two given applications in food delivery to better illustrate the main trade-offs in our delivery policies. Here, we extend the model to a batch size greater than two and conduct numerical studies. Namely, we consider that each batch has a size of three or more. In Figure 5(a), we provide the empirical cumulative distribution functions on the courier's actual travel time per order when using batch with different sizes. As we can see, increasing the batch size can reduce the chance of experiencing long travel times per order. As a result, the overall service time is also reduced if we do not consider the time to form a batch and the in-line delay from queueing aspects. As we can see in Figure 5(b), just by adding the batch accumulation time for different sizes, increasing the batch size may not always be beneficial in improving efficiency. Even if we only consider the courier's travel time, this margin of improvement gets smaller as the batch size increases as shown in Figure 5(a). In addition, smaller batch sizes have higher chances of inducing a very short travel time when orders are near the vendor. In addition to these observations, we choose not to consider batch sizes greater than three for the following reasons. First, for any batch with a size greater than two, we need to consider proper routing policies in delivery, which is not the focus of this paper. When the batch size equals three, in the following, we consider that the courier delivers orders with a purely random ordering. But it is observed in Figure 5(c) that the gap between a random fulfillment policy versus a delivery

**Figure 5.** (Color online) Empirical Cumulative Distribution Functions of Travel Time per Order with Various Batch Sizes




*Notes.* (a) $r = 1$. (b) With order accumulation time $r = 1$, $\lambda = 1.5$. (c) Random versus shortest path $r = 1$.

policy based on the shortest path gets larger when the batch size increases. Second, as the batch size increases, it may be in the vendor's best interest to consider contingent policies as in Section 6.2, which we leave as a future research direction as they should be analyzed using nonstationary models. We acknowledge the potential shortcomings of our current approach for large batch sizes.

We analyze this system using an Erlang-3 arrival process. An arriving order does not enter the dispatch queue until a batch of three is formed. Then, batches have the arrival rate of $\lambda/3$ with the interbatch time following the Erlang-3 distribution. Similar to the derivations in (8), we have

$$\begin{aligned}\mathbb{E}[X_{3B}] &= \int_{\mathcal{A}} \left( \sqrt{x+y-2\sqrt{xy}\cos(\theta)} \right. \\ &\quad \left. + \sqrt{y+z-2\sqrt{yz}\cos(\phi)} + \sqrt{x} + \sqrt{z} \right) \frac{1}{\pi^2} \mathrm{d}\mathcal{A}, \\ \mathbb{E}[X_{3B}^2] &= \int_{\mathcal{A}} \left( \sqrt{x+y-2\sqrt{xy}\cos(\theta)} \right. \\ &\quad \left. + \sqrt{y+z-2\sqrt{yz}\cos(\phi)} + \sqrt{x} + \sqrt{z} \right)^2 \frac{1}{\pi^2} \mathrm{d}\mathcal{A}, \end{aligned} \tag{21}$$

where $\mathcal{A} = [0,1]^3 \times [0,\pi]^2$ and $\mathrm{d}\mathcal{A}/\pi^2$ is the measure of set $\mathcal{A}$ under the uniform distribution. Using these first and second moments, we can get the service rate and load factor as $\mu_{3B} = 1/\mathbb{E}[X_{3B}]$ and $\rho_{3B} = \lambda/(3\mu_{3B})$. Furthermore, the coefficient of variation is $C_{3B} = 1/3 + [\mathbb{E}[X_{3B}^2] - (\mathbb{E}[X_{3B}])^2)]/(\mathbb{E}[X_{3B}])^2$. As a result, the expected wait time can be calculated using

$$W_{3B}(\lambda,\mu,C) = \frac{1}{\lambda} + \frac{\lambda}{\mu(3\mu-\lambda)}\frac{C}{2} + \frac{1}{3}\left(\frac{4}{3} + \frac{128}{45\pi}\right) r, \tag{22}$$

where the first term is the average wait time an order has to wait to form a batch (an order needs to wait for zero, one, or two more orders with equal probability to form a batch), the second term is the in-line delay, and the last term is the extra delay if another order(s) in the batch needs to be delivered first. Then, $W_{3B}(\lambda,\mu_{3B},C_{3B})$ is the expected wait time.

In our numerical calculations, as shown in Figure 6, we always observe that there are single thresholds in the service radius $r$ and wait cost $c$, respectively, such that below which the vendor should serve with batches and above which dedicated delivery generates more revenue. This is consistent with our findings in the case with a batch size of two. Thus, our main insights are not limited by the simplification of considering batches with the size of two.

Consistent with Section 6.3, for general arrival rates, when the radius is large enough, the vendor should use dedicated delivery. We can again find numerical examples such that two thresholds exist on the service radius. Between these thresholds, serving batch outperforms dedicated delivery in terms of revenue maximization. However, we can also find extreme parameters such that batch delivery is completely dominated by dedicated delivery for all service radii. We believe that the possible inferior performance of batch delivery with a size greater than two is contributed to the random routing policy and lack of contingent policies as mentioned earlier, and they are beyond the scope of this paper and left for future research.

### 6.5. Circular Service Area

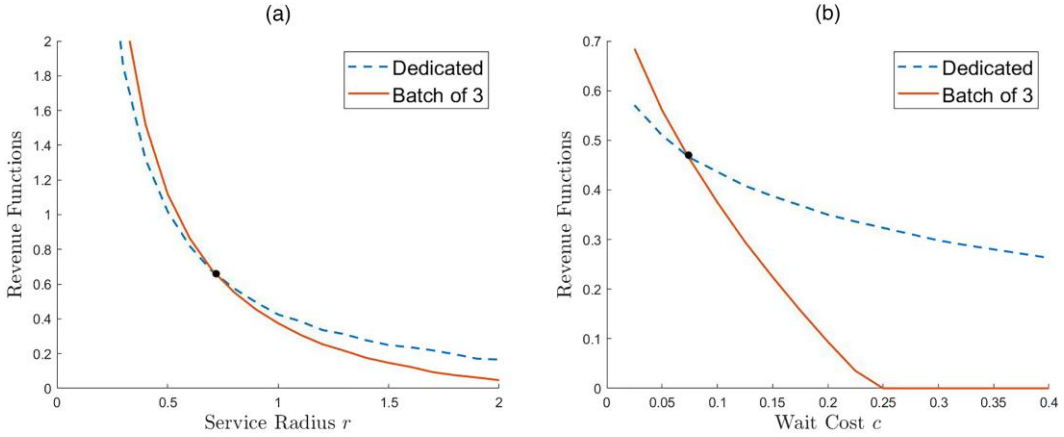
In this section, we consider a service area that only constitutes the edge of the disk, that is, the circumference of the circle. That is, we still have the facility located at the center of the disk, but orders are only coming from locations that are uniformly distributed on the edge of the disk with a radius $r$. This type of city structure has been examined by many researchers before, most notably by Salop (1979). The so-called circular city model has a lot of practical implications because many major cities have this kind of circular or ring structure (e.g., Beijing and Moscow). These cities have massive business areas in the inner rings with residential areas surrounding the city center in an outer ring. The circular city model also captures scenarios in which the storage warehouse is in a relatively remote area, and couriers have to travel long distances in each direction to reach the nearest residential area. Furthermore, it also serves as an extreme case in which customers' locations are not uniformly distributed inside the service area.

As in the previous sections, we propose an appropriate queueing system and analyze the delivery strategies. When serving dedicated, the service time for each order is deterministic because the time travel from the center to any point on the edge of the circle is fixed. Thus, this delivery system can be treated as an M/D/1 system under dedicated delivery. We still denote the demand rate for this system by $\lambda$ under the arrival rate $\Lambda r^2$. Then, this M/D/1 queue has the service rate and load factor as $\mu_{D,C} = 1/(2r)$, and $\rho_{D,C} = \lambda/\mu_{D,C} = 2\lambda r$, respectively, in which the subscript represents dedicated in a circular service area. The wait time for a customer in this system is simply characterized by function $W_D(\lambda, \mu_{D,C}, C_{D,C})$ in (5) with $C_{D,C} = 1$ because the arrival process is Poisson and the service process is deterministic. Furthermore, the revenue function can be written as

$$\begin{aligned} &V_D(\lambda, W_D(\lambda,\mu_{D,C},C_{D,C})) \\ &\qquad = \lambda\left[F^{-1}\left(1-\frac{\lambda}{\Lambda r^2}\right) - cW_D(\lambda,\mu_{D,C},C_{D,C})\right]. \end{aligned}$$

Next, we consider serving batch in a circular service area with radius $r$. Again, the arrival process has the interarrival time following an Erlang-2 distribution. The service time needs to include three parts: first, the travel time from the center of the circle to a random point on

**Figure 6.** (Color online) Revenue Functions When Serving Dedicated vs. Batch of Size Three Under a Large Market




*Notes.* (a) $c = 0.1$. (b) $r = 1$.

its edge; second, the travel time between two uniformly distributed points on the edge of the circle; and finally, the travel time from the edge of the circle back to the center. Denote by random variable $Y$ the distance a courier needs to travel per trip. Then, we have random variable $Y$ following a uniform distribution on $[2r, 2r + \pi r]$ with $\mathbb{E}[Y] = 2r + \pi r/2$, and $\sigma_Y^2 = (\pi r)^2/12$. Thus, this queueing system has the service rate and load factor as $\mu_{B,C} = 1/\mathbb{E}[Y] = 1/(2r + \pi r/2)$, and $\rho_{B,C} = \lambda/(2\mu_{B,C}) = \lambda r(4 + \pi)/4$, respectively, in which the subscript represents batch in the circular service area.

Next, again, we use Kingman's formula to approximate the average wait time for each customer. The average wait time for each order follows from $W_B(\lambda, \mu_{B,C}, C_{B,C})$ with $C_{B,C} = 1/2 + \sigma_Y^2/(\mathbb{E}[Y])^2 = 1/2 + \pi^2/[3(4 + \pi)^2]$. Furthermore, the revenue function in (13) incorporates the adjusted wait time as

$$V_B(\lambda, W_B(\lambda, \mu_{B,C}, C_{B,C})) = \lambda\left[F^{-1}\left(1 - \frac{\lambda}{\Lambda r^2}\right) - cW_B(\lambda, \mu_{B,C}, C_{B,C})\right], \ \lambda \in (0, 2\mu_{B,C}).$$

We find that all the major results in Sections 4 and 5 still hold even if we change the service area from a disk to a circle. We relegate the formal statements and detailed derivations to Online Appendix C.3 to avoid repetition. Because of the change in the city's geometry, no orders are coming from areas inside the disk. Thus, the courier must always travel a fair distance before reaching the delivery area (the outer ring). As a result, the thresholds for switching delivery strategy also change, though the threshold structure remains. The next proposition provides the relationship between thresholds in a circular city and those in the base model.

**Proposition 9.** *Assume a crowded market and suppose that the demand rate can be endogenized.*

i. *In a circular service area, there exist thresholds $\hat{c}_\infty$ and $\hat{r}_\infty$ such that the vendor should serve batch if the wait cost and radius parameters $c$ and $r$ fall below the thresholds, respectively. Otherwise, the vendor should serve dedicated.*

ii. *We have*

$$c_\infty > \hat{c}_\infty, \text{ and } r_\infty > \hat{r}_\infty, \tag{23}$$

*where $c_\infty$ and $r_\infty$ are thresholds in Propositions 3 and 4, respectively, for the counterpart of the base model serving the entire disk.*

Proposition 9 shows that both thresholds on the wait cost coefficient and service radius are lower if orders only come from the edge of the disk. The reason is that the courier needs to travel to the edge of the disk before benefiting from the pooling effect of batch delivery. Thus, serving dedicated has more advantages in this setting. This implies that dedicated delivery more likely is beneficial when the orders tend to be distributed on the outskirts of a service region than when they have a more uniform distribution inside the region.

### 6.6. Distance-Dependent Wait Time

Now, we consider an extension to the base model in which customers are sensitive not only to the in-line delay, but also to the courier's actual travel time. Because customers take the courier's traveling time into their wait time estimation, the order distribution on the disk is no longer uniform because customers from different locations may have different expected wait times.

**Proposition 10.** *With endogenized demand under a single price, the demand rate is nonincreasing with respect to the distance from the hub at the origin of the disk. Furthermore, the optimal pricing strategy may lead to a shrinkage in the service area. In other words, the vendor may abandon customers who live far away from the hub on purpose.*

Proposition 10 shows that, because of customers' sensitivity to the courier's traveling time, their demand is not uniformly distributed over the disk. Instead, customers

located further away from the hub order less than those who are closer to the hub. Further, when the service region is too large to begin with, because of a single price imposed over the entire disk, it may not be in the vendor's best interest to serve those who are too far away from the hub, resulting in that the service region is effectively shrunk to a smaller disk.

Unfortunately, it is difficult to provide any further analytical characterization of the delivery systems under this setting. Even numerical analysis or simulation is challenging to conduct under this setting because the demand rate is a market equilibrium outcome.[7] Aiming to get more analytical and numerical characterizations of the delivery systems, we consider a simplified service region.

Consider a service region in which customers' locations are uniformly distributed on two rings, an inner ring with radius $\underline{r}$ and an outer ring with radius $\overline{r} > \underline{r}$. The vendor is still located at the center of the rings. Thus, the courier needs to travel a longer distance serving customers who are located on the outer ring. We leave the detailed derivations under both dedicated and batch services to Online Appendix D as they are very cumbersome.

In Online Appendix D, we first analytically recover the major insights under exogenous demand and then conduct numerical analysis with endogenous demand under the simplified two-rings service region. Because the service region is separated into two rings, exogenous demands can potentially become combinatorial and, thus, we only focus on a single exogenous demand rate on both rings instead. We leave the formal statements to Online Appendix D. For endogenous demand, we compare the dedicated and batch service systems and numerically recover the insights in Section 5. In particular, when fixing the ratio between the radii of the rings, we observe a threshold on rings' radius above which serving dedicated delivery is better, as shown in Figure 7(a). We also observe that there is still a threshold on the wait cost coefficient above which the vendor should use dedicated delivery and below which the vendor should use batch delivery as shown in Figure 7(b).

### 6.7. Multiple Couriers

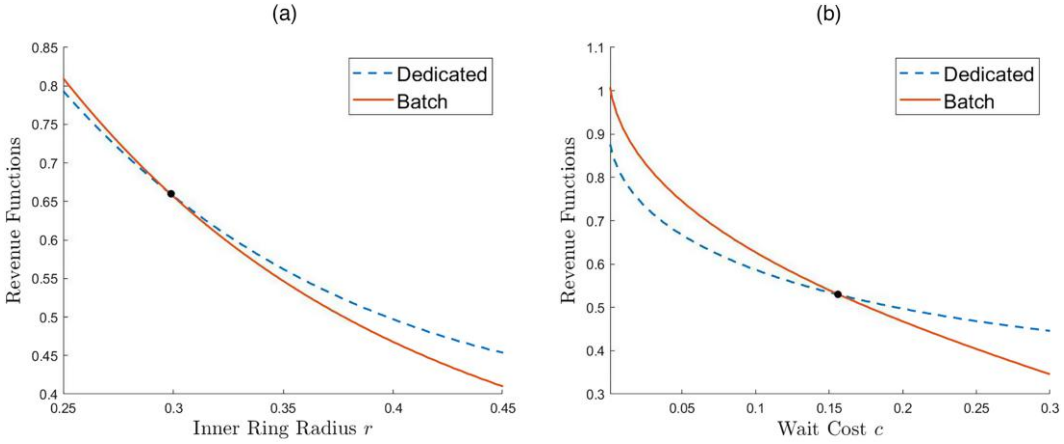
So far, we have focused on cases with a single courier. Suppose the vendor hires $k$ couriers to serve the disk-shaped area at the same time. We reassess the performance of dedicated versus batch delivery.

First, note that having $k$ couriers does not change the service process for each courier individually. Thus, when serving dedicated, the arrival process is still Poisson, and the service rate remains $\mu_D$ for each courier. However, the load factor is different because we have $k$ couriers instead of one. That is, we have $\rho_{D,k} = \lambda/(k\mu_D) = 4\lambda r/(3k)$, where the subscript of the load factor represents serving dedicated with $k$ couriers. Thus, this service system can be analyzed through an M/G/$k$ queue. To obtain a tractable expected wait time, we utilize two approximations together. Recall that the summation of coefficients of variation of the arrival and service processes is $C_D = 9/8$. We approximate the in-line delay of an M/G/$k$ queue as

$$W_q\{\mathrm{M/G}/k\} \approx \frac{C_D}{2} W_q\{\mathrm{M/M}/k\} \approx \frac{C_D}{2} \frac{\rho_{D,k}^{\sqrt{2(k+1)}}}{\lambda(1-\rho_{D,k})}, \tag{24}$$

where we first use an M/M/$k$ queue with the same input to approximate the in-line delay of the M/G/$k$ counterpart (see, e.g., Gross et al. 2008) and then use a well-studied approximation for the M/M/$k$ queue itself (see, e.g., Sakasegawa 1977). This approximation is consistent with recent studies on on-demand economy; see, for example, Bai et al. (2019), Taylor (2018), and Benjaafar et al. (2022). As the result of such an approximation,

**Figure 7.** (Color online) Revenue Functions Under Dedicated and Batch Delivery




*Notes.* (a) $\overline{r}/\underline{r} = 2$, $\Lambda = 25$. (b) $c = 0.2$, $\Lambda = 25$.

the expected wait time for each customer is

$$W_{D,k}(\lambda) \approx \frac{C_D}{2}\frac{\rho_{D,k}^{\sqrt{2(k+1)}}}{\lambda(1-\rho_{D,k})} = \frac{C_D}{2(k\mu_D-\lambda)}\left(\frac{\lambda}{k\mu_D}\right)^{\sqrt{2(k+1)}-1}. \tag{25}$$

After these setups, the revenue function is simply $V(\lambda, W_{F,k}(\lambda))$.

Next, we consider serving batch. Similar to serving dedicated, each courier's service rate $\mu_B = \frac{45\pi}{4r(32+15\pi)}$ remains the same, but the load factor needs to take $k$ couriers into consideration. That is, we have $\rho_{B,k} = \frac{\lambda}{2k\mu_B} = \frac{2\lambda r(32+15\pi)}{45k\pi}$. Using the same approximation method as in (24), which can be applied to G/G/$k$ systems as well, the expected wait time for each customer is

$$W_{B,k}(\lambda) \approx \frac{1}{2\lambda} + \frac{C_B}{2(2k\mu_B-\lambda)}\left(\frac{\lambda}{2k\mu_B}\right)^{\sqrt{2(k+1)}-1} + \frac{r}{3}. \tag{26}$$

It is worth pointing out that, under this approximation scheme, the queueing systems reduce to those in Section 3 in which $k = 1$ for both dedicated and batch.

With multiple couriers, we can show analytically that there is a threshold on the exogenous demand rate below which serving dedicated generates higher revenue and above which serving batch is more profitable. This result is consistent with Proposition 1. If the vendor can endogenize the demand, there is still a threshold on customers' wait cost parameter above which it is optimal to serve dedicated, consistent with Proposition 7. We leave the formal statements of these analytical results to Online Appendix C.4. Other results in Sections 4 and 5, such as Propositions 2 and 8, are very difficult to prove analytically with multiple couriers. However, we still observe these results in our numerical experiments.

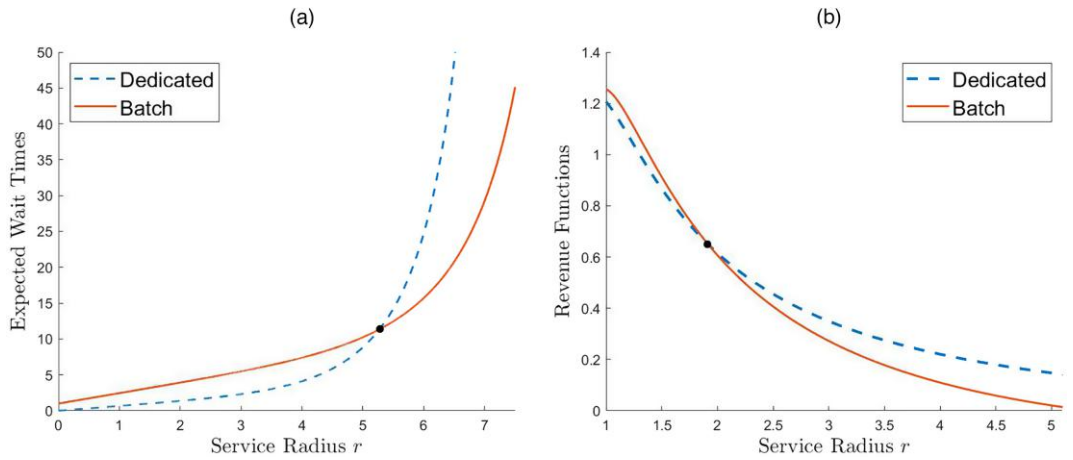
Figure 8(a) shows the relationship between the expected wait time when serving dedicated and batch. Similar to Proposition 2(i) and Figure 3(a), there appears to be a threshold on the service radius $r$ below which serving dedicated leads to a shorter wait time than serving batch when the demand rate $\lambda$ is fixed and above which it is the other way around. Furthermore, Figure 8(b) gives an example of the revenue function with service radius. As we can see, serving dedicated still outperforms batch when the service radius is large enough, just as in Proposition 8 in Section 5.

## 7. Conclusion

This paper compares and contrasts the fundamentals of using dedicated versus pooling delivery strategy. We model the two strategies as queueing systems serving dedicated and batch, respectively. In addition, we incorporate a spatial feature in these systems using a generalized circular city model. This spatial feature makes our service system relevant to the daily operations of the on-demand delivery industry. We highlight the scenarios in which dedicated or pooling delivery strategy is optimal, and our results remain robust in various extensions.

Our research contributes to the literature on innovative operations and smart cities. One of the major managerial insights is that, contrary to the common belief, temporal pooling, such as serving batch, may not always increase delivery efficiency in a large service area and lead to higher revenue for the vendor. When the vendor can endogenize the demand, a vendor should only use pooling when it can profitably sustain a relatively large demand rate. With impatient customers or a large service area, the vendor should use the dedicated strategy but charge a relatively high delivery price. We also contribute to the spatial queueing literature by providing an analytically tractable framework using a generalized circular city model, which is relevant to many practical applications. Our model accurately depicts delivery systems with a small number of orders per trip.

**Figure 8.** (Color online) Expected Wait Time and Revenue Functions with Multiple Couriers




*Notes.* (a) $k = 5$, and $\lambda = 0.5$. (b) $k = 3$, $c = 0.1$, and $\Lambda = 10$.

This paper can shed light on operational policies for on-demand delivery services in the emerging markets such as food or grocery delivery for which dedicated couriers or robots/drones are deployed to make deliveries, for warehouses where humans or robots/pods are pickers, and for an on-demand transportation service such as a micro-transit service from/to a subway station. Although our focus is mainly on investigating the benefit of temporal and spatial pooling in delivery, there are many other interesting research questions in the delivery business. Our modeling framework can potentially serve as a building block for future research in areas such as but not limited to contracting and compensation for couriers and incentive management with freelancers, for example, by endogenizing the number of dedicated couriers or freelancers in a shift through a wage decision or a payout contract, which is currently missing in the model.

Our work is not without limitations. First, given a general arrival rate, for some parameters, for example, when the wait cost parameter or the service radius has a sufficiently low value, we cannot obtain an unambiguous preference for the dedicated or pooling delivery strategy. For these parameters and a general valuation distribution, one needs to resort to a numerical comparison. Second, the empirical demand distribution is most likely not a uniform distribution. A data-driven approach needs to be adopted to prescribe the best strategy for a specific practical setting. Third, we assume that the couriers are employees of the vendor, and their delivery speed is independent of their workload. As mentioned, the compensation and behavioral issues for couriers may also need to be examined. Finally, we assume the firm commits to either the dedicated or pooling strategy as the resulting pricing and response time could be easily conveyed to consumers. In practice, the firm can improve its performance by making optimal contingent decisions about dispatching and routing depending on the realized locations of outstanding orders, which are outside the scope of our stylized model.

## Acknowledgments

The authors thank the department editor, the associate editor, and three anonymous reviewers for valuable comments and suggestions on the manuscript throughout the whole review process.

## Endnotes

[1] Gorillas employs a fleet of full-time bike couriers; see https://www.businessinsider.com/ultra-fast-grocery-delivery-startup-gorillas-to-launch-in-us-at-end-of-may-2021-5 and https://gorillas.io/en-us/ride-with-us. Gopuff hires deliver workers in advance to staff multihour shifts; see https://www.indeed.com/cmp/Gopuff/faq/working-hours.

[2] See also https://gadallon.substack.com/p/premature-scaling-will-gorillas-go.

[3] We can also assume that the arrival rate scales with the circumference of the circle, which is linear in $r$. That is, the arrival rate is $\Lambda r$. Our results still hold.

[4] According to an internal study by one of the largest delivery platforms in China, their couriers carry fewer than two orders per trip on average; see Figure 1.

[5] Note that the only approximation in Equation (8) is on the coefficient in the second moment, which is computed accurately using numerical integration.

[6] In Online Appendix E, we also consider a contingent policy based on order locations.

[7] Without analytical characterizations on the service system, such as the service rate and coefficient of variations, one needs to search for the equilibrium demand distribution over the disk. This process involves searching for an invariant function, which is very challenging computationally.

## References

Azi N, Gendreau M, Potvin J-Y (2012) A dynamic vehicle routing problem with multiple delivery routes. *Ann. Oper. Res.* 199(1): 103–112.

Bai J, So KC, Tang CS, Chen X, Wang H (2019) Coordinating supply and demand on an on-demand service platform with impatient customers. *Manufacturing Service Oper. Management* 21(3):556–570.

Benjaafar S, Ding J-Y, Kong G, Taylor T (2022) Labor welfare in on-demand service platforms. *Manufacturing Service Oper. Management* 24(1):110–124.

Berman O, Larson RC, Chiu S (1985) Optimal server location on network operating as an M/G/1 queue. *Oper. Res.* 33(4):746–771.

Berman O, Larson RC, Parkan C (1987) The stochastic queue $p$-median problem. *Transportation Sci.* 21(3):207–216.

Bertsimas DJ, van Ryzin G (1990) A stochastic and dynamic vehicle routing problem in the Euclidean plane. *Oper. Res.* 39(4):601–615.

Bertsimas DJ, van Ryzin G (1992) Stochastic and dynamic vehicle routing in the Euclidean plane with multiple capacitated vehicles. *Oper. Res.* 41(1):60–76.

Besbes O, Cachon G (2021) Do robots always win? Operational trade-offs in robot-driven fulfillment centers. *MSOM Conf.*

Buzzell RD, Gale BT, Sultan RGM (1975) Market share—A key to profitability. *Harvard Bus. Rev.* 97–106.

Cachon G (2014) Retail store density and the cost of greenhouse gas emissions. *Management Sci.* 60(8):1907–1925.

Cao J, Qi W (2022) Stall economy: The value of mobility in retail on wheels. *Oper. Res.* 71(2):708–726.

Cao J, Olvera-Cravioto M, Shen Z-JM (2020) Last-mile shared delivery: A discrete sequential packing approach. *Math. Oper. Res.* 45(4):1466–1497.

Carlsson JG, Song S (2017) Coordinated logistics with a truck and a drone. *Management Sci.* 64(9):4052–4069.

Chen M, Hu M, Wang J (2022) Food delivery service and restaurant: Friend or foe? *Management Sci.* 68(9):6539–6551.

Chen M, Sun P, Wan Z (2021) Matching supply and demand with mismatch-sensitive players. Preprint, submitted November 9, https://dx.doi.org/10.2139/ssrn.3458673.

Chen M, Chen M, Hu M, Wang J (2023) Distance-based fee design of on-demand delivery. Preprint, submitted July 9, https://dx.doi.org/10.2139/ssrn.4504691.

Cui S, Wang Z, Yang L (2020) The economics of line-sitting. *Management Sci.* 66(1):227–242.

Cui S, Wang Z, Yang L (2021) A model of queue-scalping. *Management Sci.* 67(11):6803–6821.

Daniels K, Turcic D (2021) Matching technology and competition in ride-hailing marketplaces. Preprint, submitted September 7, https://dx.doi.org/10.2139/ssrn.3918009.

Farahani MH, Dawande M, Janakiraman G (2022) Order now, pickup in 30 minutes: Managing queues with static delivery guarantees. *Oper. Res.* 70(4):2013–2031.

Feldman P, Frazelle AE, Swinney R (2022) Managing relationships between restaurants and food delivery platforms: Conflict, contracts, and coordination. *Management Sci.* 69(2):812–823.

Feng G, Kong G, Wang Z (2021) We are on the way: Analysis of on-demand ride-hailing. *Manufacturing Service Oper. Management* 23(5):1237–1256.

Gross D, Shortle JF, Thompson JM, Harris CM (2008) *Fundamentals of Queueing Theory*, 4th ed. (John Wiley & Sons, Hoboken, NJ).

Haddon H (2021) The pizza business is divided on delivery. *The Wall Street Journal* (May 30), https://www.wsj.com/articles/the-pizza-business-is-divided-on-delivery-11622367002.

He EJ, Goh J (2022) Profit or growth? Dynamic order allocation in a hybrid workforce. *Management Sci.* 68(8):5891–5906.

He L, Liu S, Shen Z-JM (2021) On-time last-mile delivery: Order assignment with travel-time predictors. *Management Sci.* 67(7):4095–4119.

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

Jargon J (2018) Starbucks to offer coffee delivery across U.S. *The Wall Street Journal* (December 14), https://www.wsj.com/articles/starbucks-to-offer-coffee-delivery-across-u-s-as-it-seeks-to-reach-more-customers-11544729828.

Kingman JFC (1962) Some inequalities for the queue GI/G/1. *Biometrika* 49(3/4):315–324.

Klapp MA, Erera AL, Toriello A (2018a) The dynamic dispatch waves problem for same-day delivery. *Eur. J. Oper. Res.* 271(2):519–534.

Klapp MA, Erera AL, Toriello A (2018b) The one-dimensional dynamic dispatch waves problem. *Transportation Sci.* 52(2):402–415.

Mak H (2022) Enabling smarter cities with operations management. *Manufacturing Service Oper. Management* 24(1):24–39.

Mao W, Ming L, Rong Y, Tang CS, Zheng H (2022) On-demand meal delivery platforms: Operational level data and research opportunities. *Manufacturing Service Oper. Management* 24(5):2535–2542.

Qi W, Li L, Shen Z-JM (2018) Shared mobility for last-mile delivery: Design, operational prescriptions, and environmental impact. *Manufacturing Service Oper. Management* 20(4):737–751.

Rana P, Haddon H (2021a) DoorDash and Uber Eats are hot. They're still not making money. *The Wall Street Journal* (May 28), https://www.wsj.com/articles/doordash-and-uber-eats-are-hot-theyre-still-not-making-money-11622194203.

Rana P, Haddon H (2021b) Restaurants and startups try to outrun Uber Eats and DoorDash. *The Wall Street Journal* (February 21), https://www.wsj.com/articles/restaurants-and-startups-try-to-outrun-uber-eats-and-doordash-11613903401.

Rana P, Kang J (2021) For DoorDash and Uber Eats, the future is everything in about an hour. *The Wall Street Journal* (May 31), https://www.wsj.com/articles/for-doordash-and-uber-eats-the-future-is-everything-in-about-an-hour-11622453401.

Sakasegawa H (1977) An approximation formula $L_q \approx \alpha\rho^{\beta}/(1-\rho)$. *Ann. Inst. Statist. Math.* 29(1):67–75.

Salop SC (1979) Monopolistic competition with outside goods. *Bell J. Econom.* 10(1):141–156.

Solomon H (1978) *Geometric Probability* (SIAM, Philadelphia).

Szymanski DM, Bharadwaj SG, Varadarajan PR (1993) An analysis of the market share-profitability relationship. *J. Marketing* 57(3):1–18.

Taylor T (2018) On-demand service platforms. *Manufacturing Service Oper. Management* 20(4):704–720.

Ulmer MW, Thomas BW, Mattfeld DC (2019) Preemptive depot returns for dynamic same-day delivery. *EURO J. Transportation Logist.* 8:327–361.

Ulmer MW, Thomas BW, Campell AM, Woyak N (2021) The restaurant meal delivery problem: Dynamic pickup and delivery with deadlines and random ready times. *Transportation Sci.* 55(1):75–100.

Voccia SA, Campbell AM, Thomas BW (2019) The same-day delivery problem for online purchases. *Transportation Sci.* 53(1):167–184.

Yildiz B, Savelsbergh M (2019) Service and capacity planning in crowd-sourced delivery. *Transportation Res. Part C.* 100:177–199.