Online Appendix to

"Proximity-Based Delivery Policies: Efficiency, Coverage, and Fairness"
by Zhongbin Wang, Shiliang Cui, Yongjian Li, and Qian Zhang

Table A.1 Detailed comparison between Wang et al. (2023) and our study

	Wang et al. (2023)	This paper
Model	$M/M/1$ queue where service time is independent of distance	$M/G/1$ queue where service time grows with distance
Context	Humanitarian services (e.g., healthcare, vaccination)	Commercial on-demand delivery services
Priority Rule	Prioritize distant customers	Prioritize nearby customers
Focal Question	How to expand service coverage and social welfare	How to balance efficiency, coverage (profit), and fairness
Main Results	Distant-priority boosts throughput (up to 50%)	Proximity-priority reduces average delay (up to 50%)

Appendix A: Extensions

In this Appendix, we present two extensions to our main model as robustness checks. Appendix A.1 relaxes the assumption of uniform customer proximity, and Appendix A.2 discusses the effects of a non-zero order preparation time prior to delivery.

Appendix A.1: Non-Uniform Customer Proximity

Our base model assumes a uniform distribution of delivery proximity over $[0, \bar{r}]$ for analytical tractability. In practice, however, demand may be non-uniform, clustering around restaurants or business centers, or spreading more broadly. To simulate how such non-uniform delivery patterns affect the effectiveness of our policies, this appendix considers that delivery proximities follow a general distribution over $[0, \bar{r}]$ with probability density function (PDF) $f(x)$ and cumulative distribution function (CDF) $F(x)$. For the effective delivery range $[0, r_e] \subseteq [0, \bar{r}]$, we define

$$E_1(r_e) \equiv \frac{1}{F(r_e)} \int_0^{r_e} x f(x)\, dx, \quad E_2(r_e) \equiv \frac{1}{F(r_e)} \int_0^{r_e} x^2 f(x)\, dx$$

as the first and second moments of the delivery proximity distribution within effective radius r_e. The service-time moments are given by

$$\begin{aligned}
\mathbb{E}[S] &= \frac{1}{F(r_e)} \int_0^{r_e} \mathbb{E}[2T(x)]\, dF(x) = 2E_1(r_e), \\
\mathbb{E}[S^2] &= \frac{1}{F(r_e)} \int_0^{r_e} \mathbb{E}\left[(2T(x))^2 \,\middle|\, x\right] dF(x) = 4(1+\xi)E_2(r_e), \\
\mathbb{E}[S_e] &= \frac{\mathbb{E}[S^2]}{2\mathbb{E}[S]} = \frac{(1+\xi)E_2(r_e)}{E_1(r_e)}.
\end{aligned}$$

Then the expected in-line delay under the $\mathcal{FCFS}$ policy can be calculated as

$$W_{r_e}^{\mathcal{F}} = \frac{\rho \mathbb{E}[S_e]}{1-\rho},$$

where $\rho = \lambda \mathbb{E}[S] = \Lambda F(r_e) \cdot 2E_1(r_e)$ denotes the system load.

Under a K-grade $\mathcal{PBD}$ policy, the scheduling policy is characterized by a proximity partition structure $\mathbf{r} = (r_1, r_2, \ldots, r_{K-1})$, where $0 = r_0 \le r_1 \le \cdots \le r_{K-1} \le r_K = r_e$ are the $K-1$ grade thresholds. Specifically, in grade $j \in \{1, 2, \ldots, K\}$, customer arrivals form a Poisson process with effective rate $\lambda_j \equiv \Lambda \cdot p_j = \Lambda \cdot [F(r_j) - F(r_{j-1})]$. The service time distribution for grade j has the first moment

$$\mathbb{E}[S_j] = \frac{1}{F(r_j) - F(r_{j-1})} \int_{r_{j-1}}^{r_j} \mathbb{E}[2T(x)]\, f(x)\, dx.$$

This yields a load and the corresponding expected in-line delay for grade j as follows:

$$\rho_j \equiv \lambda_j \mathbb{E}[S_j] = \Lambda \int_{r_{j-1}}^{r_j} \mathbb{E}[2T(x)] f(x)\, dx, \qquad W^K_{j,r_e} = \frac{\rho\, \mathbb{E}[S_e]}{(1 - \sum_{i=0}^{j-1} \rho_i)(1 - \sum_{i=0}^{j} \rho_i)}.$$

Then, the expected in-line delay of a customer located at x when $K \to \infty$ is

$$W^\infty_{r_e}(x) = \lim_{K \to \infty} W^K_{r_e}(x) = \frac{\rho \mathbb{E}[S_e]}{(1 - \Lambda \int_0^x 2t f(t)\, dt)^2},$$

and the overall average in-line delay of the system is

$$\overline{W}^\infty_{r_e} = \frac{1}{F(r_e)} \int_0^{r_e} W^\infty_{r_e}(x) f(x)\, dx.$$

In particular, we consider three representative Beta distributions with PDF

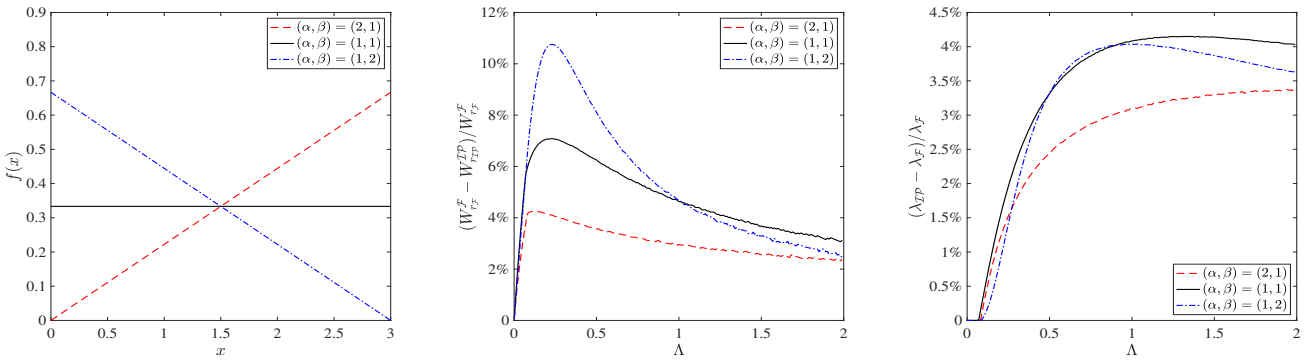
$$f(x) = \frac{x^{\alpha-1}(\bar{r} - x)^{\beta-1}}{\bar{r}^{\alpha+\beta-1} \mathrm{B}(\alpha, \beta)}, \quad x \in [0, \bar{r}].$$

When $\alpha = 2$ and $\beta = 1$ (or $\alpha = 1$ and $\beta = 2$), $f(x)$ is increasing (or decreasing) in x, capturing scenarios with a higher (or lower) density of distant customers. When $\alpha = \beta = 1$, the Beta distribution reduces to the uniform distribution. Figure A.1(a) illustrates the shapes of these three PDFs, and Table A.2 summarizes the parameters. Given $W^{\mathcal{IP}}_{r_{\mathcal{IP}}} = \overline{W}^\infty_{r_{\mathcal{IP}}}$, solving $u_{\mathcal{X}}(r_{\mathcal{X}}) = V - P - c(W^{\mathcal{X}}_{r_{\mathcal{X}}} + r_{\mathcal{X}}) = 0$ further derives the effective coverage $r_{\mathcal{X}}$ and corresponding demand $\lambda_{\mathcal{X}} = \Lambda \int_0^{r_{\mathcal{X}}} f(x)\, dx$ under policy $\mathcal{X} \in \{\mathcal{F}, \mathcal{IP}\}$. Notably, when demand is non-uniform, the $\mathcal{IP}$ policy's value on market expansion should be assessed by the relative demand gain $(\lambda_{\mathcal{IP}} - \lambda_{\mathcal{F}})/\lambda_{\mathcal{F}}$, rather than by coverage alone.

Table A.2 System parameters/performances with different (α, β) values

(α, β)	$f(x)$	λ	ρ	$\mathbb{E}[S_e]$
(2,1)	$\frac{2x}{\bar{r}^2}$	$\frac{\Lambda r_e^2}{\bar{r}^2}$	$\frac{4\Lambda r_e^3}{3\bar{r}^2}$	$\frac{3r_e}{4}$
(1,1)	$\frac{1}{\bar{r}}$	$\frac{\Lambda r_e}{\bar{r}}$	$\frac{\Lambda r_e^2}{\bar{r}}$	$\frac{2r_e}{3}$
(1,2)	$\frac{2(\bar{r}-x)}{\bar{r}^2}$	$\frac{\Lambda r_e(2\bar{r}-r_e)}{\bar{r}^2}$	$\frac{2(3\bar{r}-2r_e)\Lambda r_e^2}{3\bar{r}^2}$	$\frac{r_e(4\bar{r}-3r_e)}{2(3\bar{r}-2r_e)}$

Figure A.1 The value of $\mathcal{IP}$ policy under different demand distributions



(a) demand distribution (b) efficiency improvement (c) demand expansion

Note. $v=(V-P)/c=3$, $\bar{r}=3$, $\xi=0$.

Figure A.1(b)-(c) numerically demonstrate that the $\mathcal{IP}$ policy consistently improves efficiency and expands demand across all distributions, though the magnitude varies. The results of sensitivity analysis are as follows. The case $(\alpha,\beta)=(2,1)$ (i.e., red curve) yields the smallest gains, suggesting that in demand environments with sparse nearby but dense distant customers (i.e., dispersed demand), the benefits of the $\mathcal{IP}$ policy are relatively limited. By contrast, $(\alpha,\beta)=(1,2)$ (i.e., blue curve) achieves the largest improvements in both efficiency and demand when the potential arrival rate is moderate (i.e., in non-extreme scenarios). This implies that in clustered regions with dense nearby but sparse distant demand, the $\mathcal{IP}$ policy can generate substantial value.

Appendix A.2: Non-Zero Order Preparation Time

In the main model, we assumed that order preparation time was negligible, reflecting the common practice that preparation typically begins as soon as an order is placed, so it is mostly ready when the courier arrives. In practice, however, couriers may occasionally need to wait for an order at the vendor store due to large order volumes. In such cases, the courier remains idle until the order is ready, affecting the effective service time. To capture this phenomenon and enable practical implementation of a SEPT-based policy, we extend our model by introducing a random variable $t_p \geq 0$ representing the residual preparation time upon courier arrival. We denote its first and second moments by $\mathbb{E}[t_p]=E_1$ and $\mathbb{E}[t_p^2]=E_2$. The service time of an order is then $S=t_p+2x$, where $x\in[0,r_e]$ is the customer proximity. Under this specification, the service-time moments can be computed as $\mathbb{E}[S]=\frac{1}{r_e}\int_0^{r_e}\mathbb{E}[t_p+2x]\,dx=E_1+r_e$, and $\mathbb{E}[S^2]=\frac{1}{r_e}\int_0^{r_e}\mathbb{E}[(t_p+2x)^2]\,dx=E_2+2E_1r_e+\frac{4(1+\xi)}{3}r_e^2$, which gives $\mathbb{E}[S_e]=\frac{\mathbb{E}[S^2]}{2\mathbb{E}[S]}=\frac{E_2+2E_1r_e+4(1+\xi)r_e^2/3}{2(E_1+r_e)}$.

Under a K-grade preparatin-plus-travel-time-based scheme, the scheduling policy is characterized by a partition structure $\mathbf{s}=(s_1,s_2,\ldots,s_{K-1})$, where $s_j=t_p+r_j$ is the threshold for

grade $j \in \{1,2,...,K\}$, and $0 = r_0 \le r_1 \le \cdots \le r_{K-1} \le r_K = r_e$ are the corresponding thresholds for proximity. Since t_p is independent of travel time, the effective arrival rate in grade j remains $\lambda_j \equiv \lambda \cdot p_j = \lambda \cdot (r_j - r_{j-1})/r_e$, while the service time distribution has the first moment $\mathbb{E}[S_j] = \mathbb{E}[t_p] + \frac{1}{r_j - r_{j-1}} \int_{r_{j-1}}^{r_j} 2x\,dx = E_1 + r_{j-1} + r_j$. This leads to a traffic load $\rho_j \equiv \lambda_j \mathbb{E}[S_j] = \lambda(r_j - r_{j-1})(E_1 + r_j + r_{j-1})/r_e$. Given the system load $\rho = \lambda \mathbb{E}[S] = \delta r_e(E_1 + r_e)$, we can calculate the in-line delay under policy $\mathcal{X} \in \{\mathcal{F}, \mathcal{IP}\}$ by the following expressions:

$$W_{r_e}^{\mathcal{F}} = \frac{\rho \mathbb{E}[S_e]}{1-\rho}, \quad W_{r_e}^{\mathcal{IP}} = \frac{1}{r_e}\int_0^{r_e} W_{r_e}^{\infty}(x)dx = \frac{1}{r_e}\int_0^{r_e} \frac{\rho \mathbb{E}[S_e]}{(1-\delta x(E_1+x))^2}dx.$$

Substituting $r_e = r_{\mathcal{X}}$ into $W_{r_e}^{\mathcal{X}}$ and solving $u_{\mathcal{X}}(r_{\mathcal{X}}) = V - P - c\big(W_{r_{\mathcal{X}}}^{\mathcal{X}} + r_{\mathcal{X}}\big) = 0$ yield the effective coverage $r_{\mathcal{X}}$ under each policy.

We can prove that $r_{\mathcal{IP}} > r_{\mathcal{F}}$ holds. We first show that

$$\begin{aligned} W_{r_e}^{\mathcal{F}} = \frac{\rho \mathbb{E}[S_e]}{1-\rho} &> W_{r_e}^{\mathcal{IP}} = \frac{1}{r_e}\int_0^{r_e} \frac{\rho \mathbb{E}[S_e]}{(1-\delta x(E_1+x))^2}dx \\ \Leftrightarrow \frac{r_e}{1-\delta r_e(E_1+r_e)} &> \int_0^{r_e} \frac{1}{(1-\delta x(E_1+x))^2}dx \end{aligned}$$

holds for any $r_e > 0$ that satisfies $\rho = \delta r_e(E_1 + r_e) < 1$. The above inequality follows from a direct comparison argument: defining

$$F(x) = \frac{x}{1-\delta x(E_1+x)}, \quad G(x) = \frac{1}{(1-\delta x(E_1+x))^2},$$

we have

$$F'(x) = \frac{1+\delta x^2}{(1-\delta x(E_1+x))^2} > G(x)$$

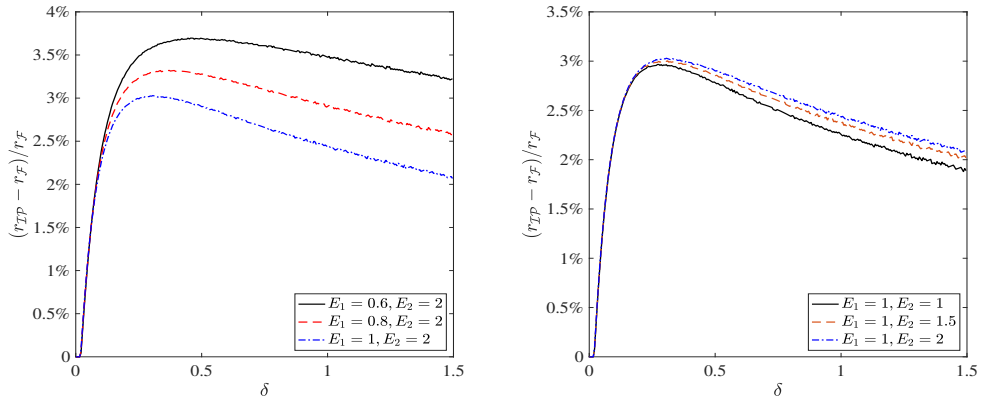
for all $x \in [0, r_e]$, which implies

$$\int_0^{r_e} G(x)\,dx < \int_0^{r_e} F'(x)\,dx = F(r_e) - F(0) = \frac{r_e}{1-\delta r_e(E_1+r_e)}.$$

Furthermore, since the effective coverage $r_{\mathcal{X}}$ under each policy $\mathcal{X} \in \{\mathcal{F}, \mathcal{IP}\}$ is determined by the indifference condition $u_{\mathcal{X}}(r_{\mathcal{X}}) = V - P - c\big(W_{r_{\mathcal{X}}}^{\mathcal{X}} + r_{\mathcal{X}}\big) = 0$, and the disutility $W_{r_{\mathcal{X}}}^{\mathcal{X}} + r_{\mathcal{X}}$ is increasing in $r_{\mathcal{X}}$, the inequality $W_{r_e}^{\mathcal{F}} > W_{r_e}^{\mathcal{IP}}$ immediately implies $r_{\mathcal{IP}} > r_{\mathcal{F}}$. Because if $u_{\mathcal{F}}(r_{\mathcal{F}}) = 0$ and $r_{\mathcal{IP}} < r_{\mathcal{F}}$, then $W_{r_{\mathcal{F}}}^{\mathcal{F}} > W_{r_{\mathcal{F}}}^{\mathcal{IP}} > W_{r_{\mathcal{IP}}}^{\mathcal{IP}}$ holds, implying $u_{\mathcal{IP}}(r_{\mathcal{IP}}) > 0$.

Figure A.2 further illustrates how non-zero order preparation times affect the relative value of the $\mathcal{IP}$ policy compared with the $\mathcal{F}$ policy. Relative to the main model (i.e., when $E_1 = E_2 = 0$) which achieves up to a 4.25% increase in market coverage, $\mathcal{IP}$'s value diminishes when preparation times are present. Specifically, Figure A.2(a) shows that, for a fixed second moment $E_2 = 2$, a larger mean residual preparation time E_1 reduces the relative advantage of the $\mathcal{IP}$ policy. Intuitively, when couriers spend more time idly waiting at vendors, the effective service

Figure A.2 The value of the $\mathcal{IP}$ policy under non-zero order preparation time



(a) Impact of E_1 (b) Impact of E_2

Note. $v \equiv (V - P)/c = 3, \xi = 0.$

rate decreases, diminishing the benefits of scheduling. This finding highlights the importance of actively reducing residual preparation times—for example, by optimizing kitchen workflows—to preserve dispatching effectiveness. Nevertheless, we can demonstrate the $\mathcal{IP}$ policy's value in market expansion, i.e., $\Delta_M = (r_{\mathcal{IP}} - r_{\mathcal{F}})/r_{\mathcal{F}} \geq 0$ always holds. Figure A.2(b) further examines the case where the mean is fixed at $E_1 = 1$ while the second moment E_2 varies. Here, increasing E_2—which reflects higher variability in residual preparation times—further amplifies the value of the $\mathcal{IP}$ policy, as prioritization helps mitigate extreme delays. In other words, in environments with unpredictable or highly variable preparation durations—such as during peak hours—implementing the $\mathcal{IP}$ policy becomes especially critical for improving system efficiency. This is consistent with the rationale that higher travel-time variability similarly enhances the effectiveness of the $\mathcal{IP}$ policy.

Notably, the above policy—based on the total expected service time defined as the sum of residual preparation time and delivery time—is non-preemptive. That is, once a courier is assigned to an order, the assignment remains fixed even if a newly arriving order would have a shorter service time while the courier is still waiting for the original order at the restaurant. This non-preemptive interpretation is consistent with many on-demand delivery settings, where reassignment after a pickup commitment is operationally difficult due to coordination frictions between vendors and couriers. If preemptive assignment is allowed, the introduced additional operational flexibility would substantially complicate the analytical characterization of the system. Under such a mechanism, the priority index of each order would depend on total service time of all orders, thus evolving continuously as preparation progresses and as new orders arrive. Consequently, the system state must track the residual preparation times and delivery times of all pending orders, together with the service commitments of couriers.

This leads to a high-dimensional state space and a state-dependent priority structure that changes over time. In addition, preparation and delivery occur in two operationally distinct stages—orders are first processed by the vendor and subsequently delivered by couriers—so courier assignment decisions effectively couple the preparation process with the delivery queue. The resulting system therefore departs from the class of tractable priority queues commonly analyzed in the queueing literature and generally does not admit a Markovian representation with low-dimensional sufficient statistics, particularly when service times are non-exponential. As a result, obtaining closed-form performance characterizations for the optimal dynamic dispatch policy becomes analytically challenging. For this reason, we focused on a non-preemptive priority rule in the paper that would apply at courier assignment epochs, which preserves tractability while still capturing the preparation delay. Importantly, even under this non-preemptive implementation, the proposed $\mathcal{IP}$ policy continues to outperform the benchmark $\mathcal{FCFS}$ policy, indicating that the mechanism retains meaningful operational value.

Appendix B: Model Calibration Using Other Data

In this Appendix, we conduct a direct comparison of key parameters from our mini case with those reported in Epstein et al. (2024) and further use their parameters to calibrate the performance of our policy. Epstein et al. (2024) used data from an on-demand grocery delivery company in Latin America to investigate how customer locations affect actual delivery times and customers purchasing behavior. This company operates its own fulfillment centers (FCs) and offers deliveries in as fast as 15 minutes. Because of the tight delivery promises, the company operates with hired labor (pickers and drivers) instead of gig workers.

Their data and summary statistics show that the average time from order placement to delivery completion (including both in-line delay and en-route travel time) was 12.38 minutes, with an average delivery distance of 1.27 km and a maximum delivery distance of $r_{\mathcal{F}} = 3.5$ km. Regression results further suggest that each additional kilometer of distance is associated with a 3.5-minute increase in delivery time, which corresponds to a delivery pace of $s = 3.5/60 = 0.058$ hours/km. Using these statistics, the implied average in-line delay is $W^{\mathcal{F}}_{r_{\mathcal{F}}} = 12.38/60 - (1.27 \times 0.058) = 0.132$ hours. We then calibrated our model by solving

$$W^{\mathcal{F}}_{r_{\mathcal{F}}} = \frac{2sr_{\mathcal{F}} \cdot s\delta r^2_{\mathcal{F}}}{3(1 - s\delta r^2_{\mathcal{F}})} = 0.132,$$

which yields a market density parameter of $\delta = 0.695$ orders/(hour$\times$km). Next, by solving the indifference condition $u_{\mathcal{F}}(r_{\mathcal{F}}) = V - P - c\big(W^{\mathcal{F}}_{r_{\mathcal{F}}} + s \cdot r_{\mathcal{F}}\big) = 0$, we obtained a reward-to-delay sensitivity ratio of $v = (V - P)/c = 0.335$ hours.

Finally, given these calibrated parameters, solving $u_{\mathcal{IP}}(r_{\mathcal{IP}}) = V - P - c\big(\overline{W}^{\infty}_{r_{\mathcal{IP}}} + s \cdot r_{\mathcal{IP}}\big) = 0$ derives $r_{\mathcal{IP}} = 3.62$ km under the Incentive-Compatible $\mathcal{PBD}$ ($\mathcal{IP}$) policy. This implies that adopting the $\mathcal{IP}$ policy would generate tangible performance improvements for this on-demand grocery delivery company: a 5.12% reduction in in-line delays, a 3.43% expansion in market coverage, and a 6.97% increase in consumer surplus.

The differences in key parameters as shown in Table 3 can be explained by the distinct delivery arrangements: in our case, Tastien relies on platform-based couriers, whereas Epstein et al.'s grocery delivery company employs its own drivers, enabling faster delivery speeds and thereby amplifying customers' expectations for instant delivery (reflected in a lower reward-to-delay sensitivity ratio).

Appendix C: Proofs

Proof of Proposition 1 To demonstrate this proposition, we begin with the simplest implementation of a $\mathcal{PBD}$ policy, namely a 2-grade $\mathcal{PBD}$ policy. Under a near-far threshold $\hat{r}_p$, the courier prioritizes delivering orders to customers closer to the vendor (i.e., $x \le \hat{r}_p$) over those farther away (i.e., $\hat{r}_p < x \le r_e$). The expected in-line delay for near and far orders are given as follows:

$$W^2_{1,r_e} = \frac{\rho \mathbb{E}[S_e]}{1-\rho_1} = \frac{2(1+\xi)\delta r_e^3}{3(1-\delta \hat{r}_p^2)} \quad \text{and} \quad W^2_{2,r_e} = \frac{\rho \mathbb{E}[S_e]}{(1-\rho_1)(1-\rho)} = \frac{2(1+\xi)\delta r_e^3}{3\left(1-\delta r_e^2\right)\left(1-\delta \hat{r}_p^2\right)}.$$

The system's average delay is then expressed as

$$\overline{W}^2_{r_e} = p_1 W^2_{1,r_e} + p_2 W^2_{2,r_e} = \frac{2(1+\xi)\delta r_e^3 (1-\delta r_e \hat{r}_p)}{3\left(1-\delta r_e^2\right)\left(1-\delta \hat{r}_p^2\right)},$$

which is directly influenced by the proximity partition threshold $\hat{r}_p$. Note that $\rho = \delta r_e^2/\bar{r} < 1$, we can easily verify that

$$\overline{W}^2_{r_e}(\hat{r}_p) - W^{\mathcal{F}}_{r_e} = \frac{2(1+\xi)\delta^2 r_e^3 \hat{r}_p (\hat{r}_p - r_e)}{3\left(1-\delta r_e^2\right)\left(1-\delta \hat{r}_p^2\right)} < 0$$

holds for any $\hat{r}_p \in (0, r_e)$. In addition, we can show that

$$\frac{\mathrm{d}\overline{W}^2_{r_e}(\hat{r}_p)}{\mathrm{d}\hat{r}_p} = -\frac{2(1+\xi)\delta^2 r_e^3 \left((r_e - 2\hat{r}_p) + \delta r_e \hat{r}_p^2\right)}{3\left(1-\delta r_e^2\right)\left(1-\delta \hat{r}_p^2\right)^2}.$$

Thus, $\overline{W}^2_{r_e}(\hat{r}_p)$ is decreasing first and then increasing in $\hat{r}_p$, which is minimized at $\hat{r}_p = \frac{1-\sqrt{1-\delta r_e^2}}{\delta r_e}$.

Next, we prove that $(K+1)$-grade proximity-based delivery can strictly dominates the K-grade proximity-based delivery. Specifically, under the K-grade proximity-based delivery, type-j customers are located at $x \in [r_{j-1}, r_j)$, where $r_0 = 0$, $r_K = r_e$, and $j = 1, 2, ..., K$. Thus, the expected in-line delay for type-j customers is

$$W^K_{j,r_e} = \frac{\rho \mathbb{E}[S_e]}{(1-\sum_{i=0}^{j-1}\rho_i)(1-\sum_{i=0}^{j}\rho_i)}.$$

Given $\mathbb{E}[S_j] = r_{j-1} + r_j$, $\rho_j = \lambda(r_j^2 - r_{j-1}^2)/r_e$ and $p_j = (r_j - r_{j-1})/r_e$, the average in-line delay is

$$\begin{aligned}\overline{W}_{r_e}^K &= \sum_{j=1}^{K} \frac{p_j \rho \mathbb{E}[S_e]}{(1-\sum_{i=0}^{j-1}\rho_i)(1-\sum_{i=0}^{j}\rho_i)} \\ &= \sum_{j=1}^{K} \frac{2(r_j - r_{j-1})\lambda r_e}{3(1-r_{j-1}^2\lambda/r_e)(1-r_j^2\lambda/r_e)} \\ &= \frac{2(1+\xi)\sqrt{\lambda} r_e^{3/2}}{3} \sum_{j=1}^{K} \frac{y_j - y_{j-1}}{(1-y_{j-1}^2)(1-y_j^2)},\end{aligned}$$

where $y_j = r_j\sqrt{\lambda/r_e}$. Thus, under policy $(y_1, y_2, \ldots, y_{K-1})$, we have

$$\overline{W}_{r_e}^K(y_1, y_2, \ldots, y_{K-1}) = \frac{2(1+\xi)\sqrt{\lambda} r_e^{3/2}}{3} \sum_{j=1}^{K} \frac{y_j - y_{j-1}}{(1-y_{j-1}^2)(1-y_j^2)}.$$

If we provide additional grade between thresholds y_0 and y_1, it becomes

$$\overline{W}_{r_e}^{K+1}(y_\epsilon, y_1, y_2, \ldots, y_{K-1}) = \frac{2(1+\xi)\sqrt{\lambda} r_e^{3/2}}{3} \left(\frac{y_\epsilon}{1-y_\epsilon^2} + \frac{y_1 - y_\epsilon}{(1-y_\epsilon^2)(1-y_1^2)} + \sum_{j=2}^{K} \frac{y_j - y_{j-1}}{(1-y_{j-1}^2)(1-y_j^2)} \right).$$

Then it is sufficient to prove that for any $y_\epsilon \in (0, y_1)$, we have

$$\frac{y_\epsilon}{(1-y_\epsilon^2)} + \frac{y_1 - y_\epsilon}{(1-y_\epsilon^2)(1-y_1^2)} < \frac{y_1}{1-y_1^2}.$$

Note that

$$\frac{y_\epsilon}{(1-y_\epsilon^2)} + \frac{y_1 - y_\epsilon}{(1-y_\epsilon^2)(1-y_1^2)} - \frac{y_1}{1-y_1^2} = \frac{y_1 y_\epsilon (y_\epsilon - y_1)}{(y_1^2-1)(y_\epsilon^2-1)} < 0$$

holds for any $y_\epsilon \in (0, y_1)$, which follows that $\overline{W}_{r_e}^{K+1}(y_\epsilon, y_1, y_2, \ldots, y_{K-1}) < \overline{W}_{r_e}^K(y_1, y_2, \ldots, y_{K-1})$. Given $(y_1^*, y_2^*, \ldots, y_{K-1}^*)$ to minimize $\overline{W}_{r_e}^K$, we have

$$\min \overline{W}_{r_e}^{K+1} \le \overline{W}_{r_e}^{K+1}(y_\epsilon, y_1^*, y_2^*, \ldots, y_{K-1}^*) < \min \overline{W}_{r_e}^K = \overline{W}_{r_e}^K(y_1^*, y_2^*, \ldots, y_{K-1}^*),$$

which proves Proposition 1(ii). Combining $\min \overline{W}_{r_e}^2 < W_{r_e}^{\mathcal{F}}$ and $\min \overline{W}_{r_e}^{K+1} < \min \overline{W}_{r_e}^K$ easily derives

$$\min \overline{W}_{r_e}^K < W_{r_e}^{\mathcal{F}}$$

for any $K \ge 2$, which proves Proposition 1(i). ■

Proof of Lemma 1 and Proposition 2 To identify the optimal scheduling policy, it is equivalent to solve Eq. (5a) - (5c).

The Lagrangian of this optimization problem is given by

$$\mathcal{L}(\mathbf{y}, \mathcal{Z}) = \sum_{j=1}^{K} \frac{y_j - y_{j-1}}{(1-y_{j-1}^2)(1-y_j^2)} + \sum_{j=1}^{K} \mathcal{Z}_j \cdot (y_{j-1} - y_j).$$

By KKT (Karush-Kuhn-Tucker) condition, we have $\partial\mathcal{L}(\mathbf{y},\mathcal{Z})/\partial y_j = 0$ and $\mathcal{Z}_j \cdot (y_{j-1} - y_j) = 0$ for $\mathcal{Z}_j \geq 0$ and $j = 1,2,...,K$. By Proposition 1(ii), we can deduce that the optimal thresholds are all attained the an interior points. This is because if there are k optimal values at the boundary, it will degrade to a $(K-k)$-grade scheduling policy, which implies that the $(K-k)$-grade policy dominates the K-grade policy, leading to a contradiction with Proposition 1(ii). Thus, $\mathcal{Z}_j = 0$ holds for $j = 1,2,...,K$ and the optimal values for $y_1, y_2, \ldots, y_{K-1}$ should adhere to their respective First-Order Conditions (FOCs), resulting in

$$y_j(y_{j-1} + y_{j+1}) = 1 + y_{j-1}y_{j+1} - \sqrt{(1-y_{j-1}^2)(1-y_{j+1}^2)}$$

for $j = 1,2,\ldots,K-1$. It follows that

$$\frac{y_{j+1} - y_j}{y_j - y_{j-1}} = \sqrt{\frac{1-y_{j+1}^2}{1-y_{j-1}^2}}. \tag{A.1}$$

Since $y_{j+1} > y_{j-1}$, it follows that $\frac{y_{j+1}-y_j}{y_j-y_{j-1}} < 1$, that is, $y_j - y_{j-1}$ is decreasing in j, which completes the proof of Lemma 1(ii).

Note that Eq. (A.1) can be rewritten as

$$\frac{y_{j+1} - y_j}{y_j - y_{j-1}} = \sqrt{\frac{(1-y_{j+1}^2)(1-y_j^2)}{(1-y_j^2)(1-y_{j-1}^2)}} \Leftrightarrow \frac{\sqrt{(1-y_j^2)(1-y_{j-1}^2)}}{y_j - y_{j-1}} = \frac{\sqrt{(1-y_{j+1}^2)(1-y_j^2)}}{y_{j+1} - y_j}.$$

That is, $\frac{\sqrt{(1-y_j^2)(1-y_{j-1}^2)}}{y_j-y_{j-1}}$ is a constant, denoted by $\mathcal{A}$, for $j = 1,2,\ldots,K-1$. This completes the proof of Lemma 1(ii). In addition, we can verify that $y_1 = \sqrt{1/(1+\mathcal{A}^2)}$, which decreases in $\mathcal{A}$. Assume that y_{j-1} is decreasing in $\mathcal{A}$. Then, given

$$\phi(y_j, y_{j-1}, \mathcal{A}) = \frac{\sqrt{(1-y_j^2)(1-y_{j-1}^2)}}{y_j - y_{j-1}} - \mathcal{A} = 0,$$

$\frac{\partial\phi(y_j,y_{j-1},\mathcal{A})}{\partial y_j} < 0$, and $\frac{\partial\phi(y_j,y_{j-1},\mathcal{A})}{\partial y_{j-1}} > 0$, we can conclude that, when y_j is fixed at $y_j(\mathcal{A}_1)$, for any $\mathcal{A}_1 < \mathcal{A}_2$, $\phi(y_j(\mathcal{A}_1), y_{j-1}(\mathcal{A}_1), \mathcal{A}_1) = 0 = \phi(y_j(\mathcal{A}_2), y_{j-1}(\mathcal{A}_2), \mathcal{A}_2) > \phi(y_j(\mathcal{A}_1), y_{j-1}(\mathcal{A}_2), \mathcal{A}_2)$ holds, which follows that $y_j(\mathcal{A}_2) < y_j(\mathcal{A}_1)$. This implies that y_j is decreasing in $\mathcal{A}$, i.e., Lemma 1(iii). Thus, we have

$$y_j = \frac{\mathcal{A}^2 y_{j-1} + \sqrt{\left(1+y_{j-1}^2\right)^2 + \mathcal{A}^2\left(1-y_{j-1}^4\right)}}{\mathcal{A}^2 + y_{j-1}^2 + 1},$$

with $y_0 = 0$ and $y_K = \sqrt{\lambda r_e}$, which completes the proof of Proposition 2. ■

Proof of Corollary 1 Note that $y_1 = \sqrt{1/(1+\mathcal{A}^2)}$, since $y_1 < y_K = \sqrt{\lambda r_e}$ and $y_1 > y_K/K$ by Lemma 1/(2), we can identify the upper bound and lower bound of $\mathcal{A}$ immediately. ■

Proof of Proposition 3 Note that for $x \in [r_{j-1}, r_j)$, we have

$$\frac{\rho \mathbb{E}[S_e]}{(1-\sum_{i=0}^{j-1}\rho_i)^2} < W_{r_e}^K(x) < \frac{\rho \mathbb{E}[S_e]}{(1-\sum_{i=0}^{j}\rho_i)^2}.$$

Recall by Lemma 1 that there exists a positive number $\mathcal{A}$ such that the optimal thresholds satisfy

$$\frac{\sqrt{(1-y_j^2)(1-y_{j-1}^2)}}{y_j - y_{j-1}} = \mathcal{A} \quad \text{for} \quad j = 1, 2, \ldots, K.$$

We can show that $\lim_{K\to\infty} \mathcal{A} = +\infty$. To substantiate this conclusion, it suffices to demonstrate that as K approaches infinity, the distance between at least two adjacent threshold values tends to zero. Otherwise, if all adjacent threshold values are greater than a positive number x, then $1 = \sum_{i=1}^{\infty} y_i - y_{i-1} > \sum_{i=1}^{\lceil 1/x \rceil + 1} y_i - y_{i-1} = y_{\lceil 1/x \rceil + 1} > 1$, leading to a contradiction. It indicates that

$$\lim_{K\to\infty} \frac{\sqrt{(1-y_j^2)(1-y_{j-1}^2)}}{y_j - y_{j-1}} = +\infty \quad \text{for} \quad j = 1, 2, \ldots, K,$$

which implies that $\lim_{K\to\infty} y_j^2 - y_{j-1}^2 = \lim_{K\to\infty} (y_j - y_{j-1})(y_j + y_{j-1}) = \lim_{K\to\infty} \rho_j = 0$ for all $j \in \mathbb{N}$. Thus, by Squeeze Theorem, the expected waiting time of customer at location x is

$$W_{r_e}^{\infty}(x) = \lim_{K\to\infty} W_{r_e}^K(x) = \frac{\rho \mathbb{E}[S_e]}{(1-\lambda x^2/r_e)^2}.$$

The average inline-delay is given by

$$\overline{W}_{r_e}^{\infty} = \frac{1}{r_e}\int_0^{r_e} W_{r_e}^{\infty}(x)dx = \frac{(1+\xi)r_e}{3}\left(\frac{\lambda r_e}{1-\lambda r_e} + \frac{\sqrt{\lambda r_e}}{2}\cdot \log\left(\frac{1+\sqrt{\lambda r_e}}{1-\sqrt{\lambda r_e}}\right)\right),$$

which completes the proof. ■

Proof of Proposition 4 Given $\rho = \lambda r_e \in (0,1)$, the $\mathcal{PBD}$ policy can reduce the system's average delay by

$$\Delta_1(\rho) = \frac{W_{r_e}^{\mathcal{F}} - \overline{W}_{r_e}^{\infty}}{W_{r_e}^{\mathcal{F}}} = \frac{1}{2} - \frac{(1-\rho)\log\left(\frac{1+\sqrt{\rho}}{1-\sqrt{\rho}}\right)}{4\sqrt{\rho}}.$$

Note that

$$\frac{d\Delta_1(\rho)}{d\rho} = \frac{\log\left(\frac{1+\sqrt{\rho}}{1-\sqrt{\rho}}\right) - 2\sqrt{\rho}/(1+\rho)}{8\rho^{3/2}/(1+\rho)} > 0$$

for all $\rho \in (0,1)$, it follows that $\Delta_1(\rho)$ is maximized at $\rho \to 1^-$. Thus, we have $\lim_{\rho\to 1^-} \Delta_1(\rho) = 0.5$ by L'Hospital's rule. In other words, the total delay can be reduced by up to half, which completes the proof of Proposition 4 (i).

While for individual customers, whose location is at no more than $\tilde{x}$ can enjoy lower expected in-line delay under the $\mathcal{PBD}$ policy than the $\mathcal{FCFS}$ policy. Note that

$$W_{r_e}^{\infty}(x) < W_{r_e}^{\mathcal{F}} \Leftrightarrow x < \tilde{x} = r_e\sqrt{\frac{1-\sqrt{1-\rho}}{\rho}},$$

the $\mathcal{PBD}$ policy can reduce the expected delay for $\Delta_2(\rho)$ of customers, where

$$\Delta_2(\rho) = \frac{\tilde{x}}{r_e} = \frac{1}{\sqrt{1+\sqrt{1-\rho}}}.$$

Similarly, given that $\frac{d\Delta_2(\rho)}{d\rho} > 0$, it follows that $\Delta_2(\rho)$ is minimized at $\rho \to 0^+$ and maximized at $\rho \to 1^-$, which completes the proof of Proposition 4 (ii). ■

Proof of Lemma 2 Note that $r_{\mathcal{F}}$ and $r_{\mathcal{P}}$ are the unique solutions for $u_{\mathcal{F}}(r_{\mathcal{F}}) = 0$ and $u_{\mathcal{P}}(r_{\mathcal{P}}) = 0$, respectively. Assume $r_{\mathcal{P}} \geq r_{\mathcal{F}}$, then we have

$$W^{\infty}_{r_{\mathcal{P}}}(x = r_{\mathcal{F}}) > W^{\mathcal{F}}_{r_{\mathcal{F}}} \Leftrightarrow \Lambda\left(\bar{r} - \Lambda r_{\mathcal{F}}^2\right)^2\left(\bar{r}\left(r_{\mathcal{P}}^3 - r_{\mathcal{F}}^3\right) + \Lambda r_{\mathcal{F}}^5\right) > 0,$$

which leads to

$$u_{\mathcal{P}}(r_{\mathcal{F}}) < u_{\mathcal{F}}(r_{\mathcal{F}}) = 0.$$

Furthermore, given that $\frac{du_{\mathcal{P}}(x)}{dx} = -c - \frac{8(1+\xi)\Lambda^2\bar{r}r_{\mathcal{P}}^3 x}{3\left(r-\Lambda x^2\right)^3} < \frac{du_{\mathcal{F}}(x)}{dx} = -c < 0$ and $u_{\mathcal{P}}(r_{\mathcal{P}}) = 0$, we can deduce $u_{\mathcal{P}}(r_{\mathcal{F}}) > 0$, which makes a contradiction. Then, $r_{\mathcal{P}} < r_{\mathcal{F}}$ can be proved. Besides, combining this result with $\frac{d\overline{W}^{\infty}_{r_e}}{dr_e} > 0$ and $\overline{W}^{\infty}_{r_{\mathcal{F}}} < W^{\mathcal{F}}_{r_{\mathcal{F}}}$ further deduces $\overline{W}^{\infty}_{r_{\mathcal{P}}} < \overline{W}^{\infty}_{r_{\mathcal{F}}} < W^{\mathcal{F}}_{r_{\mathcal{F}}}$. ■

Proof of Theorem 1 First, the IC constraint (7c) implies

$$\begin{aligned}
&u_{\mathcal{IP}}(x) > \bar{u}_{\mathcal{IP}}(x) \\
&\Leftrightarrow V - P - c\left(W^{\infty}_{r_{\mathcal{IP}}}(x) + x\right) - D(x; r_{\mathcal{IP}}) > V - P - c\left(\overline{W}^{\infty}_{r_{\mathcal{IP}}} + x\right) \\
&\Leftrightarrow D(x; r_{\mathcal{IP}}) < c\left[\overline{W}^{\infty}_{r_{\mathcal{IP}}} - W^{\infty}_{r_{\mathcal{IP}}}(x)\right].
\end{aligned}$$

Then, according to fairness constraint (7d), the proximity-based priority fee should be set at $D(x; r_{\mathcal{IP}}) = c\left[\overline{W}^{\infty}_{r_{\mathcal{IP}}} - W^{\infty}_{r_{\mathcal{IP}}}(x)\right] - \epsilon$, where $\epsilon > 0$ is independent of x. Consequently, we can obtain the form of proximity-based priority fee as follow:

$$D(x; r_{\mathcal{IP}}) = \begin{cases} c\left[\overline{W}^{\infty}_{r_{\mathcal{IP}}} - W^{\infty}_{r_{\mathcal{IP}}}(x)\right] - \epsilon > 0, & \text{if } x \in [0, \tilde{r}_{\mathcal{IP}}(\epsilon)], \\ c\left[\overline{W}^{\infty}_{r_{\mathcal{IP}}} - W^{\infty}_{r_{\mathcal{IP}}}(x)\right] - \epsilon < 0, & \text{if } x \in (\tilde{r}_{\mathcal{IP}}(\epsilon), r_{\mathcal{IP}}], \end{cases}$$

where $\tilde{r}_{\mathcal{IP}}(\epsilon) = \max\{x \mid D(x; r_{\mathcal{IP}}) \geq 0\}$.

Also, the internal transfer constraint (7e) then requires

$$\int_0^{r_{\mathcal{IP}}} D(x; r_{\mathcal{IP}})dx = c\left[\overline{W}^{\infty}_{r_{\mathcal{IP}}} - \int_0^{r_{\mathcal{IP}}} W^{\infty}_{r_{\mathcal{IP}}}(x)dx\right] - \epsilon = -\epsilon = 0,$$

implying $\epsilon \to 0^+$. By letting $\epsilon \to 0^+$, we have $\lim_{\epsilon \to 0^+} \tilde{r}_{\mathcal{IP}}(\epsilon) = \max\{x \mid W^{\infty}_{r_{\mathcal{IP}}}(x) \leq \overline{W}^{\infty}_{r_{\mathcal{IP}}}\} \equiv \tilde{r}_{\mathcal{IP}}$. This threshold $\tilde{r}_{\mathcal{IP}}$ partitions customers into two groups, each facing two service options:

- For **nearby customers** $x \in [0, \tilde{r}_{\mathcal{IP}}]$, the vendor offers the choice to (i) pay an additional amount $D(x; r_{\mathcal{IP}})$ to obtain higher priority, or (ii) reject this option, and be served as if randomly drawn from $[0, r_{\mathcal{IP}}]$.
- For **distant customers** $x \in (\tilde{r}_{\mathcal{IP}}, r_{\mathcal{IP}}]$, the vendor offers the choice to (i) accept a bonus $-D(x; r_{\mathcal{IP}})$ and receive a relatively lower priority, or (ii) decline and be randomly served.

Next, substituting $D(x; r_{\mathcal{IP}})$ into the IR constraint (7b) derives

$$u_{\mathcal{IP}}(r_{\mathcal{IP}}) = V - P - c\left(\overline{W}^{\infty}_{r_{\mathcal{IP}}} + r_{\mathcal{IP}}\right) = 0, \tag{A.2}$$

which yields the market coverage $r_{\mathcal{IP}}$. Since SEPT minimizes the average delay for a given effective demand (Wolff 1989, Ayhan and Olsen 2000) and increased demand exacerbates the delay, it follows that the market coverage $r_{\mathcal{IP}}$ satisfying Eq. (A.2) is maximized.

Finally, we can verify that the IC constraint (7c) holds since $u_{\mathcal{IP}}(x) = \bar{u}_{\mathcal{IP}}(x) + \epsilon > \bar{u}_{\mathcal{IP}}(x)$ for any $\epsilon > 0$, and remains valid when $\epsilon = 0^+$ by the sign-preserving property. Consequently, $D(x; r_{\mathcal{IP}})$ induces all customers to participate in the $\mathcal{IP}$ policy voluntarily. ■

Proof of Lemma 3 We assume $r_{\mathcal{IP}} \le r_{\mathcal{F}}$. According to Proposition 1, we can conclude that for any given r_e, $\mathcal{PBD}$ must result in a lower average delay than $\mathcal{FCFS}$, that is, $\overline{W}^{\infty}_{r_e} < W^{\mathcal{F}}_{r_e}$. Besides, note that

$$\frac{dW^{\mathcal{F}}_{r_{\mathcal{F}}}}{dr_{\mathcal{F}}} = \frac{2(1+\xi)\rho_{\mathcal{F}}(3-\rho_{\mathcal{F}})}{3(1-\rho_{\mathcal{F}})^2} > 0$$

and

$$\frac{d\overline{W}^{\infty}_{r_{\mathcal{IP}}}}{dr_{\mathcal{IP}}} = (1+\xi)\left[2\left(\frac{1}{1-\rho_{\mathcal{IP}}}\right)^2 + \sqrt{\rho_{\mathcal{IP}}}\log\left(\frac{1+\sqrt{\rho_{\mathcal{IP}}}}{1-\sqrt{\rho_{\mathcal{IP}}}}\right) - 2\right] > 0,$$

where $\rho_{\mathcal{F}} = \delta r^2_{\mathcal{F}} < 1$ and $\rho_{\mathcal{IP}} = \delta r^2_{\mathcal{IP}} < 1$, thus $\overline{W}^{\infty}_{r_{\mathcal{IP}}} < W^{\mathcal{F}}_{r_{\mathcal{F}}}$ must hold when $r_{\mathcal{IP}} \le r_{\mathcal{F}}$. Then, combining $u_{\mathcal{IP}}(x) - u_{\mathcal{F}}(x) = c(W^{\mathcal{F}}_{r_{\mathcal{F}}} - \overline{W}^{\infty}_{r_{\mathcal{IP}}}) > 0$ and $u_{\mathcal{F}}(r_{\mathcal{F}}) = 0$ can deduce $u_{\mathcal{IP}}(r_{\mathcal{F}}) > 0$. Furthermore, given that $\frac{du_{\mathcal{IP}}(x)}{dx} = \frac{du_{\mathcal{F}}(x)}{dx} = -c < 0$ and $r_{\mathcal{IP}} \le r_{\mathcal{F}}$, we have $u_{\mathcal{IP}}(r_{\mathcal{IP}}) \ge u_{\mathcal{IP}}(r_{\mathcal{F}}) > 0$, which contradicts $u_{\mathcal{IP}}(r_{\mathcal{IP}}) = 0$. Then, $r_{\mathcal{IP}} > r_{\mathcal{F}}$ can be proved. Besides, combining this result with $\frac{du_{\mathcal{IP}}(x)}{dx} = \frac{du_{\mathcal{F}}(x)}{dx} = -c < 0$, we have $u_{\mathcal{IP}}(x) > u_{\mathcal{F}}(x)$, which follows that $\overline{W}^{\infty}_{r_{\mathcal{IP}}} < W^{\mathcal{F}}_{r_{\mathcal{F}}}$. ■

Proof of Theorem 2 Since we have proved $r_{\mathcal{IP}} > r_{\mathcal{F}}$ and $\frac{du_{\mathcal{IP}}(x)}{dx} = \frac{du_{\mathcal{F}}(x)}{dx} < 0$, $u_{\mathcal{IP}}(x) > u_{\mathcal{F}}(x)$ must hold. It means that each customer at location x can improve its utility. And $r_{\mathcal{IP}} > r_{\mathcal{F}}$ indicates that more customers will enter the system, thus the vendor can benefit from the increased system throughput. ■

Proof of Proposition 5 This proposition quantifies the benefits of $\mathcal{IP}$ in comparison with $\mathcal{F}$. Specifically, the market coverage can be improved by

$$\Delta_M = \frac{r_{\mathcal{IP}} - r_{\mathcal{F}}}{r_{\mathcal{F}}}.$$

We examine two extreme cases of δ.

(i) $\delta \to \infty$. Since $\rho_{\mathcal{F}} = \delta r_{\mathcal{F}}^2 < 1$ and $\rho_{\mathcal{IP}} = \delta r_{\mathcal{IP}}^2 < 1$ must be satisfied, we have $r_{\mathcal{IP}} \to 0$ and $r_{\mathcal{F}} \to 0$ in this case. It follows that $\Delta_M = 0$.

(ii) $\delta \to 0^+$. Note that $r_{\mathcal{IP}}$ and $r_{\mathcal{F}}$ are the unique solutions of $V - P - c(\overline{W}^{\infty}_{r_{\mathcal{IP}}} + r_{\mathcal{IP}}) = 0$ and $V - P - c(W^{\mathcal{F}}_{r_{\mathcal{F}}} + r_{\mathcal{F}}) = 0$, we have $\overline{W}^{\infty}_{r_{\mathcal{IP}}} \to 0$ and $W^{\mathcal{F}}_{r_{\mathcal{F}}} \to 0$ in this case. It follows that $r_{\mathcal{IP}} \to (V-P)/c$ and $r_{\mathcal{F}} \to (V-P)/c$, implying that $\Delta_M = 0$.

Define $v \equiv (V-P)/c$, $V - P - c(\overline{W}^{\infty}_{r_{\mathcal{X}}} + r_{\mathcal{X}}) = 0$ is equivalent to $v = \overline{W}^{\infty}_{r_{\mathcal{X}}} + r_{\mathcal{X}}$ for $\mathcal{X} \in \{\mathcal{F}, \mathcal{IP}\}$. Similar to two cases of δ, we have (i) $r_{\mathcal{X}} \to 0$ and $\Delta_M = 0$ hold when $v \to 0^+$; (ii) $r_{\mathcal{X}} \to \bar{r}$ and $\Delta_M = 0$ hold when $v \to \infty$.

Combining the above cases and the proved result of $r_{\mathcal{IP}} > r_{\mathcal{F}}$ (i.e., $\Delta_M > 0$), we can conclude that Δ_M is non-monotone in δ and v. Note that $\frac{du_{\mathcal{IP}}(x)}{dx} = \frac{du_{\mathcal{F}}(x)}{dx} = -c < 0$, the improvement of customer surplus $\Delta_{CS} = \Delta_M^2 + \Delta_M$ is also non-monotone in δ and v. ■

Proof of Corollary 2 Note that

$$\frac{r_{\mathcal{IP}}}{r_{\mathcal{F}}} > \frac{r_{\mathcal{IP}}}{\tilde{r}_{\mathcal{F}}} = \frac{r_{\mathcal{IP}} \frac{dW^{\mathcal{F}}_{r_e}}{dr_e}|_{r_e = r_{\mathcal{IP}}}}{r_{\mathcal{IP}} \frac{dW^{\mathcal{F}}_{r_e}}{dr_e}|_{r_e = r_{\mathcal{IP}}} - \left(W^{\mathcal{F}}_{r_{\mathcal{IP}}} + r_{\mathcal{IP}}\right) + \left(\overline{W}^{\infty}_{r_{\mathcal{IP}}} + r_{\mathcal{IP}}\right)} = \frac{1}{1 - F(x)},$$

where

$$F(x) = \frac{(1-x^2)\left(2x - (1-x^2)\log(\frac{1+x}{1-x})\right)}{4(3-x^2)}$$

and $x = r_{\mathcal{IP}}\sqrt{\delta} \in (0,1)$, we have $\sup F(x) = 0.039$. Given that $\Delta_M > \frac{1}{1-F(x)} - 1$, then $\sup \Delta_M$ can be at least 0.04. ■

Proof of Proposition 6 Note that

$$u_{\mathcal{IP}}(r_{\mathcal{IP}}) = 0 \Leftrightarrow P = V - c(W^{\mathcal{IP}}(r_{\mathcal{IP}}) + r_{\mathcal{IP}}),$$

where $W^{\mathcal{IP}}(r_{\mathcal{IP}}) = \overline{W}^{\infty}_{r_{\mathcal{IP}}}$ and $dP/dr_{\mathcal{IP}} < 0$. Thus, to obtain the price implication, it suffices to solve for the optimal $r_{\mathcal{IP}}$ that maximizes the vendor profit. Given

$$\Pi^{\mathcal{IP}}(r_{\mathcal{IP}}) = \Lambda\left(V - c\left(W^{\mathcal{IP}}(r_{\mathcal{IP}}) + r_{\mathcal{IP}}\right)\right) \cdot r_{\mathcal{IP}}/\bar{r},$$

$$\frac{d^2\Pi^{\mathcal{IP}}(r_{\mathcal{IP}})}{dr_{\mathcal{IP}}^2} = -\frac{c \cdot \rho_{\mathcal{IP}}\left(\frac{6+2\rho_{\mathcal{IP}}^2+2\rho_{\mathcal{IP}}\left(9-8\rho_{\mathcal{IP}}+3\rho_{\mathcal{IP}}^2\right)\xi}{3(1-\rho_{\mathcal{IP}})^3} + (1+\xi)\sqrt{\rho_{\mathcal{IP}}}\log\left(\frac{1+\sqrt{\rho_{\mathcal{IP}}}}{1-\sqrt{\rho_{\mathcal{IP}}}}\right)\right)}{r_{\mathcal{IP}}^2} < 0,$$

where $\rho_{\mathcal{IP}} = \Lambda r_{\mathcal{IP}}^2/r < 1$, we can deduce that the optimal $r^*_{\mathcal{IP}}$ uniquely solves $\frac{d\Pi^{\mathcal{IP}}(r_{\mathcal{IP}})}{dr_{\mathcal{IP}}} = 0$. It follows that

$$6V - c \cdot r^*_{\mathcal{IP}}\left(\frac{2\left(3\rho_{\mathcal{IP}}^2 - 7\rho_{\mathcal{IP}} + 6 + (5-3\rho_{\mathcal{IP}})\rho_{\mathcal{IP}}\xi\right)}{\left(1-\rho_{\mathcal{IP}}\right)^2} + 3(1+\xi)\sqrt{\rho_{\mathcal{IP}}} \cdot \log\left(\frac{1+\sqrt{\rho_{\mathcal{IP}}}}{1-\sqrt{\rho_{\mathcal{IP}}}}\right)\right) = 0,$$

Then, the optimal price is $P^*_{\mathcal{IP}} = V - c(W^{\mathcal{IP}}(r^*_{\mathcal{IP}}) + r^*_{\mathcal{IP}})$, and we have

$$\frac{dr^*_{\mathcal{IP}}}{d\delta} = -\frac{(1+\xi)(r^*_{\mathcal{IP}})^3\left(2\left(3(\rho^*_{\mathcal{IP}})^2 - 8\rho^*_{\mathcal{IP}} + 13\right)\sqrt{\rho^*_{\mathcal{IP}}} + 3(1-\rho^*_{\mathcal{IP}})^3\log\left(\frac{1+\sqrt{\rho^*_{\mathcal{IP}}}}{1-\sqrt{\rho^*_{\mathcal{IP}}}}\right)\right)}{4\left(2\sqrt{\rho^*_{\mathcal{IP}}}\left(\xi\rho^*_{\mathcal{IP}}\left(3(\rho^*_{\mathcal{IP}})^2 - 8\rho^*_{\mathcal{IP}} + 9\right) + (\rho^*_{\mathcal{IP}})^2 + 3\right) + 3(1+\xi)\rho^*_{\mathcal{IP}}(1-\rho^*_{\mathcal{IP}})^3\log\left(\frac{1+\sqrt{\rho^*_{\mathcal{IP}}}}{1-\sqrt{\rho^*_{\mathcal{IP}}}}\right)\right)} < 0,$$

$$\frac{dr^*_{\mathcal{IP}}}{dV} = \frac{3(1-\rho^*_{\mathcal{IP}})^3}{c\left(2(\rho^*_{\mathcal{IP}})^2 + 6 + 2\xi\rho^*_{\mathcal{IP}}\left(3(\rho^*_{\mathcal{IP}})^2 - 8\rho^*_{\mathcal{IP}} + 9\right) + 3(1+\xi)\sqrt{\rho^*_{\mathcal{IP}}}(1-\rho^*_{\mathcal{IP}})^3\log\left(\frac{1+\sqrt{\rho^*_{\mathcal{IP}}}}{1-\sqrt{\rho^*_{\mathcal{IP}}}}\right)\right)} > 0,$$

$$\frac{dr^*_{\mathcal{IP}}}{dc} = -\frac{r^*_{\mathcal{IP}}(1-\rho^*_{\mathcal{IP}})\left(12 + 2\rho^*_{\mathcal{IP}}(3\rho^*_{\mathcal{IP}} - 7 + (5-3\rho^*_{\mathcal{IP}})\xi) + 3(1+\xi)\sqrt{\rho^*_{\mathcal{IP}}}(1-\rho^*_{\mathcal{IP}})^2\log\left(\frac{1+\sqrt{\rho^*_{\mathcal{IP}}}}{1-\sqrt{\rho^*_{\mathcal{IP}}}}\right)\right)}{2c\left(2(\rho^*_{\mathcal{IP}})^2 + 6 + 2\xi\rho^*_{\mathcal{IP}}\left(3(\rho^*_{\mathcal{IP}})^2 - 8\rho^*_{\mathcal{IP}} + 9\right) + 3(1+\xi)\sqrt{\rho^*_{\mathcal{IP}}}(1-\rho^*_{\mathcal{IP}})^3\log\left(\frac{1+\sqrt{\rho^*_{\mathcal{IP}}}}{1-\sqrt{\rho^*_{\mathcal{IP}}}}\right)\right)} < 0.$$

Furthermore, note that $\frac{dP^*_{\mathcal{IP}}}{dr^*_{\mathcal{IP}}} = -\frac{c\left((\rho^*_{\mathcal{IP}})^2 - 2\rho^*_{\mathcal{IP}} + 2\xi\rho^*_{\mathcal{IP}}(2-\rho^*_{\mathcal{IP}}) + (1+\xi)\sqrt{\rho^*_{\mathcal{IP}}}(1-\rho^*_{\mathcal{IP}})^2\log\left(\frac{1+\sqrt{\rho^*_{\mathcal{IP}}}}{1-\sqrt{\rho^*_{\mathcal{IP}}}}\right) + 3\right)}{3(1-\rho^*_{\mathcal{IP}})^2} < 0$, we have

$$\frac{dP^*_{\mathcal{IP}}}{d\delta} = \frac{dP^*_{\mathcal{IP}}}{dr^*_{\mathcal{IP}}} \cdot \frac{dr^*_{\mathcal{IP}}}{d\delta} > 0,$$

$$\begin{aligned}\frac{dP^*_{\mathcal{IP}}}{dV} =& 1 + \frac{dP^*_{\mathcal{IP}}}{dr^*_{\mathcal{IP}}} \cdot \frac{dr^*_{\mathcal{IP}}}{dV} \\ =& \frac{3 + \rho^*_{\mathcal{IP}}\left(2 + (\rho^*_{\mathcal{IP}})^2 - \rho^*_{\mathcal{IP}} + \left(14 - 10\rho^*_{\mathcal{IP}} + 4(\rho^*_{\mathcal{IP}})^2\right)\xi\right) + 2(1+\xi)\sqrt{\rho^*_{\mathcal{IP}}}(1-\rho^*_{\mathcal{IP}})^3\log\left(\frac{1+\sqrt{\rho^*_{\mathcal{IP}}}}{1-\sqrt{\rho^*_{\mathcal{IP}}}}\right)}{6 + 2(\rho^*_{\mathcal{IP}})^2 + 2\rho^*_{\mathcal{IP}}(9 - 8\rho^*_{\mathcal{IP}} + 3(\rho^*_{\mathcal{IP}})^2)\xi + 3(1+\xi)\sqrt{\rho^*_{\mathcal{IP}}}(1-\rho^*_{\mathcal{IP}})^3\log\left(\frac{1+\sqrt{\rho^*_{\mathcal{IP}}}}{1-\sqrt{\rho^*_{\mathcal{IP}}}}\right)} > 0,\end{aligned}$$

$$\begin{aligned}\frac{dP^*_{\mathcal{IP}}}{dc} =& -(W^{\mathcal{IP}}(r^*_{\mathcal{IP}}) + r^*_{\mathcal{IP}}) + \frac{dP^*_{\mathcal{IP}}}{dr^*_{\mathcal{IP}}} \cdot \frac{dr^*_{\mathcal{IP}}}{dc} \\ =& -\frac{r^*_{\mathcal{IP}}(1+\xi)\left(\frac{2\rho^*_{\mathcal{IP}}\left(-3(\rho^*_{\mathcal{IP}})^3 + 9(\rho^*_{\mathcal{IP}})^2 - 23\rho^*_{\mathcal{IP}} + 21 + 2\xi\left(3(\rho^*_{\mathcal{IP}})^2 - \rho^*_{\mathcal{IP}}\right)\right)}{1-\rho^*_{\mathcal{IP}}} + \left(-3(\rho^*_{\mathcal{IP}})^3 + 15(\rho^*_{\mathcal{IP}})^2 - 7\rho^*_{\mathcal{IP}} + 3 + 2\xi\left(3(\rho^*_{\mathcal{IP}})^2 + \rho^*_{\mathcal{IP}}\right)\right)\sqrt{\rho^*_{\mathcal{IP}}}\log\left(\frac{1+\sqrt{\rho^*_{\mathcal{IP}}}}{1-\sqrt{\rho^*_{\mathcal{IP}}}}\right)\right)}{12\left((\rho^*_{\mathcal{IP}})^2 + 3\xi\rho^*_{\mathcal{IP}}\left(3(\rho^*_{\mathcal{IP}})^2 - 8\rho^*_{\mathcal{IP}} + 9\right)\right) + 18(1+\xi)\sqrt{\rho^*_{\mathcal{IP}}}(1-\rho^*_{\mathcal{IP}})^3\log\left(\frac{1+\sqrt{\rho^*_{\mathcal{IP}}}}{1-\sqrt{\rho^*_{\mathcal{IP}}}}\right)} < 0.\end{aligned}$$

Thus, $P^*_{\mathcal{IP}}$ increases in δ and V, while decreases in c.

Similarly, with $\mathcal{FCFS}$, it can be easily deduce that, to maximize the profit rate given by $\Pi^{\mathcal{F}}(r_{\mathcal{F}}) = \Lambda\left(V - c\left(W^{\mathcal{F}}\left(r_{\mathcal{F}}\right) + r_{\mathcal{F}}\right)\right) \cdot r_{\mathcal{F}}/\bar{r}$, the optimal price should be set at

$$P^*_{\mathcal{F}} = V - c(W^{\mathcal{F}}(r^*_{\mathcal{F}}) + r^*_{\mathcal{F}}),$$

where $r^*_{\mathcal{F}}$ uniquely solves $3V - 2c \cdot r_{\mathcal{F}}\left(1 + \frac{2+2\xi\rho_{\mathcal{F}}(2-\rho_{\mathcal{F}})}{(1-\rho_{\mathcal{F}})^2}\right) = 0$, and $\rho_{\mathcal{F}} = \Lambda r^2_{\mathcal{F}}/\bar{r} < 1$. Specifically, we have $\frac{dr^*_{\mathcal{F}}}{d\delta} = -\frac{4(1+\xi)(r^*_{\mathcal{F}})^3}{3+\rho^*_{\mathcal{F}}\left(3+\rho^*_{\mathcal{F}}\left(3-\rho^*_{\mathcal{F}}\right)+2\xi\left((\rho^*_{\mathcal{F}})^2-3\rho^*_{\mathcal{F}}+6\right)\right)} < 0$, $\frac{dr^*_{\mathcal{F}}}{dV} = \frac{3(1-\rho^*_{\mathcal{F}})^3}{2c\left(3+\rho^*_{\mathcal{F}}\left(3+\rho^*_{\mathcal{F}}\left(3-\rho^*_{\mathcal{F}}\right)+2\xi\left((\rho^*_{\mathcal{F}})^2-3\rho^*_{\mathcal{F}}+6\right)\right)\right)} > 0$, $\frac{dr^*_{\mathcal{F}}}{dc} = -\frac{r^*_{\mathcal{F}}(1-\rho^*_{\mathcal{F}})\left(3-2\rho^*_{\mathcal{F}}+(\rho^*_{\mathcal{F}})^2+2\xi(2-\rho^*_{\mathcal{F}})\rho^*_{\mathcal{F}}\right)}{c\left(3+\rho^*_{\mathcal{F}}\left(3+\rho^*_{\mathcal{F}}\left(3-\rho^*_{\mathcal{F}}\right)+2\xi\left((\rho^*_{\mathcal{F}})^2-3\rho^*_{\mathcal{F}}+6\right)\right)\right)} < 0$, $\frac{dP^*_{\mathcal{F}}}{dr^*_{\mathcal{F}}} = -\frac{c\left(3+(\rho^*_{\mathcal{F}})^2+2\xi\rho^*_{\mathcal{F}}(3-\rho^*_{\mathcal{F}})\right)}{3\left(1-\rho^*_{\mathcal{F}}\right)^2} < 0$, $\frac{dP^*_{\mathcal{F}}}{d\delta} = \frac{dP^*_{\mathcal{F}}}{dr^*_{\mathcal{F}}} \cdot \frac{dr^*_{\mathcal{F}}}{d\delta} > 0$, $\frac{dP^*_{\mathcal{F}}}{dV} = \frac{1}{2} + \frac{(1+\xi)\rho^*_{\mathcal{F}}(\rho^*_{\mathcal{F}}+3)}{3+\rho^*_{\mathcal{F}}\left(3+\rho^*_{\mathcal{F}}\left(3-\rho^*_{\mathcal{F}}\right)+2\xi\left((\rho^*_{\mathcal{F}})^2-3\rho^*_{\mathcal{F}}+6\right)\right)} > 0$, and $\frac{dP^*_{\mathcal{F}}}{dc} = -\frac{4r^*_{\mathcal{F}}\rho^*_{\mathcal{F}}(1+\xi)\left(3+(\rho^*_{\mathcal{F}})^2(2\xi-1)\right)}{9+3(\rho^*_{\mathcal{F}}-4)(\rho^*_{\mathcal{F}})^3+6\xi\rho^*_{\mathcal{F}}(1-\rho^*_{\mathcal{F}})\left(6+(\rho^*_{\mathcal{F}})^2-3\rho^*_{\mathcal{F}}\right)} < 0$.
Thus, $P^*_{\mathcal{F}}$ increases in δ and V, while decreases in c, which complete this proof. ■

Proof of Theorem 3 Denote by $\Omega_{\mathcal{S}}$ the set of all scheduling policies, for $\mathcal{X} \in \Omega_{\mathcal{S}}$, the vendor revenue is given by

$$\Pi^{\mathcal{X}}(P) = \Lambda\left(V - c\left(W^{\mathcal{X}}(\tilde{r}) + \tilde{r}\right)\right) \cdot \tilde{r}/r,$$

where $\tilde{r}$ is a function of P. Define $\bar{\Pi}^{\mathcal{X}}$ as the maximum profit by setting the optimal price.

Since SEPT minimizes the average delay (Wolff 1989, Ayhan and Olsen 2000), we can easily conclude that when the price is set at the optimal price under $\mathcal{X}$ policy (i.e., $P_{\mathcal{X}}$),

$$\Pi^{\mathcal{IP}}(P_{\mathcal{X}}) - \bar{\Pi}^{\mathcal{X}} = c\Lambda \cdot \tilde{r}(P_{\mathcal{X}}) \cdot \left(W^{\mathcal{X}}\left(\tilde{r}\left(P_{\mathcal{X}}\right)\right) - W^{\mathcal{IP}}\left(\tilde{r}\left(P_{\mathcal{X}}\right)\right)\right)/r > 0 \text{ and } \bar{\Pi}^{\mathcal{IP}} \geq \Pi^{\mathcal{IP}}(P_{\mathcal{X}})$$

hold for any $\mathcal{X} \neq \mathcal{IP}$. It follows that $\bar{\Pi}^{\mathcal{IP}} \geq \bar{\Pi}^{\mathcal{X}}$ holds for all $\mathcal{X} \in \Omega_{\mathcal{S}}$, completing this proof. ■

Proof of Proposition 7 Recall that $\frac{d\Pi^{\mathcal{F}}}{d\tilde{r}}$ and $\frac{d\Pi^{\mathcal{IP}}}{d\tilde{r}}$ decrease in $\tilde{r}$, and $r^*_{\mathcal{F}}$ (resp. $r^*_{\mathcal{IP}}$) uniquely solves $\frac{d\Pi^{\mathcal{F}}}{d\tilde{r}} = 0$ (resp. $\frac{d\Pi^{\mathcal{IP}}}{d\tilde{r}} = 0$). Furthermore, note that

$$\frac{d\Pi^{\mathcal{F}}}{d\tilde{r}} < \frac{d\Pi^{\mathcal{IP}}}{d\tilde{r}} \Leftrightarrow 3\left(1-x^2\right)^2 \log\left(\frac{1+x}{1-x}\right) + 2\left(x^3 - 3x\right) < 0$$

holds for all $x = \tilde{r}\sqrt{\delta} \in (0,1)$, it follows that $r^*_{\mathcal{F}} < r^*_{\mathcal{IP}}$. Thus,

$$\bar{CS}^{\mathcal{F}} = \frac{\Lambda c (r^*_{\mathcal{F}})^2}{2\bar{r}} < \bar{CS}^{\mathcal{IP}} = \frac{\Lambda c (r^*_{\mathcal{IP}})^2}{2\bar{r}}$$

can be obtained. ■

Appendix D: Other Discussions

This section discusses the impacts of driving distance cost, bundled delivery, and wait-time limit on proximity-based priority policy, thereby highlighting its practical relevance.

1. Driving Distance Cost

In practice, each additional unit of distance driven incurs tangible costs, such as fuel consumption. In this section, we discuss how driving distance costs (DDC) may influence policy evaluation and vendor profit. Specifically, in our model, the total driving distance depends on both each order's travel distance and the number of orders served per unit of time (e.g., per day). For instance, under policy $\mathcal{X}$ with equilibrium service radius $r_{\mathcal{X}}$, the effective arrival rate—that is, the total number of orders served per day—is given by $\lambda_{\mathcal{X}} = \delta r_{\mathcal{X}}$. Hence, the total DDC is

$$DDC_{\mathcal{X}} = c_d \cdot \lambda_{\mathcal{X}} \cdot \mathbb{E}[S_{\mathcal{X}}] = c_d \cdot \delta r_{\mathcal{X}} \cdot \frac{1}{r_{\mathcal{X}}} \int_0^{r_{\mathcal{X}}} 2\mathbb{E}[T(x)]\, dx = c_d \delta r^2_{\mathcal{X}},$$

where c_d denotes the driving cost per unit distance. This expression indicates that DDC mirrors the pattern of market coverage, which is intuitive and suggests that our analysis of market coverage across the three policies can naturally extend to incorporate DDC as well.

By incorporating DDC, the vendor's profit can be expressed as

$$\Pi^{\mathcal{X}} = P\delta r_{\mathcal{X}} - c_d \delta r_{\mathcal{X}}^2 = \delta r_{\mathcal{X}} \big(P - c_d \cdot r_{\mathcal{X}}\big),$$

where $c_d \cdot r_{\mathcal{X}}$ can be interpreted as the marginal delivery cost per order. In the Chinese food delivery context of our mini-case study (see §7), electric scooters are the dominant delivery vehicle. The typical electricity price is around 0.6 CNY/kWh, and scooters consume approximately 15 kWh per 100 km, resulting in a unit driving cost of $c_d = 0.6 \times 15/100 = 0.09$ CNY/km. Relative to the average order price $P = 34$ CNY in our mini-case study, this cost is negligible. After accounting for DDC, the profit improvement of the $\mathcal{IP}$ policy over the $\mathcal{FCFS}$ policy is 3.46%—almost identical to the 3.5% increase in market coverage—suggesting that DDC plays only a limited role in shaping the $\mathcal{IP}$ policy's value and does not alter our core insights.

2. Bundled Delivery

This section discusses the $\mathcal{PBD}$ policy's value under bundled delivery. In particular, we analytically examine an extreme scenario in which a courier can carry an unlimited number of orders, thereby clearing all pending orders upon each return to the vendor. Let the set of pending orders at the vendor be $\mathbf{x} = (x_1, \ldots, x_n)$, the orders currently carried by the courier be $\mathbf{y} = (y_1, \ldots, y_m)$, and the courier's current location be z, so that the system state is $(\mathbf{x}, \mathbf{y}, z)$. Customer orders arrive according to a Poisson process with rate Λ, and each order's delivery proximity is independently and uniformly distributed on $[0, r_e]$ just like the main model.

Under the $\mathcal{PBD}$ policy, we denote by T_i the i-th bundle's delivery time, and the next bundle consists of all orders arriving during the interval of length T_i. Given $T_i = t$, let $N(t) \sim \text{Poisson}(\Lambda t)$ denote the number of arrivals during this interval. Define $M(t) = \max\{X_1, \ldots, X_{N(t)}\}$, with the convention that $M(t) = 0$ if $N(t) = 0$. Since the courier must travel to the farthest order and return, the next bundle's service time is $T_{i+1} = 2\,M(t)$. The conditional distribution can be derived as

$$\mathbb{P}(M(t) \le m \mid T_i = t) = \sum_{n=0}^{\infty} \left(\frac{m}{r_e}\right)^n \frac{e^{-\Lambda t}(\Lambda t)^n}{n!} = \exp\big(-\Lambda t\,(1 - m/r_e)\big)$$

for $m \in [0, r_e]$. Therefore, for $s \in [0, 2r_e]$,

$$F_{T_{i+1}|T_i}(s \mid t) = \mathbb{P}(T_{i+1} \le s \mid T_i = t) = \exp\big(-\Lambda t\,[1 - s/(2r_e)]\big)$$

and there is an atom at zero with mass $\mathbb{P}(T_{i+1} = 0 \mid T_i = t) = e^{-\Lambda t}$. For $s \in (0, 2r_e)$, the conditional density is $f_{T_{i+1}|T_i}(s \mid t) = \frac{\Lambda t}{2r_e} \exp\big(-\Lambda t\,[1 - s/(2r_e)]\big)$. Denote by T the steady-state service time with density f_T and distribution function F_T. For $s \in [0, 2r_e]$, the distribution can be determined by the following fixed-point integral equation.

$$F_T(s) = \int_0^{2r_e} \exp\big(-\Lambda t[1 - s/(2r_e)]\big)\, dF_T(t). \tag{A.3}$$

Indeed, from the integral equation (A.3), the kernel $\exp\{-\Lambda t[1 - s/(2r_e)]\}$ is decreasing in Λ and in r_e for each $s < 2r_e$. Hence $F_T(s)$ is decreasing in (Λ, r_e), implying that T is stochastically increasing in Λ or r_e. Unfortunately, the fixed-point equation admits no elementary closed form in general. Nevertheless, the solution is *unique* in the class of distribution functions on $[0, 2r_e]$ and can be computed numerically. By letting $\Phi(u) = \int_0^{2r_e} e^{-us}\, dF_T(s) = \mathbb{E}[e^{-uT}]$ be the Laplace-Stieltjes transform of T, we have $F_T(s) = \Phi\left(\Lambda\left(1 - \frac{s}{2r_e}\right)\right)$. Next, we can compute the residual service time through $\mathbb{E}[T_e] = \frac{\mathbb{E}[T^2]}{2\,\mathbb{E}[T]} = -\frac{\Phi''(0)}{2\Phi'(0)}$.

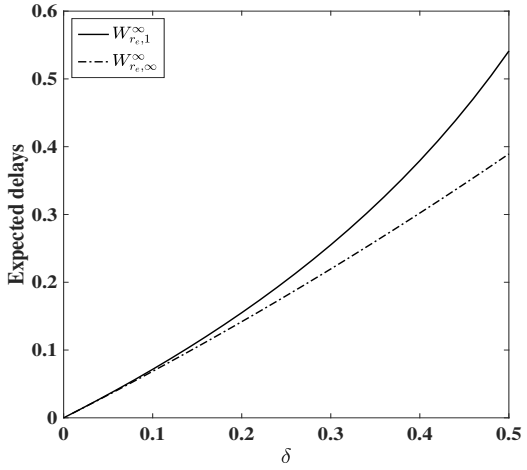
Compared with our main model where the courier carries a single order at a time, the "infinite-capacity batching" regime yields a larger mean residual time, since each cycle involves a round trip to the farthest order in the batch, which lengthens both the cycle time and the residual time observed by new arrivals. Nevertheless, it minimizes the number of dispatches, as each cycle clears all accumulated orders. Let N denote the maximum number of orders that can be carried at one time. Our baseline model corresponds to $N = 1$, while the model discussed above corresponds to $N = \infty$. Consequently, for a customer located at proximity $x \in [0, r_e]$, her expected sojourn time is $\mathbb{E}[W^{\infty}_{r_e,\infty}(x)] = \rho\,\mathbb{E}[T_e] + x$, where ρ is the steady-state probability that the system is busy.

Clearly, the expected in-line delays under market coverage r_e with infinite priority grades satisfy $W^{\infty}_{r_e,\infty} < W^{\infty}_{r_e,N} < W^{\infty}_{r_e,1}$, where $W^{\infty}_{r_e,\infty} = \rho\mathbb{E}[T_e]$ and $W^{\infty}_{r_e,1} = \overline{W}^{\infty}_{r_e}$. To illustrate this, we numerically plot $W^{\infty}_{r_e,1}$ and $W^{\infty}_{r_e,\infty}$ against market density in Figure A.3, which confirms that the expected delay under bundled delivery is lower than that under dedicated delivery. Moreover, we conjecture that the advantage of the $\mathcal{PBD}$ policy over the $\mathcal{FCFS}$ policy may become even more pronounced with bundled delivery. Thus, Figure A.3 also identifies the upper and lower bounds of the expected waiting time under bundled delivery. The intuition is that when N orders are bundled, couriers deliver them sequentially along their routes, effectively applying the $\mathcal{PBD}$ principle both at the trip-planning level and within each trip. This suggests that allowing couriers to carry multiple orders per trip does not diminish the benefit of the $\mathcal{PBD}$ policy; rather, it amplifies it.

3. Wait-Time Limit

In practice, distant customers may face the risk of repeated deprioritization under the $\mathcal{PBD}$ policy (i.e., their orders are repeatedly preempted by those from nearby customers). This increases randomness in their service experience and can result in particularly long realized waiting times, even when the expected waiting time remains acceptable. Such variability is a general limitation of priority queueing systems with stochastic arrivals. To mitigate this issue, we extend our model by introducing a **wait-time limit (WTL)**: *If an order is not dispatched within a pre-specified threshold $\bar{w}$, the vendor either (i) loses the customer or must discard the food due to quality constraints, or (ii) hires gig labor to fulfill the order immediately, incurring a fixed (penalty) cost l in both cases*. This mechanism

Figure A.3 Illustration of expected delays under different bundle constraints



Note. $r_e = 1$.

establishes an upper bound on waiting times, prevents excessive order delays, and captures the interplay between delay management and customer update behavior.

Under this setting, when the effective delivery region is $r_e \in [0, \bar{r}]$, we denote the waiting time under the $\mathcal{FCFS}$ policy as $\mathcal{W}^{\mathcal{F}}_{\bar{w}}$, which is a random variable. Clearly, $\mathcal{W}^{\mathcal{F}}_{\bar{w}}$ is stochastically increasing in $\bar{w}$ because a higher $\bar{w}$ involves more customers. Consequently, the penalty cost under $\mathcal{FCFS}$ is $l^{\mathcal{F}}_{\bar{w}} = \frac{\Lambda}{\bar{r}} \int_0^{r_e} Prob\left[\mathcal{W}^{\mathcal{F}}_{\bar{w}} > \bar{w}\right] \cdot l x dx$. By contrast, under the $\mathcal{PBD}$ policy, let the wait time for a type-x customer be $\mathcal{W}^{\mathcal{P}}_{\bar{w}}(x)$. The penalty cost under $\mathcal{PBD}$ is $l^{\mathcal{P}}_{\bar{w}} = \frac{\Lambda}{\bar{r}} \int_0^{r_e} Prob\left[\mathcal{W}^{\mathcal{P}}_{\bar{w}}(x) > \bar{w}\right] \cdot l x dx$. It follows that there exists a threshold $\check{x}$ such that for $x \in [0, \check{x}]$, the $\mathcal{FCFS}$ policy incurs higher costs, while for $x \in [\check{x}, r_e]$ the $\mathcal{PBD}$ incurs higher costs. Therefore, two cases arise: (1) if $l^{\mathcal{P}}_{\bar{w}} \geq l^{\mathcal{F}}_{\bar{w}}$, then the delay reduction effect of $\mathcal{PBD}$ is more pronounced, but at the expense of higher penalty costs; (2) if $l^{\mathcal{P}}_{\bar{w}} < l^{\mathcal{F}}_{\bar{w}}$, the delay reduction effect of $\mathcal{PBD}$ is weakened, but it lowers penalty costs.

Overall, the vendor can balance delay and penalty cost under different modes by adjusting WTL $\bar{w}$, i.e., choosing $\bar{w}^{\mathcal{P}}$ and $\bar{w}^{\mathcal{F}}$. Since we have already shown that $\mathcal{PBD}$ reduces the average delay, we conjecture that under $\mathcal{PBD}$ the vendor will adopt a relatively higher WTL, namely $\bar{w}^{\mathcal{P}} > \bar{w}^{\mathcal{F}}$. However, determining these thresholds and conducting further optimization remains challenging, as the exact distributions of waiting times (i.e., $\mathcal{W}^{\mathcal{F}}_{\bar{w}}$ and $\mathcal{W}^{\mathcal{P}}_{\bar{w}}(x)$) are analytically intractable. We therefore leave this as a promising direction for future research.