NORTHWESTERN UNIVERSITY

Essays on System Efficiency in Service Operations

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Operations Management

By

Can Ozkan

EVANSTON, ILLINOIS

September 2017

ABSTRACT

Essays on System Efficiency in Service Operations

Can Ozkan

Services have been constantly evolving and operational efficiency has been a key initiative for progress. In this collection of academic papers, we investigate the efficiency of three different industry practices, to each of which we dedicate a chapter. Chapter 1 and Chapter 2 cover completed research, while the research covered in Chapter 3 is at preliminary stage.

In Chapter 1, we study priority queues to understand the determinants of social efficiency. Many service providers utilize priority queues. Many consumers revile priority queues. However, some form of priority service may be necessary to maximize social welfare. Consequently, it is useful to understand how the priority scheme chosen by a revenue-maximizing firm differs from the one a social planner would use. We examine this in a single server-queue with customers that draw their valuation from a continuous distribution and have a per-period waiting cost that is proportional to their realized valuation. The decision maker must post a menu offering a finite number of waiting time-price pairs. There are then three dimensions on which a revenue maximizer and social planner can differ: coverage (i.e., how many customers in total

to serve), coarseness (i.e., how many classes of service to offer), and classification (i.e., how to map customers to priority levels).

We show that differences between the decision makers priority policies are all about classification. Both are content to offer very coarse schemes with just two priority levels, and they will have negligible differences in coverage. However, differences in classification are persistent. Further, a revenue maximizer may — relative to the social planner — have too few or too many high priority customers. Whether the revenue maximizer over- or under-stuffs the high priority class depends on a measure of consumer surplus that is captured by the mean residual life function of the valuation distribution. In addition, we show that there is a large class of valuation distributions for which a move from first-in, first-out service to a priority scheme that places those with higher waiting costs at the front of the line reduces consumer surplus.

In Chapter 2, we study the impact of the increased availability of real-time information on the behavior of strategic agents and the implications of this phenomenon for service efficiency. The use of real-time information in on-demand services provides agents with access to an unprecedented amount of information about their competitors. We use data from one of the leading e-hailing taxi platforms in South America to study the real-time reactions of agents to the dynamic entry of new competitors in their serving zone. Information about competitor locations could potentially induce herding behaviour (because competitors' actions may convey information about market opportunity) or scattering (because the entry of competitors reduces the expected market share and the appeal of a serving zone). We find that the net response to the real-time information indicating entry of new competitors in a service zone is an increase in the scattering of the agents previously in the serving zone. The response is not homogenous and some agents are more likely to respond to entry. We find that those agents who are more

likely to react to the real-time presence of competitors by scaterring achieve higher utilization. We investigate the consequences of these behaviors for the efficiency of service systems.

Finally, in Chapter 3, we analyze the effect of carry-on bag policy on the system efficiency. Air-carriers want to utilize their airplanes as much as possible. One of the obstacles against a high utilization is the delay due to the boarding process. Some of the low-cost carriers started to apply fees on carry-on bags so that passengers would be encouraged to check in their bags instead of taking these bags with theirselves to the board. In this study, we investigate the effects of this new policy on the air-carrier delay. We use the available data of U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) which contains flights of Frontier Airlines that applies this new policy. We observe that this policy change was successful in decreasing the departure delays. Furthermore, we propose the requirement of robustness analysis with additional factors that capture the dynamics of the industry more realistically.

Acknowledgements

I am grateful to my academic advisors Professors Martin Lariviere, Itai Gurvich and Antonio Moreno-Garcia for their guidance and advice throughout my PhD years. None of what is presented in this thesis would be possible without them. My sincere thanks also goes to Prof. Ohad Perry for serving on my dissertation committee.

I would like to thank to my family for their greatest love and support. My father Sefer, a sea-man who spent the majority of his life on seas for his family, showed me the importance of sacrifice in life. My mother Muki taught me how to be planned and motivated to achieve success in life. My sister Deniz showed me the joy of intellectual pursuit. My wife Jamie taught me how to handle hard situations and turn them into advantages in life. I am very grateful to her. Last but not least, my son Theodore Ediz (Teddy) filled me with strength and love. I dedicate this work to him.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Coverage, coarseness and classification: Determinants of social efficiency in priority queues

Joint work with Martin Lariviere and Itai Gurvich

1.1. Introduction

In 2013 Walibi, a Belgian amusement park, introduced Speedy-Pass, a premium service that allowed purchasers to jump to the front of the line at park rides. These shorter waits did not come cheap; the service more than doubled the price of an adult ticket to the park. The announcement was met with an outpouring of opprobrium. An educator asked, "How in the name of God do you explain to a child that he has to wait in line in a long queue, while other children can go straight to the front, just because their parents have got more money?" Government ministers also chimed in denouncing the program (Flandersnews.be 2013).

Angst over priority queues is not limited to Europe. In the United States, travelers have petitioned the Transportation Security Administration not to allow airlines to profit from selling priority access to airport security screening[1] while the implementation of tolled express lanes – so called Lexus Lanes – has faced opposition in Georgia and Colorado (Walker 2012, Whaley 2015).

[1] See *TSA: Don't Allow "Priority" Airport Screening Lines*, www.change.org/p/tsa-don-t-allow-priority-airport-screening-lines. Accessed Sep 2, 2015.

On the one hand, such outrage is puzzling. Many firms offer multiple products or service formats and consumers presumably benefit from the increased range of options. A customer who buys a Chevrolet instead of a Cadillac must believe that the former offers better value. On the other hand, one must acknowledge that capacity constrained service providers differ from firms selling physical goods. Queuing creates externalities between customers in different classes of service so that increased sales to one class can reduce the value obtained by customers of the other classes. If General Motors sells more Cadillacs, Chevy buyers are unharmed, but the more Speedy-Passes Walibi sells, the worse service regular customers receive.

This, however, does not mean that priority schemes necessarily compromise social welfare. A social planner would use priorities if customers have different waiting costs. Consequently, the question is not whether a revenue-maximizing firm such as Walibi should use a priority scheme but whether Walibi's implementation differs dramatically from what the social planner would do.

Such a comparison of the revenue maximizer's and the social planner's actions is the subject of this paper. We consider a service system modeled as an $M/M/1$ queue. The decision maker may or may not offer multiple priority classes. Arriving customers all have the same average service time but differ in their valuations and waiting costs: they independently draw valuations for the service from a common, continuous distribution and have a per-unit-time waiting cost proportional to their valuation. The state of the queue is not observable to the customers, so they must choose which class of service to purchase based on a posted menu of expected delays and prices.

The decision maker, whether seeking to maximize revenue or social welfare, must make three decisions.

- *Coverage.* If the arrival rate of customers exceeds available capacity, the decision maker must turn away some customers. The revenue maximizer might choose to serve, in total, more or fewer customers than the social planner.
- *Coarseness*. We do not assume that the market is exogenously and *a priori* divided into different classes. It is up to the decision maker to route arrivals into distinct priority levels. With valuations and waiting costs drawn from a continuous distribution, both types of decision makers would benefit from posting a continuous menu of prices and delays. In reality it is more common for service providers to use coarse priority schemes that split arrivals into a finite number of discrete priority levels. A revenue maximizer might opt for a coarser or for a more refined division than the social planner.[2]
- *Classification*. Given a level of coarseness, the decision maker must still determine how to classify customers into priority levels: how many customers should go into each class and who these customers are. Even if the revenue maximizer chooses the same coverage and coarseness as the social planner, social efficiency might suffer. The revenue maximizer might benefit from pursuing an *ultra-luxury strategy* where a smaller high priority class than is socially optimal is charged higher prices. In other scenarios, the revenue maximizer might benefit from pursuing a *mass-luxury strategy* with a high priority that is larger than socially optimal.

Our analysis shows that the loss of societal efficiency resulting from the revenue maximizer's actions is largely a question of classification. We employ a limiting regime akin to

[2]The term "coarse priorities" is common in the the economic literature, specifically in the study of optimal allocation. Coarse priorities mean there that "rankings are coarse, that is, they rank classes of agents, and everyone within the same class is deemed in a tie" Ehlers and Erdil (2010). Nazerzadeh and Randhawa (2015) appear to be the first to use the term in the context of designing a queuing priority scheme.

Nazerzadeh and Randhawa (2015) in which the arrival rate of customers and the processing rate of the server are scaled up together. We show that both the social planner and revenue maximizer are content to use extremely coarse priority schemes; for either type of decision maker, the loss from using just two classes is negligible. Further, the level of coverage that both offer is essentially the same. Thus, revenue maximization is socially optimal as far as coverage and coarseness are concerned.

Differences, however, remain in classification. The revenue maximizer may opt for a mass-luxury strategy and admit more customers to the high priority class than is socially optimal or for an ultra-luxury strategy and admit fewer customers to the high priority class. Which approach she chooses depends on how consumer surplus changes with the level of coverage. We show that this is related to the mean residual life (MRL) of the valuation distribution. If the MRL function is convex, the elasticity of consumer surplus is decreasing and revenue maximizer opts for a mass-luxury strategy. When the MRL is concave, the elasticity is increasing and the revenue maximizer follows an ultra-luxury strategy.

Additional intuition follows from considering the problem of maximizing consumer surplus. Maximizing either revenue or social welfare calls for putting those with high valuations (and thus high waiting costs) at the front of the line. This is not necessarily the case with consumer welfare. If the valuation distribution has a decreasing failure rate, those with high valuations should be served first. However, with an increasing failure rate distribution, serving those with low valuations first maximizes consumer welfare. Note that this implies that consumers whose valuations follow an increasing failure rate distribution are better off under first-in, first-out service than under a priority scheme that favors those with high valuations. Further, a decreasing, convex MRL function implies an increasing failure rate, suggesting that the social planner is

less aggressive than the revenue maximizer in routing customers to the high priority class in order to reduce the impact on consumer welfare.

1.2. Literature Review

The comparison of revenue maximization and social optimization in queues has been a topic of interest since Naor's seminal paper (Naor 1969). Hassin and Haviv (2003) and Hassin (2016) provide excellent surveys. In this brief literature review, we focus on research that allows for a continuum, rather than an exogenously given number, of customer types.

Much of the work in this area builds on Kleinrock (1967), which considers customers who arrive to an unobservable queue and bid for priority. The customer bids are drawn from a common continuous distribution. The service provider offers, in turn, a continuum of priority levels and a customer is given priority over any customer who has bid less.

While the distribution of bids in Kleinrock (1967) is exogenously determined, subsequent work ties the distribution of bids to an equilibrium outcome between customers. Afèche and Mendelson (2004) consider customers who draw their valuation for service from a common distribution and who have a delay cost with two components: one that is proportional to the realized valuation and one that is independent of the valuation. They show that with a single first-in-first-out (FIFO) queue—that is, with uniform pricing—a revenue maximizer may offer greater or smaller coverage than socially optimal depending on the valuation distribution and the delay-cost structure. Studying priority auctions, they establish conditions for the revenue maximizer to achieve social efficiency. Katta and Sethuraman (2005) provide conditions under which the priority auctions in Afèche and Mendelson (2004) are in fact revenue maximizing.

Specializing the results of Afèche and Mendelson (2004) to our setting, we would have that the revenue maximizer admits fewer people than is socially optimal under a uniform price but admits customers at the socially optimal rate when places in line are auctioned off under preemptive priorities. If non-preemptive priorities are used, the revenue maximizer does not achieve social efficiency.

In Afèche and Mendelson (2004) as well as most other work in this vein, types are never pooled, so small differences in waiting costs may result in absolute differences in priorities. Depending on the waiting cost structure, the revenue maximizer may pool some types together, impose a common price and offer the same expected wait (Katta and Sethuraman 2005). Additionally, Afèche and Pavlin (2016) show that a revenue maximizer facing customers with a utility structure different than ours may use a complex service discipline that may pool customers or exclude some with intermediate valuations or impose strategic delay. Note that in these papers, the service provider still offers a continuum of priority levels despite pooling some types: customers who are not pooled are still given distinct priority levels despite small differences in waiting costs.

A limited set of papers consider how to map a continuum of customers to a coarse set of priority levels. Ghanem (1975) considers customers that differ only on their per-period waiting costs drawn from a common distribution, and examines how they should be classified into a predetermined number of priority classes in order to minimize total delay costs. Gilland and Warsing (2009) consider a similar problem from the perspective of revenue maximization but restrict their analysis to uniformly distributed delay costs and assume that all customers must be served.

In Gavirneni and Kulkarni (2014), customers differ only in their waiting costs which follow a Burr Type XII distribution and the service provider only offers two levels of priority. They present examples in which the revenue maximizer classifies more customers as high priority than the social planner would.

Doroudi et al. (2015) consider the same valuation and cost structure as we do; arriving customers draw a valuation from a common distribution and have waiting costs that are proportional to their realized valuations. The bulk of their analysis focuses on offering a continuum of priorities but they do demonstrate numerically that a coarse priority scheme with a limited number of priority classes performs very well.

Finally, Nazerzadeh and Randhawa (2015) examine how a revenue-maximizing service provider should manage coverage, coarseness, and classification when customer draw valuations from a common distribution and have per-period waiting costs that are a function of their realized valuations. They use an asymptotic analysis to show that a very coarse priority scheme is sufficient; two levels of priority capture nearly all of the possible system value. In examining coarseness, we take a similar approach. Indeed, their results partly provide the revenue maximizer side of our comparison. We not only expand the analysis to the social planner but also strengthen it. To demonstrate the near optimality of two classes for the revenue maximizer (as done in Nazerzadeh and Randhawa (2015)), it suffices to identify one solution (among possibly many solutions). In order to compare the decisions (i.e, the price and waiting-time menus) of the revenue maximizer and social planner, however, we must assure the uniqueness of the asymptotically optimal prescriptions.

1.3. Model Formulation

We consider a service modeled as a queuing system. There is a (potential) arrival stream that is Poisson with rate Λ. The queue is served by a single server, and the service time is exponential with rate μ and independent across customers. Without loss of generality, we fix μ to 1. We further assume that $\Lambda \geq 1$ so that not all customers can be served. How much of the market the decision maker choses to cover is then a non-trivial question.

Customer valuation for the service is drawn from a distribution F with support (a, b) with $0 \leq a < b$. Valuations are independent across customers. Customers are also adverse to delay. A customer's delay cost is linear in her waiting time (which includes the delay in the queue and the service time) with a coefficient that is proportional to her valuation: a customer with service valuation v incurs a cost of αv per unit of delay where $\alpha < 1$. This specification provides heterogeneity in both the cost coefficient and the valuation. We generalize this structure in Section 1.6.2.

F is assumed to be differentiable with differentiable density f. Let $\bar{F}(v) = 1 - F(v)$, let $h(v) = f(v)/\bar{F}(v)$ be the distribution's failure rate and $g(v) = vh(v)$ be the generalized failure rate of $F(v)$. We make the following assumption on F and g.

Assumption 1: F has a finite mean.

Assumption 2: F has an increasing generalized failure rate, i.e., $g'(v) \geq 0$ on (a, b).

The IGFR assumption is satisfied by many common distributions. It implies that f is strictly positive on (a, b) so both F and $\bar{F}$ are invertible. Let MRL be the mean residual life of F, i.e.,

$$MRL(v) = E[X - v | X \geq v] = \frac{\int_v^\infty t f(t)\, dt}{\bar{F}(v)} - v.$$

In the reliability literature the mean residual life represents the expected remaining life of a component given that it has survived to a given age. Here, $MRL\,(v)$ represents the expected value beyond a base level v created by serving a customer conditional on that customer having a value of at least v. Since F has a finite mean, $MRL(v)$ is defined for all v in (a, b) .

Customer actions. The service provider offers $K \in \mathbb{Z}_+$ different bundles of price and waiting time, (p_i, W_i). A customer must consequently choose whether to join the queue and, if she does, at which grade of service. A customer with valuation v that chooses menu item i obtains the utility

$$U(v; i) = v - \alpha v W_i - p_i. \tag{1.1}$$

The service provider must pair higher prices with shorter waiting times as no customer would buy a bundle that charges a higher price and imposes a longer delay than another menu item. We must, in particular, have $W_i \neq W_j$ and $p_i \neq p_j$ for $i \neq j$.

Customers who select a particular menu item have valuations that fall within an interval:

Lemma 1. *Suppose that a customer with valuation $\widetilde{v}$ optimally chooses menu item (p_i, W_i) and another customer with valuation $\widehat{v}$ optimally chooses menu item (p_j, W_j) . If $W_i > W_j$ and $p_i < p_j$, then $\widetilde{v} \leq \widehat{v}$. Therefore if two customers with valuations $v < u$ choose the same menu item (p_k, W_k) , a customer with valuations $w \in (v, u)$ must choose the same menu item.*

Thus, a price-delay menu segments the valuation space into intervals and we may, without loss of generality, number the offerings such that a higher index corresponds to a higher price and a shorter wait. Customers with the highest valuations thus choose item K. Since the set of customer valuations for a given menu item is an interval and F is strictly increasing, there

exists a unique cutoff valuation v_i such that $U(v_i; i) = U(v_i; i-1)$, i.e., a customer with the valuation v_i is indifferent between menu items i and $i-1$. Let $\mathbf{v}$ be the vector of cut off values. The fraction of all customers choosing menu item $i \in \{1, \ldots, K\}$ is given by $\bar{F}(v_i) - \bar{F}(v_{i+1})$ (we define $v_{K+1} \equiv \infty$) and the rate of such customers is

$$\lambda_i(\mathbf{v}) = \Lambda(\bar{F}(v_i) - \bar{F}(v_{i+1})).$$

We write $\lambda(\mathbf{v})$ (dropping the subscript) for the vector of arrival rates. Let

$$\bar{\lambda}_i(\mathbf{v}) = \sum_{j=i}^{K} \lambda_j(\mathbf{v}) = \Lambda \bar{F}(v_i)$$

be the rate of customers that choose menu items i or higher. With this notation $\bar{\lambda}_1(\mathbf{v}) = \Lambda \bar{F}(v_1)$ is the service provider's *coverage*: the volume of customers who enter the system per time unit. For stability this volume must be strictly smaller than the service rate—$\bar{\lambda}_1(\mathbf{v}) < 1$—so that no customer with valuation less than

(vbar)
$$\bar{v} = \bar{F}^{-1}\left(\frac{1}{\Lambda}\right),$$

enters the service (i.e., $v_1 > \overline{v}$).

The menu prices are uniquely determined by the cutoffs and the waiting times. Indeed, since the customer with valuation v_i is indifferent between i and $i-1$, we have

$$U(v_i; i-1) = v_i - p_{i-1} - \alpha v_i W_{i-1} = v_i - p_i - \alpha v_i W_i = U(v_i; i),$$

so that

$$p_i - p_{i-1} = \alpha v_i (W_{i-1} - W_i)\ . \tag{1.2}$$

The lowest-valuation customer who patronizes the service is indifferent between joining and not joining. Assuming that all customers have an outside option of zero, we then have

$$v_1 - W_1 \alpha v_1 - p_1 = 0 \implies p_1 = \ v_1 \left(1 - \alpha W_1\right). \tag{1.3}$$

p_1 can be interpreted as an *entrance fee* to the system – anyone that enters has to pay at least p_1. A customer with valuation $v > v_1$ enters the system and obtains a strictly positive utility since

$$U(v; i) \geq v(1 - \alpha W_1) - p_1 > v_1(1 - \alpha W_1) - p_1 = U(v_1; 1) = 0.$$

Thus, given cutoff values $v_1 < v_2 < \ldots < v_K$ and waiting times $W_1 > W_2 > \cdots > W_K$, the prices menu is uniquely determined by (1.2) and (1.3) regardless of whether the decision maker seeks to maximize revenue or social welfare. The objective functions of the two providers will determine how the vectors $\mathbf{v}$ and $\mathbf{W}$ are set.

The social planner's problem. For a fixed number K of classes, the social planner must choose a vector $\mathbf{v}$ of cut-off values and a vector $\mathbf{W}$ of waiting times to maximize social welfare, i.e., the aggregate utility of arriving customers. The expected utility of a class-i customer (a customer that chooses menu item i) with valuation v is $v - \alpha v W_i$. The social welfare following from a given $\mathbf{v}$ and $\mathbf{W}$ can be written as

$$\begin{aligned} S_K(\mathbf{v},\mathbf{W}) &= \Lambda\sum_{i=1}^{K}\left(\int_{v_i}^{v_{i+1}} uf(u)du\right)(1-\alpha W_i) \\ &= \Lambda V(v_1) - \alpha\sum_{i=1}^{K}\lambda_i(\mathbf{v})\,\frac{(V(v_i)-V(v_{i+1})}{\bar{F}(v_i)-\bar{F}(v_{i+1})}W_i \\ &= \Lambda V(v_1) - \sum_{i=1}^{K}\lambda_i(\mathbf{v})\,c_i^S(\mathbf{v})W_i \end{aligned}$$

where $V(x) := \int_x^\infty uf(u)\,dv$ so that $(V(v_i) - V(v_{i+1}))/\left(\bar{F}(v_i) - \bar{F}(v_{i+1})\right)$ is the average contribution to social welfare ignoring delay costs of a customer conditional on her selecting class i. Similarly, $c_i^S(\mathbf{v}) := \alpha\frac{(V(v_i)-V(v_{i+1})}{\bar{F}(v_i)-\bar{F}(v_{i+1})}$ is the average cost of delay among class-i customers. In maximizing $S_K(\mathbf{v},\mathbf{W})$, the constraints on the social planner's actions are that the cut-off values $\mathbf{v}$ are increasing while the waiting-time vector $\mathbf{W}$ must be decreasing and feasible given the induced arrival rates. We allow the use of preemptive priority schemes; see §1.6.1 for the case of non-preemptive policies. Let $\mathcal{W}(\mathbf{v})$ be the set of feasible waiting times given $\mathbf{v}$ and the resulting arrival rates. $\mathcal{W}(\mathbf{v})$ is determined by the achievable region; see e.g. Theorem A.10 of Stidham (2009).

The social planner's problem can then be written as

$$S_K^* = \max_{\mathbf{v}\uparrow,\mathbf{W}\downarrow} S_K(\mathbf{v},\mathbf{W}) \text{ s.t. } \mathbf{W}\in\mathcal{W}(\mathbf{v}).$$

where $\mathbf{v}\uparrow$ denotes $v_1 < v_2 < \cdots < v_K$ and , $\mathbf{W}\downarrow$ denotes $W_1 > W_2 > \cdots > W_K$. Because F is strictly increasing, $V(v_i) - V(v_{i+1}) > 0$ for all i so that work conservation is optimal (rather than, say, inserting strategic delay). Given $\mathbf{v}$, the cost coefficients $c_i^S(\mathbf{v})$ are increasing in i so that it is optimal to preemptively prioritize customers in decreasing order of $c_i^S(\mathbf{v})$ and

the social planner's problem reduces to a search over cutoff values. We thus have (recall that $\lambda_i(\mathbf{v}) = \Lambda(\bar{F}(v_i) - \bar{F}(v_{i+1}))$)

$$S_K^* = \max_{\mathbf{v}\uparrow} S_K(\mathbf{v}) := \Lambda V(v_1) - \alpha\Lambda \sum_{i=1}^{K} (V(v_i) - V(v_{i+1}))W_i(\lambda(\mathbf{v})) \tag{1.4}$$

where, for each $i = 1, \ldots, K$

$$W_i(\lambda(\mathbf{v})) = \frac{1}{(1-\bar{\lambda}_{i+1}(\mathbf{v}))(1-\bar{\lambda}_i(\mathbf{v}))} = \frac{1}{\left(1-\Lambda\bar{F}\left(v_i\right)\right)\left(1-\Lambda\bar{F}\left(v_{i+1}\right)\right)} \tag{1.5}$$

is the preemptive static priority waiting time of class i with arrival vector $\lambda(v)$ and service rate equal to one.

The revenue maximizer's problem. As with the social planner, the revenue maximizer sets cutoff values $\mathbf{v}$ and expected waits $\mathbf{W}$. The relationships (1.2) and (1.3) map prices and waits to cutoffs and waits. The firm's revenue is then given by

$$\begin{aligned}
R_K(\mathbf{v}, \mathbf{W}) &= \sum_{i=1}^{K} \lambda_i\left(\mathbf{v}\right) p_i = \Lambda p_1 \bar{F}\left(v_1\right) + \Lambda \sum_{i=1}^{K} \left(p_{i+1}(\mathbf{v}) - p_i(\mathbf{v})\right) \bar{F}\left(v_{i+1}\right) \\
&= \Lambda v_1 \bar{F}\left(v_1\right)\left(1 - \alpha W_1\right) + \alpha\Lambda \sum_{i=1}^{K} \left(W_i - W_{i+1}\right) v_{i+1} \bar{F}\left(v_{i+1}\right) \\
&= \Lambda\rho\left(v_1\right) - \alpha \sum_{i=1}^{K} \lambda_i(\mathbf{v}) \frac{\rho\left(v_i\right) - \rho\left(v_{i+1}\right)}{\bar{F}(v_i) - \bar{F}(v_{i+1})} W_i \\
&= \Lambda\rho\left(v_1\right) - \sum_{i=1}^{K} \lambda_i(\mathbf{v}) c_i^R(\mathbf{v}) W_i,
\end{aligned}$$

where $\rho\left(v\right) := v\bar{F}\left(v\right)$ and $\rho\left(v_{K+1}\right) = 0$. The coefficient $c_i^R(\mathbf{v}) = \alpha \frac{(\rho(v_i) - \rho(v_{i+1}))}{\bar{F}(v_i) - \bar{F}(v_{i+1})}$, captures the discount given to customers of class i to compensate them for their delay. Using $\bar{\lambda}_i(\mathbf{v}) = \Lambda\bar{F}(v_i)$, these can also be written as $c_i^R(\mathbf{v}) = \alpha\Lambda \frac{\widetilde{\rho}(\bar{\lambda}_i(\mathbf{v})) - \widetilde{\rho}(\bar{\lambda}_{i+1}(\mathbf{v}))}{\bar{\lambda}_i(\mathbf{v}) - \bar{\lambda}_{i+1}(\mathbf{v})}$, where $\widetilde{\rho}(\lambda) := \rho(\bar{F}^{-1}(\lambda/\Lambda))$

for $\lambda \geq 0$. If ρ is decreasing and $\widetilde{\rho}$ is convex then $c_i^R(\mathbf{v})$ is positive and increasing in i. Then much like the social planner, work conservation is optimal and the revenue maximizer will want to preemptively prioritize customers in decreasing order of $c_i^R(\mathbf{v})$. In that case, we can state the revenue maximizer's problem as choosing cutoff values knowing that customers will have expected waits given by $W_i\left(\lambda\left(\mathbf{v}\right)\right)$.

$$R_K^* = \max_{\mathbf{v}\uparrow} R_K\left(\mathbf{v}\right) := \Lambda\rho\left(v_1\right) - \alpha\Lambda\sum_{i=1}^{K}(\rho(v_i) - \rho(v_{i+1}))W_i(\lambda(\mathbf{v})), \tag{1.6}$$

which should be contrasted with (1.4). The relationship between $V(v_i) - V(v_{i+1})$ and $\rho(v_i) - \rho(v_{i+1})$ will play a key role in our results.

Notice that if customers were not delay sensitive (i.e., $\alpha = 0$) the revenue maximizer's problem reduces to choosing the cutoff that maximizes $\Lambda v\bar{F}(v) = \Lambda\rho(v)$ or equivalently an arrival rate that maximizes $\lambda p(\lambda)$ where $p(\lambda) = \bar{F}^{-1}(\lambda/\Lambda)$. If ρ is a decreasing function, the revenue can be read as the delay-insensitive revenue minus a delay-discount of $\alpha(\rho\left(v_i\right) - \rho\left(v_{i+1}\right))$ for class-i customers.

We conclude this section with a lemma that allows us to conclude that, indeed, $c_i^R(\mathbf{v}) > 0$ and increasing in i.

Lemma 2. *Given Assumptions 1 and 2:*

(1) $v_0^* := \inf\{v : g\left(v\right) \geq 1\}$ *is finite.*

(2) $\rho\left(v\right)$ *is maximized at* v_0^*. *It is increasing and concave for* $v < v_0^*$ *and decreasing for* $v \geq v_0^*$.

(3) $\widetilde{\rho}\left(\lambda\right)$ *is maximized at* $\lambda_0^* := \Lambda\bar{F}\left(v_0^*\right)$. *For* $\lambda < \lambda_0^*$, $\widetilde{\rho}\left(\lambda\right)$ *is increasing and concave.*

(4) $\varepsilon\left(\lambda\right) := -p\left(\lambda\right)/(\lambda p'\left(\lambda\right))$ *is increasing in* λ.

In what follows we will assume that the system is capacity constrained in the sense that

$$v_0^* < \overline{v},$$

where recall that $\overline{v} = \bar{F}^{-1}\left(\frac{1}{\Lambda}\right)$. In words, the revenue maximizer can not admit as many customers as she would want to if customers were not delay sensitive. This restricts us to the range of λ such that $\widetilde{\rho}(\lambda)$ is increasing and concave in λ and the range of v such that $\rho(v)$ is decreasing in v.

1.4. The Tension between Social Planning and Revenue Maximization

There is, of course, a link – or perhaps more accurately, a gap – between the objective of the social planner and that of the revenue maximizer. The social planner's objective is the sum of the firm's revenue and the consumers' surplus, i.e.,

$$S_K(\mathbf{v}, \mathbf{W}) = R_K(\mathbf{v}, \mathbf{W}) + C_K(\mathbf{v}, \mathbf{W})$$

where $C_K(\mathbf{v}, \mathbf{W})$ is the consumer surplus with K classes assuming that incentive compatible prices are used. It is written as the surplus from admitting all customers with valuations greater than v_1 less the waiting costs they incur:

$$\begin{aligned} C_K(\mathbf{v}, \mathbf{W}) &= \Lambda CS(v_1) - \sum_{i=1}^{K} \alpha \left(CS(v_i) - CS(v_{i+1})\right) W_i \\ &= \Lambda CS(v_1) - \sum_{i=1}^{K} \lambda_i(\mathbf{v}) c_i^{CS}(\mathbf{v}) W_i, \end{aligned}$$

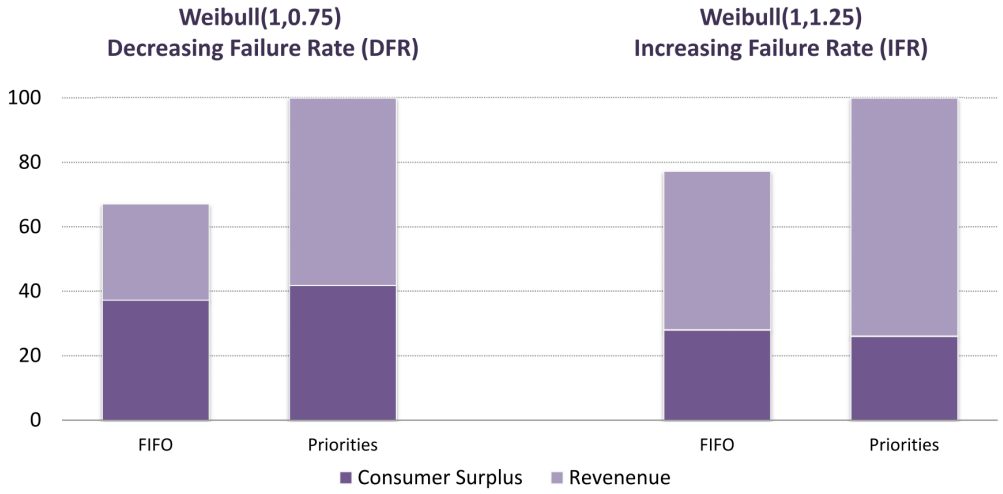


Figure 1.1. Priority to high valuation customers decreases consumer surplus if F is IFR. For ease of comparison, the results are scaled so that social welfare under priorities is equal to 100. Weibull(η, k) is a Weibull distribution with scale parameter η shape parameter k.

where $CS(v) = V(v) - \rho(v) = \int_v^\infty (u - v) f(u) du$ for $v \geq 0$ and $c_i^{CS}(\mathbf{v}) = \alpha \frac{CS(v_i) - CS(v_{i+1})}{\bar{F}(v_i) - \bar{F}(v_{i+1})}$ is the loss of surplus per unit of delay of a class i customer.

Recalling that $\bar{\lambda}_i(\mathbf{v}) = \Lambda \bar{F}(v_i)$, we have that

$$c_i^{CS}(\mathbf{v}) = \alpha \Lambda \frac{\widetilde{CS}(\bar{\lambda}_i(\mathbf{v})) - \widetilde{CS}(\bar{\lambda}_{i+1}(\mathbf{v}))}{\bar{\lambda}_i(\mathbf{v}) - \bar{\lambda}_{i+1}(\mathbf{v})},$$

where

$$\widetilde{CS}(\lambda) = CS\left(\bar{F}^{-1}\left(\frac{\lambda}{\Lambda}\right)\right). \tag{1.7}$$

Given $\mathbf{v}$, $c_i^{CS}(\mathbf{v})$ is increasing in i if $\widetilde{CS}$ is convex in λ.

The tension between R_K and C_K is important to the results of this paper and, as it turns out, the nature of this tension depends to a great extent on the failure rate of the distribution.

In Figure 1.1 we plot, for two classes, both the revenue and consumer surplus for different distributions and service disciplines. The total height of a bar corresponds to the social welfare under a particular priority scheme while the shading indicates how much is captured by the firm versus by customers. We compare FIFO service with a priority scheme in which those with higher valuations – and hence higher waiting costs – are served first.

On the left-hand side of the figure, we consider a Weibull distribution with a shape parameter of 0.75. In this case, both consumers and the firm are better off when priorities are used, and the increase in social welfare comes from making both parties better off. The right-hand side of the figure, however, presents a contrasting story. Here we have a Weibull distribution with a shape parameter of 1.25. Now consumer surplus falls as we move from FIFO service to priorities. Social welfare only increases because the service provider's revenue rises by more than consumer welfare drops. This is a generalizable phenomenon. There is a large class of valuation distributions – those with an increasing failure rate (IFR) – for which consumer surplus falls if those with higher valuations are given priority. It is straightforward that $CS(v_i) - CS(v_i) > 0$ so that work conservation is optimal. As with our analysis of the social planner and the revenue maximizer, if one wants to maximize consumer surplus it is optimal to preemptively prioritize customers in decreasing orders of $c_i^{CS}(v)$. However, the values $c_i^{CS}(v)$ may not be ordered as one expects. Specifically, if F has a decreasing failure rate (DFR) then $\widetilde{CS}$ is concave, implying that $c_i^{CS}(v)$ is increasing in i. Conversely, if F is IFR, $\widetilde{CS}$ is convex, implying that $c_i^{CS}(v)$ is decreasing.

Lemma 3. *$c_i^{CS}(\mathbf{v}) \leq c_{i+1}^{CS}(\mathbf{v})$ if the failure rate $h(\cdot)$ is decreasing and $c_i^{CS}(\mathbf{v}) \geq c_{i+1}^{CS}(\mathbf{v})$ if it is increasing. With constant failure rate, $c_i^{CS}(\mathbf{v}) = c_{i+1}^{CS}(\mathbf{v})$.*

When the valuation distribution is DFR[3], the priority scheme that maximizes consumer surplus aligns with that used by the social planner or the revenue maximizer: Customers with high valuations (and thus high waiting costs) are placed at the front of the line while those with low valuations (and thus low waiting costs) are relegated to the rear. However, the situation is reversed when valuations are governed by an IFR distribution. Now it would optimal to let those with low waiting costs enjoy shorter waits. This explains the phenomenon illustrated in Figure 1.1. The Weibull distribution is DFR for shape parameters less than one but IFR for shape parameters greater than one.

Prioritizing customers with low valuations is not implementable if valuations are private information. The prices and waiting times must satisfy (1.2) and (1.3) and a menu where customers with low valuations wait less cannot be incentive compatible. Yet, the decision maker could then maximize consumer surplus by using a service discipline such as FIFO that is independent of customer valuations.

For intuition on the role of the failure rate, note that when a customer with valuation v opts for class i, the price she pays to join that class is determined by the waiting cost of the lowest type to join that class, i.e., αv_i. However, her waiting costs are αv. Her contribution to consumer surplus is proportional to $v - v_i$. Consequently, it is worth considering the distribution of $v - v_i$ as a function of v_i. If v is IFR [DFR], then $v - v_i$ is stochastically decreasing [increasing] in v_i (Lai and Xie (2006)). Thus under an IFR distribution, customers who choose a higher class contribute less in expectation to consumer surplus and to waiting costs than those who select a lower class. That relationship is reversed with a DFR distribution.

[3]There are IGFR distributions such as the log-normal that have non-monotone failure rates. However, one cannot easily characterize how a non-monotone failure rate impacts the ordering of $c_i^{CS}(v)$. It may be that all change points of $h(v)$ fall below $\bar{v}$ so $h(v)$ is monotone over the relevant range. Otherwise, we cannot rule out that the value of $c_i^{CS}(v)$ may not be monotone.

Additionally, consider the elasticity of $\widetilde{CS}(\lambda)$:

$$\eta(\lambda) = \frac{\lambda \widetilde{CS}'(\lambda)}{\widetilde{CS}(\lambda)} = \frac{1}{MRL\left(\bar{F}^{-1}\left(\frac{\lambda}{\Lambda}\right)\right) h\left(\bar{F}^{-1}\left(\frac{\lambda}{\Lambda}\right)\right)}.$$

We have two ways of interpreting $\eta(\lambda)$. The first is just the relative rate of change in consumer surplus. The second is to tie it to waiting costs. The revenue maximizer charges one price to all customers and that depends on the lowest admitted type. The social planner, of course, cares about the average waiting cost. $\alpha\widetilde{CS}(\lambda)$ is then the gap between these values and $\eta(\lambda)$ is the elasticity of this gap.

To examine the behavior of $\eta(\lambda)$, standard results give that $MRL'(v) = MRL(v)h(v) - 1$ and $MRL'(v) \geq -1$. Additionally, all IFR[DFR] distributions have decreasing[increasing] mean residual life functions Lai and Xie (2006). The next lemma is then immediate.

Lemma 4.

$$\eta(\lambda) = \frac{1}{MRL'\left(\bar{F}^{-1}\left(\frac{\lambda}{\Lambda}\right)\right) + 1}.$$

If $F(v)$ is strictly IFR[DFR], then $\eta(\lambda) > [<]1$. If $MRL(v)$ is convex[concave], $\eta(\lambda)$ is decreasing[increasing].

When $\widetilde{CS}(\lambda)$ is elastic (i.e., $\eta(\lambda) > 1$), a small change in the throughput results in a relatively large change in surplus. Since a larger λ corresponds to a lower v, it is those with lower valuations that would then have an oversized waiting cost relative to the price they have paid and thus should be given priority. When $\widetilde{CS}(\lambda)$ is inelastic (i.e., $\eta(\lambda) < 1$), the situation is reversed and those with lower valuations should be placed at the back of the line. We will see below that the last part of the lemma will have implications for classification.

We close this section by noting that our model, while obviously stylized, has some implications for how priorities impact consumers. In particular, it suggests that there are some settings in which customers as a whole are better off under FIFO service than under incentive compatible priority schemes. Any IFR distribution has a coefficient of variation less than one. Conversely, any DFR distribution has a coefficient of variation greater than one (Barlow and Proschan (1965)). That suggests that consumers may welcome the introduction of a priority scheme when there is significant dispersion in values and waiting costs. Conversely, when there is not much dispersion, the average consumer is likely better off under FIFO.

This last assertion implicitly assumes that moving from FIFO to priorities does not increase coverage. Once coverage is in play, consumer surplus is potentially pulled in two ways. Priorities may harm customers but expanded coverage would benefit them. We examine this question in Table 1.1.

Shape Parameters.	Coefficient of Variation	% Admitted FIFO	% Admitted Priority	Priority Surplus/ FIFO Surplus
1.0	1.00	13.56%	15.62%	1.0543
1.5	0.68	14.79%	16.12%	1.0088
2.0	0.52	15.33%	16.29%	0.9871
2.5	0.43	15.63%	16.38%	0.9743
3.0	0.36	15.82%	16.44%	0.9659

Table 1.1. Weibull valuation with scale parameter set so that v_0^* remains at 10. The other parameters are set to $\alpha = 0.25$ and $\Lambda = 3$.

We consider customers whose valuations follow a Weibull distribution and vary the shape parameter from one to three. This takes the coefficient of variation from 1 down to to 0.36. For each value, we compute the optimal coverage (reported as the fraction of customers admitted) offered by a revenue maximizer under both FIFO and a priority scheme with two levels (i.e., $K = 2$). We see that in all examples, the revenue maximizer expands coverage when offering priorities. However, the increase in coverage falls as valuations become less dispersed. Looking

at the last column, one sees that the ratio of consumer surplus under the priority scheme to the surplus under FIFO falls as valuations become less dispersed. For sufficiently low coefficients of variation consumers are indeed worse off under a priority scheme.

Implications for classification: With IFR valuation distributions, $c_i^{CS}(\mathbf{v})$ is decreasing in i and the social planner faces a tension as it seeks to maximize the sum of firm revenue and consumer surplus: Revenue maximization dictates prioritizing customer with higher valuations while consumer surplus is compromised by such priority. We will see that the social planner will alleviate some of this tension by having a smaller high priority class than the revenue maximizer; see Remark 1 further below.

1.5. Coverage, Coarseness and Classification

If the providers are restricted to use a single menu item – an admission fee and a delay to go with it, the revenue maximizer's problem reduces to

$$R_1^* = \max_{v>\bar{v}} \Lambda\rho(v)\left(1-\alpha W_1(\lambda(v))\right) \tag{1.8}$$

where $W_1\left(\lambda(v)\right) = (1-\Lambda\bar{F}(v))^{-1}$ is the delay under FIFO service. The social planner's problem is

$$S_1^* = \max_{v>\bar{v}} \Lambda V(v)(1-\alpha W_1(\lambda(v))), \tag{1.9}$$

The function $(1-W_1(\lambda(v)))$ is concave increasing and $V(v)$ is concave and decreasing in v. The social planner's objective function is thus strictly concave and has a unique maximizer

v_S^*. By Lemma 2 $\rho(v)$ is concave and decreasing for all $v > \bar{v}$ so that the revenue has a unique maximizer v_R^*.

These are special instances of the problems in Afèche and Mendelson (2004). It follows from their Proposition 1 and our Lemma 2 (particularly, part 3) that with an IGFR distribution $v_S^* \leq v_R^*$ so that $\lambda_1(v_S^*) \geq \lambda_1(v_R^*)$. The revenue maximizer offers a smaller coverage than the social planner. If F has a constant generalized failure rate (i.e., it is a Pareto distribution), then $\lambda_1(v_R^*) = \lambda_1(v_S^*)$.

Contrast this with a setting in which the providers can tailor a different price and delay to each valuation v. The welfare from a customer with valuation v is $v - \alpha v W(v)$. The higher the customer's valuation, the higher her priority so that a customer with valuation v has an expected waiting time $W_c(v) = \left(1 - \Lambda \bar{F}(v)\right)^{-2}$ (Kleinrock 1967, Theorem 2) which is a continuous segmentation analogue of (1.5). The social planner solves the problem

$$S_\infty^* = \max_{v_S \geq \bar{v}} \Lambda \int_{v_S}^{\infty} (v - \alpha v W_c(v)) f(v) \, dv. \tag{1.10}$$

The revenue maximizer chooses similarly an admission valuation v_R^* such that a customer enters if and only if her valuation is $v \geq v_R^*$. The price is determined so that the first customer to enter is indifferent between entering or not, i.e., $p(v_R^*) = v_R^* - \alpha v_R^* W_c(v_R^*)$. All other entering customers pay the entering price plus a premium for shorter delays, i.e.,

$$p(v) = p(v_R^*) - \alpha \int_{v_R^*}^{v} u W_c'(u) du.$$

Since $W_c'(v) \leq 0$ the premium is positive. The revenue maximizer's problem reduces then to

$$R_\infty^* = \max_{v_R \geq \bar{v}} \Lambda \left(-p\left(v_R\right) \bar{F}\left(v_R\right) + \int_{v_R^*}^{\infty} p'\left(v\right) \bar{F}(v) dv \right) \tag{1.11}$$

Problems (1.10) and (1.11) are instances of the priority auctions described in Afèche and Mendelson (2004). By Proposition 3 there, $v_R^* = v_S^*$ (i.e., the two providers choose the same entry cutoff) provided that $\epsilon(\lambda)$ (recall Lemma 2) is increasing in λ.

In summary, with a single class, for strictly IGFR valuation distributions, the revenue and social optimizers choose different coverage levels and the social planner serves more customers. With a continuum of classes, however, they make identical decisions. We turn to study the intermediate and practical case in which there is a finite number of classes.

High-volume analysis. When dealing with multiple (but finite number of) classes, the social and revenue maximizer problems are rather intractable. Fortunately, when the volume is high we can characterize the decisions of the providers and compare them. We build on the approach of Nazerzadeh and Randhawa (2015) and study the asymptotic performance, as the arrival rate and service rate are scaled up by a multiplier n: the nominal arrival rate is Λn and the service rate is n. We superscript all relevant notation with n to capture the dependence on this multiplier. Thus, for example, $v_{i,R}^{n*}$ [$v_{i,S}^{n*}$] is the optimal class-i cutoff of the revenue maximizer [social planner]. Since both the nominal arrival rate and the service rate are multiplied by n, $\bar{v}$ does not depend on n.

For large values of n, it turns out, the differences between subsequent optimal cutoff values, $v_{i+1}^{n*} - v_i^{n*}$ become small. Roughly speaking we can then replace $V(v_{i+1}^n) - V(v_i^n)$ in the objective function (1.4) of the social planner with a Taylor expansion around v_i^n. It is such Taylor expansions that enable the analysis of an otherwise intractable problem. Our analysis

allows us to nail down the optimal arrival-rate vectors (and revenue/welfare outcomes) up to an error that is negligible relative to the square root of the multiplier n. Accordingly, two coverage/classification decisions are distinguishable if they are $\sqrt{n}$ apart.

We first re-visit the single-class FIFO queue. From Afèche and Mendelson (2004) it follows that, in our setting, the revenue maximizer admits fewer customers than socially optimal. We can now characterize more fully the difference in coverage.

Following standard notation, we say that $\xi(n) = o(\sqrt{n})$ if $\xi(n)/\sqrt{n} \to 0$ as $n \to \infty$.

Lemma 5 (coverage difference with FIFO)**.** *With a single (FIFO) class the cut-off entry valuations of the social planner and the revenue maximizer satisfy*

$$v_{1,S}^{n*} = \bar{v} + \sqrt{\alpha}\sqrt{-\frac{\bar{F}(\bar{v})\,V(\bar{v})}{f(\bar{v})\,V'(\bar{v})}}n^{-\frac{1}{2}} + o(n^{-\frac{1}{2}}),\ v_{1,R}^{n*} = \bar{v} + \sqrt{\alpha}\sqrt{-\frac{\bar{F}(\bar{v})\,\rho(\bar{v})}{f(\bar{v})\,\rho'(\bar{v})}}n^{-\frac{1}{2}} + o(n^{-\frac{1}{2}}).$$

Optimal admission rates consequently satisfy,

$$\bar{\lambda}_{1,S}^{n*} = n - \Lambda f(\bar{v})\sqrt{\alpha}\sqrt{-\tfrac{\bar{F}(\bar{v})V(\bar{v})}{f(\bar{v})V'(\bar{v})}}\sqrt{n} + o(\sqrt{n}),\ \bar{\lambda}_{1,R}^{n*} = n - \Lambda f(\bar{v})\sqrt{\alpha}\sqrt{-\tfrac{\bar{F}(\bar{v})\rho(\bar{v})}{f(\bar{v})\rho'(\bar{v})}}\sqrt{n} + o(\sqrt{n})$$

If F is IGFR then $V(\bar{v})/V'(\bar{v}) > \rho(\bar{v})/\rho'(\bar{v})$ and, consequently, the social planner has a larger coverage. Furthermore, the difference in coverage is non-negligible, $\bar{\lambda}_{1,S}^{n*} - \bar{\lambda}_{1,R}^{n*} \neq o(\sqrt{n})$ if the inequality is strict.

We next show that with multiple customer classes the coverage gap disappears. Classification, however, differs depending on the mean residual life of the distribution. Theorem 1 below is the main result of this paper and focuses on the comparison of the actions of the social planner and the revenue maximizer. The full characterization of their decisions appears in Theorems 4 and 5 in the appendix.

Theorem 1. (**coverage, coarseness and classification: SP Vs. RM**) *With $K > 1$ levels of service, the coverage of the social planner and the revenue maximizer are asymptotically identical in the sense*

(Coverage) $$\bar{\lambda}_{1,R}^{n*} - \bar{\lambda}_{1,S}^{n*} = o(\sqrt{n}).$$

For both, two classes are sufficient:

(Coarseness) $$S_K^* = S_2^{n*} + o(\sqrt{n}), \text{ and } R_K^* = R_2^{n*} + o(\sqrt{n}).$$

Classification is asymptotically different except for linear MRL:

(Classification) $$\bar{\lambda}_{2,R}^{n*} - \bar{\lambda}_{2,S}^{n*} = \gamma n^{3/4} + o(n^{3/4}),$$

where $\gamma \geq 0$ (resp. $\gamma \leq 0$) if the valuation distribution has a convex (resp. concave) MRL and $\gamma = 0$ if the MRL is linear. In particular, the revenue maximizer directs more volume to the high priority when F has a convex MRL. Further, if $\gamma \neq 0$ (classification is asymptotically different), the social cost of revenue maximization grows (at least) as fast as the square root of the arrival rate:

(Social welfare gap) $$\liminf_{n\to\infty} \frac{S_2^{n*} - S_2^n(\mathbf{v}_R^{n*})}{\sqrt{n}} > 0.$$

We find then that both providers choose very coarse priority schemes and offer identical coverage (up to a small difference) but that their admission to the high priority class differs. Table 1.2 is a numerical illustration of this result. The example serves to illustrate three additional points: (i) that the asymptotics-based comparison holds also for small values of n, (ii)

that the difference in classification can be rather large: for Weibull(1,0.3) which has a concave MRL $\gamma = -140 < 0$. It can also be rather small: for the Weibull(1,2) which has convex MRL, $\gamma = 0.011$, and (iii) that both the social planner and the revenue maximizer have a high priority class that is much larger than the low priority class. This latter observation is supported in Theorems 4 and 5 in the appendix.

	Coverage		High Priority	
n	$\bar{\lambda}^{n*}_{1,S}$	$\bar{\lambda}^{n*}_{1,R}$	$\lambda^{n*}_{2,S}$	$\lambda^{n*}_{2,R}$
10	8.573	8.568	5.552	5.57
100	95.495	95.485	73.683	73.861
1000	985.789	985.767	847.413	848.764
	Coverage		High Priority	
n	$\bar{\lambda}^{n*}_{1,S}$	$\bar{\lambda}^{n*}_{1,R}$	$\lambda^{n*}_{2,S}$	$\lambda^{n*}_{2,R}$
10	8.44	7.9	4.487	3.548
100	95.177	92.365	64.783	53.451
1000	985.144	975.720	786.348	695.097

Table 1.2. $\Lambda = 30$, $\alpha = 0.2$: (LHS) Weibull(1,2) for convex MRL (RHS) Weibull(1,0.3) for concave MRL.

The differences in classification matter. They suffice to make the revenue maximizer's actions socially inefficient. Figure 1.2(LHS) displays $(S_2^{n*} - S_2^n(\mathbf{v}_R^{n*})/\sqrt{n}$ as a function of n for a Weibull(1,0.3) distribution (concave MRL). Thus, while a continuum of classes means that maximizing revenue is equivalent to maximizing welfare, for each finite K revenue maximization leads to non-negligible social inefficiency. Additional numerical examples appear in section A.6 of the e-companion.

Further, any number of classes larger than two has little impact on either revenue or social welfare. This was proved already for the revenue maximizer by Nazerzadeh and Randhawa (2015) and we extend this result to the social planner. Figure 1.2(RHS) provides numerical evidence for this result.

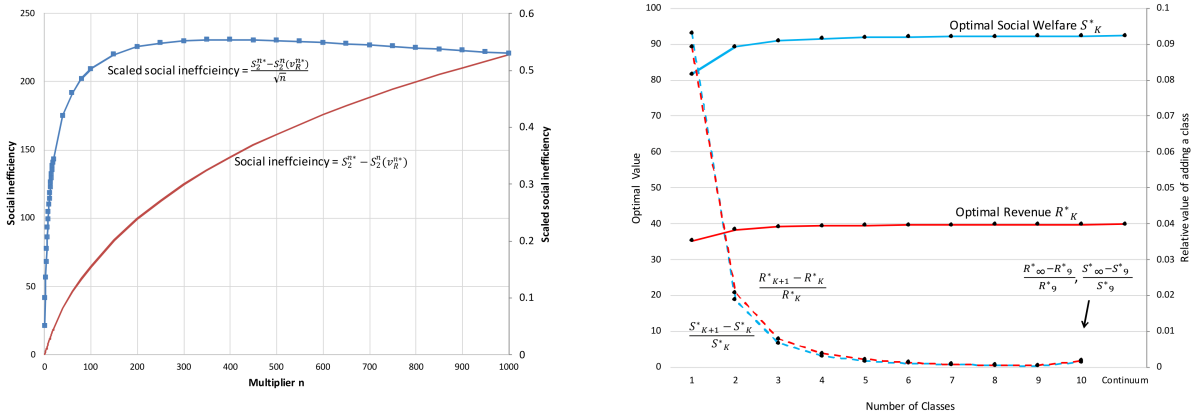


Figure 1.2. F =Weibull(1,0.3), $\alpha = 0.2$, $\Lambda = 30$ (LHS) The social inefficiency of the RM actions as a function of the multiplier n. (RHS) The effect of the number of classes on social welfare and revenue $(n = 1)$.

Remark 1 (The Role of Mean Residual Life)**.** *The mean residual life function plays a decisive role in the comparison between the decisions of the social planner and the revenue maximizer because it is intimately linked to consumer surplus. If we impose a bit of structure, we can apply Lemma 3. If the MRL is both convex and decreasing, the failure rate is increasing. The social planner then faces a trade off: Maximizing revenue calls for putting those with high valuations at the front of the line but maximizing consumer surplus requires putting those with low valuations first. The former dominates but limiting the size of the high priority class minimizes the impact on low-valuation customers. The revenue maximizer has no such compunctions and thus opts for a larger high priority class. The situation is reversed when the MRL is both concave and increasing. The failure rate would now be decreasing. From the social planner's perspective, giving high value customers high priority boosts both firm revenue and consumer surplus. She then has an incentive to classify a large number of customers as high priority. The revenue maximizer, of course, does not see any benefit to increasing consumer surplus and therefore is more conservative than the social planner in expanding the higher priority class.*

From Lemma 4, we have that when the MRL is convex, adding more customers to the high priority class increases consumer surplus at a decreasing rate. Conversely, a concave MRL implies adding more customers to the high priority class increases the marginal rate at which surplus increases. In the social planner's eyes, the former favors a limited high priority class while the latter favors a large high priority class.

Additionally, recall that we can interpret $\eta(\lambda)$ as the elasticity of the gap between average waiting costs (the social planner's concern) and the wait-driven component of the price (the revenue maximizer's concern). We then have that if $\eta(\lambda)$ is decreasing, the revenue maximizer's objective underestimates the social cost of waiting as v increases (causing λ to decrease). Conversely, the revenue maximizer's objective will overestimate the cost of waiting as v increases if $\eta(\lambda)$ is increasing. ■

Remark 2 (Who pays for social inefficiency)**.** *When the social planner and the revenue maximizer's classification decisions diverge, consumer surplus may be smaller than socially optimal. Consider the case of $K = 2$ (two classes). While the average customer stands to lose from revenue maximization, some customers may gain. In fact, with a concave MRL, the revenue maximizer has a smaller high priority class and, consequently, offers a shorter delay to both classes and a higher customer utility (per class); see Table 1.3 for a numerical example. The total consumer surplus is higher under social planning. Who gains depends on the convexity/concavity of the MRL.*

Figure 2.8 provides a schematic view of how the utility $U(v; i)$ changes with the customer valuation v. Informally representing Theorem 1, the figure has both service providers offering

	(p_i, W_i)	*Avg. Utility*
SP - Low	$(53.396, 1.163)$	16.046
SP - High	$(76.921, 0.181)$	198.197

	(p_i, W_i)	*Avg. Utility*
RM - Low	$(63.071, 0.737)$	23.499
RM - High	$(79.804, 0.155)$	236.390

Table 1.3. Waiting and pricing menus for $F =$ Weibull(1,0.3) (concave MRL), $\alpha = 0.2$, and nominal arrival rate $\Lambda n = 300$: (LHS) Social planner (RHS) Revenue Maximizer

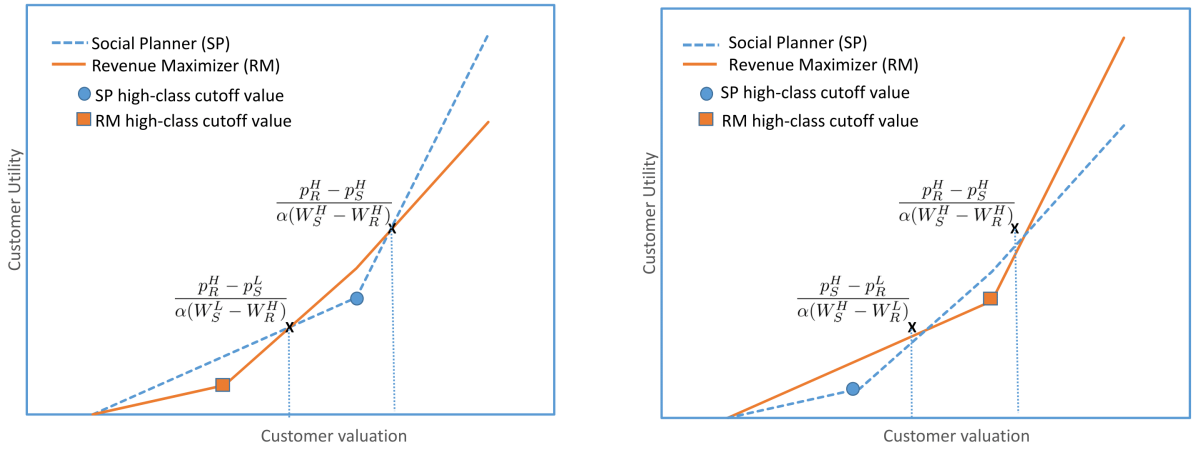


Figure 1.3. Utility of customers under SP and RM: (LHS) Convex MRL, (RHS) Concave MRL

identical coverage. With a convex MRL (LHS), the revenue maximizer has a smaller high priority cutoff (represented by a square) than that of the social planner (represented by a circle) and, in turn, a larger high priority class. The story is reversed with a concave MRL (RHS).

More in detail: let (P_S^L, W_S^L) *and* (P_S^H, W_S^H) *be the two menu items offered by the social planner (where* L *stands for low priority, i.e.,* $W_S^L \geq W_S^H$*). Define similar notation for the revenue maximizer (with the subscript* R*). Because of the equal coverage, we have* $v_1 - p_R^L - \alpha v_1 W_L^R = v_1 - p_S^L - \alpha v_1 W_S^L = 0$ *so that* $(p_S^L - p_R^L) = \alpha v_1 (W_R^L - W_S^L)$.

For a convex MRL, we have $W_R^L \geq W_S^L$ *so that the revenue maximizer charges a lower price than the social planner. However, that price reduction is insufficient to compensate low priority*

customers. Any customer with valuation greater than v_1 who would be classified as low priority by either decision maker is worse off under the revenue maximizer. One gets similar results at the other extreme. High priority customers also get a price break from the revenue maximizer but for those with very high valuations it is does not adequately compensate for increased waits. Thus, with a convex MRL, those with very high or very low valuations are certain to lose as we move from the social planning to the revenue maximization.

With a concave MRL, the story is reversed. Customers with valuation greater than than $(p_R^H - p_S^H)/(\alpha(W_S^H - W_R^H))$ or smaller than $(p_S^H - p_R^L)/(\alpha(W_S^H - W_R^L))$ are better off under revenue maximization. These are the customers that absorb all the gain in the class utilities seen in Table 1.3. The burden of social inefficiency is all carried by the "middle class": the customers with intermediate valuations. In particular those who would have been classified as high priority under the social planner but are moved to low priority under the revenue maximizer are worse off. With convex MRL, in contrast, those with very high or very low valuations are certain to lose as we move from the social planning to the revenue maximization. However, we cannot guarantee that those with intermediate values are better off. ■

1.6. Extensions

1.6.1. Non-Preemptive Policies

Thus far we allowed the provider to use preemptive policies and, indeed, in their optimal solutions both providers use preemptive static priorities. It is often unrealistic to preempt customers in the middle of their service. A restriction to non-preemptive policies entails a change in the menus offered to customers. We prove, however, that our results do not change: whereas the preemptive and non-preemptive menus might differ, the differences are relatively negligible.

The objective functions $S_K(\mathbf{v}, \mathbf{W})$ and $R_K(\mathbf{v}, \mathbf{W})$ are the same as in §1.3. The difference is that, in optimizing, $\mathbf{W}$ must be taken from the *non-preemptive* achievable region $\mathcal{W}_{NP}(\mathbf{v})$. Let $S^{n*}_{2,NP}$ be the optimal social welfare (optimized over all valuation cutoffs and non-preemptive disciplines). Since the providers are now constrained now to use non-preemptive policies it holds that $S^{n*}_{2,NP} \leq S^{n*}_2$ and $R^{n*}_{2,NP} \leq R^{n*}_2$. Our next theorem shows that the loss is minimal.

Let $S^n_{2,NP}(\mathbf{v}^n)$ be the social planner's objective function value when two-class are used, the cut-off vector $\mathbf{v}^n$ is used for classification and static *non-preemptive* priority is used with the highest priority provided to the highest valuations customers. Similarly define $R^n_{2,NP}(\mathbf{v}^n)$ for the revenue maximizer.

Theorem 2. (**Optimality of preemptive cut-offs with non-preemptive service**) *Using two non-preemptive priority classes with the optimal preemptive cut-off valuations is nearly optimal in the $\sqrt{n}$ scale for both the social planner and the revenue maximizer. That is,*

$$S^n_{2,NP}(\mathbf{v}^{n*}_S) - S^{n*}_2 = o(\sqrt{n}) \text{ and } R^n_{2,NP}(\mathbf{v}^{n*}_R) - R^{n*}_2 = o(\sqrt{n}).$$

respectively.

The non-preemptive provider can use the same coarseness, coverage and classification as the preemptive provider with negligible compromise to optimality. In turn, the comparisons that apply to the preemptive case, apply to the non-preemptive restriction.

1.6.2. General delay costs

In our base model, a customer's delay cost rate is proportional to her valuation. Here we show that our results generalize to a broader class of delay cost functions. Specifically, we consider delay costs $d(v)$ which are differentiable and increasing at a sub-linear rate, i.e., functions for which $d(v)/v$ is non-increasing.[4] The sub-linearity of $d(v)$ preserves the ordering of customer types in the sense that if a customer with valuation v chooses to patronize the service, any customer with valuation $v' > v$ will also purchase the service (although she may choose a different priority level). Coverage then will be determined by a minimum valuation to admit.

The optimality of preemptive static priorities for social planning follows from the monotonicity of $d(v)$. The sub-linearity, as in Nazerzadeh and Randhawa (2015), ensures the optimality of preemptive priority also for the revenue maximizer.

Theorem 1 can be extended to this more general setting with a replacement of the MRL concavity/convexity conditions with one stated in terms of a suitable elasticity measure. We define

$$\zeta(\lambda) = -\frac{M(\lambda)}{M'(\lambda)\lambda} \tag{1.12}$$

where

$$N(v) = D(v) - d(v)\bar{F}(v), \text{ with } D(v) = \int_v^\infty d(u)f(u)du,$$

and

$$M(\lambda) = N\left(\bar{F}^{-1}\left(\frac{\lambda}{\Lambda}\right)\right). \tag{1.13}$$

[4]This does not necessarily imply that $d(\cdot)$ is a concave function—take $d(v) = v + 1/(v+1)$ as a case in point.

Theorem 3. (**coverage, coarseness and classification for generalized delay**) *With $K > 1$ levels of service, the coverage of the social planner and the revenue maximizer are asymptotically identical in the sense*

(Coverage) $$\bar{\lambda}_{1,R}^{n*} - \bar{\lambda}_{1,S}^{n*} = o(\sqrt{n}).$$

For both, two classes are sufficient:

(Coarseness) $$S_K^* = S_2^{n*} + o(\sqrt{n}), \text{ and } R_K^* = R_2^{n*} + o(\sqrt{n}).$$

Classification is asymptotically different except for constant $\zeta(\cdot)$:

(Classification) $$\bar{\lambda}_{2,R}^{n*} - \bar{\lambda}_{2,S}^{n*} = \gamma n^{3/4} + o(n^{3/4}),$$

where $\gamma \geq 0$ (resp. $\gamma \leq 0$) if $\zeta(\cdot)$ is increasing (resp. decreasing) and $\gamma = 0$ if the $\zeta(\cdot)$ is constant. In particular, the revenue maximizer directs more volume to the high priority when $\zeta(\cdot)$ is increasing. Further, if $\gamma \neq 0$ (classification is different), the social cost of revenue maximization grows at least with an order of $\sqrt{n}$,

(Social welfare gap) $$\liminf_{n\to\infty} \frac{S_2^{n*} - S_2^n(\mathbf{v}_R^{n*})}{\sqrt{n}} > 0.$$

When specializing the delay-cost to be linear — $d(v) = \alpha v$ — as in our base model, the elasticity requirements reduce to convexity/concavity requirements on the MRL. For completeness, the formal derivation appears at the end of the e-companion.

Remark 3 (the role of elasticity)**.** *In our base model, whether the revenue maximizer pursued a mass-luxury or ultra-luxury strategy depended on whether the elasticity of consumer surplus $\eta(\lambda)$ was increasing or decreasing. That intuition carries over to our generalized cost structure. Comparing $\zeta(\lambda)$ and $\eta(\lambda)$, one sees that the former is the natural generalization of the latter. Both can be interpreted the elasticity of the gap between average waiting costs and the wait-driven component of the price. We again have that a decreasing elasticity results in the revenue maximizer's objective function underestimating the consequences of waits while an increasing elasticity results in the revenue maximizer overestimating their impact.* ■

Remark 4 (what may happen with decreasing $d(v)$)**.** *To this point we have an only considered having waiting costs that are positively correlated with values, making it optimal for both the revenue maximizer and the social planner to move those with high valuations to the front of the line. We now briefly consider what happens when valuations and waiting costs are negatively correlated. Specifically, we suppose that the per unit time waiting cost function $d(v)$ is decreasing in v. Since the priority scheme depends only on the waiting cost and not on the consumer's value of the service, it would now be optimal to give high priority to those with low valuations.*

Table 1.4 reports results for when $d(v) = \frac{1}{v^2}$. Here we consider a range of market sizes Λ and degrees of coarseness K (i.e., number of priority classes). For each market size-coarseness level pair, we determine the optimal incentive-compatible scheme for both types of decision maker and report their optimal objective value.[5] *Some results are not surprising. For example, both types of decision makers are better off in large markets in which patient, high-value*

[5]Note that we do not allow a decision maker to offer a degenerate class. Thus for each $K > 1$, we constrain the decision maker offering price and wait menus that result in positive traffic for each class.

customers are plentiful. Other results are not as straightforward. In particular, the two decision makers offer very different pricing and priority schemes as the market grows. The revenue maximizer always benefits from offering as many priority classes as possible. However, the social planner may prefer coarser and coarser schemes. For $\Lambda = 3,$ she prefers to offer four classes but drops to three classes when $\Lambda = 4$ and goes down to two classes when $\Lambda = 5.$ For Λ sufficiently high, the social planner opts for FIFO service, pooling all customers into a single class.

	1 Class		*2 Classes*		*3 Classes*		*4 Classes*		*Max**	
Λ	*RM*	*SP*	*RM*	*SP*	*RM*	*SP*	*RM*	*SP*	*RM*	*SP*
2.9	*0.555*	*3.895*	*0.573*	*4.573*	*0.577*	*4.663*	*0.578*	*4.686*	*4 Classes*	*4 Classes*
3	*0.571*	*4.127*	*0.589*	*4.838*	*0.593*	*4.930*	*0.595*	*4.951*	*4 Classes*	*4 Classes*
3.2	*0.602*	*4.609*	*0.622*	*5.385*	*0.627*	*5.477*	*0.628*	*5.492*	*4 Classes*	*4 Classes*
3.5	*0.647*	*5.376*	*0.670*	*6.237*	*0.675*	*6.326*	*0.677*	*6.325*	*4 Classes*	*3 Classes*
4	*0.718*	*6.760*	*0.746*	*7.728*	*0.752*	*7.795*	*0.755*	*7.760*	*4 Classes*	*3 Classes*
5	*0.848*	*9.861*	*0.885*	*10.899*	*0.893*	*10.853*	*0.896*	*10.706*	*4 Classes*	*2 Classes*
6	*0.963*	*13.296*	*1.008*	*14.227*	*1.019*	*13.970*	*1.023*	*13.665*	*4 Classes*	*2 Classes*
7	*1.067*	*16.961*	*1.118*	*17.620*	*1.131*	*17.088*	*1.135*	*16.564*	*4 Classes*	*2 Classes*
8	*1.160*	*20.783*	*1.217*	*21.026*	*1.231*	*20.171*	*1.237*	*19.450*	*4 Classes*	*2 Classes*
9	*1.245*	*24.707*	*1.307*	*24.419*	*1.323*	*23.191*	*1.329*	*22.156*	*4 Classes*	*1 Classes*
10	*1.323*	*28.697*	*1.390*	*27.794*	*1.406*	*26.117*	*1.412*	*24.838*	*4 Classes*	*1 Classes*

Table 1.4. Decreasing delay-cost function $d(v) = 1/v^2$. The valuation distribution is Weibull(1,1.1).

Even if the two decision makers both prefer offering more priority classes, our results with positively correlated waiting costs may not carry over to a setting with negatively correlated costs. Figure 1.4 presents examples using $d\left(v\right) = \frac{1}{v}$ under which both decision makers prefer offering multiple priority classes. The top graphs show that on a scaled basis, both decision makers benefit from offering more classes. That is, the loss from going from arbitrarily many classes to a very coarse scheme with just two classes is no longer negligible. Further, the

bottom graph shows that the coverage differences between the social planner and the revenue maximizer grows on the scale of $\sqrt{n}$. We conclude that while with positively correlated waiting costs the differences between revenue and welfare maximization are largely questions of classification, with negatively correlated coverage and classification are also in play. ■

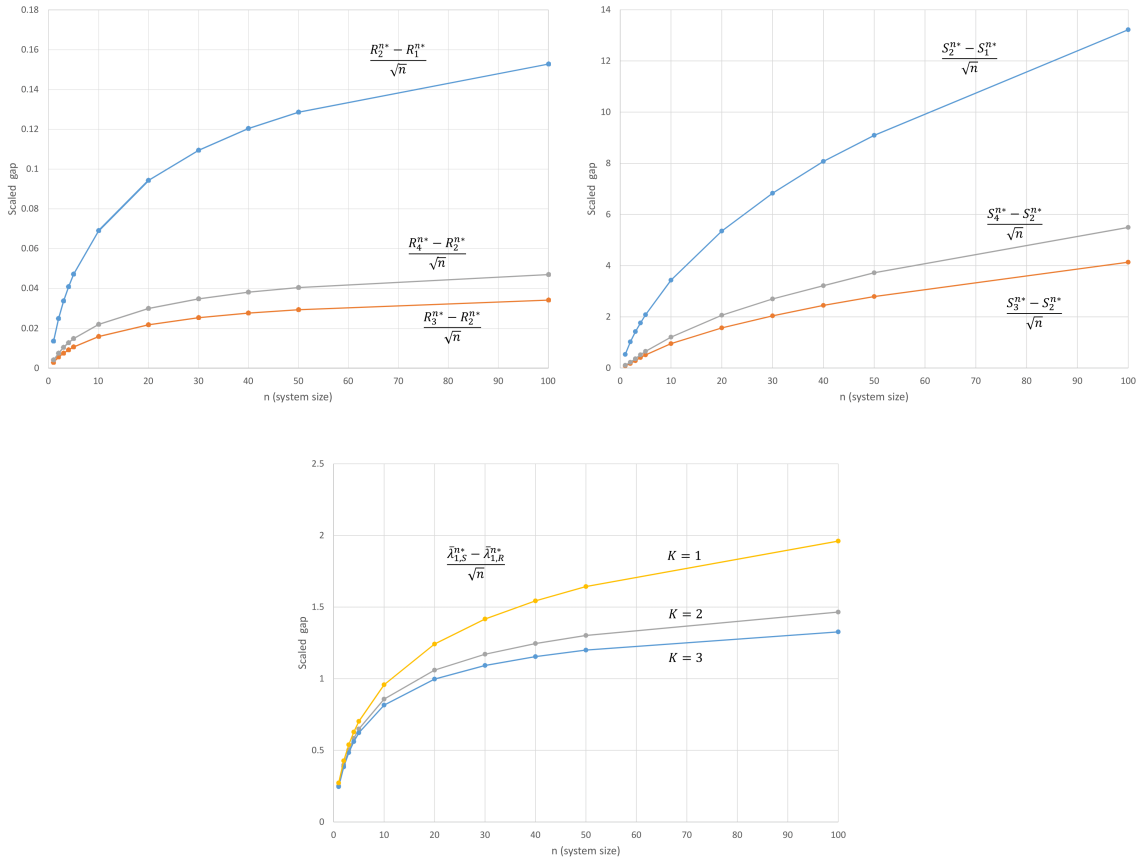


Figure 1.4. (TOP) The effect of increasing the number of classes K beyond 2 is non-negligible in the $\sqrt{n}$ scale as captured by the series $(R_4^{n*} - R_2^{n*})/\sqrt{n}$ for the revenue maximizer and $(S_4^{n*} - S_2^{n*})/\sqrt{n}$ for the social planner. (BOTTOM) the difference in coverage persist (for each K) as n grows.

1.7. Conclusion

Managing a service system as a priority queue is a challenging endeavor. To map a continuum of customers into a finite number of priority classes requires multiple decisions. One must determine how much of the market to cover, how coarse a set of priorities to offer, and how to classify customers into specific grades of service. Given such complexity, it is remarkable that decision makers pursuing distinct goals – maximizing social welfare and maximizing revenue – can agree on two out of three of these dimensions. We show in a limiting regime that a social planner and a revenue maximizer choose essentially identical levels of coverage while being content to offer very coarse priority schemes.

Meaningful differences exist, however, in how customers are classified. The revenue maximizer may pursue an ultra-luxury strategy and admit too few customers to the high priority class (in comparison to the social optimal) or a mass-luxury strategy and admit too many to the high priority class. These differences in classification are driven by differences in the behavior of consumer surplus, which is captured by the mean residual life of the customer valuation distribution. A concave MRL implies that consumer surplus becomes more elastic as more customers are admitted. The revenue maximizer does not consider this change and consequently follows an ultra-luxury strategy with a limited high priority class. Conversely a convex MRL means that consumer surplus becomes less elastic as the throughput increases. Again, the revenue maximizer ignores this factor and opts for a mass-luxury strategy that places more customers than is socially optimal in the high priority class. These results are robust to generalizations of the delay cost structure that maintain the positive correlation between a customer's valuation

of the service and her waiting cost. However, we show that if values and delay costs are negatively correlated, the decision makers may differ on every dimension, choosing different levels of coarseness, coverage and classification.

We also address the more general question of how priority schemes affect consumer surplus. We show that if the valuation distribution has a decreasing failure rate, consumer surplus is maximized by the most natural priority scheme, the one that puts those with high costs at the front of the line (which coincidently is the scheme that both the social planner and revenue maximizer would use). However, when the valuation distribution has an increasing failure rate, that natural priority scheme results in a lower consumer surplus than simple first-in, first-out service. Since a decreasing, convex MRL implies an increasing failure rate, this suggests that a convex MRL should cause the social planner to limit deviations from FIFO waits to protect consumer surplus; the revenue maximizer has no such concerns.

There are several ways this work can be extended. In particular, introducing competition could be fruitful. Competition generally results in consumers capturing more of the value the system creates. Here value can be shifted to customers through lower prices or more efficient classification. Which is the better means for rewarding customers is an interesting question.

CHAPTER 2

The Effect of Real-Time Information on Service Efficiency

Joint work with Toni Moreno and Nil Karacaoglu Garro

2.1. Introduction

Real-time information is becoming available in many services, and both customers and agents can easily access a wealth of information that may affect their decisions. For example, customers can use the GasBuddy application to monitor the price of gas in stations close to their location, or the Apple Store application to check the real-time availability of Apple products at Apple's brick-and-mortar stores.

The availability of real-time information is particularly important in two-sided markets, where a platform connects service providers with customers. In those markets, real-time information not only affects customers' actions, but also that of service providers. For example, Airbnb connects individuals looking for accommodation with private property owners, and Uber connects individuals looking for a ride with drivers able to offer one. The supply side of these markets consists of a large number of individual agents who try to maximize their individual profits. When doing so, they have access to very detailed, often real-time information. The focus of this paper is to study how hypergranular spatial real-time information affects the decisions of individual service providers and to explore the consequences of the availability of such detailed information for the efficiency of service platforms.

One of the leading applications of service platforms can be found in e-hailing. E-hailing is the process of requesting a taxi or another form of transportation by using a computer or mobile device. E-hailing taxi platforms based on ride-sharing or professional capacity providers are becoming an important alternative to traditional taxis and the number of drivers on e-hailing platforms is increasing significantly, with e-hailing platforms becoming one of the fastest growing business trends. As of December 2014, Uber had 162,037 active drivers in United States who had completed at least four or more hours for service, and it has continued to grow since then. The number of such drivers in such markets as Los Angeles, San Francisco and New York tripled during 2014.[1]

Since the launch of Uber in 2009, e-hailing platforms have helped match supply with demand in a very convenient way for both sides and also have brought more business information to participants in the market. For example, drivers in traditional taxi services are limited in the amount of information about their competitors that they can observe. However, the location of drivers is available to any e-hailing platform user, including competing drivers. Customers often can see the location of drivers in their mobile application, and drivers can access this information by using an additional device with the customer-side application on (see Figure 2.1).[2] Therefore, the use of e-hailing platforms provides drivers with access to an unprecedented amount of information about their competitors. Some of these platforms provide even more information to their drivers, such as heat maps indicating locations with higher potential demand. These new forms of real-time information bring new opportunities for the agents to be

[1]For details, see *The Washington Post* "Now we know how many drivers Uber has and have a better idea of what they're making," January 22, 2015. https://www.washingtonpost.com/news/wonk/wp/2016/01/20/now-we-know-how-many-drivers-uber-has-and-have-a-better-idea-of-what-theyre-making

[2]While some companies do not offer accurate data of driver location, many of them do, including the company we collaborated with.

Figure 2.1. An Uber driver with multiple devices to check both customer and driver applications simultaneously

more *strategic* and can lead to changes in agent behavior. For example, agents may interpret the arrival of another idle agent into their service zone in two different ways.[3] On one hand, they may think the new agent is following a high sales opportunity in the zone and therefore agents may stick to their service zone. On the other hand, agents may see the new agent as a threat to their business and believe their sales potential decreases significantly because of this arrival and thus they may decide to move to another zone.

Agents are heterogeneous in how they react to information indicating new arrivals. Different interpretations of real-time information, as discussed above, can affect the performance of the agents differently as well. More specifically, heterogeneity in decisions may affect server utilization. Since monitoring and reacting to the information is costly for agents, it is interesting to study whether such behavior substantially increases sales.

Changes in individuals' behavior in response to real-time information can potentially affect the quality of service as well. Through better balancing of capacity, platforms potentially can serve more customers and/or respond to their requests in a short time. Previous research

[3]We call the area surrounding the location of an agent a "service zone." We describe how we split the city into service zones in Section 2.4.1.

has studied the consequences of customers' strategic behavior on the efficiency of service (see Lariviere and Van Mieghem 2004). Recent work has also considered capacity management problems where a service provider achieves a required service level by giving incentives to its profit-maximizing agents (see Gurvich et al. 2015). We complement this literature by analyzing the strategic behavior of servers in an increasingly important empirical setting.

To study these questions, we obtained data from one of the leading e-hailing apps in South America, with more than 100,000 drivers working in the platform in 2015.[4] Through this collaboration, we have been able to assemble a novel, high-frequency, spatial data set that contains very granular data about the movements of drivers affiliated with the e-hailing platform. Using this data set, we study how agents respond to the availability of rich real-time information about spatial competition. Our work makes the following contributions:

First, we show that agents tend to scatter more when presented with real-time information about their competitors. More specifically, the probability that agents will leave a service zone increases when a competitor enters their vicinity. We refer to this phenomenon as "server scattering."[5]

Second, we document that there is heterogeneity in how agents respond to the availability of real-time information. We show that agents who are more likely to react to this information achieve a higher average utilization than those who are less sensitive to the availability of real-time information. Finally, we complement our empirical results with agent-based simulations informed by the empirical parameters. We estimate and show that an increase in the frequency of scattering due to the arrival of another agent makes the whole system more efficient, reduces

[4] 99Taxis has over 100,000 drivers using the application in Brazil and 30,000 alone in Sao Paulo. For details, see http://techcrunch.com/2015/02/02/99taxis-raises-significant-new-cash-from-tiger-global/

[5] We conduct a placebo test that demonstrates in the absence of real-time information "server scattering" is not observed.

the likelihood customers will abandon, and decreases customers' waiting time to get service. Taken together, our results highlight the importance of real-time information for service efficiency in distributed settings. It has been documented that the on-demand business model is often associated with increased costs arising from the platform's inability to dictate when agents should work (Gurvich et al. 2015). We empirically show that sharing real-time information with agents can increase the efficiency of services, offsetting some of these costs.

The structure of the rest of the paper is as follows: In Section 2.2, we review the relevant literature. In Section 2.3, we develop our hypotheses. In Section 2.4, we describe our empirical setting and data set. Section 2.5 analyzes how agents respond to real-time information about competition. In Section 2.6, we discuss how the (heterogeneous) response to the real-time information about competition affects sales. In Section 2.7, we discuss the alternative explanations that might lead to our results. In Section 2.8, we analyze the efficiency of the system under such agent-based decisions. Finally, we provide some concluding remarks in Section 2.9.

2.2. Literature Review

The operations management community has given increasing attention, both theoretical and empirical, to information availability and its effect on systems. On the theory side, there is a wide range of papers focusing on models in settings ranging from retailer operations (e.g., Allon and Bassamboo 2011, Su and Zhang 2009) to service operations (e.g., Veeraraghavan and Debo 2009, Allon et al. 2011, Jouini et al. 2011). For example, Allon and Bassamboo (2011) investigate a game-theoretic framework for retailer operations where retailer can share

non-verifiable information and Jouini et al. (2011) consider a queuing model for call centers to analyze the effect of delay information on system performance.

On the empirical side, information availability has been investigated in several different contexts. For example, Gallino and Moreno (2014) analyze the effect of product availability information on store sales and Bell et al. (2016) consider the effects of product information in online retail by studying the introduction of offline showrooms.

In contrast to these research studies that consider the information available for *customers*, we focus on the information available to *agents* and how such information affects agent behavior and system performance. One of the few papers that empirically study the effect of information available to agents is Song et al. (2016), which examines the effect of information about agents' performance on productivity and service quality in a complex service organization. The authors focus on how the way performance-related information is shared (publicly vs privately) affects the whole system's productivity. They show that public information leads to higher productivity without a significant decrease in service quality. Similar to Song et al. (2016), our analysis also considers the information available to agents, but our setting allows us to study different questions. Our data set contains hypergranular spatial real-time information that must be processed quickly by agents when choosing which location to serve in a competitive environment. We can observe the actions that agents take in response to the information available to them.

Furthermore, competition plays a central role in our setting. Competition between agents takes place in many different contexts. For example, agents compete with their peers wherever agents are managed based on their performance, as noted in Netessine and Yakubovich (2012). Kalai et al. (1992) analyze a system where agents compete with each other by choosing appropriate service rates, and Anand et al. (2011) consider a more general model where the service

provider sets an admission price in order to maximize its own profit and agents can choose their own service speed and quality. To our knowledge, no empirical results have been described on the analysis of competitive environments for agents when they have real-time hypergranular information.

Our work is related to the emerging literature exploring the on-demand economy and the sharing economy, including recent work by Cramer and Krueger (2016), Kabra et al. (2016b), and Li et al. (2015). Cramer and Krueger (2016), for example, find that Uber drivers have higher utilization than taxi drivers. In our analysis, we show that the utilization of drivers from an e-hailing platform varies based on how sophisticated drivers are in using the information provided by the platform. Kabra et al. (2016b) investigate driver and passenger responses to the incentives given by the platform. They show that the increase in the number of trips completed is higher for each dollar spent on incentives given to drivers rather than incentives given to passengers.

Our work is also connected to previous research in economics that has explored the taxi industry (e.g., see the theoretical work of Cairns and Liston-Heyes 1996 for taxi regulation and empirical work of Camerer et al. 1997 for an analysis on the relation between working hours and wages in the economics literature). Recent research includes analyses by Zhang et al. (2016) on how taxi drivers learn from contextual and spatial information and Buchholz (2015) on the impact of search frictions on the efficiency of the taxi industry.

2.3. Hypothesis Development

In this section, we develop our hypotheses about agents' decisions given the availability of real-time information and how these decisions affect their sales performance. Since we test our hypotheses using a data set describing the behavior of drivers in an e-hailing platform, we use the terms "driver" and "agent" interchangeably, and we refer to the area surrounding the agent as the "service zone." In other applications, one can think more generally of service classes instead of service zones.

We first consider how the behavior of agents is affected by the availability of new agents in their service zone. A priori, it is unclear whether agents will tend to move to another service zone or will stay in their zone when new agents become available in that same zone. On one hand, agents may interpret new arrivals to their zone as an increase in the service zone's popularity. Agents may think the new agent arrives because the potential of a sale is (or will be) high in their zone and they trust their new peers' "experience of sales" and therefore decide to stay. For example, herding behavior happens in queues where each queue has different service qualities. Customers prefer longer queues since the length of the queue may signal the quality, as described in Veeraraghavan and Debo (2009). Similar herding behavior may occur in the case of competing agents.

On the other hand, agents may interpret newly arriving agents as a sign that competition is increasing and they may think their service zone has more agents than needed. This scenario is similar to spatial competition with entry. For example, Palfrey (1984) considers entry of a third political party in an ongoing competition of two other parties where each voter prefers the party that is close to herself. He shows that the entrance of a third party results in spatial

separation of the other two parties. Similarly, agents may think their chance of making a sale in their zone decreases significantly after the arrival of another competitor. Such an interpretation may lead them to change their service zone so that they stay away from this increased level of competition. Thus, we test the following pair of competing hypotheses:

Hypothesis 1A (Herding). If a new competitor enters the service zone, the probability that current agents will leave the service zone decreases.

Hypothesis 1B (Scattering). If a new competitor becomes available in a service zone, the probability that current agents will leave the service zone increases.

We illustrate the theoretical ambiguity with a model in Appendix B.1.1. We test Hypothesis 1A and Hypothesis 1B in Section 5.

With the pair of competing hypotheses, 1A and 1B, we test whether agents are interpreting the arrival of another agent as a signal of a high sale opportunity in their service zone or an increased level of competition. However, it is highly possible that agents are heterogeneous in terms of their probability of changing zones following the arrival of new competitors. Some agents may be more likely than others to scatter. The question, then, is how this heterogeneity in scattering affects their performance — namely, whether it allows drivers to achieve a higher utilization rate. Responding to new entrants in a service zone by increasing the probability of leaving could hurt or help sales. The arrival of a competitor could really be a sign of a high-sales opportunity. If that is the case, scattering may result in a low utilization rate. Moreover, agents will spend time in moving to other service zones, which may further decrease the utilization of the agents. On the opposite side, the arrival of a competitor may increase the level of competition significantly. Therefore, scattering would be beneficial for the agent if she moves to

another service zone where she has higher chance of making sales. Thus, we test the following pair of competing hypotheses:

Hypothesis 2A. When agents are more likely to change their service zone following the entry of a new competitor, they have a higher utilization rate.

Hypothesis 2B. When agents are more likely to change their service zone following the entry of a new competitor, they have a lower utilization rate.

We illustrate the ambiguity in Hypothesis 2 in Appendix B.1.2. We test this hypothesis and compare the magnitude of the effects of scattering behavior in the performance of the individual driver and the whole system in Section 6. Our empirical analysis resolves this theoretical ambiguity. In the next section, we provide the details and descriptive statistics of the data set that we have gathered to test these hypotheses.

2.4. Empirical Setting and Data

2.4.1. Empirical Setting

In this study, we have collaborated with 99Taxis, one of the leading e-hailing platforms in South America. 99Taxis was founded in 2012 and operates in over 399 cities in Brazil. At the time of our analysis (September 2014), the company had over 19,000 taxi drivers in Sao Paulo. Similar to many other e-hailing taxi platforms, customers could see all available taxi drivers around their location via the mobile phone application, and they could use the application to make a ride request.

We obtained access to anonymized, high-frequency, hyperlocal information from the GPS logs of all the drivers in the network working in Sao Paulo during our period of analysis. Our

data set records the status and location of each driver (latitude and longitude) at a given time whenever they are logged in. Figure 2.2 shows a sample path followed by a driver during a 10-minute period.

Such data is also available to customers and drivers in real time through the e-hailing application and can potentially influence the behavior of drivers. One of the company's investors noted that over time, drivers realized how their revenue could increase if they used smartphone with 3G services instead of an old phone with a prepaid services plan (see TechCrunch 2015). The reason is that a smartphone allowed them to access richer real-time spatial information, such as the location of competing drivers on a map. Therefore, this data set is very appropriate to analyze the central question addressed in this paper: how the information on the entry of a new competitor to a zone affects the behavior of the other drivers in that zone. Moreover, this data set is a great source to investigate differences in the degree of driver sophistication in terms of how drivers react to available information and how these differences translate into sales and system efficiency.

While the GPS logs provide very rich information, they require substantial preprocessing. As Figure 2.2 indicates, the different observations for a driver occur at different interval lengths. Moreover, observations are not recorded at the same time for different drivers—i.e., we may have a record at 12:02:17 on Monday for one driver and a record at 12:02:18 on Monday for another driver. Therefore, we transform our data to track the drivers' movements in hour and minute levels. The following tables illustrate how we transform the data. The first table is the original data:

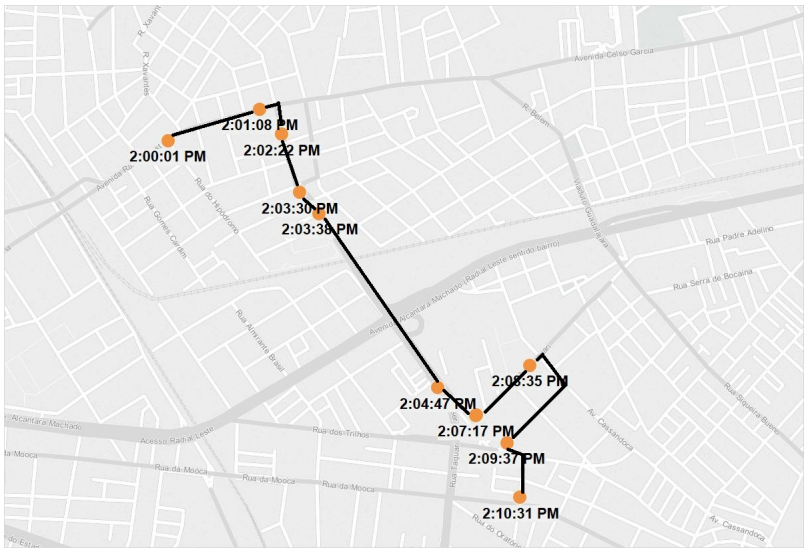


Figure 2.2. A data record from a driver on September 11 between 2:00 p.m. and 2:10 p.m

Time	Longitude	Latitude	Vacant Status
12:01:17	g_{11}	g_{12}	1
12:02:34	g_{21}	g_{22}	0
12:03:10	g_{31}	g_{32}	0
12:03:30	g_{41}	g_{42}	1

Then we transform this data into the following information set:

Time	Longitude	Latitude	Vacant Status
12:01	g_{11}	g_{12}	1
12:02	$\frac{34\times g_{11}+(60-17)\times g_{21}}{34+(60-17)}$	$\frac{34\times g_{12}+(60-17)\times g_{22}}{34+(60-17)}$	$round\left(\frac{34\times 1+(60-17)\times 0}{34+(60-17)}\right)$
12:03	$\frac{10\times g_{21}+(60-34)\times g_{31}}{10+(60-34)}$	$\frac{10\times g_{22}+(60-34)\times g_{32}}{10+(60-34)}$	$round\left(\frac{10\times 0+(60-34)\times 0}{10+(60-34)}\right)$
12:04	g_{41}	g_{42}	1

Note that *Vacant Status* denotes the availability of the driver according to the platform. However, a driver may have picked up a passenger on the street without using the platform and may

be still logged in or may have used a different e-hailing service. In such a case, other drivers who use the application will see this driver as vacant; hence, they may think that this driver is still a competitor for them.[6]

Our data covers seven consecutive days in the month of September 2014. In order to understand spatial effects, we divide the space into a grid of square service zones. Our base analysis uses 400 square zones of $500 \times 500\ m^2$. We also conduct robustness checks by considering different square sizes (e.g., $250 \times 250\ m^2$), as we explain in detail in Section 2.7. We use subscript *z* to denote the zone of a given driver. Our data set allows us to track the evolution of drivers and zones over time. In our empirical analysis, we include hourly weather data and other drivers of zone-level user interest, such as locations of public transit stops, and over 18,000 points of interest in the city center of Sao Paulo, such as hotels, restaurants, cinema and theaters, schools, and hospitals, to check whether our results are robust to prevailing demand levels. Moreover, we construct a detailed set of dependent, independent, and control variables that capture this evolution, which we describe next.

There are two types of variables that capture information relevant to a service zone: variables that capture predictable variability in sales based on the historical evolution, and variables that capture the real-time status and transitions in a zone. We also construct driver-level variables that encode some important aspects of driver behavior and outcomes.

2.4.1.1. Predictable Variability in Sales. We want to measure the overall level of sales in a day d, hour h, and zone z. First, note that we know the status of taxi drivers whenever they

[6]Therefore, in addition to considering all drivers logged into the platform, we also make a robustness analysis where we analyze the drivers with the utilization higher than some specific thresholds, which are explained in detail in Section 2.7. Note that Easy Taxi was the biggest competitor of 99Taxis; Uber had not entered Brazil's e-hailing market at the time.

are logged in. Therefore, any change from vacant to busy means that a sale has been initiated through the platform.

We first calculate the number of sales occurring in each zone z, day d and hour h. We denote this value by $s_{d,h,z}$. We use the following model to estimate time and location fixed effects:

$$s_{d,h,z} = SalesTime_{d,h} + SalesZone_z + \epsilon_{d,h,z} \tag{2.1}$$

where $SalesTime_{d,h}$ is time fixed effect, $SalesZone_z$ is location fixed effect and $\epsilon_{d,k,z}$ is the error term.[7] Note that $SalesTime_{d,h}$ can be interpreted as predictable temporal (hourly and daily) variability in sales and $SalesZone_z$ can be interpreted as predictable spatial variability in sales based on the historical evolution.

2.4.1.2. Real-time Information. In this section, we define variables identifying the transition from a given minute *t* to the next one, *t+1*, and some variables capturing the status at a given minute *t* (in a given day and hour).

We have several independent and control variables that are constructed through observing changes from minute t to $t+1$. First, we define *NewDrivers*$_{z,t}$ as the number of new drivers entering zone z from other zones between times t and $t+1$. Second, we denote *Vacant*$_{z,t}$ as the number of vacant taxis according to the platform in zone z, minute t. This is not a transition variable but individual drivers may observe five possible changes in the status of those vacant taxis at $t+1$:

(1) Some of them may be still vacant in the same zone. Number of such drivers is denoted by *VacantIn*$_{z,t}$.

[7] Mean value of estimated *SalesTime* (*SalesZone* respectively) is 3.172 (3.222 respectively), and standard deviation is 3.672 (6.673 respectively).

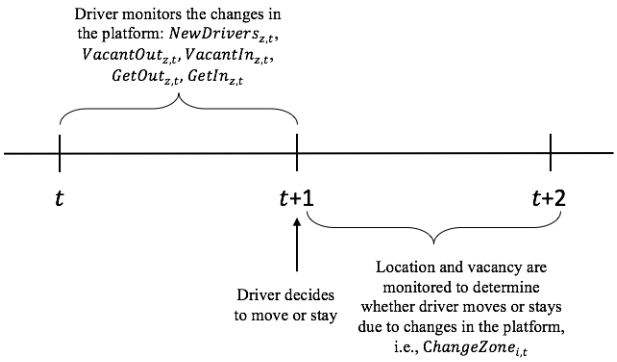


Figure 2.3. Sequence of events for a given driver

(2) Some of them may be still vacant in a different zone— i.e., they may have changed their zone and be still vacant. Denote the number of such drivers by $VacantOut_{z,t}$.

(3) Some of them may find a passenger during this one-minute observation period in the same zone. Number of such drivers is denoted by $GetIn_{z,t}$.

(4) Some of them may first change their zone and find a passenger in the new zone by minute $t+1$. Denote the number of such drivers by $GetOut_{z,t}$.

(5) The remaining drivers may log out from the system, so they suddenly disappear. We summarize the sequence of events in Figure 2.3.

2.4.1.3. Driver Variables. In this subsection, we define some driver-level variables. We start with a dependent variable. We define the binary variable $ChangeZone_{i,t}$, where 1 denotes that individual driver i changes her service zone and 0 denotes that she stays in her zone at minute t.[8] We define $PercentageBusy_{i,d,h}$ as ratio of minutes that a driver i is busy— i.e., the minutes

[8] We assume that drivers monitor changes in the platform from minute t to $t+1$ and they either move to another zone or stay in the same zone. To identify whether they move or stay, we compare the zone of the driver at minute $t+1$ and $t+2$ (where the one-minute difference is referred to as implementation time, and for robustness purposes we also compare the zone of the driver at $t+1$ and $t+3$— i.e., implementation time is set to two minutes). If

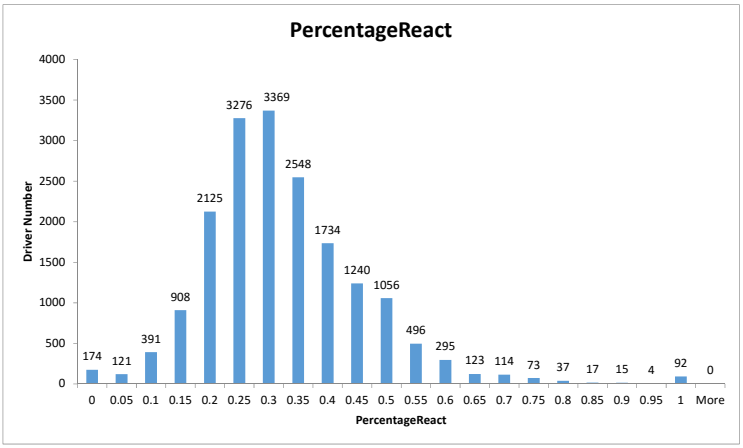


Figure 2.4. Heterogeneity in PercentageReact

that a driver is carrying a passenger hailed through the platform to the total number of minutes that she is logged into the system in a given hour h and day d. Note that a driver may carry a passenger from other platforms, so we carry out robustness checks where we consider this concern. We calculate *PercentageReact*$_i$, which is defined as the ratio of times a driver changes zones to the total number of times the driver is presented with the choice of leave or stay at the service zone following a competitor's entry; i.e., this ratio presents the scattering frequency of the driver.[9] As demonstrated in Figure 2.4, drivers are heterogeneous in how they respond to the entry of new competitors. In Section 2.6, we use quartiles to categorize each driver.

---

the zones are different, we assume the driver changes her zone (see Figure 2.3 for details). We are interested in the behavior of vacant drivers since busy drivers do not need to decide to move or stay in order to find a passenger. Therefore, we only consider the drivers who are vacant at t, $t+1$, and $t+2$.

[9]Note that we can calculate this variable in i, d, h (driver, day, hour) levels and we have a robustness analysis with that level as well.

Table 3.3 provides summary definitions for all variables included in our models. We test our hypotheses in two different sections. First, we analyze the effect of new drivers on scattering/herding decisions (namely H1) in Section 5. Then, we analyze the effect of scattering on the performance of the individual driver (H2) in Section 6. Before these analyses, we provide descriptive statistics of variables that will be used for these analyses in next sections.

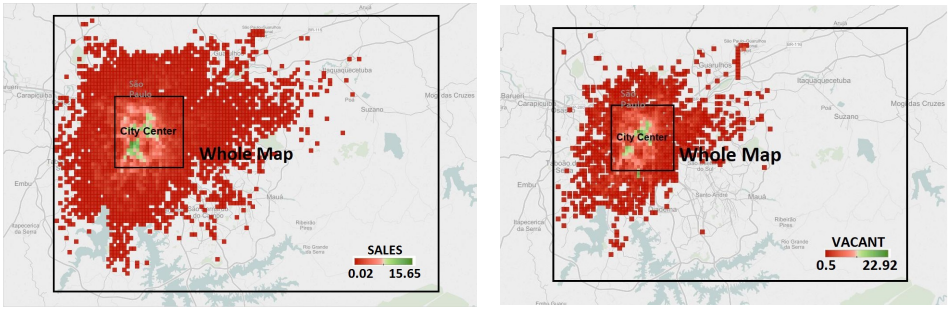
Table 2.1. Definition of Variables

Variable	Description
	Zone-level variables
NewDrivers$_{z,t}$	Number of new drivers entering zone z at time t.
Vacant$_{z,t}$	Number of vacant drivers in zone z and time t.
GetIn$_{z,t}$	Number of vacant drivers in zone z and time t who stay in the same zone after 1 minute and hail a passenger.
GetOut$_{z,t}$	Number of vacant drivers in zone z and time t who move to another zone after 1 minute and hail a passenger.
VacantIn$_{z,t}$	Number of vacant drivers in zone z and time t who are still vacant after 1 minute in the same zone.
VacantOut$_{z,t}$	Number of vacant drivers in zone z and time t who are still vacant after 1 minute in a different zone.
SalesZone$_z$	Spatial predictable variability in sales from zone z.
	Driver-level variables
ChangeZone$_{i,t}$	Indicator that shows whether driver i changes her zone time t.
PercentageBusy$_{i,d,h}$	Proportion of time that driver i is busy during hour h and day d.
PercentageReact$_i$	Fraction of time that driver i changes her zone when there is an entry to her zone.
PercentageReact_QK_i	Indicator that shows whether PercentageReact$_i$ is greater than (K-1)st quartile, but less than Kth quartile of all PercentageReact values.
	Other variables
RushHour	Indicator for rush hour defined for 6am-9am and 4pm-7pm.
Weekend	Equal to 1 if day of observation is Saturday or Sunday.
SalesTime$_{d,h}$	Temporal predictable variability in sales.

2.4.2. Descriptive Statistics

Note that we focus our analyses on the city center of Sao Paulo, which is a square-shaped region around 10 kilometers by 10 kilometers because we have a significant amount of information for each zone of the city center. On the other hand, there are many zones outside the city center where either sales occur very rarely or few vacant drivers go (see Figure 2.5).[10] For example, average sales in an hour is 6.394 per zone in the city center, but this value is only

[10]Similarly, Buchholz (2015) focuses on one borough, i.e., Manhattan, and two airports, i.e., JFK and LaGuardia, in the analysis of New York taxi data since over 90 percent of rides in New York taxi data originate from these three areas.



(a) **(b)**

Figure 2.5. Color in each zone represents the hourly average sales of that zone and its neighbors in Figure (a) and average vacant driver number per minute in Figure (b). The data for (a) is filtered so that only the zones with positive sales are colored. Similarly, the data for (b) is filtered so that only the zones with average vacant drivers more than 0.5 are displayed

0.167 per zone outside of the city center. Similarly, there are 4.299 vacant drivers in a minute per zone in the city center, but we have only 0.183 vacant drivers outside of the city center.

Table 2.2. Driver-level and zone-level variables

Driver-level variables					
Variable	Mean	SD	Min	Max	N
$ChangeZone_{i,t}$	.257	.437	0	1	11,634,626
$RushHour_{i,t}$ *(H1)*	.256	.436	0	1	11,634,626
$Weekend_{i,t}$ *(H1)*	.174	.379	0	1	11,634,626
$PercentageReact_{i}$	.301	.138	0	1	18,246
$PercentageBusy_{i,d,h}$	.263	.337	0	1	712,228
$RushHour_{i,d,h}$ *(H2)*	.215	.411	0	1	712,228
$Weekend_{i,d,h}$ *(H2)*	.204	.403	0	1	712,228
Zone-level variables					
Variable	Mean	SD	Min	Max	Observations
$NewDrivers_{z,t}$	1.593	2.228	0	38	4,442,193
$Vacant_{z,t}$	4.299	5.835	0	68	4,442,193
$GetIn_{z,t}$	.0623	.307	0	13	4,442,193
$GetOut_{z,t}$	.0358	.206	0	6	4,442,193
$VacantIn_{z,t}$	2.706	4.168	0	53	4,442,193
$VacantOut_{z,t}$	1.367	2.103	0	2	4,442,193

In Table 2.2, we provide summary statistics, including mean, standard deviation, and minimum and maximum value of each driver-level variable. We observe $18,246$ drivers make $11,634,626$ scattering or herding decisions in total throughout the observation period. The first three of the variables listed in Table 2.2 are related to these decisions. For example, 25.7 percent of these decisions are scattering. The timing of these decisions has the following statistics: 25.6 percent of these decisions are given in rush hour and 17.4 percent of them are given during the weekend.

On average, a driver chooses to scatter 30.1 percent of the time among all the scattering/herding decisions she makes. Note that these $18,246$ drivers stayed logged into the platform during $712,228$ driver hours in total, and rows 5, 6, and 7 of Table 2.2 are related to these hourly driver-level observations. For example, on average, drivers are busy 26.3 percent of the time that they are logged in, in a given hour. Of these $712,228$ observations, 21.5 percent occurred during rush hour and 20.4 percent of them during the weekend.

In Table 2.2 we report statistical summaries for zone and driver-level variables that are mostly used for the first hypothesis. We have $4,442,193$ combinations of zone (in city center) and time (minutes). On average, there are 1.59 drivers newly entering a zone in a minute. We observe on average 4.299 vacant drivers per zone in a given minute; 2.70 of these drivers stay vacant in the same zone and 1.367 of them stay vacant in another zone in the following minute.

In the next two sections, we analyze our hypotheses. The next section provides the analysis of the effect of real-time information on location choice of agents (H1). In Section 2.6, we analyze how heterogeneity in decisions affects the performance of an individual driver (H2).

2.5. The Effect of Real-Time Information on Agent Scattering

2.5.1. Econometric Model

In this section, we explain the regression models we use to analyze the effect of newly entering agents into a service zone on the decisions of the current agents, as discussed in Hypothesis 1. Our unit of observation is at taxi driver and time level. We use logit and probit specifications to model how the decision to stay in the same zone or move to another zone depends on the entry of new competitors. To test Hypothesis 1, an ideal experiment would randomly show different competition situations to drivers and track their reaction to those situations. Such an ideal experiment is not feasible in a production environment, so we have to restrict our analysis to the observational data generated during the live operation of the platform. In our study, we follow vacant drivers and study how they react to the different competition situations that naturally occur during their work shifts. Note that this could create some endogeneity concerns because the different competition situations drivers face could be correlated with unobserved factors that drive their decisions. For example, an abundance of vacant drivers could be due to an unobserved demand change that attracts more vacant drivers to a zone, which could also affect whether drivers choose to stay in or leave the zone. We discuss how this could potentially bias our results, and we supplement our analysis with the analysis of a subset of our data where such concerns are minimized.

We use logit regression with some fixed effects. We conduct robustness checks with logit regression with random effects and probit regression with fixed/random effects with robust standard errors, and we obtain consistent findings. Step by step we add controls and check our hypothesis. Let i denote the driver and t denote the time of the observation. We use control

variables including individual fixed effects of the driver (denoted by μ), time fixed effects by *hour* (denoted by ζ_h), and location fixed effects where we use the row and column of the driver's zone. We denote such column and row effects by γ_c and η_r. The family of specifications we use is the following:

$$p_{it} := Pr(Y_{it} = 1|NewDriver, C) = F(\alpha_0 + \alpha_1 NewDriver + C'_{it}\beta + \gamma_c + \eta_r + \zeta_h + \zeta_d + \mu_i)$$

where C denotes the matrix for control variables, Y_{it} is the decision of the i^{th} driver at time t, c is the column, r is the row of the driver's zone, and h denotes the *hour* and d the *day* that t belongs to. We consider predictable variability in sales based on the hour and day of the observation, *SalesTime*, as well as the location, *SalesZone*, separately under matrix C. Note that we have $F(z) = e^z/(1 + e^z)$ for the logit model and $F(z) = \Phi(z)$ for the probit model, where Φ is standard normal cumulative distribution function.

2.5.2. Results

In Table 2.3, we provide our first result for Hypothesis 1 based on the fixed effects logit model. Note that we use both individual-driver fixed effects and location fixed effects in all the models described in this table. In addition to these fixed effects, Model (5) also considers hourly and daily time fixed effects.

We observe that the coefficient of *NewDrivers* is significant and positive for all of these models. Therefore, a new competitor's entrance increases the probability a driver will change the zone. Hence, Hypothesis 1B (scattering) is supported by these models. Note that driver, location, and time fixed effects are included in Model (5) and we have the highest log-likelihood

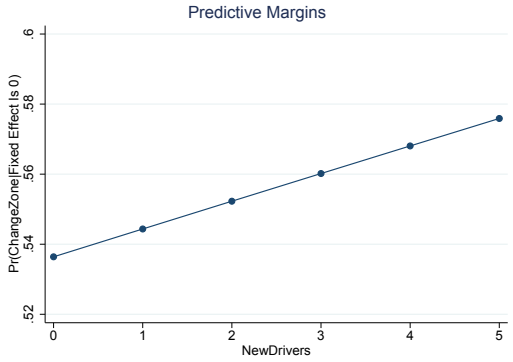


Figure 2.6. Marginal effect of NewDrivers on probability of changing zone; figure uses Model (5) from Table 2.3

among all the models listed in Table 2.3.[11] We observe that the predicted probability of changing the zone increases almost linearly as the number of *NewDrivers* increases (see Figure 2.6).[12] Each additional driver increases the probability of changing the zone almost 1 percent.

For robustness of the results reported above, we consider different sizes for each square zone. Instead of $500m \times 500m$ dimensions for each square zone, we construct zones by $250m \times 250m$ on the same city center map. We also change the implementation time from 1 minute to 2 minutes.

Specifically, we consider three additional setups: (1) same zone setup, but longer implementation time; (2) smaller zones with the same implementation time; and (3) smaller zones with longer implementation time. We observe that *NewDrivers* has a significant and positive effect on the probability of changing the zone for these setups as well, confirming our support for Hypothesis 1B. The results of these alternative specifications and additional robustness checks are described in detail in Section 2.7.1. We also perform an analysis with a binary version of

[11]We provide the results of other regression models, including random effects logit and probit, in our online appendix.

[12]For this calculation, we use Model (5) from Table 2.3 and calculate the marginal effect of *NewDrivers* on *ChangeZone*.

Table 2.3. Effect of Entry of New Drivers on Decision to Change Zone

	(1)	(2)	(3)	(4)	(5)
NewDrivers	0.0242***	0.0244***	0.0291***	0.0338***	0.0323***
	(29.75)	(29.95)	(35.35)	(35.92)	(34.09)
SalesTime	0.0341***	0.0330***	0.0380***	0.0362***	0.0346***
	(50.07)	(46.45)	(52.40)	(49.41)	(28.37)
SalesZone	-0.0291***	-0.0292***	-0.0303***	-0.0195***	-0.0194***
	(-57.03)	(-57.14)	(-59.35)	(-34.57)	(-33.47)
RushHour		-0.0293***	-0.0153**	-0.00825	
		(-5.33)	(-2.79)	(-1.49)	
Weekend			0.238***	0.165***	
			(36.61)	(24.68)	
GetIn				-0.0251***	-0.0248***
				(-5.18)	(-5.11)
GetOut				0.0200**	0.0196**
				(2.65)	(2.61)
VacantIn				-0.0294***	-0.0287***
				(-58.93)	(-54.87)
VacantOut				0.0356***	0.0344***
				(36.55)	(35.02)
Observations	1,158,965	1,158,965	1,158,965	1,158,965	1,158,965
AIC	1180087.0	1180060.5	1178738.7	1174663.0	1173607.4
BIC	1180601.4	1180586.9	1179277.0	1175249.2	1174516.6
Log-Likelihood	-590000.5	-589986.3	-589324.3	-587282.5	-586727.7
χ^2	10672.6	10701.0	12024.9	16108.6	17218.2

t statistics in parentheses
All models estimate a logit regression with row, column, and driver fixed effects.
A random 10% sample is taken to estimate the models.
Dependent variable is *ChangeZone*.
Model (5) considers hourly and daily fixed effects.
$^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

NewDrivers, where $\textit{BinaryNewDriver}_{z,t}$ is 1 if $\textit{NewDrivers}_{z,t} > 0$, and 0 otherwise. We observe positive and significant estimates for this variable too, which further supports our results.

As noted in Manski (1993), an individual's decision can be affected by her peers' decisions, which means that there can be a social interaction between each individual. The models we consider above may suppress such peer effects, which may result in an identification problem (see Manski 1993 for details). We test our first hypothesis with a different model setup and a

method where we also consider peer effects. For this purpose, we use the spatial autoregressive (SAR) model to test Hypothesis 1. Our results for the SAR model also show that agents scatter as the number of new drivers increases. The details of the model and the results are described in Appendix B.2.2.

2.6. Heterogeneity in Agent Response and Utilization

2.6.1. Econometric Models

In this section, we start by describing the model we use to analyze how heterogeneity in agent behavior affects agent utilization. In the context of e-hailing taxi platforms, agent utilization is basically the ratio of busy minutes divided by the total number of minutes that the agent is logged into the platform. We categorize each driver by using the quartile of *PercentageReact* value (fraction of times that a driver changes the zone following a competitor's entry) of all drivers. For example, *PercentageReact*$\_Q2_i = 1$ if *PercentageReact*$_i$ value of driver i is greater than the first quartile and less than the second quartile. Similarly, we define *PercentageReact*$\_Q1_i$, *PercentageReact*$\_Q3_i$, and *PercentageReact*$\_Q4_i$. To investigate the effect of scattering on *PercentageBusy*$_{i,d,h}$, namely H2, we use the following regression model

$$PercentageBusy_{i,d,h} = \alpha_0 + \alpha_1 PercentageReact\_Q2_i + \alpha_2 PercentageReact\_Q3_i + \alpha_3 PercentageReact\_Q4_i + C'_{i,d,h}\beta + \epsilon_{i,d,h}$$

where i represents the driver, h represents the hour, control variables are denoted by matrix C, and $\epsilon_{i,d,h}$ is the error term. Note that control variables include *Weekend*, *RushHour*, and

SalesTime. Since our dependent variable, namely *PercentageBusy*, is a fraction, we also use a fractional logit/probit model, which is commonly used in econometric analysis of fractional dependent variables (see Papke and Wooldridge 1996). The model has the following form:

$$E[y|x] = G(\beta x)$$

where $0 < G(z) < 1$ for any $z \in R$. The most commonly used functions are $G(z) = exp(z)/[1 + exp(z)]$ for the logit model and $G(z) = \Phi(z)$ for the probit model, where Φ is the standard normal cumulative distribution function. Similar to the binary logit model discussed in Section 2.5.1, the marginal effect of x_j on $E[y|x]$ is $\partial E[y|x]/\partial x_j = \beta_j g(z)$, where $g(z) = exp(z)/[1 + exp(z)]^2$. Therefore, sign of β_j is sufficient to declare the direction of the marginal effect.

2.6.2. Results

Table 2.4 reports the effect of heterogeneity in scattering on agent utilization. We observe that coefficients for *PercentageReact* quartiles are positive, significant, and increasing in the order of the quartile. Therefore, we can conclude that drivers with more frequent scattering behavior as a response to a newly entering competitors have a higher utilization rate than the drivers with less frequent scattering behavior.

We also observe that the difference between *PercentageReact_Qk* and *PercentageReact_Q(k-1)* is highest for the fourth quartile. Hence, increasing the scattering behavior to the highest scattering category brings the highest additional utilization.

Note that *PercentageReact* is a driver-level variable. For robustness of the results reported above, we add the time dimension to this variable. We define *PercentageReactHourly*$_{i,h}$ as the

Table 2.4. Effect of Strategic Scattering on Agent Utilization

	(OLS-1)	(OLS-2)	(Frac Logit-1)	(Frac Logit-2)
PercentageReact_Q2	0.00501***	0.00238*	0.0283***	0.0150**
	(4.87)	(2.33)	(5.01)	(2.66)
PercentageReact_Q3	0.0125***	0.00864***	0.0692***	0.0508***
	(11.73)	(8.22)	(12.05)	(8.83)
PercentageReact_Q4	0.0350***	0.0299***	0.183***	0.160***
	(27.42)	(23.62)	(27.72)	(24.13)
Vacant	-0.00408***	-0.00483***	-0.0237***	-0.0278***
	(-64.13)	(-75.90)	(-63.44)	(-73.35)
Weekend	0.0585***	0.0775***	0.305***	0.409***
	(33.70)	(45.71)	(33.77)	(45.44)
RushHour	-0.0404***	-0.0792***	-0.243***	-0.415***
	(-16.22)	(-31.65)	(-17.20)	(-29.01)
SalesTime		0.0258***		0.125***
		(123.65)		(116.63)
Constant	0.250***	0.235***	-1.093***	-1.177***
	(106.34)	(100.95)	(-86.83)	(-92.37)
Observations	713605	713605	713605	713605
R^2	0.061	0.081		
AIC	430222.5	414859.4	686458.2	678580.0
BIC	430612.8	415261.1	686848.5	678981.7
Log-Likelihood	-215077.3	-207394.7	-343195.1	-339255.0
χ^2			42461.1	57231.7

t statistics in parentheses

Dependent variable is PercentageBusy, which is a fraction.

All of the models above use daily and hourly fixed effects.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

ratio of times that a driver changes zone to the total number of times that the driver is presented with the choice of leave or stay in the service zone in hour h. This variable helps us to observe the effect of heterogeneity in the scattering behavior of the individual driver as well. We observe that *PercentageReactHourly* is positive and significant as well (see Table B.3 of the Appendix). Taken together, these results provide strong support for Hypothesis 2A: drivers who are more likely to respond to the entry of a new driver by changing zone achieve higher utilization rates.

2.7. Robustness Checks and Alternative Explanations

In the following section, we conduct a battery of robustness checks. These include considering smaller zone configurations; longer implementation times; a placebo check with a taxi data set pertaining to a time period when no real-time competition information was available to drivers; and taking into account the effects of demand shifters, street hailing, and driver experience. Overall, the results strengthen our claim that drivers use real-time information about their competitors and scatter as a response to the entry of new drivers.

2.7.1. Zone Configuration and Implementation Time Specification

In our analysis we focus on a 10x10 km^2 area at the city center of Sao Paulo. We define our zones as 500x500 m^2 squares. To show that our results are robust to the specification used to outline the zones, we repeat our analysis with increased granularity by focusing on zones composed of 250x250 m^2 squares. Considering smaller zones might impact our results, as the number of entries into and exits from a specific zone depend on the zone configuration adopted. First, under this alternative specification the average number of drivers per zone decreases. Second, the exit ratios (the percentage of drivers who change their zones) as well as the total number of drivers who change the zone will increase, since a driver whom we considered to be staying in her service zone under a bigger zone configuration might be considered as changing her zone under the smaller zone setting. Hence, we will observe the same number of drivers in total but more zone-changing decisions.

We find that the new driver entry effect on zone-changing behavior remains significant and increases in magnitude under this smaller zone configuration.[13] Thus, our main result–the finding that competition leads to scattering–is robust to zone configuration.

Next, we analyze the sensitivity of the new driver entry effect to implementation time. Under this implementation time setting, a driver is considered as having changed her zone if she moved to another zone within two minutes of observing a new driver enter her service zone. Consequently, the number of drivers who change their zones should either stay constant or increase, whereas the number of new drivers entering into the zone is not changed. Therefore, we observe an increase in the magnitude of new driver effect on zone-changing behavior under every setting.[14]

2.7.2. Placebo Test

In the preceding sections, we showed that sharing real-time information about their competitors with service agents alters agents' behavior. Agents interpret the arrival of new agents into their service zone as an increase in competition and tend to change their service zones as a response to this increased competition. This result implies that drivers monitor the competition level in their zones by leveraging visual information contained in the e-hailing application. However, one could argue that such scattering behavior could be unrelated to the availability of real-time information and that we would have observed such scattering behavior in face of increased competition even in the absence of real-time information. To understand the impact

[13]Under the smaller zone configuration, the coefficient for *NewDrivers* in the five different models that we use for Table 2.3 are 0.0723, 0.0732, 0.0733, 0.0757, and 0.0735 respectively. All these estimations are statistically significant.

[14]We observe the textitNewDrivers effect as 0.0339***, 0.0339***, 0.0385***, 0.0387***, 0.0366*** by using the same models noted in Table 2.3.

of new driver entries on agent behavior in the absence of real-time information, we conduct a placebo test on a different data set in a setting where real-time information was not available.

We use a data set from an Asian city that covers a time span prior the introduction of e-hailing applications, in May 2009.[15] E-hailing applications were introduced in early 2012 in Asia. Thus, during our period of observation (May 2009) drivers do not have access to real-time information provided by e-hailing applications. The analysis of this data set enables us to test whether the scattering decision of drivers whom we observed in the Sao Paulo data set could be caused by factors other than the real-time competition information provided by the e-hailing application.

This data set contains 2.02 millions individual trip records from 3,418 taxis during 29 days pertaining to May 2009. Each taxi GPS tracking record contains information on the taxi ID, time stamp the observation was recorded, geographic coordinates, and vacancy status. Following the procedure described in Section 2.4.1, we create 2,000 square zones of 500x500 m^2.

We follow the econometric model described in detail in Section 2.5.1 to test Hypothesis 1. Our observation unit is at taxi driver and time level, and we use a logit specification to analyze how staying in or leaving the zone depends on the entry of new drivers when none of the drivers can access real-time information. Similar to the analysis in Section 2.5.1, we include individual, time and location (row/column) fixed effects. Moreover, we analyze the predictable variability in sales based on hour, day of the week, *SalesTime* as well as the location *SalesZone*.

As observed in Table 2.5, the coefficient associated with *NewDrivers* is significant and *negative* for all these models. Hence, in the absence of real-time information sharing, there is no evidence that Hypothesis 1B (scattering) holds. The results suggest that drivers tend to move

[15] An analogous data set with information on Sao Paulo drivers' behavior before the introduction of e-hailing services was not readily available.

in a positively correlated way if no real-time information is available to them. Thus, the claim that the scattering behavior is linked to the availability of real-time information provided by the e-hailing application is consistent with the results obtained in this data set.

2.7.3. Effect of Traffic Density and Traffic Regulations

In the previous sections, we showed that the entry of new drivers into a zone induces vacant drivers to change their zones. These results indicate that competition leads to a scattering behavior. Thus, Hypothesis 1B is empirically supported. However, it could be argued that traffic conditions, not the entry of the new drivers, leads to scattering behavior.

High traffic density might impact our results in two opposite directions. First, heavy traffic might prevent drivers who are willing to change their zones from doing so. Therefore, even if drivers want to change their zones (scatter) due to the entry of new drivers, they would be unable to do so due to heavy traffic. If heavy traffic conditions prevent drivers from changing their zones, our results provide a lower bound.

Second, drivers might prefer to stay in their zones but traffic conditions might push them out. If traffic conditions are pushing drivers out of their zones, there will be more entry into and exit from zones. In this case, our results would provide an upper bound, as we would be attributing traffic-induced zone-changing behavior to the new driver effect.

We devise two methods to eliminate the concerns regarding the impact of traffic conditions on the analysis of Hypothesis 1. First, we analyze two different data samples with light and heavy traffic respectively. Second, we calculate the speed of the drivers and measure the new driver entry effect on zone-changing behavior in data samples with high and low speeds. Our analysis shows that drivers react more strongly to the entry of new drivers during periods with

Table 2.5. Effect of Entry of New Drivers on Decision to Change Zone in the Absence of Real-Time Information

	(1)	(2)	(3)	(4)
NewDrivers	-0.0419***	-0.0428***	-0.0138***	-0.0137**
	(-27.88)	(-28.36)	(-66.64)	(-66.24)
SalesTime	0.0102***	0.00569***	0.000512***	0.000761***
	(10.58)	(4.91)	(3.80)	(4.75)
SalesZone	-0.000504***	-0.000476**	-0.000577***	-0.000579***
	(-3.43)	(-3.24)	(-28.44)	(-28.52)
Weekend		-0.0464***		0.00255**
		(-6.97)		(2.88)
GetIn			-0.133***	-0.133***
			(-111.93)	(-111.91)
GetOut			.420***	0.420***
			(305.48)	(305.50)
VacantOut			0.427***	0.427***
			(853.29)	(853.03)
VacantIn			-0.0726***	-0.0725***
			(-344.82)	(-344.57)
Constant	-0.820***	-0.787***	0.238***	0.237***
	(-5.90)	(-5.66)	(12.55)	(12.44)
Observations	921591	921591	921591	921591
R^2			0.563	0.563
AIC	1072007.8	1071961.2	421486.3	421480.0
BIC	1113440.1	1113405.2	462965.5	462970.9
Log Likelihood	-532472.9	-532448.6	-207208.1	-207204.0
χ^2	61052.1	61100.8		

t statistics in parentheses

All the regressions include row, column, hour, and driver fixed effects.

A random 10% sample is taken to estimate the models.

Logit model is used in regressions (1) and (2).

Due to computational challenges, a linear probability model was used in regressions (3) and (4).

$^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

light traffic. Moreover, our results are consistent with Hypothesis 1B, with scattering being observed regardless of prevailing traffic conditions and data samples considered.

In order to observe whether heavy traffic prevents drivers from changing their zones or pushes them out of their zone, we compare time periods with light and heavy traffic conditions. If heavy traffic is pushing drivers out of their zones, we should observe an increase in the magnitude of the new driver entry effect on zone-changing behavior under heavy traffic; the coefficient measuring the new driver entry effect would reflect the combined effects of zone-changing behavior under heavy traffic and new driver entry.[16] Thus, if heavy traffic pushes drivers out of their zones, we would expect the estimated new driver effect to be bigger under heavy traffic.

To construct the subsets with light traffic conditions, we consider the time periods between 10:00 a.m. and 12:00 p.m. in the morning and 2:00 p.m. and 3:30 p.m. in the afternoon during weekdays. In Sao Paulo, the morning shift of primary schools ends at 12:20 p.m. and lunch break is around noon. Hence, the light traffic subset includes the time periods between the morning and afternoon rush and between the afternoon and evening rush. This subsample was chosen based on our analysis of average hourly traffic information provided by Google Maps. Traffic is relatively light during these periods.[17] In contrast, the heavy traffic period subsample includes the evening rush (between 4:30 p.m. and 6:30 p.m.).[18]

---

[16]It is possible that demand level under heavy traffic and light traffic are different, and that this difference in demand size impacts how drivers react to the entry of new drivers. For instance, drivers might scatter when potential demand is high and herd when potential demand is low. To make sure that the heterogeneity in drivers' reaction to new driver entry under different traffic conditions is not driven by the difference in demand size, we conduct additional analyses. Our analyses show that drivers scatter due to the entry of new drivers regardless of prevailing traffic conditions. Likewise, drivers' scattering probability increases under light traffic, even after controlling for demand. Moreover, the analyses in Section 2.7.4 indicate that although the magnitude of the new driver entry effect on zone-changing behavior is correlated with demand, entry of new drivers always results in scattering.

[17]This information is obtained from the traffic section of Google Maps. The traffic density is represented through visual maps.

[18]During the morning rush, traffic density is higher in the arterial roads not in the city center. Since we restricted our analysis to the city center, we only focused on evening rush.

The first four columns in Table 2.6 show that drivers tend to scatter as a result of new driver entry. Furthermore, the estimated magnitude of the new driver effect is smaller in the subsample with heavy traffic conditions. Thus, our results are not driven by the shoving impact of heavy traffic. However, one possible concern is that time periods with different traffic densities might also differ in terms of demand level. The results presented in the first four columns in Table 2.6 could be driven by prevailing demand characteristics. To eliminate this concern, we further subsample our data set into two time periods with similar demand size (proxied by sales volume) but different traffic densities. As observed in Figure 2.7, the demand for service and the average number of active drivers follow a similar pattern between 2 p.m. and 3 p.m. and between 6 p.m. and 7 p.m. during the weekdays.[19] Traffic densities, however, differ markedly in these two time periods; the last four columns in Table 2.6 present results in line with the previous analysis. The effect of traffic conditions on the zone-changing behavior of drivers is robust to demand levels. These results provide strong support for Hypothesis 1B—i.e., that competition, and not traffic conditions, causes the observed scattering behavior.

Next, we calculate the speed of drivers during a given hour.[20] The average speed of the driver serves as a proxy for traffic conditions; if the driver is traveling at a low speed, this is an indication that traffic is heavy. Regardless of whether we control for traffic conditions, the main results are still present and sizable. Hence, our results cannot be attributed to traffic conditions.

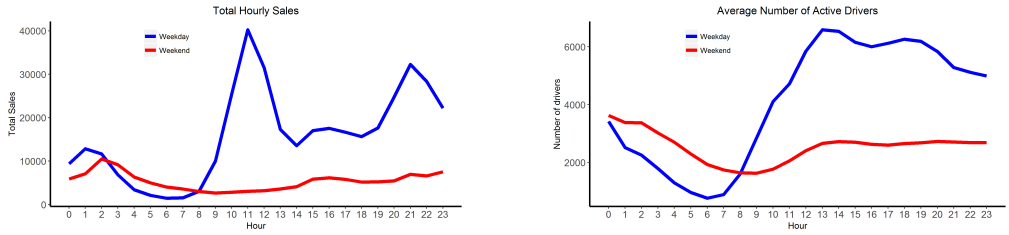
[19]The average number of active drivers includes drivers who logged into the application and changed zone at least once in a given hour.

[20]We first calculate speed by using two consecutive observations from the same driver observed in the same zone. To calculate the average speed for a given zone and hour, we take the average of these speed observations as long as there are at least five speed observations. We create a median split based on average speed information to categorize high and low speeds.

Table 2.6. Effect of Traffic on Strategic Scattering

	Light Traffic 10 am-12 pm & 2 pm-3:30 pm		Heavy Traffic 4:30 pm-6:30 pm		Light Traffic 2 pm-3 pm		Heavy Traffic 6 pm-7 pm	
NewDrivers	0.0358***	0.0329***	0.0355***	0.0322***	0.0331***	0.0290***	0.0276***	0.0259***
	(43.01)	(37.27)	(41.13)	(35.08)	(30.86)	(26.03)	(26.47)	(23.59)
SaleTime	0.0250***	0.0240***	0.0726***	0.0622***	0.0501***	0.0439***	0.0732***	0.0620***
	(9.15)	(8.79)	(19.01)	(16.24)	(8.14)	(7.11)	(17.78)	(15.01)
SalesZone	-0.0378***	-0.0293***	-0.0391***	-0.0280***	-0.0375***	-0.0235***	-0.0375***	-0.0239***
	(-77.11)	(-51.77)	(-69.08)	(-41.66)	(-51.54)	(-25.63)	(-53.12)	(-27.85)
GetIn		-0.0223***		-0.0111*		-0.0241***		-0.0185**
		(-5.19)		(-2.06)		(-3.30)		(-2.74)
GetOut		0.0223***		0.0346***		0.0204		0.0257*
		(3.41)		(3.98)		(1.69)		(2.33)
VacantIn		-0.0221***		-0.0237***		-0.0227***		-0.0229***
		(-44.70)		(-41.66)		(-33.88)		(-35.42)
VacantOut		0.0358***		0.0302***		0.0302***		0.0235***
		(39.28)		(32.15)		(26.41)		(20.89)
Observations	1776335	1776335	2245139	2245139	782413	782413	725271	725271
AIC	1817791.79	1813958.81	2206073.32	2201059.24	716322.798	714599.774	705365.304	703794.818
BIC	1818374.12	1814590.7	2206704.53	2201740.95	716831.884	715155.141	705871.053	704346.544
Log Likelihood	-908848.9	-906928.4	-1102986.7	-1100475.6	-358117.4	-357251.89	-352638.65	-351849.41
χ^2	27566.065	31407.048	33652.904	38674.981	7193.63	8924.653	6306.017	7884.503

t statistics in parentheses
All models estimate a logit regression with row, column, and driver fixed effects.
Dependent variable is *ChangeZone*.
Only weekdays are considered.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$



(a) Total hourly sales **(b)** Average number of active drivers

Figure 2.7. Total sales and total number of active drivers in each hour

2.7.4. Effect of Demand Characteristics

Another concern is that drivers' zone changing behavior might be a result of demand dynamics or street hailing. Additionally, we might be attributing drivers' zone changes to new

driver entries even when the change of zones would have taken place in the absence of driver entry. In this section, we address these concerns. First, we control for location and time-specific characteristics that might shift demand for service. Second, we explore other potential explanations of zone-changing behavior and address the issue of street hailing.

The potential demand for taxi service that originates from a zone depends on population characteristics, time of day, weather conditions, and various zone-specific characteristics. Allon et al. (2011) use U.S. Census data, and Kabra et al. (2016a) employ data from INSEE, the French national statistics bureau, to obtain population characteristics of the areas under consideration. To the best of our knowledge, a data set containing population characteristics of the Sao Paulo city center is not readily available. As a result, we are unable to control for population characteristics. Nevertheless, we collect data on locations of transportation services, points of interest, and hourly weather to account for potential demand shifters in our analysis.

Transportation options in a region might impact the demand for service and the scattering behavior of taxi drivers. Sao Paulo's broad public transportation system serves almost 14 million inhabitants daily. The subway of Sao Paulo is composed of five commuter lines and serves more than 5 million passengers during weekdays. In addition to the subway, tram, and railway systems, the city public transportation system also includes 16,000 buses. Through the Google Places API, we collect data on the locations of 44 subway, 21 train, and 3,187 bus stations in the city center of Sao Paulo, as well as the locations of taxi stations.

Apart from transportation systems, public places such as restaurants, cinema and theaters, schools and hospitals might impact the magnitude of demand in the zones in which they are located. We collect coordinates of more than 18,000 points of interest in the city center of Sao Paulo using the Google Places API. The points of interests (POI) we geocode are restaurants,

hotels, nightclubs, cafes, shopping malls, hospitals, schools, cinema and theaters, libraries, and museums. We believe that these POI data cover most of the demand generators in Sao Paulo's city center.[21]

Finally, weather conditions, such as rain or extreme temperatures, might impact demand for taxi services. We collect hourly weather data for the city center of Sao Paulo, including temperature, humidity, and weather conditions, such as rain, strong thunderstorms, mist, heavy clouds, haze, and clear skies, from a weather website for the period of analysis.

We retest Hypothesis 1 after controlling for transportation, points of interest, and weather data. Table 2.7 shows that Hypothesis 1B—i.e., that competition leads to scattering behavior—is robust to the inclusion of demand shifters.

In addition to demand shifters, changes in demand level might impact how drivers react to the entry of new drivers. Changes in potential demand could potentially induce drivers to herd instead of scatter as a result new drivers entering their zone. So that we can show that our results are robust to changes in demand size, we separately analyze two time periods with different demand sizes but an approximately equal number of average active drivers.

Demand size and distribution at a given hour can impact the zone-changing behavior of drivers in three different ways. First, fixing the number of drivers in a given zone, entry of a new driver generates a lower marginal decrease in the probability of capturing a unit demand when demand is low. Second, when demand is low, the search cost—i.e., the sum of fuel costs and the opportunity cost of time— might well exceed the expected gain obtained by cruising. Third, during low demand hours, especially during the night, demand might be concentrated in

[21]We searched for events, such as festivals and football games, which might impact demand. No such major event took place in the city center of Sao Paulo during the time period of our analysis. There was a soccer game that took place in the Arena Corinthians on September 3, but the stadium is located far from the city center.

Table 2.7. Strategic Scattering

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
NewDrivers	0.0286***	0.0309***	0.0284***	0.0307***	0.0233***	0.0261***	0.0237***	0.0257***
	(33.43)	(32.71)	(32.5)	(31.75)	(26.21)	(26.77)	(26.42)	(26.32)
SaleTime	0.0355***	0.0361***	0.0369***	0.0373***	0.0427***	0.0427***	0.0422***	0.0423***
	(28.73)	(29)	(29.38)	(29.46)	(29.02)	(28.85)	(28.67)	(28.54)
SalesZone	-0.0307***	-0.0193***	-0.0309***	-0.0196***	-0.0107***	-0.00287***	-0.0107***	-0.00316***
	(-59.59)	(-33.41)	(-58.65)	(-33.10)	(-14.16)	(-3.65)	(-14.26)	(-4.00)
Weekend	0.241***	0.181***	0.249***	0.188***	0.235***	0.183***	0.233***	0.184***
	(36.38)	(26.78)	(36.64)	(27.06)	(26.9)	(20.72)	(26.69)	(20.77)
GetIn		-0.0262***		-0.0264***		-0.0239***		-0.0217***
		(-5.38)		(-5.34)		(-4.82)		(-4.37)
GetOut		0.0325***		0.0337***		0.0291***		0.0301***
		(4.33)		(4.41)		(3.8)		(3.93)
VacantIn		-0.0293***		-0.0293***		-0.0268***		-0.0262***
		(-56.29)		(-54.85)		(-49.38)		(-47.28)
VacantOut		0.0354***		0.0355***		0.0302***		0.0297***
		(36.01)		(35.36)		(29.86)		(29.25)
Hourly Weather Data	No	No	No	No	Yes	Yes	Yes	Yes
PTS	No	No	No	No	Yes	Yes	Yes	Yes
PTS x Time-of-Day	No	No	No	No	No	No	Yes	Yes
POIs	No	No	No	No	Yes	Yes	Yes	Yes
POIs x Time-of-Day	No	No	No	No	No	No	Yes	Yes
Observations	1156044	1156044	1106963	1106963	1106963	1106963	1106963	1106963
AIC	1176400.74	1172531.43	1125296.89	1121609.76	1120702.91	1117799.62	1120065.52	1117364.81
BIC	1177202.09	1173380.62	1126095.34	1122455.88	1121930.38	1119074.75	1121936.51	1119283.47
Log Likelihood	-588133.37	-586194.71	-562581.45	-560733.88	-560248.46	-558792.81	-559875.76	-558521.41
χ^2	13523.187	17400.499	12938.058	16633.189	17604.034	20515.331	18349.424	21058.136

A random 10% sample is taken to estimate the models.

PTS : Public transportation stops

All models estimate a logit regression with row, column, hour, and driver fixed effects.

For each point of interest and locations of transportation stations category we created quartile vectors. These quartile vectors were included in the regressions to account for demand shifters. However, the magnitude of the coefficient of new driver entry does not change significantly when we include the exact number of venues, instead of their quartiles, for each point-of-interest category. Moreover, in order to control for the temporal effect POIs, we divided the day to four time-of-day periods: midnight (12am-6am), morning (6am-12pm), afternoon (12pm-6pm), and evening (6pm-12am), and created interaction variables of quartiles of POIs with time-of-day periods. These quartile of POIs and time-of-day interactions were included in the regressions (7) and (8) to account for temporal effects of demand shifters.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

a few specific zones; thus, the probability of picking up a passenger outside those specific zones by cruising is low. Consequently, during hours of low demand, we should observe that drivers are less inclined to scatter as a response to a new driver's entry.

On the other hand, taxis charge 30 percent more from 8 p.m. to 6 a.m. in Sao Paulo. Hence, even if demand is low, the increase in expected revenues obtained by picking up an additional

customer might exceed the cost of searching for a new customer during these periods. Thus, even under conditions of low demand, we might observe increased scattering behavior due to competition during the night shift. Our empirical analysis determines which of these effects dominates.

As we observe in Figure 2.7 (a), the period with the lowest demand for service is between 4 a.m. and 6 a.m. while the periods with the highest demand are 10 a.m. to 11 a.m., and 8 p.m. to 10 p.m. In Brazil, taxis typically operate in two separate shifts of 8–12 hours, but we do not know the exact time at which the shifts change. However, Figure 2.7 (b) shows that the average number of active drivers decreases between 4 a.m. and 6 a.m., hits its lowest point at 6 a.m., and increases gradually afterwards until 12 p.m. This pattern suggests that the shift change usually occurs sometime between 4 a.m. and 6 a.m. Thus, during this period drivers could be changing zones frequently due to the shift change, and not as a response to the entry of new drivers. To minimize the probability that the observed zone changes are due to shift changes, we focus on the time period from 2 a.m. to 3:30 a.m., where demand is low and shift changes are unlikely. Moreover, the average number of active drivers between 2 a.m. and 3:30 a.m. is close to the average number of active drivers between 9 a.m. and 10 a.m., but the total demand between 9 a.m. and 10 a.m. is almost double the total demand between 2 a.m. and 3:30 a.m. The results in Table 2.8 show that Hypothesis 1B—i.e., that entry of new drivers leads to scattering—finds empirical support regardless of the demand size. However, there is a significant difference in the magnitude of the new driver entry effect between low-demand and high-demand periods. During high-demand periods, drivers are much more likely to scatter as a result of competition. These results suggest that during periods with low demand, incurred search costs dominate the expected additional revenue that can be obtained by cruising.

Table 2.8. Effect of Demand and Experience on Strategic Scattering

	Low Demand (2 am-3.30 am)		High Demand (9 am-10 am)		Low Experience		High Experience	
NewDrivers	0.0133***	0.0184***	0.0556***	0.0523***	0.0180***	0.0247***	0.0259***	0.0296***
	(5.37)	(7.2)	(14.89)	(13.9)	(27.11)	(33.74)	(27.98)	(28.74)
SaleTime	0.0330***	0.0406***	0.114***	0.0759*	0.00600***	0.00951***	0.0292***	0.0323***
	(7.75)	(9.4)	(3.5)	(2.32)	(6.18)	(9.71)	(22.16)	(24.26)
SalesZone	0.00308	0.0110***	-0.0482***	-0.0342***	-0.0272***	-0.0124***	-0.0310***	-0.0196***
	(1.84)	(6.3)	(-22.46)	(-14.87)	(-70.27)	(-28.21)	(-54.81)	(-31.11)
GetIn		-0.0218		-0.0313*		-0.0299***		-0.0144**
		(-1.74)		(-2.02)		(-8.00)		(-2.72)
GetOut		0.0041		0.0646*		0.0163**		0.0345***
		(0.24)		(2.28)		(2.82)		(4.19)
VacantIn		-0.0489***		-0.0479***		-0.0339***		-0.0304***
		(-25.54)		(-20.66)		(-84.42)		(-54.44)
VacantOut		0.0203***		0.0469***		0.0297***		0.0354***
		(7.64)		(11.8)		(38.95)		(33.2)
Observations	261202	261202	284698	284698	1876984	1876984	971159	971159
AIC	247022.358	246348.608	209700.854	209144.526	2055584.46	2047793.54	996365.075	992859.03
BIC	247640.267	247008.41	210323.846	209809.754	2056405.84	2048664.7	997142.967	993684.067
Log Likelihood	-123452.18	-123111.3	-104791.43	-104509.26	-1027726.2	-1023826.8	-498116.54	-496359.52
χ^2	4250.805	4932.555	5867.803	6432.132	14004.266	21803.19	11068.91	14582.955

t statistics in parentheses

All models estimate a logit regression with row, column, hour, and driver fixed effects.

Only weekdays are considered.

Dependent variable is *ChangeZone*.

A random 20% sample is taken to evaluate the high/low experience regressions.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

There might be three other potential reasons why a driver changes zone irrespective of whether a new driver enters the zone. First, the driver may get a ride request through the platform from another zone, thereby inducing her to change zone. As we only consider instances where the driver was vacant after the observation time, this first reason is not a matter of concern for our analysis. Second, the driver may change her zone due to expected higher sales opportunities in a surrounding zone. To deal with this concern, we estimate a spatial autoregressive

(SAR) Model, where we consider *SalesZone* and *Vacant* values of surrounding zones. Reassuringly, even after controlling for these factors, we estimate a positive and significant effect of entry of new drivers on zone-changing behavior.[22]

Third, a driver may be carrying a passenger whom she found through another platform (or hailing from the street), but did not log out from our platform. Thus, we would observe this driver as vacant in our data set even if she is actually carrying a passenger. To show that our main results are robust to this last concern, we conduct two separate analyses. In the first analysis, we analyze the subsample composed by drivers with high utilization rates. High utilization rates in the 99Taxis platform indicate that the driver is less likely to use other platforms or pick up hailing passengers from the street. When we estimate the regression for this subsample, the magnitude of the new driver entry effect on zone-changing behavior increases. This result, which is in line with the results in Section 2.6, supports Hypothesis 2A— i.e., that drivers who respond more strongly to the entry of new drivers by scattering are the ones who have higher utilization.[23]

In the second analysis, we analyze the subsample composed by observations where a driver waited for at least five minutes in the same zone. Waiting in the same zone for five minutes minimizes concerns that the driver might be carrying a passenger from other platforms, or that the driver entered into this zone just to drop off a passenger. Although the effect of new driver entry upon scattering decreases in this subsample, it remains positive and significant.[24]

---

[22]SAR model analysis is provided in the Appendix. See Table B.2.

[23]We estimate the *NewDrivers* effect as 0.0276***, 0.0279***, 0.0345***, 0.0387***, 0.0383*** in the same order of the models used for Table 2.3.

[24]We estimate the *NewDrivers* effect as 0.0192***, 0.0192***, 0.0217***, 0.0143***, 0.0133*** by using the same models noted in Table 2.3.

Therefore, the entry of new drivers has a sizable impact upon the zone-changing behavior of drivers, even after accounting for street hailing.

2.7.5. Effect of Drivers' Experience

Camerer et al. (1997) show that taxi drivers become more sophisticated in adjusting their supply as they gain experience. On the other hand, Cramer and Krueger (2016) demonstrate that services such as Uber enable drivers to accommodate their supply in a sophisticated way even if they are not experienced. The experience effect could potentially impact the scattering behavior of drivers in our setting as well. Experienced drivers might be more knowledgeable about demand size and its distribution throughout the city during different time periods. They might position themselves and adjust their scattering behavior according to this extra information.

Camerer et al. (1997) use the cab-driver license number, which is assigned in chronological order, to classify drivers as experienced and inexperienced. We don't have any driver-specific information except GPS coordinates. In an ideal experiment, we would have information about the experience level of drivers and be able to assign experienced and inexperienced drivers to situations with same prevailing competition levels to test for differences in their scattering behavior. However, we do not know the drivers' experience levels. The only information we have is total active time, vacant time, and busy time pertaining to individual drivers in the system during the period under consideration. Thus, we exploit differences in total active time as a proxy of the drivers' experience with the system, and observe subsequent differences in the scattering behavior of drivers facing competition. We use a median split to classify drivers as high experience and low experience, and run independent analyses on these two subsamples. Table 2.8 shows that both driver groups react to the entry of new drivers by scattering which is

in line with Hypothesis 1B. However, drivers who are logged into the system for longer hours react to competition more strategically relative to users with less experience with the system. This analysis suggests that system usage provides drivers with a strategic leverage.

2.8. Operational Value of Visibility of the Competition

In this section, we use agent-based simulation models with multiple driver types and zones to check how agent utilization and the system's efficiency are affected by the scattering behavior we documented above. To this end, we design two simulation experiments. First, we alter the sensitivity of strategic (scattering) agents to the competition level. Second, we change the proportion of strategic (scattering) agents in the population.

We first start by providing some fundamentals of the analysis. We assume that the arrival of customers follows a Poisson process with a rate of 50 potential customers per minute. Arriving customers are spread uniformly over the zones. Whenever there is a vacant taxi in the same zone, demand occurs. Note that if there are multiple vacant drivers in the same zone as the potential customer, then each driver is equally likely to serve this customer. We also assume that a customer may abandon at the beginning of each time period with probability of 0.10; hence, customers will wait at most 10 minutes on average to receive a service. Passengers are equally likely to pick any destination zone. Travel time is calculated by the distance from the center of the originating zone to the center of the destination.

In our analysis, we assume that some of the drivers are more strategic in terms of responding to the arrival of new competitors in their zones. Those *scattering agents* change their zone with the probability of $p \times k + e$, where p is the effect of each new competitor, k is the number of

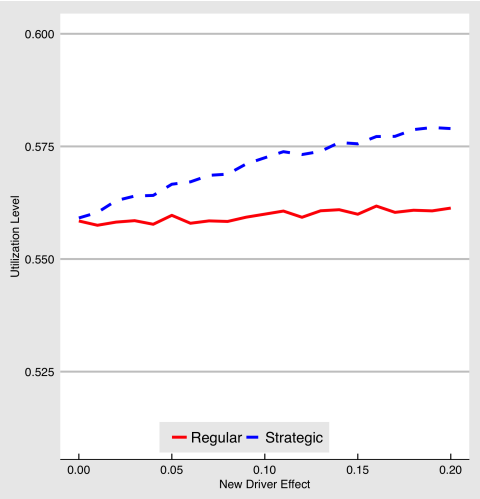
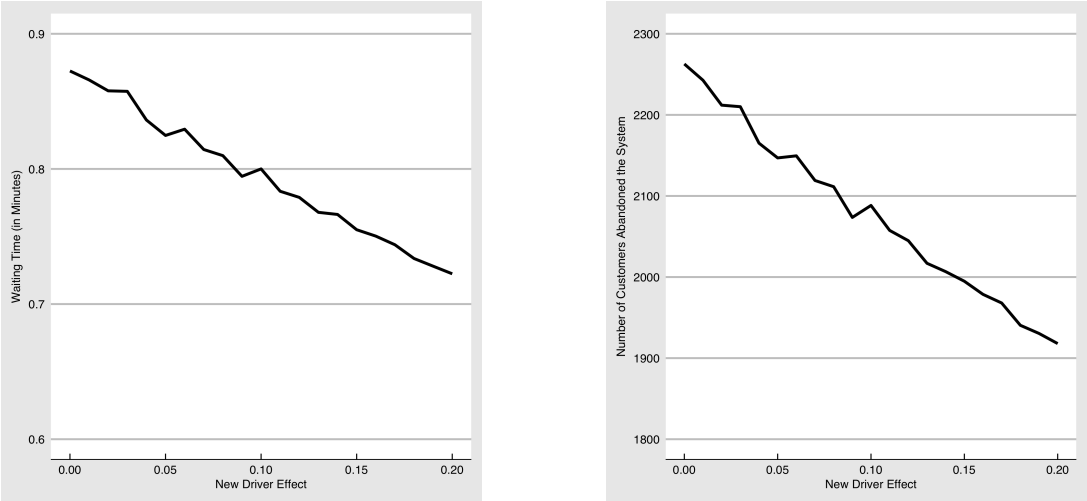


Figure 2.8. Utilization comparison of strategic and regular drivers

drivers newly entering the zone, and e is the probability of unobservable effects. Other agents change their zones only due to some unobservable factors with probability of e, and we call them *regular agents*. We start with 200 scattering and 200 regular agents. We assume that all of these agents stay logged into the platform during the whole analysis. We also assume that if an agent decides to change her zone, then she moves to a random neighboring zone without checking any information available to her. We calculate the average utilization of each type of agents and average waiting time of customers as well as the number abandoned customers.

We initially have 50 customers. Initial locations of both agents and customers are chosen randomly. We run 100 simulations where each simulation has 360 minutes to run. We assume the probability of moving due to unobservable effects is 0.10.

First, we compare the utilization of each type of agent under this system. We gradually increase the *NewDriver* effect. We observe that utilization of scattering agents is higher than regular agents, up to 2 percent (see Figure 2.8). Then we compute the average waiting time and number of abandoned customers in this system. We observe that scattering brings a significant



(a) Waiting-time comparison **(b)** Customer abandonment comparison

Figure 2.9. Efficiency of the system as NewDriver effect increases

reduction in both measures (see Figure 2.9). Average waiting-time of a passenger decreases up to 20.6 percent and the number of abandoned customers decreases up to 17.9 percent when we increase the *NewDriver* effect from 0 to 0.2. Therefore, the platform significantly benefits from the scattering behavior of drivers.

Second, we analyze how the composition of agents impacts the system utilization. As in the first part of our simulation analysis, customer arrivals follow a Poisson process with a rate 50 potential customers per minute, customers might abandon each minute with probability 0.1. Moreover, regular agents change zones with probability *e* due to unobservable effects, and the scattering agents change their zone with probability $p \times k + e$ where *p* is the effect of the new competitor and *k* is the number of new agents in the zone. We start with 400 agents and we assume that *s* percent of these agents are scattering agents and the remaining *(1-s)* percent are regular agents.

Initially, 50 customers are waiting for service and they spread uniformly over the zones. We run 100 simulations for each setting and each simulation runs for 360 minutes. We assume the probability of moving due to an unobservable event, e, is 0.1 and each new competitor increases the probability of zone changing, p, of scattering agents 10 percent. In order to understand how the proportion of scattering agents in the population impacts the system efficiency and utilization of agents, we alter the percentage of scattering agents in the population, s.

First, we calculate the utilization difference between the scattering and regular agents, depending on the percentage of scattering agents in the population. Figure 2.10 shows the scattering agents always have higher utilization than the regular agents regardless of the percentage of scattering agents in the population. However, we can observe that the utilization of regular agents increases gradually as the percentage of scattering agents in the population increases. Consequently, the presence of scattering agents helps regular agents as well, creating a positive externality. Moreover, an increase in the percentage of scattering agents in the population leads to a significant reduction in average waiting time and customer abandonments. If we move the system from a composition where there are no scattering agents to a setting where 50 percent of the agents are scattering agents, we observe decreases of 6.3 percent and 6.09 percent in the average waiting time and customer abandonments respectively. Moreover, the decreases in average waiting time go up to 12.1 percent and the customer abandonments go down up to 11 percent as we increase the percentage of scattering agents in the population from 0 percent to 90 percent (see Figure 2.11).

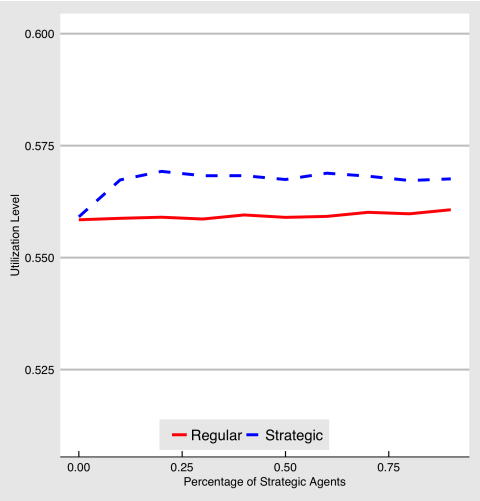
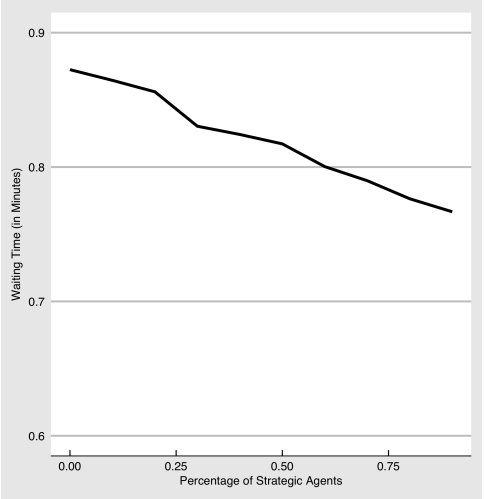
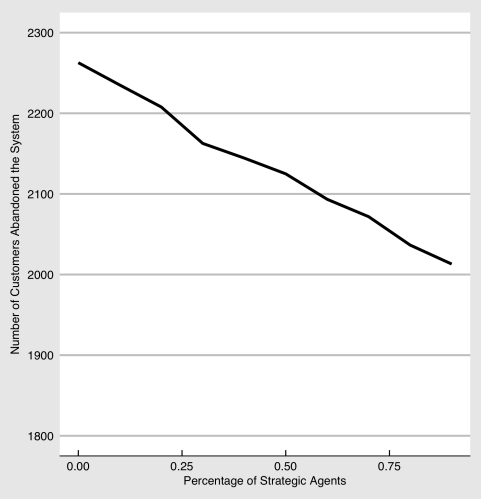


Figure 2.10. Utilization comparison of scattering and regular drivers



(a) Waiting-time comparison



(b) Abandoned customer comparison

Figure 2.11. Efficiency of the system as percentage of scattering agent increases

2.9. Conclusions

In this study, we explore how agents react to real-time information by using data from an e-hailing taxi platform. We document that agents scatter with respect to the locations of competitors; more specifically, an agent tends to move to another service zone when a competitor

enters her service zone. Therefore, such new agent arrivals are not interpreted as a signal of high sales opportunity in the zone; instead, they are considered a threat to sales.

Agents are heterogeneous in their behavior against the arrival of a competitor such that some agents change their service zone more frequently than others. We show that drivers who have a higher probability of scattering achieve higher utilization. Therefore, monitoring and reacting to the information pays off from the perspectives of both the driver and the platform. Similarly, we observe that agents achieve a higher utilization if they have a better understanding of choosing less competitive zones.

We also address the more general question of how these decisions affect the whole system. We find that the system becomes more efficient as we have more agents who engage in the following two behaviors: scattering against competitors and choosing less competitive zones. More specifically, we observe that the average utilization rate increases and both the abandonment rate and waiting time of customers decrease significantly as agents respond to competition by scattering. Moreover, the presence of scattering agents also helps the non-scattering agents, who now see less competition.

There are several ways this work can be extended. One could consider the behavior of agents as a response to real-time information about not only their competitors but also decisions of the service provider. Such information about the service provider brings more complexity because an agent needs to consider the reaction of her competitors to the service-providers decision as well as her own decision. For example, understanding how agents react to the surge pricing of e-hailing taxi platforms can be fruitful because it is not clear whether the agents would find it worthwhile to go to a zone with a higher than usual price when other competitors are already

on their way to that zone and when there is a strong chance this high price will drop due to an increase in supply.

CHAPTER 3

Impacts of Charging Carry-on Bags in Aviation Industry

3.1. Introduction

Current revenue management literature mostly focuses on seat allocation, pricing or network problems. Revenue management is a field that originates in the Airline Deregulation Act of 1978 (Talluri and van Ryzin (04 a)). There have been many studies since 1978 on different aspects of revenue management. Detailed overviews can be found in Talluri and van Ryzin (04 a) and Chiang et al. (2007). An important building block model for more complicated revenue management is single resource capacity control. It is common in airline companies to sell identical seats at different fares. The major issue is the decision process of accepting or rejecting a booking request of a certain class for a given resource. The static model in which different fare classes arrive at different, non-overlapping time stages ordered in an increasing fare class prices, is first considered by Littlewood (1972). The dynamic programming model of this problem is analyzed by Lee and Hersh (1993), and the structure of the optimality policy is investigated by Lautenbacher and Stidham (1999). For further research on single resource capacity control, see Brumelle and McGill (1993), Talluri and van Ryzin (04 b), Lan et al. (2008), Birbil et al. (2009), Özkan et al. (2013). The main focus of the literature is again the allocation problem. Since this a multi-dimensional problem, most studies focus on approximations to this problem. For example, Kunnumkal and Topaloğlu (2010) provides an approximation method for network revenue management problem with customer choice behavior by solving each flight

leg as single-leg problem. van Ryzin and Vulcano (2008) also study an approximation method for network revenue management under customer choice behavior by using a simulation-based method. Although current studies provide great insights and benefits to air-carriers, there are other factors affecting the revenue of the firm other than the optimal allocation and pricing decision.

Delays due to the boarding process play an important role on the utilization rate of the airplanes. Spirit Airlines has been applying fee on checked bag since 2007 and all US air carriers (except Southwest) charge checked bags as of June 2017 (FareCompare (2017)). This policy has two benefits. First, this is a new source of revenue for air carriers. Most carriers charge $25 for the first checked bag with increasing amount for the second and third ones as of June 2014. Second, this policy discourages passengers to carry more than they need which leads to a fuel saving and also decreases baggage handling problems (which is highly possible at connecting flights). However, there is a drawback of this second benefit. Most of these carriers do not charge for carry-on items. Carry-on items usually include one personal item and a small bag that can fit into to bins which are located above seats. Since there is a fee on the checked bag, passengers use their free right to have one small carry-on bag which is a potential source for higher departure delays. In 2010, Spirit Airlines started to charge carry-on as well (FareCompare (2017)). The company charges more for a carry-on than checked bag in order to discourage passengers to take their bag with them into the plane. Moreover, the carry-on fee is cheaper online, and increases a lot at the gate. Table 1 provides the current bag fees of Spirit Airlines. With this new change to its current bag-fee policy, Spirit airlines has $536 M revenue from non-ticket revenue which corresponds to 40% of its total revenue and 41% of this non-ticket revenue comes from bag fees (Wall Street Journal (2013)). These revenues from bags

		Booking		Check-in		Airport	Gate
	Carry-On Bag	$26*	$35	$36*	$45	$50	$100
Checked Bag	First one	$21*	$30	$31*	$40	$45	$100
	Second one	$31*	$40	$41*	$50	$55	$100
	Third or more	$76*	$85	$86*	$95	$100	$100

Table 3.1. Bag fee of Spirit Airlines as of February 2015 (* is the fee for $9 Fare Club, special program).

are increasing in each year. For example, another low-cost airline company - Frontier Airlines - made $220 M revenue from bag fees in 2015 which was $69.2 M in 2013. More importantly, this corresponds to 13% of the overall revenue for these low-cost carriers and the baggage-fee revenues were 0.2 to 4% of the overall revenue for the other major airlines (The Denver Post (2016)).

In addition to its direct effect on the revenue due to the fees, the question still remains unanswered about the operational consequences of such policy. Spirit Airlines has declared that it can board an A320 in 20 minutes which is 10 minutes less than the boarding time of large-scale carriers such as American and US Airways (Forbes (2013)). Average aircraft list price of an Airbus A320 is $99 M in 2017 (Airbus (2017)), therefore; carriers want to use their airplanes as much as possible and such reduction in boarding time increases the chance of high utilization of the plane.

As of June 2017, only two more US carriers, Allegiant Air and Frontier Airlines, use similar policies on carry-on bag as Spirit Airlines applies (TripAdvisor (2017)). On August 6th 2013, Frontier Airlines first started to charge carry-on customers that buy from 3rd party suppliers such as Orbitz and Skyscanner. On April 28th 2014, Frontier Airlines started to charge carry-on

baggage not only the customers buying from 3rd party suppliers but also customers buying from Frontier Airlines directly.

In this study, we aim to understand how the new policy on carry-on baggage affects the carrier delays. By using available data of U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS), we analyze the effects of this new policy on Frontier's delay times after the first attempt and the last attempt. We believe that pricing policy is a great opportunity for firms not only to increase their revenue but also to decrease their delay times. Moreover, this pricing policy simplifies the boarding process because carrier categorizes passengers into two groups, passengers with carry-on and passengers without carry-on (except passengers with disabilities, special program passengers etc.).

There is a significant number of studies on evaluating the performance of different the boarding processes, see Nyquist and McFadden (2008) and Van Landeghem and Beuselinck (2002) for details. For example, Van Landeghem and Beuselinck (2002) checks the boarding time performance of seven different boarding procedures by using simulation. They find that any procedure should separate consecutive passengers far enough to reduce a potential interference. Even though our study does not focus on the boarding process, the new fee on carry-on bag changes the dynamics of such processes. Therefore, we believe that our study sheds a light on a need to explore the effect of carry-on fee on boarding process.

Our study is not the first that uses data from BTS to analyze the operational efficiencies. For example, Rupp and Sayanak (2008) show that low-cost carries have slightly lower arrival delays. Another stream of research focuses on the relation between financial performance and operational efficiency (See Ramdas et al. (2013) and Phillips and Sertsios (2013)). Consumer

behavior is also analyzed by using this data set. For example, Li et al. (2014) consider a structural estimation method where they analyze the effects of strategic customers on financial performance.

To our knowledge, Nicolae et al. (2016) is the only paper that investigates effect of baggage-fee on delays. In their study, they compare policies of charging check-in bags. They find that charging only one check-in bag provides significant relative improvement in air-carrier's on-time departure performance when compared to the carriers that do not charge customers for their check-in bags. In Nicolae et al. (2016), the departure delay is recalculated due to the spill-over effect created by previous flights where they implement a technique used in Arıkan et al. (2013). In our study, we investigate another policy which has potentially more direct implications on departure delay. We also use the air-carrier delay which is part of arrival delay and reported separately than the previous delays (previous spill over delays are also reported separately). Hence, we use a different performance metric to analyze another policy.

We find that implementation of carry-on bag fees was associated with delay. We see a significant decrease in the delay when the firm charged a carry-on bag fee on every purchasing channels. In the rest of this study, our goal is to provide several robustness tests where we consider different measures of delay and other important factors affecting the dynamics of airline industry such as weather, loading factor for each flight etc.

The remainder of this study is organized as follows. In Section 2, we explain the data, variables and descriptive statistics. In Section 3, we discuss empirical specifications and initial results. In Section 4, we discuss the ideal research setting and need of additional data for robustness purposes. We have our concluding remarks in Section 5.

3.2. Data Description and Results

We use the data set of On-Time Performance of major air-carriers provided by BTS. Data set contains detailed arrival and departure information of domestic flights. More specifically, we have the origin and destination airports, scheduled and actual times for both arrival and departure, flight numbers, flight date (including which day of the week), departure delay amount, arrival delay amount which is separated into 5 components: Air Carrier delay (such as delay due to boarding, aircraft cleaning, baggage loading etc.), aircraft arrival (a previous flight with same aircraft arrived late, causing the present flight to depart late), National Aviation System (such as airport operations, non-extreme weather conditions, air traffic control, heavy traffic volume), weather delay (extreme weather conditions), and security (such as re-boarding of aircraft because of security breach).

We check the effects of the policy change by only using this available data. Please note that the new policy starts at On April 28th 2014, and we consider the data from August 2013 until February 2015. Therefore, we have 9 months for each of before and after policy periods. In total, we have 9,059,661 flights to investigate. We use i to denote the flight number. As a dependent variable we use air carrier delay which is denoted by $CarrierDelay_i$ for flight i.

Table 3.2. Before and after policy comparisons

	All Flights			Flights with a delay > 0		
	Before Policy	After Policy	Change	Before Policy	After Policy	Change
Frontier (Average delay)	2.751	3.321	$\uparrow 20.7\%$	18.922	26.540	$\uparrow 40.3\%$
Others (Average delay)	3.104	3.771	$\uparrow 21.5\%$	30.706	32.448	$\uparrow 5.7$
Frontier (Flight Numbers)	31,590	97,544		$4,593\ (14.5\%)$	$12,206\ (12.5\%)$	
Others (Flight Numbers)	2,471,822	6,458,705		$249,890\ (10.1\%)$	$750,637\ (11.6\%)$	

Percentage of flights with a delay among all flights.

Average delays for both Frontier airlines and other air-carriers are reported in Table (3.2) where we take the average of delay among all flights or only among the flights with a delay strictly greater than 0. Frontier airlines faces with an overall increases in delay; however, it is less than the industry average. On the other hand, the average delay among the flights with a delay strictly greater than 0 is much higher for Frontier than it is for other airlines. Overall, all these observations suggest that we need to take additional variables into account because delays increase across all industry which suggests considering factors related to seasonality.

In the same table, we also report the number of flights operated by Frontier v.s. other airlines where we calculate the percentage of flights with a delay. We observe that 14.5% of flights operated by Frontier have a delay before the policy change but this percentage goes down to 12.5% after the policy change. For the other airlines, this percentage increases to 11.6% from 10.1%. Therefore, Frontier Airlines improves the delay performance based on this measure whereas the other airlines faces with an increase. Overall, this suggest investigating the reason of changes and control other factors.

We define $Frontier_i$ as the dummy variable where $Frontier_i = 1$ if flight i is operated by Frontier, otherwise $Frontier_i = 0$ to consider the fixed effect of Frontier Airlines. Similarly, we define $AfterPolicy_i$ is the dummy variable where $AfterPolicy_i = 1$ if flight i is after the policy change, and $AfterPolicy_i = 0$ otherwise. Variable of interest is $Frontier_i \times AfterPolicy_i$ to understand the effect of policy change on air-carrier delay. To consider, air-carrier specific factors we define $AirCarrier_{(i,j)}$ to be 1 if the flight i is operated by air-carrier j, and 0 otherwise. There are 17 air-carriers in our data set during the observation period.

In addition to these variables, we believe that airport related factors should be taken into account. Therefore, we also consider two variables for congestion. One of them measures the

Table 3.3. Definition of Variables

Variable	Description
CarrierDelay$_i$	Amount of delay caused by the carrier that operates flight i (in minutes).
Frontier$_i$	Indicator that shows if flight i is operated by Frontier Airlines.
AfterPolicy$_i$	Indicator that shows if flight i is operated after the policy change.
LowCost$_i$	Indicator that shows if flight i is operated by a low-cost air-carrier.
Weekend$_i$	Indicator that shows if flight i is operated on a weekend.
OverallCongestion$_i$	Number of flights in the same airport at most 30 minutes before/after.
CongestionCarrier$_i$	Number of flights operated by the same air-carrier at most 30 minutes before/after.
AirCarrier$_{(i,j)}$	Indicator that shows if flight i is operated by air-carrier j.
Month$_{(i,m)}$	Indicator that shows if flight i is occurred in month m.

congestion regarding to the whole airport, and the other one is the congestion regarding to the air-carrier. Basically, $CongestionCarrier_i$ is the number of flights operated by the same air-carrier that also handles flight i at the same (departure) airport where each of these flights are at most 30 minutes before or 15 minutes after flight i. Similarly, we define $OverallCongestion_i$ where we consider all flights (independent of the carrier of flight i) at the same airport of flight i where each flight is at most 30 minutes before or 15 minutes after flight i. We also consider the distance of the flight by using 6 categorical variables which is defined for each 250 miles. For example, if a flight is from distance group two, then the distance between the origin and destination is between 250 miles and 500 miles. We denote this variable by $DistanceGroup_i$.

Note that there is a segmentation in terms of the operating cost of firms: low cost or high cost carriers. Air-carriers which define themselves as low cost air-carriers are the ones with the AirlineID of ”19393”,”20409”,”20436” (Frontier), ”20437”, or ”21171”. Other carriers declare themselves as high cost carriers. Therefore, we define $LowCost_i$ to be 1 if flight i is operated by one of these low-cost carriers , and 0 otherwise. We also control the weekend effect, i.e., $Weekend_i = 1$ if flight i is operated on a weekend, and 0 otherwise. To control other time effects, we define $Month_{(i,m)}$ is 1 if the i^{th} flight has been operated in month m, otherwise

it is 0. We also consider the effect of days by defining $WeekofDay_{(i,d)}$. For example, if $WeekofDay_{(i,2)} = 1$, then the flight i is operated on a Tuesday but if $WeekofDay_{(i,5)} = 1$, then the flight i is operated on a Friday. We summarize our variables in Table 3.3.

Table 3.4. Statistical Summary Airline Carrier Delay Data

	Mean	SD	Min	Max
CarrierDelay$_i$	3.581	21.054	0	2,402
Frontier$_i$	.0143	.119	0	1
Frontier $\times$ *AfterPolicy*$_i$	.011	.103	0	1
AfterPolicy$_i$	.724	.447	0	1
CongestionCarrier$_i$	6.610	8.174	1	67
OverallCongestion$_i$	18.135	16.404	1	96
LowCost$_i$	.279	.449	0	1
Weekend$_i$	.261	.439	0	1
Observations	9,059,661			

In Table 3.4, we provide summary statistics including mean, standard deviation, minimum and maximum value of each driver-level variable. Among 9,059,661 flights, the average delay is 3.581 minutes and the percentage of Frontier flights is 1.43%. On average, there are 18.135 flights at most 30 minutes before and 15 minutes after each flight and 6.610 of which is operated by the same airline. We also provide the correlation between the variables in Table 3.5. We

Table 3.5. Correlation between the variables.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1. CarrierDelay	1							
2. Frontier	-0.0023	1						
3. Frontier x AfterPolicy	-0.0013	0.8676	1					
4. AfterPolicy	0.0141	0.0085	0.0645	1				
5. CongestionCarrier	-0.0046	-0.0464	-0.0429	0.0050	1			
6. OverallCongestion	-0.0035	0.0138	0.0068	-0.0396	0.7300	1		
7. LowCost	-0.0025	0.1933	0.1677	0.0213	-0.1292	-0.1746	1	
8. Weekend	0.0004	-0.0001	0.0000	0.0024	-0.0207	-0.0356	-0.0007	1

observe that the correlation between $CarrierDelay$ and $Frontier \times AfterPolicy$ is small but negative.

By using these variables, we have the following model

(3.1)

$$AirCarrierDelay_i = \beta_0 + \beta_1 Frontier_i + \beta_2 AfterPolicy_i + \beta_3 \left(Frontier_i \times AfterPolicy_i\right)$$

$$+\beta_4 LowCost_i + \beta_5 Weekend_i + \beta_6 OverallCongestion_i + \beta_7 CongestionCarrier_i$$

(3.2)
$$+\sum_{m=1}^{12} \theta_m Month_{(i,m)} + \sum_{j=1}^{16} \gamma_j AirCarrier_{(i,j)} + \sum_{j=1}^{6} \zeta_j DistanceGroup_{(i,j)}$$

$$+\sum_{j=1}^{7} \nu_j WeekofDay_{(i,j)} + \epsilon_i$$

We consider variations of this model and provide our results in Table 3.6. Our variable of interest is $Frontier_i \times AfterPolicy$. Under all of the models, we observe that the coefficient for this variable is negative and significant. Therefore, the policy change caused a significant decrease in delays for Frontier airlines. For robustness purposes, the ideal research setting should consider additional variables. Even though $CarrierDelay$ is separated from weather related delay, it is important to control the weather forecast during the observation period. We also need to consider the effect of the airplane utilization for each flight. Unfortunately, air-carriers are not required to report the seat occupancy. We can only gather the information of airplane capacity by using the tail number of each flight (which gives us the model of the airplane). However, this is not sufficient to control the number of tickets sold which may both be correlated with the $CarrierDelay$ and Frontier.

Table 3.6. Carry-on Policy Effect on Carrier Delay

	(1)	(2)	(3)	(4)	(5)	(6)
Frontier	-0.269*	-0.635***	-0.635***	-0.857***	0.350	3.181**
	(-2.25)	(-5.23)	(-5.23)	(-7.07)	(0.59)	(3.29)
Frontier × **AfterPolicy**	-0.346*	-0.438**	-0.438**	-0.510***	-0.820***	-0.498***
	(-2.53)	(-3.20)	(-3.20)	(-3.73)	(-5.91)	(-3.64)
AfterPolicy	1.032***	1.002***	1.002***	0.776***	0.780***	0.778***
	(22.83)	(22.17)	(22.17)	(17.16)	(17.22)	(17.19)
LowCost			1.832***	2.219***	1.414	0.896
			(10.82)	(13.10)	(0.00)	(1.04)
Weekend			-0.199***	-0.261***	-0.265***	-0.260***
			(-7.66)	(-10.05)	(-10.22)	(-10.04)
OverallCongestion				-0.0743***	-0.0499***	-0.0738***
				(-66.83)	(-38.35)	(-66.15)
CongestionCarrier				-0.0586***	-0.163***	-0.0608***
				(-38.09)	(-60.73)	(-39.24)
Constant	19.49***	21.01***	19.17***	17.80***	24.23	19.80***
	(144.32)	(50.69)	(42.95)	(39.88)	(0.01)	(20.69)
Observations	9,059,661	9,059,661	9,059,661	9,059,661	9,059,661	9,059,661
AIC	80,811,627.9	80,795,748.1	80,795,748.1	80,784,402.7	80,761,776.6	80,780,167.3
BIC	80,812,553.2	80,801,453.9	80,801,453.9	80,790,136.6	80,784,403.9	80,787,625.6
LogLikelihood	-40,405,747.9	-40,397,467.0	-40,397,467.0	-40,391,792.3	-40,379,274.3	-40,389,551.7
chi2						

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$,*** $p < 0.001$
All models have airline, day and hour fixed effects.
Models 2, 3, 4, 5 and 6 have origin and distance-group fixed effects.
Models 5 has airline and origin interaction fixed effects.
Models 6 has airline and distance-group interaction fixed effects.

3.3. Concluding Remarks

Charging for carry-on bags is a new policy that has been adopted by a couple of low-cost air-carriers. It is an immediate revenue source for firms but it is not clear whether firms benefit or are hurt by this policy operationally. In this paper, we ask how this policy change affect the delay performance of firms. On one hand, it is highly likely to observe a decrease in the

boarding time, thereby, a decrease in the overall delay. On the other hand, passengers will prefer to use checked bags which means an increase in the handling and thereby, an increase in the overall delay. We show that the new policy caused a decrease in delay times significantly. To understand how robust our result, we also propose what additional measures we need to consider.

References

Afèche, P. and H. Mendelson (2004). Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Management Science 50*(7), 869–882.

Afèche, P. and M. Pavlin (2016). Optimal price/lead-time menus for queueing systems with customer choice: segmentation, pooling, and strategic delay. *Management Science 62*(8), 2412–2436.

Airbus (2017). 2017 price adjustment for airbus modern, fuel-efficient aircraft. http://www.airbus.com/presscentre/pressreleases/press-release-detail/detail/2017-price-adjustment-for-airbus-modern-fuel-efficient-aircraft/. Accessed: 2017-05-01.

Allon, G. and A. Bassamboo (2011). Buying from the babbling retailer? the impact of availability information on customer behavior. *Management Science 57*(4), 713–726.

Allon, G., A. Bassamboo, and I. Gurvich (2011). we will be right with you: Managing customer expectations with vague promises and cheap talk. *Operations research 59*(6), 1382–1394.

Allon, G., A. Federgruen, and M. Pierson (2011). How much is a reduction of your customers' wait worth? an empirical study of the fast-food drive-thru industry based on structural estimation methods. *Manufacturing & Service Operations Management 13*(4), 489–507.

Anand, K. S., M. F. Pac, and S. Veeraraghavan (2011). Quality-speed conundrum: trade-offs in customer-intensive services. *Management Science 57*(1), 40–56.

Arıkan, M., V. Deshpande, and M. Sohoni (2013). Building reliable air-travel infrastructure using empirical data and stochastic models of airline networks. *Operations Research 61*(1), 45–64.

Barlow, R. E. and F. Proschan (1965). *Mathematical Theory of Reliability*. SIAM.

Bell, D. R., S. Gallino, and A. Moreno (2016). Offline showrooms in omni-channel retail: Demand and operational benefits. *Management Science (Forthcoming)*.

Birbil, Ş. İ., J. Frenk, J. Gromicho, and S. Zhang (2009). The role of robust optimization in single-leg airline revenue management. *Management Science 55*(1), 148–163.

Brumelle, S. and J. McGill (1993). Airline seat allocation with multiple nested fare classes. *Operations Research 41*(1), 127–137.

Buchholz, N. (2015). Spatial equilibrium, search frictions and efficient regulation in the taxi industry. Technical report, Working paper.

Cairns, R. D. and C. Liston-Heyes (1996). Competition and regulation in the taxi industry. *Journal of Public Economics 59*(1), 1–15.

Camerer, C., L. Babcock, G. Loewenstein, and R. Thaler (1997). Labor supply of new york city cabdrivers: One day at a time. *The Quarterly Journal of Economics*, 407–441.

Chiang, W., J. Chen, and X. Xu (2007). An overview of research on revenue management: current issues and future research. *International Journal of Revenue Management 1*(1), 97–128.

Cramer, J. and A. B. Krueger (2016, May). Disruptive change in the taxi business: The case of uber. *American Economic Review 106*(5), 177–82.

Doroudi, S., M. Akan, M. Harchol-Balter, J. Karp, C. Borgs, and J. Chayes (2015). Priority pricing in queues with a continuous distribution of customer valuations. *Working paper*, Carnegie Mellon University, Tepper School of Business.

Ehlers, L. and A. Erdil (2010). Efficient assignment respecting priorities. *Journal of Economic Theory 145*(3), 1269–1282.

Elhorst, J. P. (2010). Applied spatial econometrics: raising the bar. *Spatial Economic Analysis 5.1*, 9–28.

FareCompare (2017). History of airline fees: Bags, food more. http://www.farecompare.com/travel-advice/airline-fees-bags-history/#/. Accessed: 2017-06-01.

Flandersnews.be (2013). Walibi to introduce queue jumpers pass. http://deredactie.be/cm/vrtnieuws.english/Life/1.1651691. Accessed: 2015-09-02.

Forbes (2013). Spirit airlines: We board an a320 in 20 minutes. https://www.forbes.com/sites/tedreed/2013/05/21/spirit-airlines-we-board-an-a320-in-20-minutes/#751d33e04844. Accessed: 2017-05-01.

Gallino, S. and A. Moreno (2014). Integration of online and offline channels in retail: The impact of sharing reliable inventory availability information. *Management Science 60*(6), 1434–1451.

Gavirneni, N. and V. Kulkarni (2014). Self-selecting priority queues with burr distributed waiting costs. *Working Paper, Cornell University*.

Ghanem, S. B. (1975). Computing center optimization by a pricing-priority policy. *IBM Systems Journal 14-3*, 272–291.

Gilland, W. and D. Warsing (2009). The impact of revenue-maximizing priority pricing on customer delay costs. *Decision Sciences 40*, 89–120.

Gurvich, I., M. Lariviere, and A. Moreno-Garcia (2015). Operations in the on-demand economy: staffing services with self-scheduling capacity. Technical report, Working paper.

Hassin, R. (2016). *Rational queueing*. CRC Press.

Hassin, R. and M. Haviv (2003). *Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Boston, MA: Kluwer Academic Publishers.

Jouini, O., Z. Aksin, and Y. Dallery (2011). Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management 13*(4), 534–548.

Kabra, A., E. Belavina, and K. Girotra (2016a). Bike-share systems: Accessibility and availability.

Kabra, A., E. Belavina, and K. Girotra (2016b). The efficacy of incentives in scaling marketplaces. *Working Paper*.

Kalai, E., M. I. Kamien, and M. Rubinovitch (1992). Optimal service speeds in a competitive environment. *Management Science 38*(8), 1154–1163.

Katta, A. and J. Sethuraman (2005). Pricing strategies and service differentiation in queues—a profit maximization perspective. *Working Paper, Columbia University*.

Kleinrock, L. (1967). Optimum bribing for queue position. *Operations Research 15-2*, 304–318.

Kunnumkal, S. and H. Topaloğlu (2010). A new dynamic programming decomposition method for the network revenue management problem with customer choice behavior. *Production and Operations Management 19*(5), 575–590.

Lai, C.-D. and M. Xie (2006). *Stochastic ageing and dependence for reliability*. Springer Science & Business Media.

Lan, Y., H. Gao, M. Ball, and I. Karaesmen (2008). Revenue management with limited demand information. *Management Science 54*(9), 1594–1609.

Lariviere, M. (2006). A note on probability distributions with increasing generalized failure rates. *Operations Research 54*(3), 602–604.

Lariviere, M. A. and J. A. Van Mieghem (2004). Strategically seeking service: How competition can generate poisson arrivals. *Manufacturing & Service Operations Management 6*(1), 23–40.

Lautenbacher, C. and S. Stidham (1999). The underlying Markov decision process in the single-leg airline yield-management problem. *Transportation Science 33*(2), 136–146.

Lee, T. and M. Hersh (1993). A model for dynamic airline seat inventory control with multiple seat bookings. *Transportation Science 27*, 252–265.

Li, J., N. Granados, and S. Netessine (2014). Are consumers strategic? structural estimation from the air-travel industry. *Management Science 60*(9), 2114–2137.

Li, J., A. Moreno, and D. J. Zhang (2015). Agent behavior in the sharing economy: Evidence from airbnb. *Available at SSRN 2708279*.

Littlewood, K. (1972). Forecasting and control of passengers. *In Proceedings of the 12th Annual AGIFORS Symposium*, 95–128.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies 60*(3), 531–542.

Marcus, M. and L. Lopes (1957). Inequalities for symmetric functions and hermitian matrices. *Canadian Journal of Mathematics 9*, 305–312.

Naor, P. (1969). The regulation of queue size by levying tolls. *Econometrica 37*, 15–24.

Nazerzadeh, H. and S. Randhawa (2015). Near-optimality of coarse service grades for customer differentiation in queueing systems. *Working paper*, University of Southern California, Marshall School of Business.

Netessine, S. and V. Yakubovich (2012). The darwinian workplace. *Harvard business review 90*(5), 25–6.

Nicolae, M., M. Arıkan, V. Deshpande, and M. Ferguson (2016). Do bags fly free? an empirical analysis of the operational implications of airline baggage fees. *Management Science*.

Nyquist, D. C. and K. L. McFadden (2008). A study of the airline boarding problem. *Journal of Air Transport Management 14*(4), 197–204.

Özkan, C., F. Karaesmen, and S. Özekici (2013). Structural properties of markov modulated revenue management problems. *European journal of operational research 225*(2), 324–331.

Palfrey, T. R. (1984). Spatial equilibrium with entry. *The Review of Economic Studies 51*(1), 139–156.

Papke, L. E. and J. M. Wooldridge (1996). Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of Applied Econometrics 11*, 619–632.

Phillips, G. and G. Sertsios (2013). How do firm financial conditions affect product quality and pricing? *Management Science 59*(8), 1764–1782.

Ramdas, K., J. Williams, and M. Lipson (2013). Can financial markets inform operational improvement efforts? evidence from the airline industry. *Manufacturing & Service Operations Management 15*(3), 405–422.

Rupp, N. G. and T. Sayanak (2008). Do low cost carriers provide low quality service?

Song, H., A. L. Tucker, K. L. Murrell, and D. R. Vinson (2016). Closing the productivity gap: Improving worker productivity through public relative performance feedback and validation of best practices.

Stidham, S. (2009). *Optimal design of queueing systems*. CRC Press.

Su, X. and F. Zhang (2009). On the value of commitment and availability guarantees when selling to strategic consumers. *Management Science 55*(5), 713–726.

Talluri, K. and G. van Ryzin (2004 a). *The theory and practice of revenue management*. Springer.

Talluri, K. and G. van Ryzin (2004 b). Revenue management under a general discrete choice model of consumer behavior. *Management Science 50*(1), 15–33.

TechCrunch (2015). http://techcrunch.com/2015/02/02/99taxis-raises-significant-new-cash-from-tiger-global/. Accessed: 2016-11-29.

The Denver Post (2016). Frontier airlines made $220 million on bag fees last year. http://www.denverpost.com/2016/05/04/frontier-airlines-made-220-million-on-bag-fees-last-year. Accessed: 2017-05-01.

TripAdvisor (2017). Airline baggage fees. https://www.tripadvisor.com/AirlineFees. Accessed: 2017-06-01.

Van Landeghem, H. and A. Beuselinck (2002). Reducing passenger boarding time in airplanes: A simulation based approach. *European Journal of Operational Research 142*(2), 294–308.

van Ryzin, G. and G. Vulcano (2008). Computing virtual nesting controls for network revenue management under customer choice behavior. *Manufacturing and Service Operations Management 10*(3), 448–467.

Veeraraghavan, S. and L. Debo (2009). Joining longer queues: Information externalities in queue choice. *Manufacturing & Service Operations Management 11*(4), 543–562.

Walker, B. (Oct 10, 2012). Priority queues: Paying to get to the front of the line.

Wall Street Journal (2013). Thriftiest fliers outwit fee-hungry airlines. https://www.wsj.com/articles/thriftiest-fliers-outwit-feehungry-airlines-1385407731. Accessed: 2017-05-01.

Whaley, M. (Apr 21, 2015). U.s. 36 plan hit for having ”lexus lane”.

Zhang, Y., B. Li, and K. Ramayya (2016). Learning individual behavior using sensor data: The case of gps traces and taxi drivers.

Ziya, S., H. Ayhan, and R. D. Foley (2004). Relationships among three assumptions in revenue management. *Operations Research 52*(5), 804–809.

APPENDIX A

Proofs for Chapter 1

A.1. Proofs of Lemmas

Proof of Lemma 1: The customer with valuation $\tilde{v}$ prefers to join class i over class j, meaning that

$$(1-\alpha W_i)\,\tilde{v}-p_i \geq (1-\alpha W_j)\,\tilde{v}-p_j.$$

The customer with valuation $\hat{v}$ prefers to join class j over class i, meaning that

$$(1-\alpha W_j)\,\hat{v}-p_j \geq (1-\alpha W_i)\,\hat{v}-p_i.$$

Adding up these two inequalities gives

$$(\tilde{v}-\hat{v})\alpha(W_j-W_i)\geq 0.$$

Sine $W_i > W_j$, we have $\tilde{v} \leq \hat{v}$ concluding that the customers who choose the shorter waiting time W_j have higher valuation. ■

Proof of Lemma 2: Part 1 of the lemma follows from Theorem 2 of Lariviere (2006). Parts 2 and 3 follow from Proposition 5.1 in Ziya et al. (2004). Part 4 follows noting that $\varepsilon(\lambda) = \frac{\Lambda}{\lambda}\bar{F}^{-1}(\lambda/\Lambda)\, f\left(\bar{F}^{-1}(\lambda/\Lambda)\right)$ and substituting $\bar{F}(v(\lambda))$ for λ/Λ. ■

Proof of Lemma 3: Recall that $\overline{\lambda}_i(v) = \Lambda \bar{F}(v_i)$. In what follows, we fix v and suppress the dependence of $\overline{\lambda}_i$ on v in the notation. Let $\widetilde{CS}(x) = CS\left(\bar{F}^{-1}(x/\Lambda)\right)$ and notice that

$$\widetilde{CS}'(x) = \frac{1}{\Lambda h\left(\bar{F}^{-1}(x/\Lambda)\right)}. \tag{A.1}$$

$\widetilde{CS}(x)$ is then increasing and concave in its argument [convex] if $h(v)$ is decreasing [increasing]. If $h(v)$ is constant, $\widetilde{CS}(x)$ is linear. In turn, since $\bar{\lambda}_i$ is decreasing in i, $\widetilde{CS}(\bar{\lambda}_i)$ is decreasing and convex [concave] if F is DFR [IFR]. The result then follows from noting that c_i^{CS} can be written as

$$c_i^{CS} = \frac{\widetilde{CS}(\overline{\lambda}_i) - \widetilde{CS}(\overline{\lambda}_{i+1})}{\overline{\lambda}_i - \overline{\lambda}_{i+1}}.$$

By the definition of convexity/concavity c_i^{CS} is then increasing in i if F is DFR and decreasing if F is IFR. It is constant if the hazard rate is constant. ■

Proof of Lemma 5: The expressions for $\bar{\lambda}_{1,S}^{n*}$ and $\bar{\lambda}_{1,R}^{n*}$, follow from Theorems 4 and 5 stated further below.

Here we only need to prove the comparison. The social has a strictly larger coverage (for all sufficiently large n) if $V(\bar{v})/|V'(\bar{v})| < \rho(\bar{v})/|\rho'(\bar{v})|$. Note that, since F is IGFR, $\rho'(\bar{v}) < 0$, we also have $V'(\bar{v}) = -\bar{v}f(\bar{v}) < 0$. Hence, it remains to show that

$$V(\bar{v})/V'(\bar{v}) > \rho(\bar{v})/\rho'(\bar{v}). \tag{A.2}$$

Because we assume that F has a strictly positive density there exits, for any valuation v, a unique λ such that $\lambda = \Lambda\bar{F}(v)$ or $v(\lambda) = \bar{F}^{-1}(\lambda/\Lambda)$. Let $\hat{V}(\lambda) = V(v(\lambda))$ and $\widetilde{\rho}(\lambda) = \rho(v(\lambda))$.

Then, $\hat{V}'(\lambda)) = V'(v(\lambda))v'(\lambda)$. Since $v'(\lambda) \leq 0$, (A.2) will be established if we show that

$$\frac{\hat{V}'(\lambda)}{\hat{V}(\lambda)} > \frac{\widetilde{\rho}'(\lambda)}{\widetilde{\rho}(\lambda)}. \tag{A.3}$$

Notice, to that end, that

$$\frac{\widetilde{V}'(\lambda)}{-\widetilde{V}''(\lambda)\lambda} = \Lambda \frac{\bar{F}^{-1}\left(\frac{\lambda}{\Lambda}\right) f\left(\bar{F}^{-1}\left(\frac{\lambda}{\Lambda}\right)\right)}{\lambda} = GFR(v(\lambda)).$$

Since F is IGFR and $v'(\lambda) \leq 0$ we have that $\frac{\widetilde{V}'(\lambda)}{-\widetilde{V}''(\lambda)\lambda}$ is decreasing in λ and the reciprocal $\frac{-\widetilde{V}''(\lambda)\lambda}{\widetilde{V}'(\lambda)}$ is increasing in λ. Notice that $\widetilde{V}'(\lambda) + \widetilde{V}''(\lambda)\lambda = \widetilde{\rho}'(\lambda)$. Thus, we have that $1 + \frac{\widetilde{V}''(\lambda)\lambda}{\widetilde{V}'(\lambda)} = \frac{\widetilde{\rho}'(\lambda)}{\widetilde{V}'(\lambda)}$ decreasing in λ so that, fixing λ, $\frac{\widetilde{\rho}'(\tilde{\lambda})}{\widetilde{V}'(\tilde{\lambda})} > \frac{\widetilde{\rho}'(\lambda)}{\widetilde{V}'(\lambda)}$ for any $\tilde{\lambda} \leq \lambda$. Hence we have $\widetilde{\rho}'(\tilde{\lambda})\widetilde{V}'(\lambda) > \widetilde{\rho}'(\lambda)\widetilde{V}'(\tilde{\lambda})$ and, consequently,

$$\widetilde{V}'(\lambda)\rho(\lambda) = \int_0^\lambda \widetilde{V}'(\lambda)\widetilde{\rho}'(\tilde{\lambda})d\tilde{\lambda} > \int_0^\lambda \widetilde{\rho}'(\lambda)\widetilde{V}'(\tilde{\lambda})d\tilde{\lambda} = \widetilde{\rho}'(\lambda)\widetilde{V}(\lambda)$$

which gives (A.3). All inequalities are replaced with equalities if the generalize failure rate g is constant. ■

A.2. Proof of Theorem 1

We use standard sequence notation collected in the following definition.

Definition 1 (scaling comparisons)**.** *Given a non-negative function g with $g(n) \to \infty$ as $n \to \infty$ we write, $\xi(n) = o(g(n))$ if*

$$\lim_{n\to\infty} \frac{\xi(n)}{g(n)} = 0.$$

We write $f(n) = \mathcal{O}(g(n))$ *if*

$$\limsup_{n\to\infty} \frac{|\xi(n)|}{g(n)} < \infty.$$

Finally, $\xi(n) = \Omega(g(n))$ *is the negation of* $\xi(n) = o(g(n))$*, i.e.,*

$$\xi(n) = \Omega(g(n)) \Leftrightarrow \liminf_{n\to\infty} \frac{|\xi(n)|}{g(n)} > 0.$$

The characterization of the optimal actions of the social planner and the revenue maximizer in Theorem 4 and 5 is used to prove Theorem 1. The proofs of these theorems appear then in §A.3

Theorem 4. (**optimal decisions of the SP**) *The cutoffs and admission rates*

$$\hat{v}_{i,s}^n = \bar{v} + \varphi^{\frac{i-1}{K}} \theta^{\frac{K-i+1}{K}} n^{-\frac{K-i+1}{2K}}, \text{ and } \hat{\bar{\lambda}}_{i,S}^n = n - \Lambda f(\bar{v}) \varphi^{\frac{i-1}{K}} \theta^{\frac{K-i+1}{K}} n^{\frac{K+i-1}{2K}}$$

where

$$\theta = \frac{\bar{F}(\bar{v})}{f(\bar{v})}\sqrt{\alpha}, \varphi = 2\frac{V(\bar{v}) - \rho(\bar{v})}{\bar{F}(\bar{v})} = 2MRL(\bar{v}),$$

are nearly optimal in the sense that $S_K^{n*} - S_K^n(\hat{v}_S^n) = o(\sqrt{n})$*.*

The welfare maximizing decisions **must** *be at most small perturbation of* $\hat{v}_S^n$ *and* $\hat{\lambda}_S^n$*. That is,*

$$v_{i,S}^{n*} = \hat{v}_{i,S}^n + o(n^{-\frac{K-i+1}{2K}}),\ i = 2, \ldots, K, \text{ and } \bar{\lambda}_{i,S}^{n*} = \hat{\lambda}_{i,S}^n + o(n^{\frac{K+i-1}{2K}}), \tag{A.4}$$

Finally, increasing the number of classes beyond 2 *can increase social welfare by at most* $o(\sqrt{n})$*:*

$$\Lambda n V(\bar{v}) - 2\bar{v}\sqrt{\alpha}\sqrt{n} + o(\sqrt{n}), \text{ for any } K \geq 2.$$

The following is a strengthened version of Theorem 3 in Nazerzadeh and Randhawa (2015) where we add, to their result, that the optimal actions are asymptotically unique. This is crucial for the comparison of the social-planner and revenue-maximizer actions.

Theorem 5. (**optimal decisions of the RM**) *The cutoffs and arrival rates*

$$\hat{v}_{i,R}^n = \bar{v} + \Phi^{\frac{i-1}{K}}\theta^{\frac{K-i+1}{K}}n^{-\frac{K-i+1}{2K}} + o(n^{-\frac{K-i+1}{2K}}),\ i = 2,\ldots,K,\ \textit{and}\ \hat{\bar{\lambda}}_{i,R}^{n*} = n - \Lambda f\left(\bar{v}\right)\Phi^{\frac{i-1}{K}}\theta^{\frac{K-i+1}{K}}n^{\frac{K+i-1}{2K}}$$

where

$$\theta = \frac{\bar{F}\left(\bar{v}\right)}{f\left(\bar{v}\right)}\sqrt{\alpha},\ \Phi = 2\left(\frac{f\left(\bar{v}\right)\bar{F}\left(\bar{v}\right)}{f'\left(\bar{v}\right)\bar{F}\left(\bar{v}\right) + 2\left(f\left(\bar{v}\right)\right)^2}\right) = -\frac{2}{\Lambda\widetilde{\rho}''(1)}\frac{\bar{F}(\bar{v})}{f(\bar{v})^2},$$

are nearly optimal in the sense that $R_K^{n*} - R_K^n(\hat{v}_R^n) = o(\sqrt{n})$.

The welfare maximizing decisions **must** *be at most small perturbation of* $\hat{v}_S^n$ *and* $\hat{\lambda}_S^n$*. That is,*

$$v_{i,R}^{n*} = \hat{v}_{i,R}^n + o(n^{-\frac{K-i+1}{2K}}),\ i = 2,\ldots,K,\ \textit{and}\ \bar{\lambda}_{i,R}^{n*} = \hat{\lambda}_{i,R}^n + o(n^{\frac{K+i-1}{2K}}). \tag{A.5}$$

Finally, increasing the number of classes beyond 2 *can increase revenue by at most* $o(\sqrt{n})$*:*

$$R_K^{n*} = \Lambda n\rho(\bar{v}) - 2\bar{v}\sqrt{\alpha}\sqrt{n} + o\left(\sqrt{n}\right),\ \textit{for any}\ K \geq 2.$$

In the statement of this theorem $\theta \geq 0$. Also, since $\widetilde{\rho}(\lambda)$ is concave for $\lambda < \lambda_0^*$ (see Lemma 2) and, in particular, for $\lambda = 1 = \Lambda\bar{F}(\bar{v}) < \Lambda\bar{F}(v_0^*)$, we have that $\widetilde{\rho}''(1) \geq 0$ so that $\Phi \geq 0$.

Theorems 4 and 5 provide the basis for Theorem 1: the constant γ in Theorem 1 stands for $\varphi - \Phi$. With $K = 2$, if $\gamma = \varphi - \Phi > 0$, $\bar{\lambda}_{2,R}^{n*} > \bar{\lambda}_{2,S}^{n*}$ so that the revenue maximizer has a larger high priority class. The following lemma studies, then, $\gamma = \varphi - \Phi$.

Lemma 6. *Suppose that F has strictly positive density on its support and that $m(\cdot)$ is a convex (respectively concave, linear respectively) MRL. Then,*

$$MRL(x)\left(\frac{h'(x)}{h(x)} + h(x)\right) \geq (\leq, = \ respectively)\ 1 \tag{A.6}$$

for any x in the support of F.

Recalling the definition of φ and Φ, we must show is that

$$\varphi = \frac{\left(V(\bar{v}) - \bar{v}\bar{F}(\bar{v})\right)}{\bar{F}(\bar{v})} \geq \frac{f(\bar{v})\bar{F}(\bar{v})}{f'(\bar{v})\bar{F}(\bar{v}) + 2\left(f(\bar{v})\right)^2} = \Phi.$$

and that the opposite holds for concave MRL. The left hand side of the inequality is precisely the MRL of F at the point $\bar{v}$ so that this inequality is equivalent to

$$MRL(\bar{v})\left(\frac{h'(\bar{v})}{h(\bar{v})} + h(\bar{v})\right) \geq 1, \tag{A.7}$$

which follows, with convex MRL, from Lemma 6. ■

Proof of Theorem 2: By definition, the optimal objective function value when policies are restricted to **non-preemption** is smaller than the optimal value under the larger family of pre-emptive policies. That is, $S_{K,NP}^{n*} \leq S_K^{n*}$. Then, we have

$$S_{K,NP}^n(\hat{v}_S^n) \leq S_{K,NP}^{n*} \leq S_K^{n*} = S_2^{n*} + o(\sqrt{n}) = S_2^n(\hat{v}_S^n) + o(\sqrt{n}),$$

where $\hat{v}^n_{S,i} = \bar{v} + \varphi^{\frac{i-1}{2}}\theta^{\frac{2-i+1}{2}}n^{-\frac{3-i}{4}}$ for $i = 1, 2$. The last two equalities follow from Theorem 4. This holds for all K. Corresponding arrival rates for $\hat{v}^n_{S,i}$ are

$$\hat{\lambda}^n_{S,i} = n - C_{S,i}n^{\frac{K+i-1}{2K}},$$

where $C_{S,i} = f(\bar{v})\varphi^{\frac{i-1}{2}}\theta^{\frac{2-i+1}{2}}$. Similarly for revenue maximization

$$R^n_{K,NP}(\hat{v}^n_R) \leq R^{n*}_{K,NP} \leq R^{n*}_K = R^{n*}_2 + o(\sqrt{n}) = R^n_2(\hat{v}^n_R) + o(\sqrt{n}).$$

where $\hat{v}^n_{R,i} = \bar{v} + E_i n^{-\frac{3-i}{4}}$ with $E_i = \Phi^{\frac{i-1}{K}}\theta^{\frac{K-i+1}{K}}$ for $i = 1, 2$. Corresponding cumulative arrival rates for $\hat{v}^n_{R,i}$ are

$$\hat{\lambda}^n_{R,1} = n - C_{R,i}n^{\frac{K+i-1}{2K}}.$$

where $C_{R,i} = f(\bar{v})\Phi^{\frac{i-1}{2}}\theta^{\frac{3-i}{2}}$. It then suffices to prove that

$$R^n_2(\hat{v}^n_R) - R^n_{2,NP}(\hat{v}^n_R) = o(\sqrt{n}), \text{ and } S^n_2(\hat{v}^n_S) - S^n_{2,NP}(\hat{v}^n_S) = o(\sqrt{n}).$$

Let us start with the revenue maximizer. Notice that

$$R^n_{2,NP}(v^{n*}_R) - R^n_2(v^{n*}_R) = \Lambda n\Delta R_2(W^{NP}_{R,H} - W^P_{R,H}) + \Lambda n\Delta R_1(W^{NP}_{R,L} - W^P_{R,L}),$$

where $\Delta R_2 = \hat{\lambda}_{R,2}\hat{v}^n_{R,2}$ and $\Delta R_1 = \hat{\lambda}_{R,1}\hat{v}^n_{R,1} - \hat{\lambda}_{R,2}\hat{v}^n_{R,2}$. With preemption, the steady-state sojourn times satisfy

$$\text{(A.8)} \quad W^P_{R,L} = \frac{n}{(n-\hat{\lambda}^n_{R,1})(n-\hat{\lambda}^n_{R,2})} = \frac{n}{C_{R,1}n_1^{1/2}C_{R,2}n^{3/4}}, W^P_{R,H} = \frac{1}{(n-\hat{\lambda}^n_{R,2})} = \frac{1}{C_{R,2}n^{3/4}},$$

where the subscripts H and L are for high and low priority respectively (recall that class 2 is the high priority) and t stands for either S or R. Similarly under non-preemption

$$W_{R,L}^{NP} = \frac{\hat{\lambda}_{R,1}^n}{(n-\hat{\lambda}_{R,1})(n-\hat{\lambda}_{R,2})} + \frac{1}{n} = \frac{n-C_{R,1}n^{1/2}}{C_{R,1}n_1^{1/2}C_{R,2}n^{3/4}} + \frac{1}{n},\ W_{R,H}^{NP} = \frac{\hat{\lambda}_{R,1}^n}{n(n-\hat{\lambda}_{R,2}^n)} + \frac{1}{n} = \frac{n-C_{R,1}n^{1/2}}{nC_{R,2}n^{3/4}} + \frac{1}{n}.$$

Hence

$$W_{R,L}^{NP} - W_{R,L}^{P} = n^{-1} - C_{R,2}^{-1}n^{-3/4}, \text{ and } W_{R,H}^{NP} - W_{R,H}^{P} = n^{-1} - C_{R,1}C_{R,2}^{-1}n^{-5/4}.$$

Then,

$$\Lambda n \Delta R_2 (W_{R,H}^{NP} - W_{R,H}^{P}) = \hat{\lambda}_{R,2}\hat{v}_{R,2}(n^{-1} - C_{R,1}C_{R,2}^{-1}n^{-5/4}) = \mathcal{O}(1) = o(\sqrt{n}),$$

and

$$\Lambda n \Delta R_1 (W_{R,L}^{NP} - W_{R,L}^{P}) = (\hat{\lambda}_{R,1}\hat{v}_{R,1} - \hat{\lambda}_{R,2}\hat{v}_{R,2})(n^{-1} - C_{R,1}C_{R,2}^{-1}n^{-5/4}) = \mathcal{O}(1) = o(\sqrt{n}),$$

so that

$$\begin{aligned} R_{2,NP}^n(v_R^{n*}) - R_2^n(v_R^{n*}) &= \Lambda n \Delta R_2 (W_{R,H}^{NP} - W_{R,H}^{P}) + \Lambda n \Delta R_1 (W_{R,L}^{NP} - W_{R,L}^{P}) \\ &= O(1) = o(\sqrt{n}). \end{aligned}$$

For the social planner,

$$S_2^n(\hat{v}_S^n) - S_{2,NP}^n(\hat{v}_S^n) = \Lambda n V(\hat{v}_{S,2}^n)(W_{S,H}^{NP} - W_{S,H}^{P}) + \Lambda n (V(\hat{v}_{S,1}^n) - V(\hat{v}_{S,2}^n))(W_{S,L}^{NP} - W_{S,L}^{P}),$$

where the expressions for the waiting times are the same as for the revenue maximizer with the obvious replacements of R with S everywhere. Using Taylor expansion on $V(\cdot)$ at $\bar{v}$ and that $\hat{v}^n_{S,i} = \bar{v} + D_i n^{-\frac{3-i}{4}}$, we have

$$\begin{aligned}\Lambda n V(\hat{v}^n_{S,2})(W^{NP}_{S,H} - W^{P}_{S,H}) &= \Lambda n(V(\bar{v}) - \bar{v}f(\bar{v})D_2 n^{-1/4} + \mathcal{O}(n^{-1/2}))(n^{-1} - C_{S,1}C^{-1}_{S,2}n^{-5/4}) \\ &= \mathcal{O}(1) = o(\sqrt{n}),\end{aligned}$$

and

$$\begin{aligned}\Lambda n(V(\hat{v}^n_{S,1}) - V(\hat{v}^n_{S,2}))(W^{NP}_{S,L} - W^{P}_{S,L}) &= \Lambda n(\bar{v}f(\bar{v})D_2 n^{-1/4} + \mathcal{O}(n^{-1/2}))(n^{-1} - C_{S,1}C^{-1}_{S,2}n^{-5/4}) \\ &= \mathcal{O}(n^{-1/4}) = o(\sqrt{n}),\end{aligned}$$

so that $S^n_2(\hat{v}^n_S) - S^n_{2,NP}(\hat{v}^n_S) = \mathcal{O}(1) = o(\sqrt{n})$, as stated.

In passing, it is worthwhile noticing the subtlety in the argument above. It builds on the fact that, under the optimal preemptive actions, the high-priority volume is order-of-magnitude larger than that of the low priority. The latter's is of the order of $n^{3/4}$; see equation (A.8). ■

A.3. Proofs of Theorems 4 and 5

We consider a sequence of queues indexed by the service rate n. The nominal arrival rate in the n^{th} queue is Λn.

A perturbation formulation:

We express the cutoffs as deviations from $\bar{v} \equiv \bar{F}^{-1}(1/\Lambda)$: $v_i = \bar{v} + u_i$ or, in vector notation, $v = \bar{v}\mathbf{e} + u$. As no customer with valuation smaller than $\bar{v} \equiv \bar{F}^{-1}(1/\Lambda)$ joins the queue u is a non-negative vector. Let $W^n_i(u)$ be the expected waiting time of class i under preemptive static

priority under the cutoff vector $\bar{v}\mathbf{e}+u$. The social planner's problem (1.4) with nominal arrival rate Λn and service rate n is re-written as

$$\textbf{(SP}_n\textbf{)}\quad \mathrm{S}_K^{n*}=\max_{u\uparrow} S_K^n(u):=\max_{v\uparrow}\Lambda n\left[V(\bar{v}+u_1)-\alpha\sum\nolimits_{i=1}^{K}\left(V\left(\bar{v}+u_i\right)-V\left(\bar{v}+u_{i+1}\right)\right)W_i^n(\bar{v}\mathbf{e}+u)\right],$$

and that for the revenue maximizer as

$$\textbf{(RM}_n\textbf{)}\quad \mathrm{R}_K^{n*}=\max_{u\uparrow} R_K^n(u):=\max_{u\uparrow}\Lambda n\left[\rho(\bar{v}+u_1)-\alpha\sum\nolimits_{i=1}^{K}(\rho(\bar{v}+u_1)-\rho(\bar{v}+u_{i+1}))W_i^n(\bar{v}\mathbf{e}+u)\right].$$

Given optimal solutions $u_{i,S}^{n*}$ and $u_{i,R}^{n*}$ for (**SP**$_n$) (respectively (**RM**$_n$)), the optimal cutoffs are given by $v_{i,S}^{n*}=\bar{v}+u_{i,S}^{n*}$ (respectively $v_{i,R}^{n*}=\bar{v}\mathbf{e}+u_{i,R}^{n*}$).

We first state several auxiliary lemmas, the proofs of which appear at the end of this companion. The first of these, analogous to Lemma 2 in Nazerzadeh and Randhawa (2015), shows that the optimal cut-offs v_R^{n*} and v_S^{n*} are clustered around $\bar{v}$ when the volume is high.

Lemma 7. *For each n, there exist optimal solutions v_S^{n*} and v_R^{n*} for* **SP**$_n$ *and* **RM**$_n$ *respectively. Let $\{(v_R^{n*},v_S^{n*});n=1,2,\ldots\}$ be a sequence of optimal solutions. Then, $v_R^{n*}\to\bar{v}$ and $v_S^{n*}\to\bar{v}$ as $n\to\infty$.*

That all "good" decisions must be small perturbations around $\bar{v}$ means that Taylor expansion should be useful in uncovering these perturbations.

Lemma 8. *Fix a sequence of $u^n=o(1)$ of cutoff values. Then,*

$$\text{(A.9)}\quad S_K^n(u^n)=n\Lambda V\left(\bar{v}\right)+\left(n\Lambda V'\left(\bar{v}\right)u_1^n+\tfrac{\alpha V'(\bar{v})}{\Lambda(f(\bar{v}))^2}\tfrac{1}{u_1^n}\right)-\left(\tfrac{1}{u_K^n}\beta\left(\bar{v}\right)+\gamma\left(\bar{v}\right)\sum\nolimits_{i=1}^{K-1}\frac{u_{i+1}^n}{u_i^n}\right)+\epsilon^n,$$

where

$$\beta(\bar{v}) := \frac{\alpha V(\bar{v})}{f(\bar{v})} - \frac{\alpha \bar{v}}{\Lambda f(\bar{v})},\ \gamma(\bar{v}) := \frac{\alpha \bar{F}(\bar{v})}{2f(\bar{v})},$$

and

$$\epsilon^n := \mathcal{O}\left(\sum\nolimits_{i=1}^{K-1} \frac{\left(u_{i+1}^n\right)^2}{u_i^n}\right) + \mathcal{O}(n(u_1^n)^2) + \mathcal{O}(1).$$

Let u_i^{n*} be the optimal solution to the social planner's problem (**SP**$_n$). There are two parts to this proof.

(Step 1) $u_i^{n*} = \hat{u}_i^n + o(n^{-\frac{K-i+1}{2K}})$ for $1 \leq i \leq K$ where $\hat{u}_i^n = \varphi^{\frac{i-1}{K}} \theta^{\frac{K-i+1}{K}} n^{-\frac{K-i+1}{2K}}$.

(Step 2) $S_K^{n*} - S_2^{n*} = o(\sqrt{n})$.

(Proof of Step 1) Re-write

$$S_K^n(u^n) = n\Lambda V(\bar{v}) + M(u_1^n) + B(u^n) + E(u^n), \tag{A.10}$$

where,

$$B(u^n) = -\left(\frac{1}{u_K^n}\beta(\bar{v}) + \gamma(\bar{v}) \sum\nolimits_{i=1}^{K} \frac{u_{i+1}^n}{u_i^n}\right),\ M(u_1^n) = n\Lambda V'(\bar{v})\, u_1^n + \frac{\alpha V'(\bar{v})}{\Lambda\left(f(\bar{v})\right)^2} \frac{1}{u_1^n}, \tag{A.11}$$

and, using Lemma 8,

$$E(u^n) = \sum\nolimits_{i=1}^{K-1} \mathcal{O}\left(\frac{\left(u_{i+1}^n\right)^2}{u_i^n}\right) + \mathcal{O}(n(u_1^n)^2) + \mathcal{O}(1). \tag{A.12}$$

Suppose that u^n is a sequence that has $\epsilon_{i,n} := |u_i^n - \hat{u}_i^n| = \Omega\left(n^{-\frac{K-i+1}{2K}}\right)$ for some i. Then, we will show that it must be sub-optimal.

Step 1.1. First, consider the case of $i = 1$, namely, that $\epsilon_{1,n} := |u_1^n - \hat{u}_1^n| = \Omega(n^{-1/2})$. We will show that, in this case

$$M(\hat{u_1^n}) - M(u_1^n) = \Omega(n\epsilon_{1,n}) = \Omega(\sqrt{n}), \tag{A.13}$$

but

$$B(\hat{u}^n) - B(u^n) = o(\sqrt{n}), \text{ and } E(\hat{u}^n) - E(u^n) = o(n\epsilon_{1,n}), \tag{A.14}$$

so that, overall

$$S_K(\hat{u}^n) - S_K^n(u^n) = \Omega(n\epsilon_{1,n}) = \Omega(\sqrt{n}). \tag{A.15}$$

In particular, u^n is sub-optimal for all n sufficiently large.

To prove (A.13), notice that the function $M(\cdot)$ is maximized by

$$\hat{u}_1^n = \frac{1}{\Lambda f(\bar{v})} n^{-\frac{1}{2}} \sqrt{\alpha} = \frac{\bar{F}(\bar{v})}{f(\bar{v})} n^{-\frac{1}{2}} \sqrt{\alpha} = \theta n^{-\frac{1}{2}},$$

where we used the fact that $\Lambda \bar{F}(\bar{v}) = 1$ and, recall, $\theta = \frac{\bar{F}(\bar{v})}{f(\bar{v})}\sqrt{\alpha}$. By Lemma 7 we can assume, without loss of generality, that $\epsilon_{i,n} = o(1)$. By definition

$$\begin{aligned}
\frac{M(\hat{u}_1^n) - M(\hat{u}_1^n + \epsilon_{1,n})}{n\epsilon_{1,n}} &= -\Lambda V'(\bar{v}) + \frac{\alpha V'(\bar{v})}{\Lambda (f(\bar{v}))^2} \frac{1}{n\hat{u}_1^n (\hat{u}_1^n + \epsilon_{1,n})} \\
&= -\Lambda V'(\bar{v}) + \frac{\alpha V'(\bar{v})}{\Lambda (f(\bar{v}))^2} \frac{\sqrt{n}}{n\theta \left(\frac{\theta}{\sqrt{n}} + \epsilon_{1,n}\right)} \\
&= V'(\bar{v}) \left(-\frac{1}{\bar{F}(\bar{v})} + \frac{\bar{F}(\bar{v})\alpha}{(f(\bar{v}))^2} \frac{1}{\left(\frac{\bar{F}(\bar{v})}{f(\bar{v})}\right)^2 \alpha + \theta\sqrt{n}\epsilon_{1,n}} \right),
\end{aligned}$$

where we use again the definition of θ and fact that $\Lambda = \frac{1}{\bar{F}(\bar{v})}$. Further simplification gives

$$\begin{aligned} \frac{M\left(\hat{u}_1^{1,n}\right) - M\left(\hat{u}_1^{1,n} + \epsilon_{1,n}\right)}{n\epsilon_{1,n}} &= \frac{V'\left(\bar{v}\right)}{\bar{F}\left(\bar{v}\right)} \left(-1 + \left(\frac{\bar{F}\left(\bar{v}\right)}{f\left(\bar{v}\right)} \right)^2 \alpha \frac{1}{\left(\frac{\bar{F}(\bar{v})}{f(\bar{v})} \right)^2 \alpha + \theta\sqrt{n}\epsilon_{1,n}} \right) \\ &= \frac{V'\left(\bar{v}\right)}{\bar{F}\left(\bar{v}\right)} \left(-1 + \frac{1}{1 + \frac{\sqrt{n}\epsilon_{1,n}}{\theta}} \right). \end{aligned} \tag{A.16}$$

Since $V'\left(\bar{v}\right) < 0$, $\bar{F}\left(\bar{v}\right) = \frac{1}{\Lambda} > 0$ and $\theta > 0$, we have the relation

$$\liminf_{n\to\infty} \frac{M\left(\hat{u}_1^n\right) - M\left(\hat{u}_1^n + \epsilon_{1,n}\right)}{n\epsilon_{1,n}} > 0 \Leftrightarrow \liminf_{n\to\infty} \sqrt{n}\epsilon_{1,n} > 0$$

Since, $\epsilon_{1,n} = \Omega\left(n^{-\frac{1}{2}}\right)$, we have $\liminf_{n\to\infty} \sqrt{n}\epsilon_{1,n} > 0$ and, in particular, that

$$\liminf_{n\to\infty} \frac{M\left(\hat{u}_1^n\right) - M\left(\hat{u}_1^n + \epsilon_{1,n}\right)}{n\epsilon_{1,n}} > 0,$$

equivalently,

$$M\left(\hat{u}_1^n\right) - M\left(\hat{u}_1^n + \epsilon_{1,n}\right) = \Omega(n\epsilon_{1,n}).$$

Since $\hat{u}_1^n$ is the maximizer of $M(\cdot)$ we have (A.13).

We next prove (A.14) starting with $B(\cdot)$. Since $|B(\hat{u}^n) - B(u^n)| \leq B(\hat{u}) + B(u^n)$, it suffices to prove that $B(\hat{u}^n) = o(\sqrt{n})$ and $B(u^n) = o(\sqrt{n})$. Since $\hat{u}_{i+1}^n = o(1)$ for all i and since cut-offs satisfy $\hat{u}_{i+1} \geq \hat{u}_i$, it suffices to have $\frac{1}{\hat{u}_1^n} = \mathcal{O}(\sqrt{n})$ to conclude that $\frac{\hat{u}_{i+1}}{\hat{u}_i} = o(\sqrt{n})$ for all i and, in turn, that $B(\hat{u}^n) = o(\sqrt{n})$. Similarly, it suffices to prove that $\frac{1}{\hat{u}_1^n + \epsilon_{1,n}} = \mathcal{O}(\sqrt{n})$ to have $B(u^n) = o(\sqrt{n})$.

First, by definition, $\hat{u}_1^n = \theta n^{-1/2}$ so that $1/\hat{u}_1^n = \mathcal{O}(\sqrt{n})$. Because $\theta\sqrt{n}\epsilon_{1,n} = \Omega(1)$ we have that

$$\limsup_{n\to\infty} \frac{1}{\sqrt{n}\hat{u}_1^n} = \limsup_{n\to\infty} \frac{1}{\theta\sqrt{n}\epsilon_{1,n}} < \infty.$$

Since $\sqrt{n}\epsilon_{1,n} = \Omega(1)$ by assumption, we also have that $\frac{1}{\sqrt{n}\hat{u}_1^n + \sqrt{n}\epsilon_{1,n}} = \mathcal{O}(1)$ and, in turn, that $\frac{1}{\hat{u}_1^n + \epsilon_{1,n}} = \mathcal{O}(\sqrt{n})$. We turn to $E(\cdot)$. Since $|E(\hat{u}^n) - E(u^n)| \leq E(\hat{u}^n) + E(u^n)$, it again suffices to show that $E(\hat{u}) = o(\sqrt{n})$ and $E(u^n) = o(\sqrt{n})$. Notice that

$$\text{(A.17)} \qquad \begin{aligned} E(\hat{u}^n) &= \sum\nolimits_{i=1}^{K} \mathcal{O}\left(\frac{\left(\hat{u}_{i+1}^n\right)^2}{\hat{u}_i^n}\right) + \mathcal{O}(n\,(\hat{u}_1^n)^2)) \\ E(u^n)) &= \sum\nolimits_{i=1}^{K} \mathcal{O}\left(\frac{\left(\hat{u}_{i+1}^n + \epsilon_{i+1,n}\right)^2}{\hat{u}_i^n + \epsilon_{i,n}}\right) + \mathcal{O}(n\,(\hat{u}_1^n + \epsilon_{1,n})^2) \end{aligned}$$

Since $n(\hat{u}_1^n)^2 = n\theta^2 n^{-1} = \theta = \mathcal{O}(1)$ and $n\hat{u}_1^n\epsilon_n = \theta\sqrt{n}\epsilon_n$ we have that $\mathcal{O}((n\hat{u}_1^n)^2) = \mathcal{O}(1) = o(\sqrt{n})$ and

$$\begin{aligned} \mathcal{O}(n\,(\hat{u}_1^n + \epsilon_{1,n})^2) &= \mathcal{O}(n(\hat{u}_1^n)^2 + 2n\hat{u}_1^n\epsilon_{1,n} + n\epsilon_{1,n}^2) = \mathcal{O}(1 + \sqrt{n}\epsilon_{1,n} + n\epsilon_{1,n}^2) = \mathcal{O}\left(n\epsilon_{1,n}^2\right) \\ &= o(n\epsilon_{1,n}). \end{aligned}$$

The second to last equality follows since $\epsilon_{1,n} = \Omega\left(n^{-\frac{1}{2}}\right)$, $\sqrt{n}\epsilon_{1,n} = \mathcal{O}(n\epsilon_{1,n}^2)$ and the last equality follows since $\epsilon_{1,n} = o(1)$. To take care of the other terms of $E(\hat{u}^n)$ and $E(u^n)$ notice that, since $u^n = o(1)$ and $\epsilon_{i,n} = o(1)$ for all i,

$$\mathcal{O}\left(\frac{\left(\hat{u}_{i+1}^n + \epsilon_{i,n}\right)^2}{\hat{u}_i^n + \epsilon_{i,n}}\right) = o\left(\frac{\hat{u}_{i+1}^n + \epsilon_{i+1,n}}{\hat{u}_i^n + \epsilon_{i,n}}\right) \text{ and } \mathcal{O}\left(\frac{\left(\hat{u}_{i+1}^n\right)^2}{\hat{u}_i^n}\right) = o\left(\frac{\hat{u}_{i+1}^n}{\hat{u}_i^n}\right).$$

It is therefore sufficient to show that

$$o\left(\frac{\hat{u}^n_{i+1}+\epsilon_{i+1,n}}{\hat{u}^n_i+\epsilon_{i,n}}\right)=o(\sqrt{n}),\text{ and } o\left(\frac{\hat{u}^n_{i+1}}{\hat{u}^n_i}\right)=o(\sqrt{n}).$$

Moreover, since $\hat{u}^n_{i+1}+\epsilon_{i+1,n}=o(1)$ and $\hat{u}^n_{i+1}=o(1)$, it is sufficient to show that $\frac{1}{\hat{u}^n_i+\epsilon_{i,n}}=\mathcal{O}(\sqrt{n})$ and $\frac{1}{\hat{u}^n_i}=\mathcal{O}(\sqrt{n})$. We also know that $\frac{1}{\hat{u}^n_i}<\frac{1}{\hat{u}^n_1}$ and $\frac{1}{\hat{u}^n_i+\epsilon_{i,n}}<\frac{1}{\hat{u}^n_1+\epsilon_{1,n}}$. We already showed that $\frac{1}{\hat{u}^n_1}=\mathcal{O}(\sqrt{n})$ and $\frac{1}{\hat{u}^n_1+\epsilon_{1,n}}=\mathcal{O}(\sqrt{n})$.

This completes the proof of (A.13) and (A.14) and hence of (A.15). We reached a contradiction to $|u^n_1-\hat{u}^n_1|=\Omega(1/\sqrt{n})$. Notice that we can repeat the above for any subsequence. We may thus conclude that $|u^n_1-\hat{u}^n_1|=o(1/\sqrt{n})$.

Step 1.2. We proved that any optimal sequence u^n must satisfy that $u^n_1=\hat{u}^n_1+o(n^{-1/2})$. We turn to prove that this, in turn, implies that any such sequence must have $u^n_i=\hat{u}^n_i+o(n^{-\frac{K-i+1}{2K}})$ for $i=2,\ldots,K$. We will use the following lemma where, given a vector $u\in\mathbb{R}^K_+$ and $j<K$, we write $u_{[j]}=(u_1,\ldots,u_j)$ and $u_{-[j]}=(u_{j+1},\ldots,u_K)$.

Lemma 9. *Given $u_1,\ldots,u_j$ for some $j<K$,*

$$u_i=f_{i,j}(u_j):=\varphi^{\frac{i-j}{K-j+1}}(u_j)^{\frac{K-i+1}{K-j+1}},\ i=j+1,\ldots,K,$$

is the unique solution to

$$\max_{0<u_{j+1}<\cdots<u_K} B(u_{-[j]};u_{[j]}).$$

We define $f_j(u_j)$ to be the vector $f_{i,j}(u_j)$ for $i = j+1, \ldots, K$. Notice that in the special case that $[j] = \{1\}$, we have

$$u_i = \varphi^{\frac{i-1}{K}}(u_1)^{\frac{K-i+1}{K}},\ i = 2, \ldots, K.$$

Fix a sequence of cutoffs $\tilde{u}^n$ such that $\tilde{u}_1^n = \hat{u}_1^n + o(1/\sqrt{n})$ and such that $\tilde{u}_k^n$ for $k = 1, \ldots, K$ are determined by Lemma 9 with $j = 2$ there. In particular, notice, $\tilde{u}_i^n = \varphi^{1/K}(\tilde{u}_1^n)^{\frac{K-i+1}{K}} = \hat{u}_i^n + o\left(n^{-\frac{K-i+1}{2K}}\right)$ and our goal is to prove that $u_i^{n*} - \tilde{u}_i^n = o\left(n^{-\frac{K-i+1}{2K}}\right)$. Let $u^n = o(1)$ be a sequence of cutoffs where $u_1^n = \hat{u}_1^n + o(n^{-1/2})$. We will show first that if $|u_2^n - \hat{u}_2^n| = \Omega\left(n^{-\frac{K-1}{2K}}\right)$ then u_2^n must be sub-optimal; specifically that

$$S_K^n(\tilde{u}^n) - S_K^n(u^n) = \Omega(n^{\frac{1}{2K}}) > 0.$$

Recall that $S_K^n(u) = M(u^n) + B(u^n) + E(u^n)$ where M, B and E are as defined in (A.11) and (A.12). First, because both $\tilde{u}_1^n = \hat{u}_1^n + o(n^{-1/2})$ and $u_1^n = \hat{u}_1^n + o(n^{-1/2})$ and, because $u^n = o(1)$, we have that $M(\tilde{u}_1^n) - M(u_1^n) = \mathcal{O}(1)$, $B(u_{-1}^n, u_1^n) - B(u_{-1}^n, \tilde{u}_1^n) = \mathcal{O}(1)$ and $E(u_{-1}^n, u_1^n) - E(u_{-1}^n, \tilde{u}_1^n) = \mathcal{O}(1)$. It suffices, then to consider sequence u^n with $u_1^n = \tilde{u}_1^n$. In that case,

$$S_K^n(\tilde{u}^n) - S_K^n(\tilde{u}_1^n, u_{-1}^n) = B(\tilde{u}_{-1}^n; \tilde{u}_1^n) - B(u_{-1}^n, \tilde{u}_1^n) + E(\tilde{u}_{-1}^n; \tilde{u}_1^n) - E(u_{-1}^n, \tilde{u}_1^n).$$

Since cutoff vectors are increasing ($u_{i+1}^n > u_i^n; i = 1, \ldots, K-1$), we have that $E(u^{n*}) = O(1)$. Since we are considering in this step the case that $u_1^n = \hat{u}_1^n + o(n^{-1/2})$, we also have that

$n\left(u_1^n\right)^2 = \mathcal{O}(1)$. Thus, for any sequence of cutoff u^n that has $u_1^n = \hat{u}_1^n + o(n^{-1/2})$:

$$E(u^n) = E\left(u^n\right) = \sum\nolimits_{i=1}^{K} \mathcal{O}\left(\frac{\left(u_{i+1}^n\right)^2}{u_i^n}\right) + \mathcal{O}(n\left(u_1^n\right)^2) + \mathcal{O}(1)$$

$$= o\left(\sum\nolimits_{i=1}^{K-1} \frac{u_{i+1}^n}{u_i^n}\right) + \mathcal{O}(1) = o(-B(u^n)) + \mathcal{O}(1). \tag{A.18}$$

By definition of $\tilde{u}^n$ we have $0 \geq B(\tilde{u}_{-1}^n); \tilde{u}_1^n) \geq B(u_{-1}^n; \tilde{u}_1^n)$ so that we further have

$$\begin{aligned} S_K^n(\tilde{u}^n) - S_K^n(\tilde{u}_1^n, u_{-1}^n) &= B(f_1^n; \tilde{u}_1^n) - B(u_{-1}^n, \tilde{u}_1^n) + E(\tilde{u}_{-1}^n; \tilde{u}_1^n) - E(u_{-1}^n, \tilde{u}_1^n) \\ &> B(\tilde{u}_{-1}^n; \tilde{u}_1^n) - B(u_{-1}^n, \tilde{u}_1^n) + \epsilon B((u_{-1}^n; \tilde{u}_1^n) + O(1) \\ &\geq B(\tilde{u}_{-1}^n; \tilde{u}_1^n) - (1-\epsilon)\tilde{B}(u_{-1}^n, \tilde{u}_1^n) + \mathcal{O}(1), \end{aligned}$$

where can be taken to be an arbitrarily small strictly positive constant.

By definition $0 \geq B(f_2^n(u_2^n); u_{[2]}^n) \geq B(u_{-1}^n; \tilde{u}_1^n)$, so that, further

$$S_K^n(\tilde{u}^n) - S_K^n(\tilde{u}_1^n, u_{-1}^n) \geq B(\tilde{u}_{-1}^n; \tilde{u}_1^n) - (1-\epsilon)B(f_2^n(u_2^n); u_{[2]}^n) + \mathcal{O}(1), \tag{A.19}$$

By lemma 9 with $j = 2$, we have

$$\begin{aligned} B(f_2(u_2^n), u_{[2]}^n) &= -\gamma\left(\bar{v}\right)\frac{u_2^n}{\tilde{u}_1^n} + \varphi^{\frac{1}{K-1}}\gamma\left(\bar{v}\right)(K-2)(u_2^n)^{\frac{-1}{K-1}} + \varphi^{\frac{1}{K-1}}\frac{\beta\left(\bar{v}\right)}{\varphi}(u_2^n)^{\frac{-1}{K-1}} \\ &= -\gamma\left(\bar{v}\right)\left(\frac{u_2^n}{\tilde{u}_1^n} + (K-1)\varphi^{\frac{1}{K-1}}(u_2^n)^{\frac{-1}{K-1}}\right) \\ &= -\gamma(\bar{v})\left(\frac{1}{\tilde{u}_1^n}\right)^{\frac{1}{K}}\left(\zeta^n + (K-1)\varphi^{\frac{1}{K-1}}\left(\frac{1}{\zeta^n}\right)^{\frac{1}{K-1}}\right) \end{aligned}$$

where $\zeta^n = \frac{u_2^n/\tilde{u}_1^n}{\left(1/\tilde{u}_1^n\right)^{1/K}} = u_2^n(\tilde{u}_1^n)^{-\frac{K-1}{K}}$ and we used the fact that $\frac{\beta(\bar{v})}{\gamma(\bar{v})} = \varphi$. Notice that $g(x) = x + (K-1)\varphi^{\frac{1}{K-1}}\left(\frac{1}{x}\right)^{\frac{1}{K-1}}$ is convex in x and minimized at $x^* = \varphi^{\frac{1}{K}}$ with $g(x^*) = \varphi^{\frac{1}{K}}K$.

Notice that

$$B(\tilde{u}^n) = B(f_1(\tilde{u}_1^n), \tilde{u}_1^n) = -\varphi^{\frac{1}{K}} \gamma(\bar{v}) K(\tilde{u}_1^n)^{-\frac{1}{K}} = -\gamma(\bar{v}) \left(\frac{1}{\tilde{u}_1^n}\right)^{\frac{1}{K}} g(x^*). \tag{A.20}$$

Suppose that $\eta := \liminf_{n\to\infty} \zeta^n > \varphi^{\frac{1}{K}}$ we have a constant $c_\eta > 1$ such that

$$B(f_2(u_{[2]}^n), u_2^n) \leq c_\eta B(\tilde{u}^n),$$

for all sufficiently large n in which case by (A.22) (choosing ϵ so that $\xi := c_\eta(1-\epsilon) - 1 > 0$; recall ϵ was arbitrary) we have that

$$\begin{aligned} S_K^n(\tilde{u}^n) - S_K^n(\tilde{u}_1^n, u_{-1}^n) &\geq B(\tilde{u}^n) - (1-\epsilon) B(f_2^n(u_{[2]}^n); u_{[2]}^n) + \mathcal{O}(1) \\ &\geq -\xi B(\tilde{u}^n) + \mathcal{O}(1) = \Omega(n^{\frac{1}{2K}}) > 0. \end{aligned} \tag{A.21}$$

The last equality follows since $\tilde{u}_1^n = \hat{u}_1^n + o(n^{-1/2}) = \theta n^{-1/2} + o(n^{-1/2})$ so that $-B(\tilde{u}^n) = \gamma(\bar{v}) \left(\frac{1}{\tilde{u}_1^n}\right)^{\frac{1}{K}} g(x^*) = \Omega(n^{1/2K})$ and we would conclude that, for all n sufficiently large $S_K^n(\tilde{u}^n) > S_K^n(\tilde{u}_1^n, u_{-1}^n)$ meaning that u^n is sub-optimal.

The same argument applies if, instead of $\liminf_{n\to\infty} \zeta^n > \varphi^{\frac{1}{K}}$, we have $\limsup_{n\to\infty} \zeta^n < \varphi^{\frac{1}{K}}$. It remains to prove that one of these hold if $|u_2^n - \hat{u}_2^n| = \Omega\left(n^{-\frac{K-1}{2K}}\right)$. Indeed, if $u_2^n = \hat{u}_2^n + \Omega(n^{-\frac{K-1}{2K}}) = \varphi^{\frac{1}{K}} \theta^{\frac{K-1}{K}} n^{-\frac{K-1}{2K}} + \Omega(n^{-\frac{K-1}{2K}})$, then there exist $\delta > 0$ such that, for all sufficiently large n, either

$$u_2^n > \varphi^{\frac{1}{K}} \theta^{\frac{K-1}{K}} n^{-\frac{K-1}{2K}} + \delta n^{-\frac{K-1}{2K}} \text{ , or } u_2^n < \varphi^{\frac{1}{K}} \theta^{\frac{K-1}{K}} n^{-\frac{K-1}{2K}} - \delta n^{-\frac{K-1}{2K}}.$$

Thus, there exists $\tilde{\delta} > 0$ such that $\zeta^n = u_2^n(\tilde{u}_1^n)^{-\frac{K-1}{K}} = u_2^n(\hat{u}_1^n)^{-\frac{K-1}{K}} = u_2^n(\theta n^{-1/2}+o(\sqrt{n}))^{-\frac{K-1}{K}}$ $> \varphi^{1/K} + \tilde{\delta}$ or $\zeta^n \leq \varphi^{1/K} - \tilde{\delta}$ for all sufficiently large n so that (A.21) holds. We conclude that any optimal sequence u^n must have that $|u_2^n - \hat{u}_2^n| = o(n^{-\frac{K-1}{2K}})$.

Now, one proceeds sequentially. Fixing a sequence such that $u_1^{n*} = \tilde{u}_1^n$ and $u_2^{n*} = \tilde{u}_2^n$ but $|u_3^n - \hat{u}_3^n| = \Omega(n^{-\frac{K-2}{2K}})$ we have

$$S_K^n(\tilde{u}^n) - S_K^n(\tilde{u}_{[2]}^n, u_{-[2]}^n) \geq B(\tilde{u}_{-[2]}^n; \tilde{u}_{[2]}^n) - (1-\epsilon)B(f_3^n(u_{[3]}^n); u_{[3]}^n) + \mathcal{O}(1), \tag{A.22}$$

and one proceeds similarly to our argument above to show that u^n is sub-optimal if $|u_3^n - \hat{u}_3^n| = \Omega(n^{-\frac{K-i+1}{2K}})$. One then proceeds to u_4 and so on.

(Proof of Step 2) In this step, we first calculate the optimal objective function value for the social planner by using the optimal decisions we find in step 2. To do so, we consider $S_K^n(\cdot)$ as defined in (A.10)

$$S_K^n(u^{n*}) = M\left(u_1^{n*}\right) - M(\hat{u}_1^n) + M(\hat{u}_1^n) + B(u^{n*}) - B(\hat{u}^n) + B(\hat{u}^n) + E(u^{n*})$$

where $\hat{u}^n \equiv (\hat{u}_1^n, \cdots, \hat{u}_K^n)$ and $u^{n*} \equiv (u_1^{n*}, \cdots, u_K^{n*})$. Note that $E(u^{n*}) = \mathcal{O}(1)$ and $M\left(u_1^{n*}\right) - M(\hat{u}_1^n) = o(\sqrt{n})$ by using (A.16). Similarly, we have $B(u^{n*}) - B(\hat{u}^n) = o(\sqrt{n})$. Hence we have

$$\begin{aligned} S_K^{n*} &= n\Lambda V\left(\bar{v}\right) - 2\bar{v}\sqrt{n}\sqrt{\alpha} \\ &\quad -n^{\frac{1}{2K}}\left(\beta\left(\bar{v}\right)\varphi^{\frac{K-1}{K}}\theta^{\frac{1}{K}} - \gamma\left(\bar{v}\right)(K-1)\varphi^{\frac{1}{K}}\theta^{-\frac{1}{K}}\right) + o\left(\sqrt{n}\right) + \mathcal{O}(1) \end{aligned}$$

for $K \geq 2$. Hence, two class policy is asymptotic optimal on $o\left(\sqrt{n}\right)$ scale, i.e.

$$\lim_{n \longrightarrow \infty} \frac{S_K^{n^*} - S_2^{n^*}}{\sqrt{n}} = 0.$$

Now, we check if it's worth to offer whether two classes or only single class. Therefore, we calculate the optimal objective function value for the case of single class. The social planner's problem reduces to the following when $K = 1$

$$\begin{aligned}
\max_{v_1^n \in \mathbb{R}_+} S_1^n &= n\Lambda \left[V\left(v_1^n\right) - \alpha W_K V\left(v_1^n\right)\right] \\
&= \max_{u_1^n \in \mathbb{R}_+} n\Lambda V\left(u_1^n\right) \left[1 - \alpha \frac{1}{n\Lambda f\left(\bar{v}\right) u_1^n} - \mathcal{O}\left(\frac{1}{n}\right)\right] \\
&= \max_{u_1^n \in \mathbb{R}_+} n\Lambda \left[V\left(\bar{v}\right) + V'\left(\bar{v}\right) u_1^n + \mathcal{O}\left(\left(u_1^n\right)^2\right)\right] \left[1 - \alpha \frac{1}{n\Lambda f\left(\bar{v}\right) u_1^n} - \mathcal{O}\left(\frac{1}{n}\right)\right] \\
&= n\Lambda V\left(\bar{v}\right) - \frac{\alpha V'\left(\bar{v}\right)}{f\left(\bar{v}\right)} + \max_{u_1^n \in \mathbb{R}_+} -\alpha \frac{V\left(\bar{v}\right)}{f\left(\bar{v}\right) u_1^n} + n\Lambda V'\left(\bar{v}\right) u_1^n + \mathcal{O}\left(n\left(u_1^n\right)^2 + 1\right)
\end{aligned}$$

which gives the following optimal $u_1^{n^*}$ and $S_1^{n^*}$

$$u_1^{n*} = \sqrt{\alpha} n^{-\frac{1}{2}} \sqrt{\frac{\bar{F}\left(\bar{v}\right)}{f\left(\bar{v}\right)}} \sqrt{\frac{V\left(\bar{v}\right)}{\bar{v} f\left(\bar{v}\right)}} + o\left(n^{-\frac{1}{2}}\right)$$

$$S_1^{n*} = n\Lambda V\left(\bar{v}\right) + \alpha\bar{v} - 2\sqrt{\alpha}\sqrt{n}\sqrt{\bar{v}} \sqrt{\frac{V\left(\bar{v}\right)}{\bar{F}\left(\bar{v}\right)}} + \mathcal{O}\left(1\right)$$

Then we have

$$\lim_{n \longrightarrow \infty} \frac{S_1^{n^*} - S_K^{n^*}}{\sqrt{n}} = 2\bar{v}\sqrt{\alpha} - 2\sqrt{\alpha}\sqrt{\bar{v}} \sqrt{\frac{V\left(\bar{v}\right)}{\bar{F}\left(\bar{v}\right)}}$$

where $K \geq 2$. Therefore, there is a significant benefit of offering more than 1 class and we can conclude that offering 2 classes is asymptotically optimal on $o\left(\sqrt{n}\right)$ scale. ■

Proof of Theorem 5: Similar to the proof of Theorem 4. We omit the details.

A.4. Proofs of Auxiliary Lemmas

Proof of Lemma 7: We prove this result for the social planner. The argument for the revenue maximizer requires only minor changes. We first prove that $v_1^{n*} \to \bar{v}$ and then proceed to show that $v_i^{n*} \to \bar{v}$ for all $i = 2, \ldots, K$.

Suppose $v^{n*} \to v^0$ for some vector $v^0 \in \mathbb{R}_+^K$ (if the sequence does not converge we can apply to argument below to any convergent subsequence). Suppose, further, that $v_1^0 > \bar{v}$. We will prove that this leads to a contradiction to the optimality of v^{n*}.

Since the cutoffs u_i^{n*} increase in i, we then have that $u_i^{n*} = v_i^{n*} - \bar{v} = \Omega(1)$. This, in turn, implies (recall that $\Lambda \bar{F}(\bar{v}) = 1$) the existence of $\delta < 1$ such that $\Lambda \bar{F}(v_i^{n*}) \leq \delta$ for all $i = 1, \ldots, K$, and all n sufficiently large. Consequently, for all such n,

$$W_i^n(\bar{v}\mathbf{e} + u^{n*}) = \frac{1}{n(1 - \Lambda \bar{F}(v_i^{n*}))(1 - \Lambda \bar{F}(v_{i+1}^{n*}))} \leq \frac{1}{n(1-\delta)^2} = \mathcal{O}\left(1\right), \; i = 1 \ldots, K,$$

so that $\Lambda n W_i^n(\bar{v}\mathbf{e} + u^{n*}) = \mathcal{O}(1)$. Since $0 < V(x) \leq V(\bar{v})$ for all $x \geq \bar{v}$, we then have

$$\begin{aligned} S_K^n(u^{n*}) &= \Lambda n \left[V(\bar{v} + u_1^{n*}) - \alpha \sum\nolimits_{i=1}^{K} \left(V\left(\bar{v} + u_i^{n*}\right) - V\left(\bar{v} + u_{i+1}^{n*}\right) \right) W_i^n(\bar{v}\mathbf{e} + u^{n*}) \right] \\ &= \Lambda n \left[V(\bar{v} + u_1^{n*}) \right] + \mathcal{O}(1). \end{aligned}$$

Take $\underline{u}^n = (u_1^{n*}/2, u_2^{n*}, \ldots, u_K^{n*})$ to be the vector obtained from u^{n*} by replacing u_1^{n*} with $u_1^{n*}/2$ and keeping all other entries the same. Let $\underline{v}^n = \bar{v}\mathbf{e} + \underline{u}^n$. Notice that

$$\underline{v}^n \to \underline{v}^0 = \left(v_1^0 - \frac{v_1^0 - \bar{v}}{2}, v_2^0, \ldots, v_K^0 \right).$$

Since $v_1^0 - \frac{v_1^0 - \bar{v}}{2} > \bar{v}$ we have, as before, that $W_i^n = \mathcal{O}(1/n)$ for all i and all sufficiently large n so that

$$S_K^n(\underline{v}^n) = \Lambda n \left[V\left(\bar{v} + \frac{u_1^{n*}}{2} \right) \right] + \mathcal{O}(1)$$

Therefore,

$$S_K^n(\underline{v}^n) - S_K^n(v^{n*}) = \Lambda n \left[V\left(\bar{v} + \frac{u_1^{n*}}{2} \right) - V(\bar{v} + u_1^{n*}) \right] + \mathcal{O}(1)$$

It follows from the strictly positive density of F, and from $u_1^{n*} \to v_1^0 - \bar{v}$ and $u_1^{n*}/2 \to (v_1^0 - \bar{v})/2$, that

$$\text{(A.23)} \quad V\left(\bar{v} + \frac{u_1^{n*}}{2} \right) - V(\bar{v} + u_1^{n*}) = \int_{\frac{u_1^{n*}}{2}}^{u_1^{n*}} x f(x) dx \geq \frac{u_1^{n*}}{2} \left(\bar{F}\left(\frac{u_1^{n*}}{2} \right) - \bar{F}(u_1^{n*}) \right) = \Omega(1),$$

and, in particular, that

$$S_K^n(\underline{v}^n) - S_K^n(v^{n*}) \geq \Lambda n \frac{u_1^{n*}}{2} \left(\bar{F}\left(\frac{u_1^{n*}}{2} \right) - \bar{F}(u_1^{n*}) \right) + \mathcal{O}(1) \geq Ln,$$

for some $L > 0$ contradicting the optimality of v^{n*}. We conclude that optimal cutoffs must satisfy $v_1^{n*} \to \bar{v}$.

From this first step it follows in particular that a limit v^0 of v^{n*} must have $v_1^0 = \bar{v}$. Suppose that there exists an index $i \leq K$ such that $v_i^0 > \bar{v}$. Let i_0 be the smallest such index. Then, $W_i^n(\bar{v}\mathbf{e} + u^{n*}) = \mathcal{O}(1/n)$ for all $i \geq i_0$ and, in turn,

$$S_K^n(u^{n*}) = \Lambda n \left[V(\bar{v} + u_1^{n*}) - \alpha \sum\nolimits_{i=1}^{i_0 - 1} \left(V\left(\bar{v} + u_i^{n*} \right) - V\left(\bar{v} + u_{i+1}^{n*} \right) \right) W_i^n(\bar{v}\mathbf{e} + u^{n*}) \right] + \mathcal{O}(1).$$

Replicating our arguments above, take

$$\underline{u}^n = \left(u_1^{n*}, \ldots, u_{i_0-1}^{n*}, \frac{u_{i_0}^{n*}}{2}, u_{i_0+1}^{n*}, \ldots, u_K^{n*}\right) \to \left(0, \ldots, \frac{v_{i_0}^0 - \bar{v}}{2}, \ldots, v_K^0\right).$$

Notice that, by the definition of i_0 we have $u_{i_0}^{n*}/2 > u_{i_0-1}^{n*}$ for all sufficiently large n so the monotonicity of cutoffs (in i) is maintained. Let $\underline{v}^n = \bar{v}\mathbf{e} + \underline{u}^n$. Then, also for $\underline{u}^n$,

$$S_K^n(\underline{u}^n) = \Lambda n \left[V(\bar{v} + \underline{u}_1^n) - \alpha \sum\nolimits_{i=1}^{i_0-1} \left(V\left(\bar{v} + \underline{u}_i^n\right) - V\left(\bar{v} + \underline{u}_{i+1}^n\right)\right) W_i^n(\bar{v}\mathbf{e} + \underline{u}^n)\right] + \mathcal{O}(1),$$

so that

$$\begin{aligned} S_K^n(\underline{u}^n) - S_K^n(u^{n*}) = &-\alpha\Lambda n(V(\bar{v} + u_{i_0-1}^{n*}) - V(\bar{v} + u_{i_0}^{n*}/2))W_{i_0-1}^n(\bar{v} + \underline{u}^n) \\ &+ \alpha\Lambda n(V(\bar{v} + u_{i_0-1}^{n*}) - V(\bar{v} + u_{i_0}^{n*}))W_{i_0-1}^n(\bar{v} + u^{n*}) + \mathcal{O}(1) \end{aligned}$$

Using

$$\begin{aligned} W_{i_0-1}^n(\bar{v} + u^{n*}) &= \frac{1}{n\left(1 - \Lambda\bar{F}(\bar{v} + u_{i_0-1}^{n*})\right)\left(1 - \Lambda\bar{F}(\bar{v} + u_{i_0}^{n*})\right)}, \\ W_{i_0-1}^n(\bar{v} + \underline{u}^n) &= \frac{1}{n\left(1 - \Lambda\bar{F}(\bar{v} + u_{i_0-1}^{n*})\right)\left(1 - \Lambda\bar{F}(\bar{v} + u_{i_0}^{n*}/2)\right)}, \end{aligned}$$

we get

$$S_K^n(\underline{u}^n) - S_K^n(u^{n*}) \;=\; \frac{\alpha\Lambda}{1 - \bar{F}(\bar{v} + u_{i_0-1}^{n*})}\left(g^n(u_{i_0}^{n*}) - g^n(u_{i_0}^{n*}/2)\right) + \mathcal{O}(1),$$

where, for $x \geq u_{i_0-1}^{n*}$,

$$g^n(x) := \frac{V\left(\bar{v} + u_{i_0-1}^{n*}\right) - V(\bar{v} + x)}{1 - \Lambda\bar{F}(\bar{v} + x)} = \frac{V\left(\bar{v}\right) - V(\bar{v} + x)}{1 - \Lambda\bar{F}(\bar{v} + x)} + \frac{V\left(\bar{v} + u_{i_0-1}^{n*}\right) - V(\bar{v})}{1 - \Lambda\bar{F}(\bar{v} + x)}.$$

Notice that since $u^{n*}_{i_0-1} \to 0$, we have that $(1 - \Lambda\bar{F}(\bar{v} + u^{n*}_{i_0-1}))^{-1} \to \infty$ as $n \to \infty$ so that, to prove that u^{n*} is sub-optimal, it suffices to show that $\left(g^n(u^{n*}_{i_0}) - g^n(u^{n*}_{i_0}/2)\right) = \Omega(1)$ as this will imply that $S^n_K(\underline{u}^n) - S^n_K(u^{n*}) \to \infty$ as $n \to \infty$.

To that end, let

$$\bar{g}(x) := \frac{V(\bar{v}) - V(\bar{v} + x)}{1 - \Lambda\bar{F}(\bar{v} + x)} = \frac{V(\bar{v}) - V(\bar{v} + x)}{\Lambda\bar{F}(\bar{v}) - \Lambda\bar{F}(\bar{v} + x)}.$$

($\Lambda\bar{g}(x)$ is the expected valuation conditional on it being between $\bar{v}$ and $\bar{v} + x$). Since F is assumed to have a strictly positive density, $\bar{g}(x)$ is strictly increasing in x so that, since $u^{n*}_{i_0}/2 = \Omega(1)$ and $u^{n*}_{i_0} - u^{n*}_{i_0}/2 = \Omega(1)$,

$$\bar{g}(u^{n*}_{i_0}) - \bar{g}(u^{n*}_{i_0}/2) = \Omega(1).$$

Also, since $u^{n*}_{i_0} = \Omega(1)$ but $u^{n*}_{i_0-1} \to 0$, we have that

$$\frac{V\left(\bar{v} + u^{n*}_{i_0-1}\right) - V(\bar{v})}{1 - \Lambda\bar{F}(\bar{v} + u^{n*}_{i_0})} = o(1),$$

and the same holds with $u^{n*}_{i_0}$ replaced with $u^{n*}_{i_0}/2$. Combined, we have that

$$g^n(u^{n*}_{i_0}) - g^n(u^{n*}_{i_0}/2) = \bar{g}(u^{n*}_{i_0}) - \bar{g}(u^{n*}_{i_0}/2) + o(1) = \Omega(1),$$

contradicting the optimality of v^{n*}.

Finally, in repeating the proof for the revenue maximizer, $\bar{g}(x)$ will be replaced by $(\rho(\bar{v}) - \rho(\bar{v} + x))/(\Lambda\bar{F}(\bar{v}) - \Lambda\bar{F}(\bar{v} + x))$ which is increasing in x by Lemma 2. ■

Proof of Lemma 8: The sequence u^n is fixed throughout the proof. For simplicity of notation, we write W_i^n for $W_i^n(\bar{v}\mathbf{e} + u^n)$. Recall that

$$S_K^n\left(v^n\right) = n\Lambda\left[V\left(v_1^n\right) - \alpha\sum\nolimits_{i=1}^{K-1} W_i^n\left[V\left(v_i^n\right) - V\left(v_{i+1}^n\right)\right] - \alpha W_K^n V\left(v_K^n\right)\right] \tag{A.24}$$

Taking a Taylor expansion of $V(v_i^n)$ for $i = 1, 2, \cdots, K$ around $\bar{v}$ and recalling $u_i^n = \bar{v} - v_i^n$ we have

$$\begin{aligned} \mathrm{S}_K^n\left(v^n\right) = n\Lambda\Big[& \left(V\left(\bar{v}\right) + V'\left(\bar{v}\right)u_1^n\right) - \alpha\sum\nolimits_{i=1}^{K-1} W_i^n\left(V'\left(\bar{v}\right)\left(u_i^n - u_{i+1}^n\right) + \frac{V''\left(\bar{v}\right)}{2}\left(\left(u_i^n\right)^2 - \left(u_{i+1}^n\right)^2\right)\right) \\ & + \left(-\alpha W_K^n\left(V\left(\bar{v}\right) + V'\left(\bar{v}\right)u_K^n\right)\right)\Big] + \epsilon^n \end{aligned}$$

where

$$\epsilon^n = \mathcal{O}\left(n\sum\nolimits_{i=1}^{K-1} W_i^n\left(u_{i+1}^n\right)^3\right) + \mathcal{O}\left(nW_K^n\left(u_K^n\right)^2\right) + \mathcal{O}(n\left(u_1^n\right)^2), \tag{A.25}$$

Collecting terms we have

$$\begin{aligned} S_K^n\left(v^n\right) = n\Lambda\Big[& \left(V\left(\bar{v}\right) + V'\left(\bar{v}\right)u_1^n\right) - \alpha W_K^n V\left(\bar{v}\right) - \alpha V'\left(\bar{v}\right)\sum\nolimits_{i=1}^{K-1} W_i^n\left(u_i^n - u_{i+1}^n\right) \\ & -\alpha\frac{V''\left(\bar{v}\right)}{2}\sum\nolimits_{i=1}^{K-1} W_i^n\left(\left(u_i^n\right)^2 - \left(u_{i+1}^n\right)^2\right) - \alpha W_K^n\left(V'\left(\bar{v}\right)u_K^n\right)\Big] + \epsilon^n \end{aligned} \tag{A.26}$$

We next apply Taylor expansion to W_i^n. First, using $\Lambda\bar{F}(\bar{v} + u_K^n) = \Lambda\bar{F}(\bar{v}) + \Lambda f(\bar{v})u_K^n + \Lambda\mathcal{O}((u_K^n)^2)$ and recalling that $\Lambda\bar{F}(\bar{v}) = 1$, we have

$$W_K^n = \frac{1}{n\left(1 - \Lambda\,\bar{F}\left(\bar{v} + u_K^n\right)\right)} = \frac{1}{n\Lambda f\left(\bar{v}\right)u_K^n + n\Lambda\mathcal{O}\left(\left(u_K^n\right)^2\right)} \tag{A.27}$$

Next, taking further the Taylor expansion of W_K at $n\Lambda f(\bar{v})u_K^n$ we get

$$\begin{aligned} W_K^n &= \frac{1}{n\Lambda f(\bar{v})u_K^n} + \frac{1}{n\Lambda}\sum\nolimits_{i=1}^{\infty}(-1)^i \frac{\mathcal{O}\left((u_K^n)^2\right)^i}{\left(f(\bar{v})u_K^n\right)^{i+1}} \\ &= \frac{1}{n\Lambda f(\bar{v})u_K^n} + \mathcal{O}\left(\frac{1}{n\Lambda}\right). \end{aligned}$$

In the last equality we use Lemma 7 by which $u_K^n = o(1)$ so that there exist $L, \bar{L} > 0$ such that

$$\begin{aligned} \left|\sum\nolimits_{i=1}^{\infty}(-1)^i \frac{\mathcal{O}\left((u_K^n)^2\right)^i}{\left(f(\bar{v})u_K^n\right)^{i+1}}\right| &\leq \sum\nolimits_{i=1}^{\infty}\left|\frac{\mathcal{O}\left((u_K^n)^2\right)^i}{\left(f(\bar{v})u_K^n\right)^{i+1}}\right| \leq \sum\nolimits_{i=1}^{\infty}\frac{L^i\left|(u_K^n)^{2i}\right|}{\left(f(\bar{v})|u_K^n|\right)^{i+1}} \\ &= \frac{L}{f(\bar{v})^2}\sum\nolimits_{i=1}^{\infty}\left|\frac{Lu_K^n}{f(\bar{v})}\right|^{i-1} \leq \bar{L}. \end{aligned}$$

The last inequality follows from the fact that $u_K^n = o(1)$ so that $Lu_K^n \leq 1/2$ for all n sufficiently large. For the Taylor expansion of W_i^n ($i < K$) we have

$$\begin{aligned} W_i^n &= \frac{1}{\left(1-\Lambda\bar{F}\left(\bar{v}+u_i^n\right)\right)\left(1-\Lambda\bar{F}\left(\bar{v}+u_{i+1}^n\right)\right)} \\ &= \frac{1}{\left(n\Lambda f(\bar{v})u_i^n + n\Lambda\mathcal{O}\left((u_i^n)^2\right)\right)\left(\Lambda f(\bar{v})u_{i+1}^n + \Lambda\mathcal{O}\left(\left(u_{i+1}^n\right)^2\right)\right)}. \end{aligned}$$

We take one more Taylor expansion to get

$$W_i^n = \frac{1}{n\left(\Lambda f(\bar{v})\right)^2 u_i^n u_{i+1}^n} - \frac{f'(\bar{v})}{2n\Lambda^2 f(\bar{v})^3}\left(\frac{1}{u_i^n}+\frac{1}{u_{i+1}^n}\right) + \mathcal{O}\left(\frac{1}{n}\frac{u_i^n}{u_{i+1}^n}\right) + \mathcal{O}\left(\frac{1}{n}\frac{u_{i+1}^n}{u_i^n}\right) + \mathcal{O}\left(\frac{1}{n}\right)$$

Since $u_{i+1}^n > u_i^n$, $\mathcal{O}\left(\frac{1}{n}\frac{u_i^n}{u_{i+1}^n}\right) = \mathcal{O}(1/n)$ so that

$$W_i^n = \frac{1}{n\left(\Lambda f(\bar{v})\right)^2 u_i^n u_{i+1}^n} - \frac{f'(\bar{v})}{2n\Lambda^2 f(\bar{v})^3}\left(\frac{1}{u_i^n}+\frac{1}{u_{i+1}^n}\right) + \mathcal{O}\left(\frac{1}{n}\frac{u_{i+1}^n}{u_i^n}\right)$$

Plugging these back into (A.26) we get

$$\epsilon^n = \mathcal{O}\left(\sum\nolimits_{i=1}^{K-1} \frac{\left(u_{i+1}^n\right)^2}{u_i^n}\right) + \mathcal{O}(n\,(u_1^n)^2) + \mathcal{O}(1)$$

and into (A.27), hence objective function becomes (by also using the fact that $u_{i+1}^n > u_i^n$)

$$\begin{aligned}
S_K^n\left(v^n\right) &= n\Lambda\left(\left(V\left(\bar{v}\right) + V'\left(\bar{v}\right)u_1^n\right) - \frac{\alpha V\left(\bar{v}\right)}{n\Lambda f\left(\bar{v}\right)u_K^n}\right) - \frac{\alpha V'\left(\bar{v}\right)}{\Lambda\left(f\left(\bar{v}\right)\right)^2}\sum\nolimits_{i=1}^{K-1}\left(\frac{1}{u_{i+1}^n} - \frac{1}{u_i^n}\right) \\
&\quad - \frac{\alpha V'\left(\bar{v}\right)f'\left(\bar{v}\right)}{2\Lambda f\left(\bar{v}\right)^3}\sum\nolimits_{i=1}^{K-1}\frac{u_{i+1}^n}{u_i^n} + \alpha\frac{V''\left(\bar{v}\right)}{2\Lambda\left(f\left(\bar{v}\right)\right)^2}\sum\nolimits_{i=1}^{K-1}\frac{u_{i+1}^n}{u_i^n} \\
&\quad + \epsilon^n + \mathcal{O}(1) \\
&= \left(n\Lambda V\left(\bar{v}\right) + n\Lambda V'\left(\bar{v}\right)u_1^n\right) - \frac{\alpha V\left(\bar{v}\right)}{f\left(\bar{v}\right)u_K^n} - \frac{\alpha V'\left(\bar{v}\right)}{\Lambda\left(f\left(\bar{v}\right)\right)^2}\left(\frac{1}{u_K^n} - \frac{1}{u_1^n}\right) - \gamma\left(\bar{v}\right)\sum\nolimits_{i=1}^{K-1}\left(\frac{u_{i+1}^n}{u_i^n}\right) \\
&\quad + \epsilon^n + \mathcal{O}(1) \\
&= n\Lambda V\left(\bar{v}\right) + \left(n\Lambda V'\left(\bar{v}\right)u_1^n + \frac{\alpha V'\left(\bar{v}\right)}{\Lambda\left(f\left(\bar{v}\right)\right)^2}\frac{1}{u_1^n}\right) - \left(\frac{1}{u_K^n}\beta\left(\bar{v}\right) + \gamma\left(\bar{v}\right)\sum\nolimits_{i=1}^{K-1}\frac{u_{i+1}^n}{u_i^n}\right) \\
&\quad + \epsilon^n + \mathcal{O}(1)
\end{aligned}$$

where

$$\gamma\left(\bar{v}\right) = \frac{\alpha}{2\Lambda\left(f\left(\bar{v}\right)\right)^2}\left(\frac{V'\left(\bar{v}\right)f'\left(\bar{v}\right)}{f\left(\bar{v}\right)} - V''\left(\bar{v}\right)\right) = \frac{\alpha}{2\Lambda f\left(\bar{v}\right)} = \frac{\alpha\,\bar{F}\left(\bar{v}\right)}{2f\left(\bar{v}\right)},$$

and

$$\beta\left(\bar{v}\right) = \frac{\alpha V\left(\bar{v}\right)}{f\left(\bar{v}\right)} + \frac{\alpha V'\left(\bar{v}\right)}{\Lambda\left(f\left(\bar{v}\right)\right)^2} = \alpha\frac{V\left(\bar{v}\right)}{f\left(\bar{v}\right)} - \alpha\frac{\bar{v}\bar{F}\left(\bar{v}\right)}{f\left(\bar{v}\right)} = \frac{\alpha}{f\left(\bar{v}\right)}\left(V\left(\bar{v}\right) - \bar{v}\bar{F}\left(\bar{v}\right)\right),$$

and hence

$$\frac{\beta\left(\bar{v}\right)}{\gamma\left(\bar{v}\right)} = 2\frac{V\left(\bar{v}\right) - \bar{v}\bar{F}\left(\bar{v}\right)}{\bar{F}\left(\bar{v}\right)}. \tag{A.28}$$

■

Proof of Lemma 6: To simplify notation we replace $m(x) := MRL(x)$. Using the relationship

$$h\left(x\right) = \frac{m'\left(x\right)+1}{m\left(x\right)}, \tag{A.29}$$

between the MRL and the hazard function, the inequality (A.6) is equivalent to

$$\frac{m'\left(x\right)}{m\left(x\right)} \geq -\frac{h'\left(x\right)}{h\left(x\right)} \tag{A.30}$$

Taking a derivative in (A.30) we further have

$$h'\left(x\right) = \frac{m''\left(x\right)m\left(x\right) - m'\left(x\right)\left(m'\left(x\right)+1\right)}{m^2\left(x\right)}.$$

so that inequality (A.31) is equivalent to

$$m'\left(x\right) \geq -\frac{m''\left(x\right)m\left(x\right) - m'\left(x\right)\left(m'\left(x\right)+1\right)}{\left(m'\left(x\right)+1\right)} \tag{A.31}$$

Any MRL has $m\left(x\right) \geq 0$, $m'\left(x\right) \geq -1$; see Lai and Xie (2006). Therefore, we can multiply both sides of (A.32) by $m'(x)+1 \geq 0$ and say the inequality there holds if and only if

$$m''\left(x\right)m\left(x\right) \geq 0,$$

which holds for all x if and only if $m(\cdot)$ is convex. ■

Proof of Lemma 9 We provide the detailed proof for the case that $j = 2$. The other cases follow identically.

Consider the change of variables $x_i = u_{i+1}/u_i$ and the optimization problem

$$\text{(A.32)} \qquad min \sum_{i=1}^{K-1} \gamma(\bar{v})\, x_i + \beta(\bar{v})\, x_K \quad \text{subject to } x_i \geq 0 \text{ and} \prod_{i=1}^{K} x_i \geq \frac{1}{u_1}$$

Because $\prod_{i=1}^{K} x_i$ is a jointly concave function (see Marcus and Lopes (1957)) and the objective function is linear, this is a convex minimization problem. It therefore has a unique solution which, we will show, is given by

$$x_i^* = \varphi^{1/K} \left(\frac{1}{u_1}\right)^{1/K}, \; i = 1, 2, \cdots, K-1, \text{ and } x_K^* = \varphi^{-(K-1)/K} \left(\frac{1}{u_1}\right)^{1/K}.$$

Denote the KKT multipliers for $x_i \geq 0$ by μ_i and for $\prod_{i=1}^{K} x_i \geq \frac{C}{u_1}$ by η. We claim that $\mu_i \equiv 0$ and

$$\eta := \frac{\gamma(\bar{v})}{\prod_{j \neq i} x_j^*}$$

satisfy both the complementary slackness and first-oder (stationarity) conditions. The complementary slackness conditions are

$$\eta \left(\prod_{i=1}^{K} x_i^* - \frac{1}{u_1} \right) = 0 \text{ and } \mu_i x_i = 0, \; i = 1, \ldots, K.$$

They are both satisfied under our solution. For the first-order (stationarity) conditions, under $\mu_i \equiv 0$, we must check $\gamma(\bar{v}) = \eta \prod_{j \neq i} x_j^*, \; i = 1, 2, \cdots, K-1$ (derivative with respect to x_i) and $\beta(\bar{v}) = \eta \prod_{j \neq K} x_j^*$ (derivative with respect to x_K).

Recall that (see Theorem 4 and equation (A.29))

$$\frac{\beta(\bar{v})}{\gamma(\bar{v})} = 2\frac{V(\bar{v}) - \rho(\bar{v})}{\bar{F}(\bar{v})} = \varphi.$$

Since,

$$\frac{\prod_{j\neq K} x_j^*}{\prod_{j\neq i} x_j^*} = \frac{x_i^*}{x_K^*} = \frac{x_1^*}{x_K^*} = \varphi,$$

we have that

$$\eta = \frac{\gamma(\bar{v})}{\prod_{j\neq i} x_j^*} = \frac{\beta(\bar{v})}{\prod_{j\neq K} x_j^*},$$

which means that the first order conditions are satisfied with the proposed η.

Finally, notice that with the change of variable $x_i = u_{i+1}/u_i$ the minimization problem (A.33) is equivalent to the minimization problem

$$\max_{u_{-1}} B(u_1, u_{-1}) = \max_{u_{-1}} \left\{ -\left(\frac{1}{u_K}\beta(\bar{v}) + \gamma(\bar{v}) \sum\nolimits_{i=1}^{K} \frac{u_{i+1}}{u_i} \right) \right\},$$

because by definition $\prod_{i=1}^{K} \frac{u_{i+1}}{u_i} = \frac{1}{u_1}$.

Thus, from the solution x^* we construct the solution $u_{-1}^* = \left(\frac{u_1}{\varphi}\right)^{\frac{K-i+1}{K}}$. This solution also satisfies our requirement that $u_1 < u_2^* < \cdots < u_K^*$. ■

A.5. Proof of Theorem 3

The proof is a simple adaptation of that for the case $d(v) = \alpha v$. We outline the key ingredients, the first being an analogue of Lemma 8.

Lemma 10. *Fix a sequence of $u^n = o(1)$ of cutoff values. Then,*

$$S_K^n(u^n) = n\Lambda V(\bar{v}) + \left(n\Lambda V'(\bar{v})\, u_1^n + \frac{D'(\bar{v})}{\Lambda(f(\bar{v}))^2}\frac{1}{u_1^n} \right) - \left(\frac{1}{u_K^n}\beta(\bar{v}) + \gamma(\bar{v}) \sum\nolimits_{i=1}^{K-1} \frac{u_{i+1}^n}{u_i^n} \right) + \epsilon^n, \tag{A.33}$$

where

$$\beta\left(\bar{v}\right) := \frac{D(\bar{v}) - d(\bar{v})\bar{F}(\bar{v})}{f(\bar{v})},\ \gamma\left(\bar{v}\right) := \frac{\bar{F}(\bar{v})d'(\bar{v})}{2f(\bar{v})},$$

and

$$\epsilon^n := \mathcal{O}\left(\sum\nolimits_{i=1}^{K-1} \frac{\left(u_{i+1}^n\right)^2}{u_i^n}\right) + \mathcal{O}(n\left(u_1^n\right)^2) + \mathcal{O}(1).$$

Note that proof of this lemma is similar to the proof of Lemma 8, hence we skip it here. Rest of the proof of Theorem 3 is similar to αv case, i.e., Theorem 1. We explain the main changes below.

Optimal cutoff valuations have the same structure except that some parameters are different. As a remark, optimal cutoff values have the following structure

$$u_i^{n*} = \hat{u}_i^n + o(n^{-\frac{K-i+1}{2K}})$$

for $1 \leq i \leq K$ where

$$\hat{u}_i^n = \varphi^{\frac{i-1}{K}}\theta^{\frac{K-i+1}{K}}n^{-\frac{K-i+1}{2K}}$$

for the social planner and

$$\hat{u}_i^n = \Phi^{\frac{i-1}{K}}\theta^{\frac{K-i+1}{K}}n^{-\frac{K-i+1}{2K}}$$

for the revenue maximizer.

Under the new cost structure, assuming $d(u)$ is differentiable, we have

$$\varphi = 2\frac{\left(D\left(\bar{v}\right) - d(\bar{v})\bar{F}\left(\bar{v}\right)\right)}{d'(\bar{v})\bar{F}\left(\bar{v}\right)}$$

where

$$D(v) = \int_v^{\infty} d(u)f(u)du$$

Similarly, the key parameter for the revenue maximizer's cutoff points becomes

$$\Phi = 2\frac{d'(\bar{v})f(\bar{v})\bar{F}(\bar{v})}{f'(\bar{v})d'(\bar{v})\bar{F}(\bar{v}) - d''(\bar{v})\bar{F}(\bar{v})f(\bar{v}) + 2d'(\bar{v})f^2(\bar{v})}$$

Note that θ changes as well $\theta = \frac{\bar{F}(\bar{v})}{f(\bar{v})}\sqrt{\frac{D'(\bar{v})}{V'(\bar{v})}} = \frac{\bar{F}(\bar{v})}{f(\bar{v})}\sqrt{\frac{d(\bar{v})}{\bar{v}}}$. Therefore, coverage is same when both service providers offer same number of classes.

$$\begin{aligned} S_K^{n*} &= n\Lambda V(\bar{v}) - 2\sqrt{n}\sqrt{\bar{v}d(\bar{v})} \\ &\quad -n^{\frac{1}{2K}}\left(\beta(\bar{v})\varphi^{\frac{K-1}{K}}\theta^{\frac{1}{K}} - \gamma(\bar{v})(K-1)\varphi^{\frac{1}{K}}\theta^{-\frac{1}{K}}\right) + o\left(\sqrt{n}\right) + \mathcal{O}(1) \end{aligned}$$

for $K \geq 2$. Hence, two class policy is asymptotic optimal on $o(\sqrt{n})$ scale, i.e.

$$\lim_{n\longrightarrow\infty} \frac{S_K^{n^*} - S_2^{n^*}}{\sqrt{n}} = 0.$$

The optimal value of offering a single class is

$$S_1^{n*} = n\Lambda V(\bar{v}) + d(\bar{v}) - 2\sqrt{n}\sqrt{\bar{v}}\sqrt{\frac{D(\bar{v})}{\bar{F}(\bar{v})}} + \mathcal{O}(1).$$

We then have

$$\lim_{n\longrightarrow\infty} \frac{S_1^{n^*} - S_K^{n^*}}{\sqrt{n}} = 2\bar{v}\sqrt{\bar{v}d(\bar{v})} - 2\sqrt{\bar{v}}\sqrt{\frac{D(\bar{v})}{\bar{F}(\bar{v})}} < 0,$$

where $K \geq 2$. The last inequality follows from $d(\cdot)$ being nondecreasing. Overall, we have that there is a significant benefit of offering $K \geq 2$ and that $K = 2$ is nearly optimal in $o\left(\sqrt{n}\right)$ scale.

Finally, for classification, we compare φ and Φ. The revenue maximizer admits more the the high class than the social planner does if

$$\frac{\left(D\left(\bar{v}\right) - d(\bar{v})\bar{F}\left(\bar{v}\right)\right)}{d'(\bar{v})\bar{F}\left(\bar{v}\right)}\left(\frac{f'(\bar{v})}{f(\bar{v})} - \frac{d''(\bar{v})}{d'(\bar{v})} + \frac{2f(\bar{v})}{\bar{F}(\bar{v})}\right) \geq 1.$$

This condition is equivalent to

$$\frac{d'(\bar{v})\bar{F}(\bar{v})}{d(\bar{v})\bar{F}(\bar{v}) - D\left(\bar{v}\right)} - \frac{d''(\bar{v})}{d'(\bar{v})} + \frac{f'(\bar{v})}{f(\bar{v})} + \frac{2f(\bar{v})}{\bar{F}(\bar{v})} \geq 0 \tag{A.34}$$

Since $d(\cdot)$ is non-decreasing, $\zeta(\cdot)$ (recall 1.12) is increasing in λ if

$$\frac{f'(v))\left(D(v)-\bar{F}(v)d(v)\right)}{\Lambda\bar{F}(v)^2 f(v)d'(v)} + \frac{2f(v)\left(D(v)-\bar{F}(v)d(v)\right)}{\Lambda\bar{F}(v)^3 d'(v)} - \frac{d''(v)\left(D(v)-\bar{F}(v)d(v)\right)}{\Lambda\bar{F}(v)^2 d'(v)^2} - \frac{1}{\Lambda\bar{F}(v)} \geq 0,$$

for all v (after replacing λ with $v = \bar{F}^{-1}(\lambda/\Lambda)$. Simplifying and taking the special case $v = \bar{v}$, leads to

$$\frac{f'(\bar{v})}{f\left(\bar{v}\right)} + \frac{2f\left(\bar{v}\right)}{\bar{F}(\bar{v})} - \frac{d''\left(\bar{v}\right)}{d'\left(\bar{v}\right)} - \frac{\bar{F}(\bar{v})d'\left(\bar{v}\right)}{D(\bar{v}) - \bar{F}(\bar{v})d\left(\bar{v}\right)} \geq 0, \tag{A.35}$$

which is equivalent to (A.35).

Observe that in the special case that $d(v) = \alpha v$, $\zeta(\lambda)$ is increasing in λ if and only if the MRL is convex in v. Since

$$M'(\lambda) = \frac{1}{\Lambda}d\left(\bar{F}^{-1}\left(\frac{\lambda}{\Lambda}\right)\right) - d\left(\bar{F}^{-1}\left(\frac{\lambda}{\Lambda}\right)\right)\frac{1}{\Lambda} + \frac{d'\left(\bar{F}^{-1}\left(\frac{\lambda}{\Lambda}\right)\right)}{f\left(\bar{F}^{-1}\left(\frac{\lambda}{\Lambda}\right)\right)}\frac{\lambda}{\Lambda}$$

replacing $v_\lambda = \bar{F}^{-1}(\lambda/\Lambda)$ we have

$$\zeta(\lambda) = -\frac{D\left(v_\lambda\right) - d\left(v_\lambda\right)\bar{F}(v_\lambda)}{\left(\frac{d'(v_\lambda)}{f(v_\lambda)}\Lambda\bar{F}(v_\lambda)\right)\bar{F}(v_\lambda)}.$$

With $d(v) = \alpha v$, this reduces to

$$\zeta(\lambda) = -\frac{V\left(v_\lambda\right) - v_\lambda\bar{F}(v_\lambda)}{\left(\frac{1}{f(v_\lambda)}\ \bar{F}(v_\lambda)\right)\bar{F}(v_\lambda)}.$$

Since v_λ is decreasing in λ, $\zeta(\lambda)$ is increasing [decreasing] if and only if

$$\vartheta(v) = -\frac{V\left(v\right) - v\bar{F}(v)}{\left(\frac{1}{f(v)}\ \bar{F}(v)\right)\bar{F}(v)} = \frac{vf(v)\bar{F}(v) - f(v)V(v)}{(\bar{F}(v))^2}$$

is decreasing [increasing] in v. Recall that

$$MRL(v) = E\left[X - v | X \geq v\right] = \frac{\int_v^\infty tf\left(t\right)dt}{\bar{F}\left(v\right)} - v.$$

so that

$$MRL'(v) = -\frac{\bar{F}\left(v\right)vf(v) - f(v)\int_v^\infty tf\left(t\right)dt}{(\bar{F}\left(v\right))^2} - 1 = -\frac{vf(v)\bar{F}(v) - f(v)V(v)}{(\bar{F}(v))^2} - 1 = -\vartheta(v) - 1.$$

Convexity of the MRL is equivalent then to $\vartheta(v)$ being decreasing in v. ■

A.6. Additional Numerical Experiments

We include here additional numerical evidence for the persistence of the results (derived via asymptotic analysis) for queues with moderate arrival rate (small n). Tables A.1-A.4 focus on

the value of increasing segmentation beyond two classes. Evidently, increasing the number of classes from 1 to 2 – from FIFO to priorities – brings significant benefit. In Table A.1, the RM increases revenues by 8.30% and the SP increases welfare by 4.82%. The important columns are those in bold: the benefit of offering one more class (i.e., going up to 3 classes) is only 1.91% for the revenue maximizer and 0.87% for the social planner. This is true also with the convex MRL case in Table A.2). The corresponding asymptotic statements are reflected then in Tables A.3 and A.4 where n is set to the high value of 100.

Λ	# of classes K	RM (R_i)	SP (S_i)	$(R_i - R_1)/R_1$	$(\mathbf{R_i - R_2})/\mathbf{R_2}$	$(S_i - S_1)/S_1$	$(\mathbf{S_i - S_2})/\mathbf{S_2}$
	1	0.9025	2.6356				
	2	0.9774	2.7625	8.30%		4.81%	
4.175	3	0.9961	2.7864	10.37%	**1.91%**	5.72%	**0.87%**
	4	1.0032	2.7948	11.16%	**2.64%**	6.04%	**1.17%**
	4	1.0067	2.7986	11.55%	**3.00%**	6.18%	**1.31%**
	1	1.9374	6.2072				
	2	2.0676	6.3314	6.72%		2.00%	
20	3	2.0947	6.3492	8.12%	**1.31%**	2.29%	**0.28%**
	4	2.1045	6.355	8.62%	**1.78%**	2.38%	**0.37%**
	5	2.1091	6.3576	8.86%	**2.01%**	2.42%	**0.41%**

Table A.1. The value of segmentation for a queue with $n = 1$ (small volume) and valuation distribution Weibull(1,0.7) (Concave MRL).

Λ	# of classes K	RM (R_i)	SP (S_i)	$(R_i - R_1)/R_1$	$(\mathbf{R_i - R_2})/\mathbf{R_2}$	$(S_i - S_1)/S_1$	$(\mathbf{S_i - S_2})/\mathbf{S_2}$
	1	0.3467	0.6394				
	2	0.3663	0.6622	5.65%		3.57%	
1.9	3	0.3707	0.6666	6.92%	**1.20%**	4.25%	**0.66%**
	4	0.3723	0.6682	7.38%	**1.64%**	4.50%	**0.91%**
	5	0.3731	0.669	7.61%	**1.86%**	4.63%	**1.03%**
	1	0.5218	1.3593				
	2	0.5383	1.3758	3.16%		1.21%	
10	3	0.5414	1.3784	3.76%	**0.58%**	1.41%	**0.19%**
	4	0.5426	1.3792	3.99%	**0.80%**	1.46%	**0.25%**
	5	0.5431	1.3796	4.08%	**0.89%**	1.49%	**0.28%**

Table A.2. The value of segmentation for a queue with $n = 1$ (small volume) and valuation distribution Weibull(1,2) (Convex MRL).

Λ	# of classes K	RM (R_i)	SP (S_i)	$(R_i - R_1)/R_1$	$(\mathbf{R_i - R_2})/\mathbf{R_2}$	$(S_i - S_1)/S_1$	$(\mathbf{S_i - S_2})/\mathbf{S_2}$
	1	76.390	122.781				
	2	77.431	122.893	1.363%		0.091%	
1.9	3	77.618	122.899	1.607%	**0.241%**	0.096%	**0.005%**
	4	77.681	122.901	1.690%	**0.323%**	0.098%	**0.006%**
	5	77.710	122.901	1.728%	**0.361%**	0.098%	**0.007%**
	1	139.922	179.525				
	2	140.831	179.561	0.650%		0.020%	
10	3	140.939	179.562	0.726%	**0.076%**	0.021%	**0.001%**
	4	140.972	179.562	0.750%	**0.100%**	0.021%	**0.001%**
	5	140.987	179.562	0.761%	**0.111%**	0.021%	**0.001%**

Table A.3. The value of segmentation for a queue with $n = 100$ (high volume) and valuation distribution Weibull(1,2) (Convex MRL).

Λ	# of classes K	RM (R_i)	SP (S_i)	$(R_i - R_1)/R_1$	$(\mathbf{R_i - R_2})/\mathbf{R_2}$	$(S_i - S_1)/S_1$	$(\mathbf{S_i - S_2})/\mathbf{S_2}$
	1	162.940	370.936				
	2	164.656	371.270	1.054%		0.090%	
4.175	3	164.997	371.284	1.263%	**0.207%**	0.094%	**0.004%**
	4	165.121	371.288	1.339%	**0.282%**	0.095%	**0.005%**
	5	165.180	371.289	1.375%	**0.318%**	0.095%	**0.005%**
	1	448.547	735.216				
	2	454.389	735.435	1.303%		0.030%	
20	3	455.097	735.441	1.460%	**0.156%**	0.031%	**0.001%**
	4	455.317	735.442	1.509%	**0.204%**	0.031%	**0.001%**
	5	455.414	735.443	1.531%	**0.226%**	0.031%	**0.001%**

Table A.4. The value of segmentation for a queue with $n = 100$ (high volume) and valuation distribution Weibull(1,0.7) (Concave MRL).

APPENDIX B

Proofs for Chapter 2

B.1. Appendix A

B.1.1. Theoretical Ambiguity of Hypothesis 1

In Section 2.3, we denoted the competing Hypothesis 1A and 1B. Optimal action of the driver depends on demand characteristics, demand priors of the driver, and number of drivers entering into the zone. In this section, we illustrate the theoretical ambiguity with the following setting. Assume that demand in a given zone is either high or low denoted by D_H and D_L with $D_H > D_L$. Agent has partial information about the demand by considering the number of new agents occur in her zone. More specifically, agent thinks that demand will be high with probability $\pi_H(n)$ and it will be low with probability $1 - \pi_H(n)$ where n is the number of new agents. Assume that the agent is myopic and will stay in her zone if expected revenue in the current period is greater than her reservation price. Hence, probability of staying in the zone is

$$F\left(a\, min\left(1, \left(\frac{D_H\pi_H(n) + D_L(1-\pi_H(n))}{k+n}\right)\right)\right) \tag{B.1}$$

where a is the expected revenue for the service, k is the number of agents that were already in the zone and v is the (random) reservation price which has $F(\cdot)$ as cumulative distribution function. Therefore, probability of scattering increases as number of new agent increases if and only if $min\left(1, \left(\frac{D_H\pi_H(n)+D_L(1-\pi_H(n))}{k+n}\right)\right)$ increases as n increases. However, this is not

necessarily the case. Suppose $k \geq D_H$ and $\pi_H(n) = 1 - \frac{1}{n+1}$. Then, (B.1) increases by n if $n < \frac{-D_L + \sqrt{(D_H - D_L)(D_H k - D_L)}}{D_H}$ and it decreases by n otherwise. Therefore, this framework shows us that it is not clear how the number of new agents may affect the scattering decision of the agent. Our empirical analysis allows us to resolve this theoretical ambiguity.

B.1.2. Theoretical Ambiguity of Hypothesis 2 and Derivation of Equation

In Section 2.3 we defined Hypothesis 2A and 2B. It is not a priori clear whether Hypothesis 2A or 2B is more consistent with optimal behavior. We illustrate this theoretical ambiguity by considering a setting with 2 zones and 2 agents. Assume that ride requests are only from one zone to the other zone, there is only 1 ride request per zone at each period with probability of p and it takes one period-time to change zones. If they are in the same zone, each driver can hail the passenger equally likely when there is a ride request. We also assume that when the ride request is not satisfied by any of the drivers, the demand is lost. Consider the case where only one of the drivers changes her zone with probability of q when the other driver enters to her zone. Let us call this driver as *strategic* driver. Customer arrival occurs after the driver scatters, therefore; she losses her chances to hail a passenger for that period if she decides to change her zone.

In the long-run, strategic driver has the following probability of hailing a passenger in a given period

$$R \triangleq \frac{2(-2+p)p + (-1+p)^2 pq}{2(-3+q+p(2+(-3+2p)q))} \tag{B.2}$$

which leads to

$$\frac{\partial R}{\partial q} = -\frac{(-1+p)p(1+p(-5+2p))}{2(-3+q+p(2+q(-3+2p)))^2}$$

Therefore, if $p < \frac{5-\sqrt{17}}{4}$, then scattering always increases the chances of the strategic driver's finding a passenger, otherwise; herding is a better choice to increase the probability of finding a passenger. Hence, there can be cases where scattering may payoff even though traveling from one zone to the other zone vacant takes additional time.

B.1.2.1. Derivation of the Equation (B.2)

Following state space, $S = \{1, 2, 3\}$, is sufficient to consider every possible scenario that may occur in this setting where $s = 1$ if drivers are in different zones in the current time period, $s = 2$ if drivers are in the same zone at the current time period because strategic driver did not move but non-strategic driver changed her zone to the one that strategic driver stays in the previous time period, and lastly set $s = 3$ for the remaining three scenarios at which all scenarios result in being in the same zone at the current time period: (3.1) they had been in the same zone and none of them moved, (3.2) they were in the same zone in the previous time period and moved together, hence they are in the same zone at the current time period as well, (3.3) only strategic driver moved in the previous time period and they are in the same zone at the current time period. Note that strategic driver will scatter with probability q only when $s = 2$. To find the limiting probability, π, we construct the transition probability matrix

$$P = \begin{bmatrix} p^2 + (1-p)^2 & p(1-p) & p(1-p) \\ q(1-p) + (1-q)p & 0 & qp + (1-q)(1-p) \\ p & 0 & 1-p \end{bmatrix}$$

Since each state communicates with each other and state space is finite, solution π of $\pi P = \pi$ is the unique limiting π. Solving this equation yields to

$$\pi_1 = \frac{1}{3 - 2p - q + 3pq - 2p^2q}$$
$$\pi_2 = \frac{(1-p)p}{-2p^2q + 3pq - 2p - q + 3}$$
$$\pi_3 = \frac{-2p^2q + p^2 + 3pq - 3p - q + 2}{-2p^2q + 3pq - 2p - q + 3}.$$

Note that strategic driver will hail a passenger with probability of p when $s = 1$, $(1-q)\frac{p}{2}$ (since the driver can hail a passenger only she decides to stay instead of scatter) when $s = 2$ and $\frac{p}{2}$ when $s = 3$. Therefore, using these results with the equation below yield to the result

$$R \triangleq \pi_1 p + \pi_2 \frac{p}{2}(1-q) + \pi_3 \frac{p}{2}.$$

B.2. Appendix B: Spatial Autoregressive Model and Results

B.2.1. Model

As noted in Manski (1993), the decision of the individual can be affected by the decision of her peers, which means that there can be a social interaction between each individual. Therefore, the models we consider in previous sections do not formally consider these peer effects, which may result in so-called identification problem (See Manski 1993 for details). In this section, we test our first hypothesis with a different model setup and a method where we also consider peer effects. For this purpose, we use spatial regression model to test Hypothesis 1.

Before reporting our results, we provide some basic information about this model. Specifically, we use Spatial Autoregressive Model (SAR) to capture the effect of peers. Our estimations require a model setup with balanced data sets, i.e., for each unit of observation we need to have same number of data points over time. Since this is not possible by using drivers as a unit of observation, we consider zones as our unit of observation. Therefore, we define *LeavingRatio* which is basically the fraction of the vacant drivers that decides to leave the zone. Suppose we have N zones and K regressors. The model is specified as

$$Y_t = \alpha + \rho W Y_t + \beta X_t + u_t \tag{B.3}$$

where Y_t denotes an $N \times 1$ vector consisting of one observation on the dependent variable for every zone in the sample at time t, α is $N \times 1$ vector for individual fixed or random effects, W is $N \times N$ spatial weighting matrix, X_t is $N \times K$ matrix of regressors with associated parameters β contained in $K \times 1$ vector and $u_t = (u_{1t}, \ldots, u_{Nt})$ is vector of independently and normally distributed error terms.

As noted in Elhorst (2010), this model prevents possible identification problems. In order to understand the effects of regressors on the dependent variable, we need to have the following modification on (B.3)

$$\begin{aligned} Y_t - \rho W Y_t &= \alpha + \beta X_t + u_t \\ Y_t &= (1 - \rho W)^{-1}(\alpha + \beta X_t + u_t) \end{aligned} \tag{B.4}$$

Then, we can identify the effect of regressors on dependent variable easily. The following equation gives us the effect of k^{th} regressor among X regressors.

$$\left[\begin{array}{ccc} \frac{\partial Y}{\partial x_{1k}} & \dots & \frac{\partial Y}{\partial x_{Nk}} \end{array}\right] = \left[\begin{array}{ccc} \frac{\partial y_1}{\partial x_{1k}} & \dots & \frac{\partial y_1}{\partial x_{Nk}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_N}{\partial x_{1k}} & \dots & \frac{\partial y_N}{\partial x_{Nk}} \end{array}\right] = (1-\rho W)^{-1}\beta_k \tag{B.5}$$

Diagonal elements of (B.5) are called the *direct effects* of the change in the regressors on the dependent variable of the corresponding zone. Since many zones are considered in these models, estimation methods of such models do not list all direct effects and only provide the average of these estimations. On the other side, non-diagonal elements of (B.5) reflect the *indirect effect* because they show us the effect of a change in regressor on a dependent variable in a different zone. Similar to the listing of direct effects, only one result is provided as the estimate for indirect effect of an regressor. More specifically, the average of the row sums is reported (For further information about the direct and indirect effects of spatial regression models see Elhorst 2010).

Table B.1. Statistical Summary for Spatial Data.

Variable	Mean	SD	Min	Max
LeavingRatio	.251595	.2351412	0	1
Vacant	4.299446	5.835457	0	68
GetIn	.0654602	.3126902	0	13
GetOut	.0382696	.2119667	0	6
VacantIn	2.706472	4.167689	0	53
VacantOut	1.445378	2.081628	0	26
SalesTime	2.234779	4.032085	-3.765	14.746
SalesZone	4.163825	6.662402	-2.198	28.969
RushHour	.2501737	.433113	0	1
Weekend	.2857143	.451754	0	1
Observations	4,442,193			

B.2.2. Results

In this section, we provide our results for the SAR model where we consider spatial effects. See Table B.1 for the statistical summary. Note that unit of observation is not driver for this analysis. We have one observation for each minute, day, and zone, leading to a balanced data which is required for our spatial analysis.

In Table B.2, we provide the results of SAR model with both direct and indirect effects of each variable. Note that each zone has its own direct and indirect effects as explained in the previous section but only average of these estimations are reported due to space limitations since we have 400 zones.

As we had in regression models, *NewDrivers* variable has significant and positive *direct effect* on *LeavingRatio*. Therefore, we observe that drivers tend to leave their zone as the number of new drivers increases. Hence, Hypothesis 1B is supported with this model as well. Advantage of this model compared to regression models is that SAR considers social interaction between peers. In our SAR models we find that this social interaction parameter is significant, i.e., decisions of the drivers in a given zone are affected by decisions of peers in neighboring zones.

Table B.2. Effect of Entry of New Drivers on LeavingRatio under SAR model

	(RE-1)	(RE-2)	(RE-3)	(FE-1)	(FE-2)
Spatial: ρ	0.123***	0.121***	0.0676***	0.123***	0.121***
	(147.21)	(145.78)	(78.18)	(147.14)	(145.71)
			Direct Effects		
NewDrivers	0.00133***	0.00267***	0.00157***	0.00132***	0.00266***
	(21.54)	(37.84)	(22.32)	(21.43)	(37.74)
SalesTime	0.00560***	0.00579***	0.000312***	0.00560***	0.00579***
	(233.51)	(238.01)	(8.84)	(233.92)	(238.07)
SalesZone	0.00102	0.00155*	0.00246***	7.33e-14	0.00000236
	(1.52)	(2.36)	(3.67)	(0.07)	(0.07)
Weekend	0.00631***	0.00381***		0.00631***	0.00381***
	(25.44)	(18.66)		(25.44)	(18.67)
RushHour	-0.0289***	-0.0282***		-0.0289***	-0.0282***
	(-101.35)	(-119.60)		(-101.34)	(-119.62)
GetIn		-0.00602***	-0.00643***		-0.00601***
		(-16.28)	(-17.52)		(-16.27)
GetOut		-0.000688	0.000290		-0.000694
		(-1.20)	(0.51)		(-1.21)
VacantIn		-0.00314***	-0.00434***		-0.00314***
		(-85.18)	(-114.40)		(-85.11)
VacantOut		0.00305***	0.00199***		0.00304***
		(42.37)	(27.40)		(42.32)
			Indirect Effects		
NewDrivers	0.000182***	0.000363***	0.000113***	0.000182***	0.000362***
	(22.24)	(37.29)	(20.46)	(21.13)	(36.01)
SalesTime	0.000769***	0.000786***	0.0000224***	0.000768***	0.000786***
	(118.68)	(112.30)	(8.85)	(134.37)	(118.48)
SalesZone	0.000140	0.000211*	0.000176***	1.01e-14	0.000000320
	(1.52)	(2.36)	(3.68)	(0.07)	(0.07)
Weekend	0.000866***	0.000517***		0.000866***	0.000517***
	(25.25)	(18.42)		(25.15)	(19.08)
RushHour	-0.00396***	-0.00383***		-0.00396***	-0.00383***
	(-87.29)	(-90.29)		(-84.26)	(-93.95)
GetIn		-0.000818***	-0.000462***		-0.000816***
		(-16.15)	(-16.89)		(-16.21)
GetOut		-0.0000936	0.0000207		-0.0000942
		(-1.20)	(0.50)		(-1.21)
VacantIn		-0.000427***	-0.000312***		-0.000426***
		(-72.32)	(-61.74)		(-75.69)
VacantOut		0.000414***	0.000143***		0.000413***
		(39.49)	(24.54)		(41.05)
Observations	4,442,193	4,442,193	4,442,193	4,442,193	4,442,193
AIC	-1098983.6	-1106104.1	-1176422.7	-1102714.4	-1109809.9
BIC	-1098664.2	-1105571.9	-1174400.1	-1102474.9	-1109317.6
Log-Likelihood	549515.8	553092.1	588363.3	551375.2	554942.0

t statistics in parentheses
Unit of observation is zone
Zone fixed effects are considered
Dependent variable is *LeavingRatio*
Note: RE-3 Model considers daily and hourly time fixed effects.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

B.3. Results

Table B.3. Impact of Scattering on Agent Utilization

	(OLS-1)	(OLS-2)	(Frac Logit-1)	(Frac Logit-2)
PercentageReactHourly	0.0719***	0.0725***	0.467***	0.507***
	-53.49	-55.38	-55.25	-58.51
Weekend	0.0615***	0.0826***	0.409***	0.562***
	-43.63	-60.46	-45.28	-61.65
RushHour	-0.0286***	-0.0530***	-0.229***	-0.339***
	(-14.73)	(-27.31)	(-15.49)	(-22.69)
SalesTime		0.0184***		0.109***
		-118.11		-111.96
SalesZone		0.00570***		0.0402***
		-115.52		-113.92
Constant	0.115***	0.0549***	-1.964***	-2.450***
	-65.5	-30.48	(-156.16)	(-181.53)

t statistics in parentheses

Dependent variable is PercentageBusy, which is a fraction.

All of the models above use hourly fixed effects.

$^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$