Self-Selecting Priority Queues with Burr Distributed Waiting Costs

Srinagesh Gavirneni

Johnson School
Cornell University
Ithaca, NY 14853
nagesh@cornell.edu

Vidyadhar Kulkarni

Statistics and Operations Research
University of North Carolina
Chapel Hill, NC 27599
vkulkarn@email.unc.edu

Abstract

Service providers, in the presence of congestion and heterogeneity of customer waiting costs, often introduce a fee-based premier option using which the customers self-segment themselves. Examples of this practice are found in health care (concierge medicine), amusement parks, government (consular services), and transportation. Using a single-server queuing system with customer waiting costs modeled as a Burr Distribution, we perform a detailed analysis to (i) determine the conditions (fees, cost structure, etc.) under which this strategy is profitable for the service provider, (ii) quantify benefits accrued by the premier customers; and (iii) evaluate the resulting impact on the other customers. We show that such self-selecting priority systems can be pareto-improving in the sense that they are beneficial to everyone. These benefits are larger when the variance in the customer waiting costs is high and the system utilization is high. We use income data from areas (identified by zipcode) in which MDVIP (the most popular concierge medical service in the U.S.) has been adopted to illustrate our results. Numerical results indicate that planning for a 20-40% enrollment in the high-priority option is robust in ensuring that all the stakeholders benefit from the proposed strategy.

1 Introduction

Service component of the private (non-governmental) U.S. economy, including everything from financial services to health care, has steadily increased in the past four decades and it currently accounts for 47% of the GDP. At the same time, only 24% of the GDP is centered around products with 16% being non-durable goods (e.g. food, clothing, fuel) and 7% being durable goods (e.g. autos, furniture). This disparity between services and products in the U.S. economy is continuing to grow and is further reflected in the employment numbers. In the past 40 years, employment in the service sector increased from 48.8 million to 111.5

million while the employment in the goods sector has only changed from 22.2 million to 24.1 million. This transition from goods to services did not happen independently of each other, but rather in conjunction with each other. Virtually every product is now complemented with a significant service component thus leading to a service-product continuum in place of the traditional service-product dichotomy. This servitization of products, along with the general increase in the service component of the U.S. economy necessitates the study of efficient management of service operations.

Queues, which form when there is a temporary mismatch between demand and supply, are an integral part of most service offerings. For a typical American consumer, waiting in the emergency room or a doctor's office, waiting at a store checkout, or waiting to renew a driver's license is a common occurrence. It is estimated that an average American spends (see Rushin [17]) over two years of their lifetime waiting in queues and it is not hard to imagine that the cost associated with this waiting is not insignificant. An obvious cost associated with waiting is the resulting productivity loss while the secondary effect could be the increased service requirement (especially in the health care setting) because of the delay in service. In order to improve the efficiency of queues and thus make service operations more efficient, eventually leading to a competitive economy, it is necessary to devise strategies that minimize the costs (not just the durations) of waiting. Thus, while studying queues and their performance, there is a need for a fundamental shift from focusing on waiting time to emphasizing waiting cost. While seemingly easy, this task can be quite challenging due to the heterogeneity of the customers. Some customers may have a larger cost due to the opportunities lost during waiting while others may attribute a larger cost because their situation is deteriorating with the waiting time. While developing strategies to minimize the total cost of waiting, it is necessary to recognize this heterogeneity and implement policies that effectively segment the customers along this dimension.

This can be done by introducing a concierge (high-priority) option in a service offering, the most visible example of which is Concierge Medicine (e.g. www.mdvip.com), that is getting a lot of attention not only because of its popularity, but also because of its seemingly controversial nature. While concierge medicine is grabbing a lot of headlines, such schemes are popular in other service industries (e.g. household services, immigration services, amusement parks, financial services) as well. Benjamin Franklin Plumbing, a nationally branded franchising organization offers The Ben Franklin Society program, which, for a relatively low cost (about ten dollars) per month, allows the member to move to the top of their appointment book when the need for a plumbing service arises. SeaWorld, a popular water

based amusement park, offers the “Quick Queue” program, which at a cost of about 20 to 40 dollars depending on the day, gives the patron front of line admission at its most popular rides. USCIS, the American immigration authority expedites, at a cost of 1000 dollars, the processing of H1B (work visa) applications. In order to meet the service requirements, all these services limit (e.g. Benjamin Franklin limits the membership to 22% of their customer base) the number of customers that sign up for the concierge option.

While evaluating the prospect of introducing a concierge option in a service offering, the following design questions need to be addressed:

1. What are the system conditions (customer waiting costs, system utilization, etc.) that make concierge option attractive?
2. How much should the service provider charge for the option?
3. Should the provider restrict the number of customers that sign up for this option and if so at what level?

It is also necessary to understand the impact of introducing the concierge option on the costs realized by the customers, the profits realized by the service provider, and the overall system performance. In order to answer these questions, we develop an appropriate modeling framework, solve the resulting decision problems and characterize how the system behaves with and without the concierge option.

In order to achieve the afore-mentioned research goals, we need to identify an appropriate way to capture the heterogeneity in the customer waiting costs. This would not be easy in general, because it would depend on the specific service offering that is being analyzed. In a health care setting, a customer’s waiting cost could be a function of the person’s wealth, health condition, and possibly the age. In the case of transportation and amusement parks, it could be a function of individual income and wealth. For consular services, it could be a function of ability to pay and urgency in travel plans. It would be near impossible to identify a modeling approach that captures all these idiosyncrasies. However, in order to achieve our goals, we compromise and use hourly income as a proxy for a customer’s hourly waiting cost. It has been well acknowledged in research that a Burr Distribution is effective in capturing income distribution and we will use that as well. Most of our analytical results will use the mathematical representation of the Burr distribution and we will complement that with numerical results using Burr distributions estimated using data from MDVIP adoption in the U.S.

Our analysis is divided into two different settings, depending on how the customer reacts to the implementation of a concierge option. The first one, which we call the *Captive Customers* case, is one in which the customer will not leave the service even if his/her costs are higher under the concierge setting. This would be representative of the typical primary doctor's office in the U.S. health care system, or the H1B visa processing service of USCIS. The second setting, which we call the *Non-Captive Customers* case, the customers are allowed to leave the system (we do not model where they go) if they are not happy with the concierge option. This would be typical of services such as a barber, a restaurant, or an amusement park. We first show that under *Captive Customers*, the service provider can ensure that the non-concierge customers are no worse-off, while garnering additional revenues and providing better service to the concierge customers. Under the *Non-Captive Customers* case, it is imperative that all the remaining customers in the system are no worse off under the concierge option. Even so, the service provider can significantly increase her revenues. In all these cases, we show that there is a wide range of parameter settings in which, the customers and the service provider are no worse-off and that the total system cost is significantly reduced. By comparing this to the minimum possible system cost, we show that transition to the concierge option eliminates more than 70% of the inefficiency in the system.

While the analytical results concretely establish the conditions under which the concierge option is attractive, we acknowledge that they were derived under many assumptions. In order to test whether these results are valid in real-world settings, we analyze the data associated with adoption of MDVIP service across the country. We show that concierge medicine is adopted in areas (categorized by zip codes) where the median income is larger, the population is older; and income has a larger variance. Our analysis showed that by providing the concierge option in these locations, the service providers are able to increase the revenues by about 14%. This analysis of the real-world data complements the analytical results and enables us to conclude that concierge option has the potential to significantly improve the performance of the service offering.

This paper is closely related to three streams of research, namely service operations, priority queuing systems, and customer segmentation. The importance of service operations, their role in the U.S. economy, and the challenges in efficiently managing them has already been described. Customer segmentation is the concept in which the prices are charged based on available capacity in order to better elicit what a customer is willing to pay for a product and/or service. This enables the product or service provider to maximize the revenues they

can garner from their system. In a service setting, this often leads to queues with different priorities. Using pricing to control queuing systems has a rich history and we provide a brief snapshot below.

Naor [14] first studied how to control arrivals using admission tolls for M/M/1 queues. He studied both the individually optimal policies and the socially optimal joining and balking policies. Yechiali [21], Stidham [18], Mendelson and Yechiali [13] further study customer balking behavior for G/M/s systems. Afeche and Mendelson [2] consider a similar problem under more general delay cost structure. These papers do not consider priorities.

Kleinrock [10] allows customers to pay a bribe to secure a position in the queue. One drawback of Kleinrock's model is that it allows unstable policies, that is, left to themselves, greedy customer would not stick to the policy stipulated by the model. Balachandran [4] accounts for this deficiency, and considers Nash equilibrium policies in the case where customers can pay to select a priority in a queuing system based on the congestion information so as to minimize their own costs. Glazer and Hassin [5] do a similar analysis in an M/G/1 case. Adiri and Yechiali [1] study the queue-length dependent joining behavior of single-class customers in a multi-priority M/M/1 queue. They explicitly account for the presence or absence of customer balking (or abandonment). Alperstein [3] derives further results for the same system. Mendelson and Whang [12] consider the individually optimal policies for multi-class customers using a multi-priority M/M/1 system. They assume that the classes are distinguished by fixed distinct holding cost rates, and different utilities for obtaining service. Kim and Mannino [9] extend their results to the M/G/1 queue. Nazerzadeh and Randhawa [15] show that offering two service grades with appropriate prices is asymptotically optimal for an M/M/1 system with heterogeneous customer waiting costs. Rao and Peterson [16] consider the the problem of pricing and capacity (service rate) planning in a multi-priority queuing system. The recent book by Stidham [19] deals with subject of optimal pricing in priority queues. Hassin and Haviv [7] is an excellent source of general literature on strategic queueing and Chapter 4 therein is focused on information in priority models.

In all these papers the common assumption is that the holding costs are class dependent but constant, and the service times may or may not be class-dependent. The variation among the customers comes from their differing (typically class-dependent i.i.d. random) utility for the service. While this modeling tradition has its uses, it is difficult to estimate the utility that a customer attributes to the service offering. We hypothesize that customers' heterogeneity is a result of their differing waiting costs and that the waiting costs can be

estimated using customer characteristics such as income and age. Based on this observation, we model a service setting in which the customers have varying waiting cost rates. In this respect, Mendelson and Whang [12] and Kim and Mannino [9] are the papers that are mostly closely related to our research. These two papers model customer heterogeneity in a discrete fashion while we use a continuous distribution (specifically the Burr distribution) to capture it. Further, they work towards offering a price for each exogenously given customer class with the objective of achieving incentive compatibility, while our strategy offers just two prices and the customers are endogenously segmented into two classes. Because of the endogenous nature of customer segmentation, our strategy guarantees incentive compatibility and we focus on evaluating the benefits/costs of the proposed strategies to the various stakeholders. We demonstrate the applicability and usefulness of this analysis on the Concierge Medicine offering (popularly known as MDVIP) that is fast becoming popular in the U.S.

The rest of this paper is organized as follows. In the next section, we briefly describe the data on MDVIP adoption as we will use some properties of that data to motivate our modeling and analysis. Section 3 details the modeling framework we use and details the analysis of the basic model. Section 4 introduces the concierge strategy and details the *Captive Customers* case, the impact of the concierge option on system performance, performance of the service provider, and impact on the customers. Section 5 focuses on the *Non-Captive Customers* case and shows that the concierge option is even more effective under that setting. We conclude in sections 6 and 7 with a summary of our results and ideas for future research on this topic.

2 MDVIP Adoption

Before presenting our analytical results, we briefly present and analyze the data associated with adoption of the MDVIP service across the country. MDVIP was founded in 2000 and is the national leader of facilitating the adoption of Concierge Medicine. Currently adopted by about 400 doctors across the country, this service charges the patients an annual fee of about \$1500-1800 and guarantees them the ability to see their doctor on a short notice with same-day or next-day appointments. Further, the patients receive a comprehensive health checkup annually, ability to communicate with their doctor by email and a host of other benefits. Given the fees they charge, the priority they give to the customers that sign up, this is a perfect example of a concierge option in a service offering. While the fees the MDVIP doctors charge are on a per-year basis, it is not hard to imagine that a patient might convert that cost into a per-visit cost while evaluating whether or not he/she should enroll in this

service. We will model the concierge fees as a per-visit charge in our analysis.

Using the data on where the doctors that adopt MDVIP are located, we can identify the role that the heterogeneity of customer waiting costs plays in determining the effectiveness of the concierge strategy. Needless to say, the costs that customers attribute to waiting are not easily available, but their income and age are good proxies for it. A customer with a large income is apt to attribute a large hourly cost to the time they spend waiting for service since they see significant costs associated with lost productive time. An older customer is probably more worried about his/her health and the impact of the delay on the eventual outcome and thus is more likely to attribute a higher cost to the waiting time. While it is possible (via customer focus groups or market surveys) to generate a comprehensive characterization of the waiting cost distribution, we will use the income distribution as a proxy for waiting cost distribution. To achieve this, we first describe the characteristics of the areas where MDVIP has been adopted.

2.1 The Data

In May 2011, when we gathered this data, there were 421 doctors offering the concierge option and they were located in 260 unique zip codes. For each of these zip codes, we used the website www.zipskinny.com and obtained data on the income distribution, population size, and the median age of the population. This data is from the 2000 census which is the most recent census for which this data is available. We were not able to obtain information for nine of the zip codes and thus we had to drop them from our analysis, resulting in 251 remaining zip codes. The median income for these zip codes ranged from 14,095 to 113,788 while the median age ranged from 21.2 to 63.9. The zip code with the largest median income was 20817 covering the area of Bethesda, Maryland while the zip code with the lowest median income was 02903 representing the area of Providence, Rhode Island. The zip code with the highest median age was 33437 representing the area of Boynton Beach, Florida while the zip code with lowest median age was 78705 covering the area of Austin, TX. The zip code with the smallest population (numbering only 1528) was 02462 covering the area of Newton Lower Falls, Massachusetts. The zip code with the largest population (numbering 77,666) was 94533 covering the area of Fairfield, California. Using the data on the income distribution we were able to compute the variance of the income in a zip code.

We present four cases of income distribution (shown in Figure 1) which we will use as the basis for our analysis. The first two cases are based on the overall U.S.A. income and the second one is based on the income distribution in the zip codes where MDVIP service was

adopted. The other distributions are (i) a ZIP code in which the income is very high; and (ii) a ZIP code in which the income is very low. Figure 1 shows the four cumulative functions corresponding to these income distributions. We will use them to motivate our modeling and analysis and also to illustrate the effectiveness of the proposed concierge strategies.

Figure 1: Cumulative Distribution of Incomes in overall U.S.A., MDVIP, one rich ZIP code and one poor ZIP code areas.

3 Model Setup

We model the service offering as a single-server queuing system in which customers arrive according to a $PP(\lambda)$ (a Poisson Process with rate λ). The service times are i.i.d. random variables with a common mean τ , variance σ^2 , and second moment $s^2 = \sigma^2 + \tau^2$. The service provider charges a fee of c dollars to each customer and these customers are identical in all respects except for the costs they attribute to the waiting time experienced before entering service. We assume that the waiting cost per unit time of the i th customer is H_i and $\{H_i, i \geq 1\}$ is a sequence of i.i.d. continuous non-negative random variables with a common cdf $G(\cdot)$, pdf $g(\cdot)$ and mean h .

3.1 Modeling Heterogeneous Waiting Costs

While the waiting costs of the customers can follow any distribution, it is useful to identify a specific distribution to achieve our research objectives. It is conceivable that the distribution will depend on the service offering and the customers that avail of it. While the appropriate waiting cost distribution can be constructed by using tools such as customer surveys or other marketing studies, we will work with the assumption that a customer's income is a good proxy for the customer's waiting cost. Since the income data is readily available from census surveys, determining the waiting cost distributions will be relatively easy. Further, existing research on income distribution has demonstrated, rather convincingly, that a Burr (also called Singh-Maddala) distribution (see Tadakamalla [20]) accurately captures the variability in income.

The Burr distribution over $[0, \infty)$ is defined with parameters $a > 0$, $d > 0$ and $k > 0$ as follows:

$$G(x) = 1 - (1 + (x/a)^d)^{-k}, \quad x \geq 0.$$

See Tadakamalla [20]. Here a is a scale parameter in the sense that if X has Burr distribution with parameters a, d and K the cX has Burr distribution with parameters ca, d, k . The r th moment of this distribution is finite if $r < kd$ and is given by $a^r k B(k - r/d, 1 + r/d)$, where $B(\cdot, \cdot)$ is the beta function (see Johnson et al [8].) The mode is $a((d - 1)/(kd + 1))^{1/d}$ and the median is $2(2^{1/k} - 1)^{1/d}$.

Since we will be using incomes from the various ZIP codes in the U.S., we used SAS to fit the Burr distributions to incomes from the overall USA, the ZIP codes in which MDVIP has been adopted, a ZIP code with very low incomes, and a ZIP code with very high income. The parameters of the resulting Burr distributions are:

Constituency	a	k	d	Mean (\$/hr)	Std. Dev. (\$/hr)
Overall USA	91.73	5.27	1.33	27.74	24.96
MDVIP ZIP Codes	291.38	12.88	1.25	37.48	32.15
Poor ZIP Code	52.30	4.97	1.07	13.97	16.50
Rich ZIP Code	1205.00	48.36	1.37	66.23	49.58

Table 1: Burr distribution parameters and the mean and standard deviations of income distributions for various constituencies

The corresponding probability distributions are given in Figure 2. As is clearly evident from it, the Burr distribution is very versatile and is able to capture a wide variety of distributions of hourly income.

Figure 2: The probability distribution functions determined by fitting the Burr distribution to incomes from the various constituencies.

3.2 FIFO and MinCost Models

The service provider is interested in maximizing the revenue it receives while the customers are interested in minimizing their costs. The social optimizer is interested in minimizing the total waiting cost of all waiting customers per unit time. We study this system under different operational settings of FIFO, MinCost, and Concierge Option. This section focuses on the first two settings which serve as the lower and upper benchmarks for evaluating the effectiveness of the proposed concierge strategies. The concierge option will be formally introduced and rigorously analyzed in sections 4 and 5.

Model FIFO: In this setting, the customers are served in the sequence they arrive. This is the most basic operational strategy. Using the standard results from Gross and Harris [6] we see that the system is stable if $\rho = \lambda\tau < 1$. The expected queuing time for the FIFO system is given by

$$W = \frac{\lambda s^2}{2(1 - \rho)}. \quad (1)$$

The expected total waiting cost (which is independent of the waiting cost distribution) of

all customers per unit time is given by

$$C_f = \lambda h W,$$

and the expected revenue per unit time is given by

$$R_f = \lambda c.$$

Model MinCost: In this setting, we assume that the system planner knows the waiting cost rates of the individual customers. Suppose the service provider is constrained to implement a non-preemptive, non-idling policy (i.e. service cannot be interrupted once it starts, and the server cannot be idle if there are customers waiting). Then, to minimize the total waiting costs, it is optimal to serve the customers in decreasing order of their waiting cost rates. That is, when a customer arrives, he/she will join the queue at a position in which the person ahead has a higher waiting cost rate and the person behind has a lower waiting cost rate. For this system, expected queuing time of a customer with waiting cost rate x can be shown to be (see Kleinrock [10])

$$W \frac{1 - \rho}{(1 - \rho + \rho G(x))^2},$$

where W is as given in Equation 1. Hence the total expected cost of all customers per unit time is given by

$$C_m = \lambda W (1 - \rho) \int_0^\infty \frac{x g(x)}{(1 - \rho + \rho G(x))^2} dx.$$

The difference in costs between models FIFO and MinCost can be attributed to the fact that the MinCost system tries to remove the highest cost customers first from the system within the constraint of non-preemption and non-idling.

When the waiting cost follows the Burr distribution, the expected cost of the MinCost model can be computed as:

$$C_m = \lambda W (1 - \rho) k \int_0^\infty \frac{u^d (1 + u^d)^{k-1}}{((1 + u^d)^k - \rho)^2} du.$$

Figure 3 illustrates the ratio, C_f/C_m , for the four Burr distributions shown in Figure 2. This ratio is a measure of the inefficiency associated with the FIFO model and as is evident from the figure, the inefficiency is increasing with utilization and is larger than 200% for reasonable settings. This demonstrates the need for strategies that can alleviate this inefficiency. We claim that introduction of a concierge option, which essentially segments the customers based on their waiting cost rates, is one possible approach. In the next section, we define the concierge option more precisely, analyze the customer behavior under that option, and compare its performance to these two models.

Figure 3: Inefficiency of the FIFO model (when compared with the MinCost model) as a function of utilization for the four Burr distributions shown in Figure 1.

4 Concierge Option - Captive Customers

In this section we formally introduce the concierge option. It gives each customer a choice of paying an additional fixed fee K and join an “elite” or “high priority” group. For this additional fee, the service provider gives the elite customers non-preemptive priority over the non-elite customers, which reduces the expected queuing time for the elite customers. However, the expected queuing time for non-elite customers increases. In this section we assume that all the customers are captive to the service provider and do not have an option of leaving the system. This may be because the customers need the service or are obligated to get it. Examples of such services are consular processing services that can be expedited for a fee. We analyze the effect of the extra service charge K on the customer behavior and the operational performance of the system.

Clearly customers whose holding cost rates are high would be willing to pay the extra fee and get the preferential treatment, while the others will suffer extra waiting. Thus a given fee K will induce a fraction α of the customers to join the elite group. We call α the

participation level. We first study the relation between K and α .

4.1 Concierge Fee and Participation Level

We begin with the following question: How does the concierge fee K determine the participation level α ? The answer is based on the assumption that all customers are cost minimizers as they compare the expected cost (service fee plus the expected waiting cost) of becoming an elite customer to that of not becoming an elite customer, and pick the decision that has smaller expected total cost. The main result is given in the next theorem.

Theorem 1. *Any level of participation $\alpha \in [0, 1]$ can be induced by setting*

$$K = K(\alpha) = \frac{\rho W}{1 - \alpha \rho} m(\alpha), \quad (2)$$

where

$$m(\alpha) = G^{-1}(1 - \alpha). \quad (3)$$

Proof: Suppose a fraction α of the customers decide to become elite customers. Since the holding cost rates are i.i.d., this implies that each incoming customer is an elite customer with probability α and a non-elite customer with probability $1 - \alpha$. Thus we are looking at a two priority queue where the high priority customers (the elite) arrive according to a $PP(\lambda\alpha)$ and the low priority customers arrive according to a $PP(\lambda(1 - \alpha))$ and the high priority customers are given non-preemptive priority over the low priority customers. The expected waiting time of the high priority customers is given by (see Gross and Harris [6]):

$$W_h = W_h(\alpha) = W \frac{1 - \rho}{1 - \alpha \rho}, \quad (4)$$

and that of the low priority customers is given by:

$$W_l = W_l(\alpha) = W \frac{1}{1 - \alpha \rho}. \quad (5)$$

Clearly, $W_h(\alpha) \leq W \leq W_l(\alpha)$. Consider a customer with holding cost rate H . She will choose to become an elite member if (recall that this is the *Captive Customers* case):

$$K + c + HW_h < c + HW_l, \quad (6)$$

which can be rewritten as

$$H > \frac{K}{W_l - W_h}.$$

This implies that the proportion of customers choosing to become elite members can be computed as:

$$\alpha = P(H > \frac{K}{W_l - W_h}) = 1 - G(\frac{K}{W_l - W_h}).$$

Substituting for W_h and W_l and solving for K , we get Equation 2. □

When the waiting cost rates follow the Burr distribution, we get:

$$m(\alpha) = a(\alpha^{-1/k} - 1)^{1/d},$$

$$K(\alpha) = \frac{\rho W a}{1 - \alpha \rho} (\alpha^{-1/k} - 1)^{1/d}.$$

Since the customers in the top 100α percentile of the waiting cost rate distribution choose to become elite customers, we see that customers whose waiting cost rates are greater than $m(\alpha)$ join the elite group, while others remain non-elite customers. (Recall that leaving is not an option.) It is worth noting that $m(\alpha)$ and $K(\alpha)$ are decreasing functions of α . In other words, as the premium charged for the elite service increases, fewer customers will join the elite service. This is as expected. If the cdf G has finite support, we have $K(0) < \infty$. On the other hand, when the cdf G has infinite support, as in the Burr case, $K(\alpha) \rightarrow \infty$ as $\alpha \rightarrow 0$.

4.2 Revenue Maximization

In this subsection we study the optimal fee the service provider should charge in order to maximize his revenue. We find it more convenient to address this problem indirectly: we determine the optimal participation level which maximizes the revenue. Then Theorem 1 tells us the concierge fees to charge in order to induce this participation level. Note that the total revenue per unit time in the original system (with no priorities) is λc . When we implement the concierge system, we get this revenue, plus an additional revenue of K per elite customer. Thus the incremental revenue, (revenue from the premium charge alone) as a function of the participation level α is given by

$$R(\alpha) = \lambda \alpha K(\alpha).$$

(The subscript c for is Concierge.) Thus the optimization problem for the service provider reduces to the following constrained optimization problem:

$$\begin{aligned} &\text{Maximize } R(\alpha) \\ &\text{Subject to } 0 \leq \alpha \leq 1. \end{aligned}$$

Note that $100R(\alpha)/(\lambda c)$ represents the percentage improvement in the revenue for the service provider if he/she switches from the no priority system to the concierge system and gets α level participation. Let α^* be the optimal solution for the above problem. There is no explicit expression for α^* in the general case. It has to be obtained numerically. The concierge option with

$$K^* = K(\alpha^*) \quad (7)$$

is then the optimal premium to charge. The maximum revenue is given by

$$R^* = R(\alpha^*). \quad (8)$$

When the waiting costs follow a Burr distribution, we have

$$R(\alpha) = \frac{\rho W a \lambda \alpha}{1 - \alpha \rho} (\alpha^{-1/k} - 1)^{1/d}.$$

Setting the derivative of $R(\alpha)$ to zero, we can see that the optimal α^* is the unique solution in $(0,1)$ to

$$\alpha^{1/k} - \frac{\rho}{kd} \alpha = 1 - \frac{1}{kd}.$$

Note that this solution is independent of the scale parameter a . When $k = 2$, we have an explicit solution:

$$\alpha^* = \left(\frac{d - \sqrt{d^2 - \rho(2d - 1)}}{\rho} \right)^2.$$

In the general case solution can be obtained numerically as the limit of α_n 's produced by the following recursion:

$$\alpha_0 = 0, \alpha_{n+1} = \left(\frac{\rho}{kd} \alpha_n + 1 - \frac{1}{kd} \right)^k.$$

Figure 4 shows the plots of additional revenues and concierge fees as a function of ρ . All are increasing functions of ρ . As the system gets more congested, it is optimal to charge higher and higher premiums. Paradoxically, the higher congestion in fact induces increasing participation in the elite group in spite of the higher premiums!

4.3 Social Optimization

In this subsection we consider the social point of view under which the concierge fee K is internal to the system and hence has no role in the analysis. The social objective function is to minimize the total expected waiting cost of all customers. We determine the concierge fees K that will minimize the total waiting cost. Let $C_c(\alpha)$ be the expected total waiting cost under concierge system with participation level α . The next theorem gives the main result about $C_c(\alpha)$.

Figure 4: The optimal concierge fees and the resulting additional revenues for the four income distributions shown in Figure 1

Theorem 2. The total expected waiting cost $C_c(\alpha)$ is given by

$$C_c(\alpha) = \frac{\lambda W}{1 - \alpha \rho} ((1 - \rho)h + \rho \int_0^{m(\alpha)} (1 - G(x))dx - \rho \alpha m(\alpha)). \quad (9)$$

It is a unimodal function of α that achieves its minimum at the unique solution $\alpha = \alpha_s^*$ to

$$m(\alpha) - \rho \int_0^{m(\alpha)} (1 - G(x))dx = \rho(1 - \rho)h \quad (10)$$

where h is the mean waiting cost rate. The socially optimal waiting cost is given by

$$C_s^* = C_c(\alpha_s^*) = \lambda W m(\alpha_s^*). \quad (11)$$

Proof: Let $W_h(\alpha)$ and $W_l(\alpha)$ be as in Equations 4 and 5. Furthermore, let $h_h(\alpha)$ and $h_l(\alpha)$ be the expected holding cost rates of the elite and non-elite customers respectively. They are given by

$$h_h(\alpha) = E(H|H > m(\alpha)) = \frac{1}{\alpha} \int_{m(\alpha)}^{\infty} xg(x)dx, \quad (12)$$

and

$$h_l(\alpha) = E(H|H \leq m(\alpha)) = \frac{1}{1-\alpha} \int_0^{m(\alpha)} xg(x)dx. \quad (13)$$

The total waiting cost of all customers is given by

$$C_c(\alpha) = \lambda\alpha h_h(\alpha)W_h(\alpha) + \lambda(1-\alpha)h_l(\alpha)W_l(\alpha) \quad (14)$$

$$= \lambda(W_h(\alpha) \int_{m(\alpha)}^{\infty} xg(x)dx + W_l(\alpha) \int_0^{m(\alpha)} xg(x)dx). \quad (15)$$

Using

$$\begin{aligned} \int_{m(\alpha)}^{\infty} (1-G(x))dx &= \int_0^{\infty} (1-G(x))dx - \int_0^{m(\alpha)} (1-G(x))dx = h - \int_0^{m(\alpha)} (1-G(x))dx, \\ \int_0^a xg(x)dx &= \int_0^a (1-G(x))dx - a(1-G(a)), \end{aligned}$$

and

$$G(m(\alpha)) = 1 - \alpha.$$

We can reduce the right hand side above to obtain Equation 9. Taking the derivative of $C_c(\alpha)$ in this equation with respect to α and setting it to zero yields Equation 10. Taking the derivative of the left hand side of Equation 10, we can show that it is a decreasing function of α that starts above $(1-\rho)h$ at $\alpha = 0$ and decreases to 0 at $\alpha = 1$. Hence there is a unique solution $\alpha = \alpha^*$ to Equation 10. Equation 11 follows from using Equation 10 to simplify Equation 9. $\square$

The socially optimal α_s^* is obtained by minimizing $C_c(\alpha)$ as a function of α . Using the α_s^* , we can compute the socially optimal $K_s^* = K(\alpha_s^*)$, and the socially optimal cost of the concierge system as $C_s^* = C_c(\alpha_s^*)$. The social cost of service-provider optimal (revenue maximizing) policy is given by $C^* = C_c(\alpha^*)$, where α^* is the optimal participation level as given in Section 4.2. This has to be done numerically. When the waiting cost follows a Burr distribution, there is no closed form expression for the total expected waiting cost due to the integral in right hand side of Equation 9. Figure 5 illustrates the participation levels for the revenue maximization (RM) and social optimization (SO) settings for various levels of utilization. Observe that the α_s^* is consistently lower than α^* implying the socially optimal policy induces a smaller elite class than the revenue maximizing policy. This is analogous to Naor's [14] observation in queuing systems: socially optimal policies induce lower utilization of the system than under the profit maximizing policy.

Clearly, if the social objective function is to minimize the total waiting costs, then the best non-preemptive non-idling policy is the MinCost policy mentioned before: when the

Figure 5: Participation levels for the revenue maximization and social optimization settings for the four waiting cost distributions shown in Figure 1.

server becomes available, he/she picks the customer with highest waiting cost rate to serve next. This requires the knowledge of the holding cost rates of the waiting customers, which is usually not available. The concierge policy helps customers self-identify as high cost or low cost customers, and hence provides an implementable version of the MinCost policy. Hence it is instructive to study how effective the concierge system is in lowering the waiting cost from the no-priority FIFO system and closer to the waiting cost incurred by MinCost system. To do this recall that the cost rate of the FIFO policy is C_f and that of the MinCost policy is C_m . We compare the waiting cost gap $C_f - C_s^*$ with the gap $C_f - C_m$. We can interpret the ratio

$$\frac{C_f - C_s^*}{C_f - C_m}$$

as the “efficiency” of the concierge system in bringing the waiting costs closer to the minimum possible waiting costs. The higher the ratio, the higher the efficiency. Figure 6 illustrates the potential efficiency of the proposed concierge strategy. Observe that this value is consistently above 70% and is increasing with ρ . It is amazing, and concurs with the result in Nazerzadeh and Randhawa [15], that introduction of a two priority system brings the waiting costs to

within 25% of the system with infinite priorities, which is essentially the MinCost system.

Figure 6: Efficiency of the concierge system as a function of ρ .

5 Concierge Option - Non-Captive Customers

So far we had assumed that each customer has only two options: pay K and join the high priority group or pay nothing extra and remain in the low priority group. In many service environments, if their expected total cost in the priority system is greater than what it was under the original system with no priorities, the customers are free to leave the service system. It is possible that they go to another service provider or decide not to obtain this service. They are just no longer part of this service operation. We do not explicitly model where these customers go. We analyze such a scenario in this section.

Clearly a given K will induce the customers on the higher end of the waiting cost rates to join the high priority group, those on the lower end of the waiting cost rates to join the low priority group and those in the middle to leave. Let α be the fraction of the customer who join the high priority (elite group), β , the fraction of the customers who join the low priority and $\delta = 1 - \alpha - \beta$ be the fraction who leave. Since the waiting cost rates of individual customers are i.i.d., this implies that an incoming customer joins the high priority group with

probability α , joins the low priority group with probability β and leaves with probability δ .

First we need to resolve whether if there exists a concierge system with fee K that can achieve the given probabilities α and β . If so, we call the pair (α, β) *achievable*. The following theorem comprehensively characterizes all the achievable pairs of α and β .

Theorem 3. Let

$$\beta(\alpha) = \frac{1 - \alpha(1 + \rho) + \alpha^2 \rho^2}{1 - \alpha \rho^2}, \quad (16)$$

and

$$\phi(\rho) = \frac{1 + \rho - \sqrt{(1 + \rho)^2 - 4\rho^2}}{2\rho^2}, \quad (17)$$

Then all the achievable pairs are of the form $(\alpha, \max(\beta(\alpha), 0))$, for $0 \leq \alpha \leq 1$. For

$$0 \leq \alpha < \phi(\rho)$$

$\beta(\alpha) > 0$, and the pair $(\alpha, \beta(\alpha))$ is achieved by the fee

$$K = K(\alpha) = \frac{m(\alpha)\lambda s^2 \rho(1 - \alpha\rho)}{2(1 - \rho)(1 - \alpha\rho^2)}. \quad (18)$$

For

$$\phi(\rho) \leq \alpha \leq 1,$$

$\beta(\alpha) = 0$, and the pair $(\alpha, 0)$ is achieved by the fee

$$K = K(\alpha) = \frac{m(\alpha)\lambda s^2(1 - \alpha)}{2(1 - \rho)(1 - \alpha\rho)}. \quad (19)$$

Proof: Suppose the concierge fee K causes a customer to join the high priority group with probability α , join the low priority group with probability β and leave the system with probability δ . Thus the effective arrival process of the customers is $PP(\lambda(\alpha + \beta))$, and joining customer is elite with probability $\alpha/(\alpha + \beta)$ and non-elite with probability $\beta/(\alpha + \beta)$. Then the expected waiting time in the high priority queue is given by (using $\rho = \lambda\tau$)

$$W_h^a = W_h^a(\alpha, \beta) = \frac{\lambda(\alpha + \beta)s^2}{2(1 - \alpha\rho)}, \quad (20)$$

and that in the low priority is given by

$$W_l^a = W_l^a(\alpha, \beta) = \frac{\lambda(\alpha + \beta)s^2}{2(1 - (\alpha + \beta)\rho)(1 - \alpha\rho)}. \quad (21)$$

The expected waiting time in the original system with the no priorities and no abandonment is given by W of Equation 1. Let C be the cost to the i th customer in the original system, C_h

be the expected cost in the high priority queue and C_l the expected cost in the low priority queue. Then we have

$$C = c + H_i W, \quad C_h = c + K + H_i W_h^a, \quad C_l = c + H_i W_l^a.$$

First we consider how to achieve α . We see that the i th customer will join the high priority queue if $C_h \leq C$, i.e.,

$$K \leq H_i(W - W_h^a). \quad (22)$$

Substituting for W and W_h^a we get

$$W - W_h^a = \frac{\lambda s^2(1 - \alpha - \beta(1 - \rho))}{2(1 - \rho)(1 - \alpha\rho)}.$$

Note that $W \geq W_h^a$ always. Thus inequality 22 can be achieved if it is satisfied by the least value of H_i that an elite customer can have, namely, $m(\alpha)$. Hence we get

$$K = K(\alpha, \beta) = \frac{m(\alpha)\lambda s^2(1 - \alpha - \beta(1 - \rho))}{2(1 - \rho)(1 - \alpha\rho)}. \quad (23)$$

Thus, for a fixed α, β , the above K will induce the correct α .

Next we consider what β 's can be achieved for a given α . As before, the i th customer will join the low priority queue if $C_l \leq C$, i.e.,

$$W_l^a \leq W. \quad (24)$$

Substituting for W and W_l^a we get

$$W_l^a - W = \frac{\lambda s^2(\alpha(\rho(1 - \rho) + \delta\rho^2) - \delta)}{1(1 - \rho)(1 - \alpha\rho)(1 - (\alpha + \beta)\rho)}.$$

The right hand side is negative if and only if

$$\alpha \leq \frac{1 - \alpha - \beta}{\rho(1 - \rho) + (1 - \alpha - \beta)\rho^2}. \quad (25)$$

We see that the the above inequality is satisfied if $\beta \leq \beta(\alpha)$, where $\beta(\alpha)$ is as in Equation 16. Clearly the fraction of customers who will stay is the largest β for the given α . This yields Equation 16. We can further show that $\alpha + \beta(\alpha) \leq 1$, for all $0 \leq \alpha \leq 1$.

Now, consider the case $0 \leq \alpha < \phi(\rho)$, where $\phi(\rho)$ is as given in Equation 17. One can show that in this case $\beta(\alpha) > 0$. Thus the feasible pair is $(\alpha, \beta(\alpha))$. It can be achieved by using $K = K(\alpha, \beta(\alpha))$. Substituting $\beta = \beta(\alpha)$ in Equation 23 we get $K = K(\alpha) = K(\alpha, \beta(\alpha))$ as in Equation 18.

Finally, consider the case $\phi(\rho) \leq \alpha \leq 1$. In this case we have $\beta(\alpha) < 0$. Thus there is no β for which the low priority customers are better off. Hence no one joins the low priority. Hence the only achievable pair is $(\alpha, 0)$. It is achieved by using $K = K(\alpha, 0)$. Substituting $\beta = 0$ in Equation 23 we get $K = K(\alpha) = K(\alpha, 0)$ as in Equation 19. This proves the theorem. $\square$

Remark 1. Notice that $\beta(\alpha)$ of Equation 16, as a function of α for a fixed ρ , starts at 1 when $\alpha = 0$, and decreases to zero when $\alpha = \phi(\rho)$.

Remark 2. $\phi(\rho)$ is a convex function of ρ that starts at 1 when $\rho = 0$, decreases to .75 when ρ increases to $2/3$ and then increases to 1 as ρ increases from $2/3$ to 1.

Remark 3. It is interesting to note that the common service charge c does not play a role in the results of Theorem 3.

Figure 7 shows the achievable pairs (α, β) for the four Burr distribution shown in Figure 1.

Figure 7: The achievable pairs as a function of utilization for the four waiting cost distributions shown in Figure 1.

5.1 Choosing the Optimal Fee K

In this section we study the optimal fee, or equivalently, the optimal participation level α , so as to maximize the service provider's revenue. Recall that the fee $K(\alpha)$ that induces the participation level α is given by

$$K(\alpha) = \begin{cases} \frac{m(\alpha)\lambda s^2 \rho(1-\alpha\rho)}{2(1-\rho)(1-\alpha\rho^2)} & \text{if } 0 \leq \alpha < \phi(\rho), \\ \frac{m(\alpha)\lambda s^2 (1-\alpha)}{2(1-\rho)(1-\alpha\rho)} & \text{if } \phi(\rho) \leq \alpha \leq 1. \end{cases} \quad (26)$$

The service provider's revenue per unit time in the original system is λc . The revenue rate in the new system is given by

$$R(\alpha) = \lambda c(\alpha + \max(\beta(\alpha), 0)) + \lambda \alpha K(\alpha). \quad (27)$$

Thus the service provider solves the problem:

$$\text{Maximize } R(\alpha)$$

$$\text{Subject to: } 0 \leq \alpha \leq 1.$$

Let α^{**} be the solution to the above problem. Using these values in Equation 26 we get

$$K^{**} = K(\alpha^{**}).$$

Figure 8 illustrates the optimal additional revenues and concierge fees as a function of system utilization.

6 Choosing the Optimal Concierge Participation Level

Ultimately, the service provider must determine the parameters that should be used in implementation of these strategies. In order to answer this question, we focus on the optimal concierge participation level, which as we have shown before, has a one-to-one correspondence with concierge fees. In determining these parameters, attention must be paid to the impact on two principal stakeholders. The service provider is focused on the additional revenues that could be garnered while the concierge customers are interested in the surplus utility (reduction in their cost) they can enjoy. We tabulate these two measures of performance for both the *Captive Customers* and the *Non-Captive Customers* cases for system utilizations of 0.85 and 0.95.

Figures 9 and 10 illustrate the additional revenues and concierge customer surplus for the *Captive Customers* case for the utilizations of 0.85 and 0.95 respectively. Observe that

Figure 8: Optimal concierge fees and the additional revenues in the *Non-Captive Customers* case as a function of utilization for the four waiting cost distributions shown in Figure 1.

the additional revenues steadily increase with the concierge participation level while the concierge consumer surplus first increases and then decreases with the maximum around the 30% level.

Figures 11 and 12 illustrate the additional revenues and concierge customer surplus for the *Non-Captive Customers* case for the utilizations of 0.85 and 0.95 respectively. Observe that, in this case, the additional revenues first increase and then decrease with the maximum being achieved at the 30% participation level while the concierge customer surplus is consistently increasing.

Thus in both the *Captive Customers* and the *Non-Captive Customers* cases and for the 0.85 and 0.95 utilizations, it seems reasonable to choose the concierge participation level at around 30% and choosing the appropriate concierge fees to achieve this participation level.

Figure 9: Service provider additional revenues and concierge customers' surplus as a function of the concierge participation level for the four waiting cost distributions shown in Figure 1. Captive Customers and Utilization = 0.85.

7 Conclusion and Future Research

In this paper, we modeled the introduction of a voluntary fee-based priority option in a service offering and evaluated its impact on the customers, the service provider, and the system. We modeled the heterogeneity in the customer waiting costs using the Burr distribution which has been widely acknowledged to be suitable for capturing the variation in individual incomes. We observed that the introduction of a concierge option has the ability to reduce, by as much as 73%, the inefficiency in the system. Further, we showed that the proposed strategies can be beneficial to everyone if the service provider and the elite customers are willing to share (possibly via a discount) some of their benefits with the non-elite customers. To complement these analytical results with real-world perspective, we collected data on the adoption of the MDVIP service across the U.S. and observed that it has been adopted in areas where the median incomes are larger, people are older, and the variance in the income is larger. There are a number of directions in which this research can be extended. One option is to study multiple (not just two as done here) levels of customer segmentation

Figure 10: Service provider additional revenues and concierge customers' surplus as a function of the concierge participation level for the four waiting cost distributions shown in Figure 1. Captive Customers and Utilization = 0.95.

and evaluate their impact on system performance. From an empirical perspective, it would be educational to collect data on the enrollment of patients into MDVIP and determine customer characteristics that encourage a patient to sign up for this service.

References

- [1] Adiri, I. and U. Yechiali. (1974). Optimal Priority-Purchasing and Pricing Decisions in Nonmonopoly and Monopoly Queues, *Operations Research*, **22**, 1051-1066.
- [2] Afeche, P. and H. Mendelson. (2004). Pricing and Priority Auctions in queuing Systems with a Generalized Delay Cost Structure, *Management Science*, **50**, 869-882.
- [3] Alperstein, A. (1988). Optimal Pricing Policy for the Service Facility Offering a Set of Priority Prices, *Management Science*, **34**, 666-671.
- [4] Balachandran, K. R. (1972). Purchasing Priorities in Queues, *Management Science*, **18**, 319-326.
- [5] Glazer, A. and Hassin, R. (1986). Stable Priority Purchasing in Queues, *Operations Research Letters*, **4**, 285-288.

Figure 11: Service provider additional revenues and concierge customers' surplus as a function of the concierge participation level for the four waiting cost distributions shown in Figure 1. Non-Captive Customers and Utilization = 0.85.

- [6] Gross, D. and C. M. Harris (1985). *Fundamentals of Queuing Theory*, Second Edition, Wiley, NY.
- [7] Hassin, R. and M. Haviv (2003). *To queue or not to queue: Equilibrium behavior in queueing systems*. Kluwer Academic Publishers, USA.
- [8] Johnson, N. L., S. Kotz, and N. Balakrishnan, (1994). *Continuous Univariate Distributions*, Vol. 1, Second Edition. Wiley Interscience, NY.
- [9] Kim, Y. J. and M. V. Mannino. (2003). Optimal incentive-compatible pricing for M/G/1 queues, *Operations Research Letters*, **31**, 459-461.
- [10] Kleinrock, L. (1967). Optimum Bribing for Queue Position, *Operations Research*, **15**, 304-318.
- [11] Marchand, M. G. (1974). Priority Pricing, *Management Science*, **20**, 1131-1140.
- [12] Mendelson, H. and Whang, S. (1990). Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue, *Operations Research*, **38**, 870-883.
- [13] Mendelson, H. and U. Yechiali. (1981). Controlling the GI/M/ 1 Queue by Conditional Acceptance of Customers. *Eur. J. Opnl. Res.*, **7**, 77-85.

Figure 12: Service provider additional revenues and concierge customers' surplus as a function of the concierge participation level for the four waiting cost distributions shown in Figure 1. Non-Captive Customers and Utilization = 0.95.

- [14] Naor, P. (1969). On Regulation of Queue Size by Levying Tolls, *Econometrica*, **37**, 15-24.
- [15] Nazerzadeh, H. and R.S. Randhawa. (2014). Asymptotic Optimality of Two Service Grades for Customer Differentiation in Queueing Systems, *University of Southern California Working Paper*.
- [16] Rao, S. and E. R. Peterson. (1998). Optimal Pricing of Priority Services, *Operations Research*, **46**, 46-56.
- [17] Rushin, S. The Waiting Game. Time, August 30, 2007.
- [18] Stidham, S. (1978). Socially and Individually Optimal Control of Arrivals to a G/M/1 Queue, *Management Science*, **24**, 1598-1610.
- [19] Stidham, S. (2009) *Optimal Design of Queuing Systems*, CRC press, NY.
- [20] Tadikamalla, P. R. (1980). A Look at the Burr and Related Distributions, *International Statistical Review / Revue Internationale de Statistique*, **48**, 337-344.
- [21] Yechiali, U. (1972). Customers' Optimal Joining Rules for the GI/M/s Queue, *Management Science*, **18**, 434-448.