

Cross-Selling in a Call Center with a Heterogeneous Customer Population

Itay Gurvich

Kellogg School of Management, Northwestern University, Evanston, Illinois 60208, i-gurvich@kellogg.northwestern.edu

Mor Armony

Stern School of Business, New York University, New York, New York 10012, marmony@stern.nyu.edu

Constantinos Maglaras

Columbia Business School, New York, New York 10027, c.maglaras@gsb.columbia.edu

Cross-selling is becoming an increasingly prevalent practice in call centers, due, in part, to its unique capability to allow firms to dynamically segment their callers and customize their product offerings accordingly. This paper considers a call center with cross-selling capability that serves a pool of customers that are differentiated in terms of their revenue potential and delay sensitivity. It studies the operational decisions of staffing, call routing, and cross-selling under various forms of customer segmentation. It derives near-optimal controls in each of the settings analyzed, and characterizes the impact of a more refined customer segmentation on the structure of these policies and the center's profitability.

Subject classifications: call centers; cross-selling; queueing systems; revenue management; pricing.

Area of review: Manufacturing, Service, and Supply Chain Operations.

History: Received September 2006; revisions received May 2007, September 2007, December 2007; accepted December 2007. Published online in *Articles in Advance* January 5, 2009.

1. Introduction

Many organizations consider their call centers to be one of the most important channels of interaction with their customers, acting both as a service center and a point of sales—an opportunity for the firm to generate extra revenue by offering new or existing products to their customers. The significant revenue potential of this cross-selling strategy is underscored by the nature of the interaction that takes place in a call center and the wealth of information that is available through state-of-the-art customer relationship management (CRM) systems. Together, they enable firms to segment their customer pools effectively and to tailor their product offerings to each such segment to increase the likelihood of purchase and the associated expected revenue. A familiar and successful example of cross-selling practice is in the financial services industry, where customers who call for service, such as for account balance inquiries, are often offered new financial products.¹

Alongside its potential benefits, cross-selling may substantially increase the total workload that needs to be handled by the call center's agents,² which may degrade the system's quality of service and, in turn, have an adverse effect on the overall customer experience, as well as the effectiveness of cross-selling itself. It is important to carefully select which cross-selling opportunities to pursue and when to do so, and to account for the impact of these decisions in determining the staffing level of the call center. This paper considers a call center with cross-selling

capabilities that serves a heterogeneous pool of customers, and studies the operational decisions of staffing, call routing, and cross-selling under various forms of customer segmentation. It derives near-optimal controls in each of the settings analyzed, and characterizes the impact of more refined customer segmentation on the structure of these policies and the center's profitability.

In more detail, we consider a call center with a single pool of fully flexible agents that first handle inbound call service requests, and subsequently decide whether or not to attempt to cross-sell to some of these customers a certain product or service whenever such an opportunity arises. Cross-selling attempts are handled by the same agent that has served the customer's original request, upon completion of that task. Each cross-selling attempt is preceded by an instantaneous step that captures the customer's decision of whether or not to agree to listen to the cross-selling offer. The processing times for the original service request and the cross-selling phase are exponentially distributed with potentially different parameters. Finally, the heterogeneous pool of potential customers comprises a discrete set of types or segments. (The terms "type" and "segment" are used in this paper interchangeably.) Types differ in terms of their delay sensitivity and revenue potential. These are captured through the probability that a customer will agree to listen to a cross-selling offer as a function of the waiting time that he encountered, and through a demand relation that specifies the probability that a customer decides to buy

the offered product as a function of the quoted price and the waiting time.

The ability to segment the caller population allows the call center to customize the product offered to each caller segment. In this paper, we assume that the degree of segmentation is exogenously specified, for example, as the output of an upstream marketing analysis. Depending on the application setting at hand, product customization may involve charging a different price to different segments for the same product, or could involve changing the attributes, as well as the price of the product offered to each segment. In both settings, the goal is to better exploit the preferences of each caller segment so as to increase the expected profitability from cross-selling. The output of this pricing and/or product attribute customization process is summarized by the segment-specific expected revenue per cross-selling attempt. As we show, the latter is crucial in deciding to whom to cross-sell and how to staff the call center. For purposes of the analysis in this paper, we consider the simpler of the two settings mentioned above, in which the call center only customizes the price of the product offered to each segment, keeping all other characteristics of that product common across segments. We acknowledge this fact by using the term price customization as opposed to product customization, again keeping in mind that the essential consequence of the customization capability is that it leads to different expected revenues per cross-selling attempt for each segment.

As an example of price customization, one may consider the pricing of CD (certificate of deposit) products offered by banks to different customers. It is natural to think of the price of the CD as its associated interest rate, although two important product attributes are the minimum capital contribution and the length of the time over which the promised interest rate is guaranteed. An increasingly important application of quantitative pricing and revenue management tools in the financial services industry is in deciding the terms, and more importantly the interest rate, of the CD product that is offered to existing customers to entice them to roll their expiring CD contribution from one product to another. Although pricing to initially attract customers who may be “shopping around” for such a product is quite competitive, the subsequent repricing decisions tend to be less constrained and, indeed, an area of intense activity in that industry.

We study three variants of this model with an increasing availability of information regarding customer segmentation and, as a result, increasing flexibility in terms of the aforementioned operational and pricing decisions. The simplest model is one where customers are not segmented, or equivalently, where their types are not observable. In this case, the manager is limited to making the cross-selling decisions based solely on the aggregate load in the system, and charging all customers the same price. The second model is one where types are observed sometime during their service, and this information can therefore be used

together with the actual waiting time experienced by the customer in deciding whether to cross-sell to a customer, and if so, what price to charge. The third model is one where customer types are observable upon arrival, in which case the manager can also decide how to route customers of different types to the available agents. For each of these models, the call center manager’s problem is to select its staffing, routing, cross-selling, and pricing policies to maximize the center’s expected profit rate, given by its revenues minus the staffing cost minus a linear waiting time cost that is experienced by all customers and is incurred by the center.

The controlled two-stage service sequence of each customer and the dependence of the cross-selling phase on dynamic waiting time information makes an exact analysis of this model cumbersome and difficult, even if customers are treated as one segment. Our approach considers a deterministic relaxation of this problem, which is solved in closed form. Its solution suggests different staffing and cross-selling policies for each of the model variants listed above. In each case, we show that our proposed policy is asymptotically optimal in systems with increasing call volume, and as such is appropriate for call centers with high demand volumes.

Our contribution is twofold: From a practical viewpoint—we propose a concrete, simple, and provably near-optimal solution for the complex problem of cross-selling in environments with multiple customer classes. Our solution will allow firms to extract the revenue potential embedded in their CRM systems through smart operational management of their marketing interface. From a managerial viewpoint—our tractable deterministic analysis and the asymptotic performance guarantees of the proposed policies lead to several insights. The first one is that the marketing decisions of customer segmentation and price customization are effectively decoupled from the operational decisions of staffing, routing, and cross-selling. Specifically, once the set of customer segments has been identified through an appropriate marketing and statistical analysis, and their respective characteristics have been identified using observed data,³ the firm can precompute its price customization strategy ahead of time, instead of dynamically choosing the price charged to each customer. In particular, the prices are static and are identical across customers of the same type. These prices are then fed into the operational control problem that involves staffing, routing, and cross-selling decisions.

The availability of information on customer segmentation has many important consequences, which can also be easily seen from our deterministic relaxation. To start with, roughly speaking, the center will only cross-sell to customers that generate an expected revenue that exceeds the capacity cost involved in pursuing this attempt; the expected revenue is equal to the quoted price times the probability that this customer will buy the offered product, provided that his waiting time was zero. If the center can segment its customers, then it will only cross-sell

to its profitable types; if no segmentation capability is in place, then it will either cross-sell to all customers or to none, depending again on the expected profitability of these cross-selling attempts. In each case, the center will staff so as to handle all regular service requests plus the additional nominal workload generated by its expected cross-selling activities. Because the cross-selling is controllable, it can provide enough flexibility in the use of the center's capacity, which eliminates the need to add "safety staffing" as is typically done according to the "square-root" rule to stabilize the system and guarantee moderate congestion. It is possible that even though it is profitable to cross-sell in a system that segments its customers, this is not the case without segmentation. Our analysis outlines such cases. Overall, customer segmentation increases the center's profitability in two ways: first, through a more efficient use of capacity achieved by reducing the volume of cross-selling attempts that are unlikely to be profitable, and second, by customizing the price for each customer type so as to maximize the resulting expected revenue. Finally, we note that the effect of observing the customer type upon arrival, as opposed to after service has commenced, is small. This is explained by the fact that even when the system does not differentiate between types in its routing decisions and handles all external calls through a common first-come-first-served (FCFS) queue for all these types, the resulting waiting times are small; these are moderated through the dynamic cross-selling decisions of the call center and are reinforced by the customers' delay averseness.

The remainder of this paper is organized as follows. This section concludes with a brief literature survey. Section 2 describes the two models with observable types, emphasizing mostly the model where customer type is revealed once his service starts. These two models are analyzed in §3. Section 4 shows how the pricing problem can be treated separately from all other decisions, which is then used in §5 to analyze a model with no customer segmentation. Section 6 provides results from our numerical experiments. Section 7 contains concluding remarks. The electronic companion contains all of our proofs and is available as part of the online version that can be found at <http://or.journal.informs.org/>.

Literature Review. The literature on the operational aspects of call centers is extensive and has grown rapidly over the past decade. A survey of this literature and a tutorial on the subject can be found in Gans et al. (2003). Of particular relevance to our work is the literature on staffing of call centers. The most commonly used staffing rule in the literature is the so-called square-root safety staffing rule, according to which the number of servers required to handle an offered load of size R is $R + \beta\sqrt{R}$ for some constant β . The square-root safety staffing rule dates back to Erlang in his 1923 paper (that appeared in Erlang 1948). This rule was formalized by Halfin and Whitt (1981), who showed that this square-root safety staffing rule guarantees very short delays in an appropriate

asymptotic regime, and was shown to be nearly optimal for a pure service center that handles a homogeneous customer population in Borst et al. (2004). Square-root safety staffing has been observed to be fairly robust with respect to changes in model assumptions to include features such as customer abandonment (Garnett et al. 2002, Mandelbaum and Zeltyn 2009), multiple customer classes (Armony and Maglaras 2004a, b; Gurvich et al. 2008), multiple server pools (Armony 2005), and nonstationary arrival rates (Feldman et al. 2008). In contrast to the above set of papers, our work shows that the issue of safety staffing is of lesser importance in call centers with significant cross-selling activity because by adjusting the latter the manager can also control congestion.

There is a small but growing portion of the recent literature on call centers that in broad terms studies how to best manage the cross-selling capability of such systems. In more detail, the cross-selling control problem, i.e., the question of when and to whom the center should try to cross-sell, has been studied by several authors, including Akşin in a series of papers with Akşin and Harker (1999), Güneş and Akşin (2004), and Örmeci and Akşin (2007), and by Byers and So in two papers (Byers and So 2007a, b). These papers consider various aspects of the above dynamic control problem under three assumptions: (a) the staffing levels are exogenously fixed; (b) the products and prices offered to the various customers are homogeneous even though the center may be able to segment its customer pool according to their preferences; and (c) a simplified model of the service system that treats customers that go through the cross-selling phase as a separate class of service requests with longer service times, as opposed to as a two-phase service. This latter restriction implies that cross-selling decisions have to be made in the beginning of the interaction with the customer, and it cannot use updated state information that may be available at the completion of a customer's nominal service request. The service facility is either modeled as a single-server queue, a multi-server queue, or a multiserver loss system (i.e., customers that do not find an idle server upon their arrival are lost). For the single-server model, Byers and So (2007a) showed that the optimal cross-selling policy is of a threshold type; the center cross-sells as long as the number of customers in the system is below a certain threshold. The optimality of the threshold policy in the multiserver case has not been established. Despite the restrictive assumptions listed above, these papers made significant contributions to the literature by being the first to address the important motivating questions mentioned earlier, and by deriving insights that seem to be fairly intuitive and, to some extent, robust. They also raised interesting questions: Are these insights robust to more representative models of the service delivery process? What is their impact on staffing decisions? In what way would the staffing decision affect the structure of the cross-selling policy and the profitability of cross-selling? And, finally, what is the impact of customer segmentation on all of the above?

Recently, Armony and Gurvich (2006) proposed a more realistic stochastic model for the cross-selling process, whereby the service time of each customer comprises two distinct phases—the first captures the handling of the customer’s nominal service request, and the second, which is optional, captures the duration of the cross-selling attempt. The main analytical contribution of Armony and Gurvich (2006) is to rigorously show that a threshold-type cross-selling policy is asymptotically optimal for this more complex service model as the nominal demand and the size of the call center grow large. Armony and Gurvich (2006) also conducted a preliminary analysis of the joint staffing and cross-selling control problem for the case where the entire pool of customers is either homogeneous in terms of its preferences, or is treated as such by the system; the latter would correspond to settings where the customers are heterogeneous, but the system does not have segmentation capability.

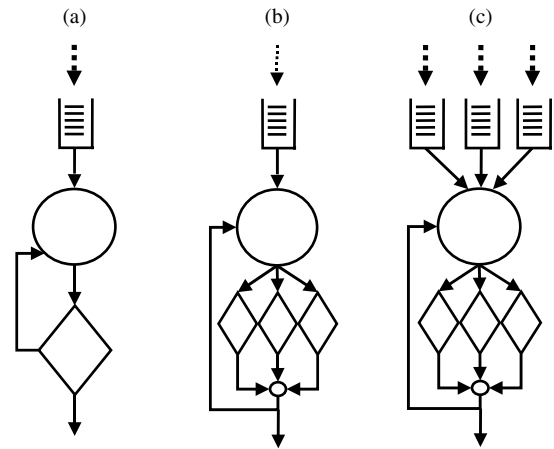
Our paper applies the service model proposed in Armony and Gurvich (2006) to a setting with heterogeneous and delay-sensitive customers to address the joint price customization, staffing, and cross-selling control problem. Our economic model is more general than those used in earlier papers, and the consideration of customer delay sensitivity is new. Our model allows for an insightful analysis of the joint pricing, staffing, and cross-selling problems, which emphasizes the trade-offs among customer segmentation, price customization, staffing costs, and the system profitability. Our work reinforces the insights derived in the various papers listed thus far. It also highlights that the ability to segment the customer pool and customize the respective prices leads to significantly different staffing and cross-selling policy recommendation from those derived in the papers mentioned above.

An important ingredient of our solution methodology hinges on the use of a deterministic relaxation for the original joint pricing, staffing, and cross-selling dynamic optimization problem, which is motivated from the work of Maglaras and Zeevi (2005). Finally, the economic model that we adopt and the notion of price discrimination that underlies our work are related to a vast literature in economics, marketing, and revenue management. We refer the reader to the book by Talluri and Van Ryzin (2004) for an introduction to these subjects.

2. Model Formulation

We consider a call center with a single pool of N fully flexible agents that serves a heterogeneous customer population, comprising K distinct segments, or types, or classes. We study three model variants depending on the extent to which the customer types are observable by the system. These are graphically depicted in Figure 1. Model (a) assumes that types are unobservable, or that the call center does not segment its customers. In model (b), the type of a customer is observed when s/he is being served, and

Figure 1. Three cross-selling models.



this information is subsequently used in the center’s cross-selling decisions. Finally, model (c) is one where the customer type is immediately observed upon arrival, e.g., by requiring customers to enter an account number, and can therefore be used in routing as well as in cross-selling decisions. We will focus on model (b), and treat model (c) as an extension and model (a) as a one-segment special case of this multisegment model.

Basic Service. Type- i customers call the center according to a Poisson process, $\{A_i(t), t \geq 0\}$, with rate λ_i . Let $A(t) = \sum_{i=1}^K A_i(t)$, and define $\Lambda = \sum_{i=1}^K \lambda_i$ to be the total arrival rate into the system. All customers require the same type of service and the processing requirement is exponentially distributed with rate μ^s , independent of the customer type. Under the assumption that types are unobservable before service begins (model (b)), all customers join a single queue and get processed in an FCFS manner.

Cross-Selling. Once regular service is completed, a customer either leaves the system or enters a cross-selling phase that is handled by the same agent. A cross-selling attempt is preceded by an instantaneous step in which the customer is asked to listen to the actual offer. The length of time required for the cross-selling attempt may depend on the customer segment and is assumed to be exponentially distributed with rate μ_i^{cs} for type- i customers. All processing times (regular service and cross-selling) and interarrival times are assumed to be independent.

The probability that a type- i customer will agree to listen to the cross-selling offer after experiencing waiting time w is given by an arbitrary nonincreasing continuous function $q_i(w): \mathbb{R}_+ \mapsto [0, 1]$, with $\lim_{w \rightarrow \infty} q_i(w) = 0$. We set $q_i := q_i(0)$ and note that it is possible to have $q_i < 1$. This allows us to model cases where some customers may always decline to listen to the cross-selling offer.

If a customer of class- i agrees to listen to a cross-selling offer, he will be offered the product at a certain price that might depend on both his class and his actual waiting time. Class- i customers have i.i.d. valuations for

this product, denoted by v_i , drawn from a continuous distribution function $F_i(\cdot)$. The perceived “cost” of the offered product may also depend on the waiting time s /he has experienced. This dependence may arise in some practical settings, such as when signing up for help desk services where the waiting acts as a proxy for the future quality of service. In other applications, the cost of the offered product should not depend on the waiting time, and this is also allowed by our model. Specifically, we assume that class- i customers have a delay-sensitivity constant $c_i \geq 0$. Then, conditional on agreeing to listen to a cross-selling offer, a class- i customer who has waited for w time units before starting her service will buy the product with probability $\bar{F}_i(p_i, w) := \bar{F}_i(p_i + c_i w) = P\{v_i > p_i + c_i w\}$. Applications where the cost of the offered product is independent of the waiting time are captured by setting $c_i = 0$. The resulting conditional expected revenue from a customer of class- i who waited w time units is given by $r_i(p, w) := p_i \bar{F}_i(p_i + c_i w)$. For simplicity of notation, we let $r_i(p) := r_i(p, 0) = \lim_{w \downarrow 0} r_i(p, w)$. We will also assume that the functions $r_i(p_i)$ are unimodal in the p_i s for each i ; this is satisfied by many commonly used demand functions (see Talluri and van Ryzin 2004). For the first few sections, we will assume a fixed vector of prices $p = (p_1, \dots, p_K)$. Hence, we will use the simplified notation $r_i(w)$ instead of $r_i(p_i, w)$ and the notation r_i for $r_i(0)$. We will return to the more general notation in §4, in which we consider the pricing problem.

Control Decisions. The call center manager selects the number of agents N for the system and has discretion with respect to the cross-selling and pricing decisions. We will consider policies, π , that decide whether to cross-sell to the j th type- i customer and which price to charge him as a function of all the information available up to the decision point. In particular, the cross-selling and pricing decisions are dynamic and may depend on the customer’s type, the waiting time encountered by this customer prior to his service, which we denote by $w_{i,j}^\pi$, the number of customers in the queue, and the number of customers of each type- i that are currently in service, denoted by $Q_i^\pi(t)$ and $Z_i^\pi(t)$, respectively. We let $Q^\pi(t) = \sum_{i=1}^K Q_i^\pi(t)$ be the total queue length at time t under π . To guarantee the existence of steady state or at least the existence of long-run averages for various quantities of interest, we will restrict the set of admissible controls as follows.

DEFINITION 1 (ADMISSIBLE CONTROLS). Given a staffing level N , and parameters $\lambda_1, \dots, \lambda_K, \mu^s, \mu_1^{cs}, \dots, \mu_K^{cs}$, we say that π is an admissible policy if it is nonpreemptive, nonanticipative, and $\lim_{t \rightarrow \infty} E[Q^\pi(t)]/t \rightarrow 0$. We denote the family of admissible policies by $\mathcal{A}(\lambda_1, \dots, \lambda_K, \mu^s, \mu_1^{cs}, \dots, \mu_K^{cs}, N)$.

Loosely speaking, $\mathcal{A}(\lambda_1, \dots, \lambda_K, \mu^s, \mu_1^{cs}, \dots, \mu_K^{cs}, N)$ is the set of stabilizing policies under the given parameters. Definition 1 takes into account the fact that the set of

admissible policies depends on the parameters of the model through the stability conditions of the system. To simplify notation, we will omit the parameters $\lambda_1, \dots, \lambda_K, \mu^s$ and $\mu_i^{cs}, i = 1, \dots, K$, whenever these are exogenously fixed, and write $\mathcal{A}(N)$ or simply \mathcal{A} whenever the staffing level is clear from the context. Note that the above definition implies that our system must be able to handle all of the nominal demand, at least when no cross-selling is exercised; that is, the staffing choice must satisfy the constraint $N > R := \Lambda/\mu^s$.

Performance Criterion. We first define two system quantities that will play an important role in the call center’s cost and revenue terms, respectively. Observe that a steady state need not exist for any $\pi \in \mathcal{A}(N)$. With that in mind, for some $\pi \in \mathcal{A}(N)$ and $i = 1, \dots, K$, we define

$$EW_i^\pi(\infty) := E \left[\limsup_{t \rightarrow \infty} \frac{\sum_{j=1}^{A_i(t)} w_{i,j}^\pi}{A_i(t)} \right],$$

$$x_i(\pi) := E \left[\liminf_{t \rightarrow \infty} \frac{\sum_{j=1}^{A_i(t)} x_{i,j}^\pi}{A_i(t)} \right],$$

and

$$(r_i x_i)(\pi) := E \left[\liminf_{t \rightarrow \infty} \frac{\sum_{j=1}^{A_i(t)} r_i(w_{i,j}^\pi) x_{i,j}^\pi}{A_i(t)} \right],$$

where $x_{i,j}^\pi$ is an indicator that is set to one whenever the j th class- i customer goes through a cross-selling phase, and $x_{i,j}^\pi$ equals zero otherwise. The performance measure $(r_i x_i)(\pi)$ should be interpreted as the long-run average revenue per class- i customer under the policy π . When a steady state exists, $EW_i^\pi(\infty)$ and $x_i(\pi)$ coincide with the expected steady-state waiting time experienced by type- i customers, and the steady-state fraction of class- i customers that are asked *and* agree to listen to a cross-selling offer under π , respectively. $(r_i x_i)(\pi)$ will then coincide with the steady-state revenue from class- i customers. Because customers are processed FCFS, it must be that $EW_i^\pi(\infty) = EW_k^\pi(\infty)$ for all i, k , which will also be denoted by $EW^\pi(\infty)$.

The call center incurs linear staffing and waiting time costs per unit time, given by $c \cdot N$ and $\Lambda h EW^\pi(\infty)$, respectively. The latter assumes that the waiting time cost is type independent. The waiting time cost can be thought of as a penalty that the system incurs in terms of lost goodwill from the customers. The type independence of the waiting cost can be relaxed with no effect on any of our results. Under an FCFS discipline it seems reasonable, however, to assign a common cost to all customers. The call center manager’s optimization problem is the following:

$$\sup_{N \in \mathbb{Z}_+, \pi \in \mathcal{A}(N)} \sum_{i=1}^K \lambda_i \cdot (r_i x_i)(\pi) - cN - \Lambda h EW^\pi(\infty). \quad (1)$$

Note that although it is not guaranteed that there exists a control that actually achieves the optimal profit rate,

it is easy to establish the existence of an optimal N^* because N is discrete, the profit rate is bounded above by $\sum_i \lambda_i r_i - c \cdot R$, and it decreases to $-\infty$ as N grows large.

An alternate formulation to (1) would replace the waiting time cost by an upper-bound constraint on the expected waiting time, typically in the order of 30 seconds, and consider the following problem:

$$\sup_{N \in \mathbb{Z}_+, \pi \in \mathcal{A}(N)} \left\{ \sum_{i=1}^K \lambda_i \cdot (r_i x_i)(\pi) - cN: EW^\pi(\infty) \leq \bar{W} \right\}. \quad (2)$$

Indeed, one can view (2) as a more natural starting point, and (1) as a “dualized” version of the problem that is perhaps simpler to address. We will refer to (1) and (2) as the *waiting cost* and *constrained formulations*, respectively. We will also make the following assumption:

ASSUMPTION 1. *Types are labeled so that*

$$r_1 - c/\mu_1^{cs} \geq \dots \geq r_K - c/\mu_K^{cs}$$

and $r_1 - c/\mu_1^{cs} > 0$.

The labeling assumption is innocuous. The condition $r_1 - c/\mu_1^{cs} > 0$ means that it is profitable to cross-sell to at least type-1 customers. As will be shown later, $r_1 - c/\mu_1^{cs}$ is roughly the expected revenue from cross-selling to a class-1 customer minus the marginal staffing costs associated with it. In the absence of this assumption, it makes sense not to invest in extra capacity for cross-selling and to only attempt to cross-sell to a negligible fraction of the customers.

3. Observable Types: Analysis Based on a Deterministic Relaxation

A direct analysis of the problems formulated above is very difficult due to their multiclass nature and the dependence of the cross-selling success probability on state-dependent information. Our approach looks at relaxations of the above problems, where in addition to the staffing and cross-selling decisions, the manager can also select the waiting times experienced by its callers, which in reality are random variables that depend on the system dynamics. These relaxations are tractable, deterministic optimization problems that have insightful solutions and give rise to near-optimal heuristics. Focusing on model (b) (cf. Figure 1) first, §3.1 studies the waiting cost formulation of (1). These results are extended to the constrained formulation of (2) in §3.2, whereas §3.3 extends our work to model (c), where the customer types are observable upon arrival. All proofs are relegated to the online appendix.

3.1. The Waiting Cost Formulation

Throughout this section, we focus on model (b) and the waiting cost formulation (1).

Deterministic Relaxation. Starting with (1), we formulate the following linear program:

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^K \lambda_i r_i(w_i) x_i - c \cdot R(1+z) - h \sum_{i=1}^K \lambda_i w_i \\ \text{s.t.} \quad & x_i \leq q_i(w_i), \quad \sum_{i=1}^K \frac{\lambda_i x_i}{\mu_i^{cs}} \leq Rz, \\ & z \geq 0, \quad x_i \geq 0, \quad w_i \geq 0, \\ & \text{for all } i = 1, \dots, K, \end{aligned} \quad (3)$$

where x_i is interpreted as the fraction of class- i customers that are being asked *and* agree to listen to a cross-selling offer; w_i is the “fictitious” waiting time experienced by class- i customers in this formulation; and z is the excess (normalized) staffing level beyond the nominal requirement of the offered load R ($:= \Lambda/\mu^s$) as a fraction of R . The condition $z \geq 0$ implies that the staffing level is sufficiently large to handle all basic service requests (i.e., $N \geq R$). The name “deterministic relaxation” comes with a slight abuse of terminology. As to whether or not this is indeed a relaxation for (1)—the answer to this question depends on the actual form of the function $q_i(\cdot)$ and, more specifically, on its concavity or lack thereof. It is a matter of a simple observation, however, that any optimal solution to (3) will have $w_i = 0$ for all i and, consequently, that an optimal solution to (3) is necessarily an upper bound for any optimal solutions to (1) if such solutions exist. Hence, we choose to refer to (3) as the deterministic relaxation.

Recall the labeling convention in Assumption 1. Denoting the optimal solution to the knapsack problem in (3) with an overbar, we have the following: set $\bar{w}_i = 0$ for all $i = 1, \dots, K$,

$$\bar{x}_i = \begin{cases} q_i & i \leq \bar{k}, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad \bar{z} = \sum_{i=1}^{\bar{k}} \frac{\lambda_i q_i}{R \mu_i^{cs}}, \quad (4)$$

where $\bar{k} = \max\{i: r_i - c/\mu_i^{cs} \geq 0, q_i(0) > 0\}$. In fact, we will assume throughout that $r_{\bar{k}} - c/\mu_{\bar{k}}^{cs} > 0$, which is equivalent to assuming that the deterministic relaxation has a unique solution. In the presence of multiple solutions to the deterministic relaxation, our approach might lead to multiple asymptotically optimal solutions. By Assumption 1, \bar{z} is guaranteed to be strictly positive. The resulting staffing level is $R + \sum_{i=1}^{\bar{k}} \lambda_i q_i / \mu_i^{cs}$. Note that the structure of the deterministic relaxation is such that as long as λ_i/Λ is known and is kept constant (which we will assume henceforth), the normalized quantities \bar{x} , \bar{z} do not change with Λ . Therefore, the relevant profit depends on the entire vector $\lambda_1, \dots, \lambda_K$ through their sum Λ only. Specifically, the profit rate associated with solution (4) is

$$\begin{aligned} \bar{\Pi}(\Lambda) &= -cR + \sum_{i=1}^{\bar{k}} \lambda_i q_i (r_i - c/\mu_i^{cs}) \\ &= -cR + \sum_{i=1}^{\bar{k}} \lambda_i q_i [(r_i - c/\mu_i^{cs}) \vee 0], \end{aligned} \quad (5)$$

which is an upper bound for the optimal profit in (1). (Here and elsewhere $x \vee y = \max\{x, y\}$.)

A Staffing and Cross-Selling Proposal. The nested structure of (4) is intuitive: we cross-sell to all types i for which their marginal revenue contribution, $\lambda_i r_i q_i$, exceeds the increase in staffing cost, $c \lambda_i q_i / \mu_i^{cs}$, resulting from the additional cross-selling workload; this reduces to the condition $r_i - c / \mu_i^{cs} > 0$. The solution to the deterministic relaxation suggests the following pair of policies for the original stochastic system:

(S) *Staffing*: Staff with $N = R(1 + \bar{z})$.

(C) *Cross-selling*: Given a sequence of thresholds $\eta_{\bar{k}} \leq \eta_{\bar{k}-1} \leq \dots \leq \eta_1$: cross-sell to a customer of type $i \leq \bar{k}$ that completes service at time t if and only if $Q(t) < \eta_i$.

The cross-selling policy (C) follows the solution of the deterministic relaxation when the queue length is modest, and then starts to reduce the amount of cross-selling activity as the system gets increasingly congested. The asymptotic performance analysis that will follow does not use the precise values of the above thresholds, and in fact only makes use of the smallest threshold $\eta_{\bar{k}}$. Consequently, one may prefer to use a simpler policy that uses only this smallest threshold $\eta_{\bar{k}}$. This single-threshold policy always cross-sells to classes $1, \dots, \bar{k} - 1$ and stops cross-selling to class \bar{k} when the queue length exceeds the threshold. In our setting, in which the arrival rates, λ_i , are known and stationary, this single-threshold policy will be asymptotically equivalent to (C) in terms of the profits it generates. Still, we choose to present the results for the more elaborate control (C). We motivate the use of multiple thresholds in a nonstationary environment in §7.

Asymptotic Optimality of (S)-(C). Despite its simple structure, (S)-(C) performs very well in the stochastic system under consideration, and is, in fact, asymptotically optimal in large-scale systems, i.e., where Λ is large. As a starting point, we will establish that the system is always stable under (S)-(C) and that it admits a unique stationary distribution. We do that by showing the stronger result that the system will be stable under (C) as long as $N > R$, even if $N < R(1 + \bar{z})$.

PROPOSITION 1 (STABILITY). *Fix Λ and assume that (C) is used for some set of thresholds: $\eta_{\bar{k}} \leq \eta_{\bar{k}-1} \leq \dots \leq \eta_1 \leq \infty$. Then, $N > R$ is a sufficient condition for stability. Moreover, for any $N > R$, the underlying Markov process admits a unique stationary distribution that is also its limiting distribution.*

This proposition illustrates the self-stabilizing nature of the cross-selling system. Note that the use of thresholds η_i is not necessary for this result to hold. Indeed, they may all be set equal to ∞ ; the stabilizing force stems from the delay sensitivity of the customers. Intuitively, when the system is heavily loaded, the queue and the resulting waiting time will grow large. In turn, fewer customers will agree to listen to cross-selling offers, thus reducing the load.

The remainder of this subsection will characterize the asymptotic performance of the original stochastic call center system under (S)-(C) in settings with large call volumes, as measured by Λ . One naturally expects that with a threshold policy, the best threshold values will be a function of the system size and in particular of Λ , the overall arrival rate. Let $\eta_{\bar{k}}^\Lambda, \dots, \eta_1^\Lambda$ be the threshold values corresponding to a system with arrival rate Λ . Then, we will show in our subsequent results that, indeed, there is a dependence of the threshold values on the system size and, moreover, that asymptotically optimal performance implies that these threshold values scale according to

$$\eta_i^\Lambda = \hat{\eta}_i \sqrt{\Lambda} \quad \text{for } i = 1, \dots, \bar{k} \quad (6)$$

and appropriate constants $\hat{\eta}_{\bar{k}} \leq \dots \leq \hat{\eta}_1$. Let $N^*(\Lambda)$, $x_i^*(\Lambda)$, and $\Pi^*(\Lambda)$ denote the (unknown) optimal staffing level, realized long-run average cross-selling rates, and the corresponding profit rate for (1), respectively, when the aggregate demand is Λ . Also, let $\hat{\Pi}(\Lambda)$ be the profit obtained when using (S)-(C) in the stochastic system.

In the sequel, we will make use of the following notation: for two positive sequences we say that x^Λ is $o(y^\Lambda)$ if $x^\Lambda / y^\Lambda \rightarrow 0$ as $\Lambda \rightarrow \infty$.

THEOREM 1 (ASYMPTOTIC OPTIMALITY). *Let Λ grow large, keeping λ_i / Λ constant for all i . Then, with thresholds satisfying (6), (S)-(C) is asymptotically optimal in the sense that*

$$\hat{\Pi}(\Lambda) = \Pi^*(\Lambda) - o(\Lambda). \quad (7)$$

Alternatively, one could write (7) in the form $\hat{\Pi}(\Lambda) / \Pi^*(\Lambda) \rightarrow 1$ as $\Lambda \rightarrow \infty$. The proof of the above result follows by showing the stronger result that $\hat{\Pi}(\Lambda)$ approaches $\bar{\Pi}(\Lambda)$, which itself is an upper bound for $\Pi^*(\Lambda)$. Because $\Pi^*(\Lambda)$ is sandwiched between $\hat{\Pi}(\Lambda)$ and $\bar{\Pi}(\Lambda)$, it must also be close to $\bar{\Pi}(\Lambda)$. This leads to a partial characterization of the unknown optimal policy in large-scale systems.

THEOREM 2 (ESTIMATES OF THE OPTIMAL SOLUTION). *Let Λ grow large, keeping λ_i / Λ constant for all i . Then, (a) $\Pi^*(\Lambda) = \bar{\Pi}(\Lambda) - o(\Lambda)$, (b) $N^*(\Lambda) = R(1 + \bar{z}) \pm o(\Lambda)$, and (c) $x_i^*(\Lambda) = \bar{x}_i + o(1)$.*

Theorems 1 and 2 together demonstrate how the solution of the deterministic relaxation captures the first-order behavior of the optimal policy for (1), both in terms of its staffing and cross-selling decisions as well as its resulting profits. A key component of the asymptotic optimality proof is the next lemma that shows that if the thresholds η are of order $\sqrt{\Lambda}$ (as in (6)), then the steady-state waiting times that characterize the system are of order $1/\sqrt{\Lambda}$ and in particular of order $o(1)$; this is the nominal time it takes an order Λ servers to clear a queue length of order $\sqrt{\Lambda}$. Thresholds of smaller magnitudes would result in even smaller waiting times.

LEMMA 1. Let Λ grow large, keeping λ_i/Λ constant for all i . Denote by $E[W^\Lambda]$ the steady-state expected waiting time under policy (S)-(C). Then, with thresholds satisfying (6), $E[W^\Lambda] = O(1/\sqrt{\Lambda})$, or equivalently, $\limsup_{\Lambda \rightarrow \infty} \sqrt{\Lambda}E[W^\Lambda] < \infty$. In particular, $E[W^\Lambda] \rightarrow 0$ as $\Lambda \rightarrow \infty$.

The next lemma then shows that, actually, it would be always optimal to staff and cross-sell so that the waiting times are very small. We denote by $E[W^{\Lambda,*}]$ the expected steady-state waiting time under the optimal control $(N^*(\Lambda), x^*(\Lambda))$.

LEMMA 2. Let Λ grow large, keeping λ_i/Λ constant for all i . If an optimal policy $(N^*(\Lambda), x^*(\Lambda))$ exists for all Λ large enough, then $\limsup_{\Lambda \rightarrow \infty} E[W^{\Lambda,*}] = 0$.

REMARK 1 (STRENGTHENING THE NOTION OF ASYMPTOTIC OPTIMALITY). The main technical problem in proving Theorems 1 and 2 lies in the so-called *limit interchange problem*. Specifically, although it might be relatively simple to get performance guarantees on finite time intervals, it is much harder to characterize the asymptotic performance, as $\Lambda \rightarrow \infty$, of the system’s steady state. The technical arguments in that respect are quite complex, as the online appendix illustrates. The interested reader is referred to part B of the online appendix for a further discussion of the underlying complexities. Consequently, refining the performance bounds by showing, for example, an $O(\sqrt{\Lambda})$ deviation from optimality, is complicated even in much simpler settings than the system we consider—especially when one wants to establish convergence of moments.

REMARK 2 (CHOOSING THE THRESHOLD VALUES). For the cost formulation, the values of the thresholds η_i can be selected via simulation. In most call centers, however, the constrained formulation (considered in the next section) is more natural. Fortunately, for the constrained formulation we have a very simple rule to determine the threshold value.

3.2. The Constrained Formulation

Lemmas 1 and 2 illustrate that the waiting times experienced in an optimally controlled call center will be of order $o(1)$. With that in mind, a waiting time constraint of the form $E[W^\Lambda] \leq \bar{W}$ will become irrelevant as Λ grows large because the actual waiting times will be much smaller than the desired target \bar{W} . A more appropriate formulation that is meaningful as Λ grows large replaces the upper-bound constraint by a quantity that itself changes with Λ such as $\bar{W}^\Lambda = \hat{W}/\sqrt{\Lambda}$ for an appropriate choice of \hat{W} .⁴ This would result in the following problem:

$$\sup_{N \in \mathbb{Z}_+, \pi \in \mathcal{M}(N)} \left\{ \sum_{i=1}^K \lambda_i \cdot (r_i x_i)(\pi) - cN : EW^\pi(\infty) \leq \bar{W}^\Lambda \right\}, \quad (8)$$

where $\bar{W}^\Lambda = \hat{W}/\sqrt{\Lambda}$ for an appropriate choice of \hat{W} . Along the lines of (3), the following is a *deterministic relaxation* of (8):

$$\begin{aligned} &\text{maximize} \quad \sum_{i=1}^K \lambda_i r_i x_i - c \cdot R(1+z) \\ &\text{s.t.} \quad \sum_{i=1}^K \frac{\lambda_i}{\Lambda} w_i \leq \bar{W}^\Lambda, \\ &\quad \quad x_i \leq q_i(w_i), \quad \sum_{i=1}^K \frac{\lambda_i x_i}{\mu_i^{cs}} \leq Rz, \\ &\quad \quad x_i \geq 0, \quad w_i \geq 0, \quad \text{for all } i = 1, \dots, K. \end{aligned} \quad (9)$$

The linear program described above has the same optimal solution as (3), making our solution insensitive to the precise articulation of the effect of customer waiting times. The resulting staffing and cross-selling heuristics are again the ones described by (S)-(C) in the previous subsection. In the case of the constrained formulation, one can also get a crude estimate for the threshold η_k to be $\eta_k := \Lambda \bar{W}^\Lambda$, which is consistent with (6). Intuitively, if the queue length is maintained below that threshold, then by a heuristic application of Little’s law, one would expect the waiting times to be below \bar{W}^Λ . The next theorem establishes this result in an asymptotic sense as Λ grows large. With a slight abuse of notation, we use $\hat{\Pi}(\Lambda)$ and $\Pi^*(\Lambda)$ to denote the profit rate for the constrained formulation under (S)-(C) and the optimal policy, respectively.

THEOREM 3 (ASYMPTOTIC OPTIMALITY). Let Λ grow large, keeping λ_i/Λ constant for all i . Then, with thresholds satisfying (6), and such that $\eta_k^\Lambda = \Lambda \bar{W}^\Lambda$, (a) $\hat{\Pi}(\Lambda) = \Pi^*(\Lambda) + o(\Lambda)$ and (b) $E[W^\Lambda] \leq \bar{W}^\Lambda + o(\bar{W}^\Lambda)$.

Theorem 3 shows that the waiting time constraint will be violated only by a negligible amount if one sets $\eta_k := \Lambda \bar{W}^\Lambda$. Of course, if one is interested in strict satisfaction of the threshold, one may start with the recommended threshold and fine-tune it in real time with small perturbations around the recommended value.

3.3. The Value of Customer Type Identification Upon Arrival

We complete the analysis of the model with observable types by comparing the model analyzed thus far (model (b) in Figure 1) with the one where the type of each customer is observed at the time of his arrival to the system (model (c)). The latter could be achieved by requiring callers to identify themselves through a PIN or an account number.

Routing Capability. Once the call center observes the type of each arriving customer, it can maintain different (virtual) queues for customers of each type, and use that added flexibility in routing calls to available agents. This will eventually trade off the delay sensitivity and waiting time cost of each type against its potential revenue

contribution. It is clear that this added element of control can only improve the call center's profitability. The question is by how much. The main result of this section shows that the performance difference between FCFS routing (used when types are unobservable upon arrival) and any other routing policy that makes use of the type information, including the optimal one, is small and asymptotically negligible. The crude asymptotic analysis of this subsection uses a sandwich argument, similar to the one applied in Theorem 2, and does not need a detailed articulation of the set of admissible routing policies. We refer the reader to Bassamboo et al. (2006) for one possible definition of these controls.

We henceforth drop the distinction between the waiting cost and constrained formulations. The results in the remainder of this section as well as those in §§4 and 5 hold for both formulations. Let $\Pi^{**}(\Lambda)$ be the optimal achievable profit for the system where customer types are observable upon their arrival, and note that $\Pi^{**}(\Lambda) \geq \Pi^*(\Lambda)$. The key to our analysis is that the deterministic relaxations for models (b) and (c) are identical. The routing capability of model (c) can only serve to improve the vector of expected waiting times $E[W_i]$. Because the relaxation treats these as free optimization variables, denoted by w_i , and sets them equal to zero, its solution will coincide with that of (3). It follows that $\bar{\Pi}(\Lambda) \geq \Pi^{**}(\Lambda) \geq \Pi^*(\Lambda)$. From Theorem 2 we have that $\Pi^*(\Lambda) = \bar{\Pi}(\Lambda) - o(\Lambda)$, which leads to the following conclusion:

PROPOSITION 2. *Let Λ grow large, keeping λ_i/Λ constant for all i . Then, $\Pi^{**}(\Lambda) - \Pi^*(\Lambda) = o(\Lambda)$.*

Therefore, although routing control capability may improve the quality of service enjoyed by some types and potentially simultaneously increase the revenue extracted from them, it will not lead to a significant overall profit gain. Moreover, the asymptotically optimal staffing and cross-selling recommendations that emerge from our analysis are insensitive (up to first order) to the use of this information.

The question that arises is whether segmentation at the cross-selling stage leads to significantly different results in comparison to no segmentation at all. To address this question, we first study the issue of type-dependent price customization in §4, and then assess the value of customer segmentation in §5.

4. The Price Customization Problem

Customer segmentation in a call center setting allows firms to customize their products to better match the characteristics of each customer type and extract higher revenues. In our model, the product offered to all customers is assumed to be the same, but the firm can customize the price quoted to each customer type. In this section, we show that the optimal prices can be computed separately from the operational decisions of staffing and cross-selling.

Towards this end, note that due to the dependence of the willingness to pay on the waiting times of customers, one expects the true optimal pricing mechanism to be a dynamic one that takes into account these realized waiting times. Hence, the pricing mechanism should be regarded as a mapping from waiting times to prices. Specifically, we assume that prices may assume values in the space $\mathcal{P} = \mathcal{P}_1 \otimes \mathcal{P}_2 \otimes \cdots \otimes \mathcal{P}_K$, where for $i = 1, \dots, K$, \mathcal{P}_i is assumed to be a compact interval in \mathbb{R}_+ . The pricing mechanism is then a function $p(\cdot) = (p_1(\cdot), \dots, p_K(\cdot)): \mathbb{R}_+ \mapsto \mathcal{P}$; we let \mathcal{S} be the space of these functions. Accordingly, we expand the notation used earlier to let $\Pi^*(\Lambda; p(\cdot))$ and $N^*(\Lambda; p(\cdot))$ be the optimal profit rate and staffing level, respectively, for (1) for a given Λ and pricing function $p(w)$. We then redefine $\Pi^*(\Lambda) := \sup_{p(\cdot) \in \mathcal{S}} \Pi^*(\Lambda, p)$ to be the optimal achievable profit rate when the call center is allowed to optimize over its price function over the set \mathcal{S} . Let $p^*(\Lambda) := p^*(\Lambda)(\cdot)$ be the optimal price function, which is assumed to exist, and $N^*(\Lambda)$ the corresponding staffing level. We also let $\bar{\Pi}(\Lambda, p)$ be the profit rate achieved in the deterministic relaxation of (3) for a given constant value of p , $\bar{\Pi}(\Lambda) := \max_{p \in \mathcal{P}} \bar{\Pi}(\Lambda, p)$ be the profit rate when optimizing over the price, and let \bar{p} denote the corresponding optimizer, which will most likely be different than the function $p^*(\Lambda)$. Whereas identifying $p^*(\Lambda)$ is hard, the deterministic price vector \bar{p} is easy to characterize by rewriting the objective function as

$$\bar{\Pi}(\Lambda, p) = -cR + \sum_{i=1}^K \lambda_i q_i [(r_i(p_i) - c/\mu_i^{cs}) \vee 0], \quad (10)$$

where $r_i(p_i) = p_i \bar{F}_i(p_i, 0)$; this expression reflects the fact that the center only cross-sells to and receives revenue from types for which $r_i(p_i) \geq c/\mu_i^{cs}$, and that it staffs accordingly. It follows that the corresponding optimal price in (10) is static (waiting time independent) and satisfies

$$\bar{p}_i = \arg \max_{p_i \in \mathcal{P}_i} p_i \bar{F}_i(p_i, 0), \quad (11)$$

and $\bar{\Pi}(\Lambda) = -cR + \sum_{i=1}^K \lambda_i q_i [(r_i(\bar{p}_i) - c/\mu_i^{cs}) \vee 0] = \bar{\Pi}(\Lambda, \bar{p})$. The corresponding staffing level is $R(1 + \bar{z}(\bar{p}))$, where

$$\bar{z}(\bar{p}) = \sum_{i=1}^{\bar{k}(\bar{p})} \frac{\lambda_i q_i}{R \mu_i^{cs}} \quad \text{and} \quad (12)$$

$$\bar{k}(\bar{p}) = \max\{i: r_i(\bar{p}_i) \geq c/\mu_i^{cs}\};$$

the above expressions assume w.l.o.g that types are relabelled so that $r_1(\bar{p}_1) \geq \dots \geq r_K(\bar{p}_K)$. We also assume that $r_1(\bar{p}_1) > c/\mu_1^{cs}$ and that $r_{\bar{k}(\bar{p})}(\bar{p}_{\bar{k}(\bar{p})}) > c/\mu_{\bar{k}(\bar{p})}^{cs}$, which guarantee, respectively, that Assumption 1 holds and that the solution of the deterministic relaxation given \bar{p} is unique. It is straightforward to show that \bar{p} , $\bar{z}(\bar{p})$, and $\bar{k}(\bar{p})$ jointly characterize the optimal solution of the deterministic relaxation, and that this solution does not change if one were

to scale Λ large, while keeping λ_i/Λ constant (this is the asymptotic setup adopted thus far). Note that although \bar{p} may be different than $p^*(\Lambda)$, $\bar{\Pi}(\Lambda, \bar{p})$ is still an upper bound for $\Pi^*(\Lambda, p^*(\Lambda))$. Using this observation and applying Theorem 2 (with the fixed-price vector \bar{p}), we find the following:

PROPOSITION 3. Define \bar{p} , $\bar{z}(\bar{p})$ through (11) and (12), respectively. Let Λ grow large, keeping λ_i/Λ constant for all i . Then: (a) $\Pi^*(\Lambda, p^*) = \bar{\Pi}(\Lambda, \bar{p}) - o(\Lambda)$, (b) $N^*(\Lambda, p^*) = R(1 + \bar{z}(\bar{p})) \pm o(\Lambda)$, and (c) $p^*(\Lambda)(0) = \bar{p} + o(1)$.

Consequently, we recommend adding the static price vector \bar{p} to the staffing and cross-selling rules proposed in §3. By Theorem 1 and Proposition 3 above, the resulting joint pricing, staffing, and cross-selling solution is asymptotically optimal for the original stochastic system.

Decoupling of Pricing and Staffing. An important consequence of the above result is that the pricing decisions can be made independently of the operational ones of staffing and cross-selling. This insight is valid in the system where types are observed upon arrival (model (c)), as well as in settings where products are customized along other nonprice attributes that do not involve capacity and quality-of-service specifications. This decoupling trivially follows in settings where the perceived cost of a product is independent of the waiting time encountered by the customer, but need not be true in the more general model considered in our paper. Moreover, because the waiting time of the customer is known to the agent, the center may want to invoke a dynamic pricing policy to optimize the expected revenue per customer. The fact that a static pricing policy is shown to perform very close to optimal is an appealing characteristic of our solution that allows the system manager to make the pricing and operational decision in a hierarchical sequence.

5. The Effect of Customer Segmentation

This section compares the profitability and behavior of the system studied in §§3 and 4 against one that does not use a segmentation mechanism and instead treats its entire customer pool as one segment. The latter is offered a common product, i.e., at the same price, and cross-selling decisions are made without the customer type information; this is model (a) in Figure 1.

A System with No Customer Segmentation. The characteristics of this combined segment are a single delay sensitivity function $q(\cdot)$ and a corresponding willingness-to-pay distribution $F(\cdot)$ that are appropriate mixtures of the corresponding quantities for the various types. The delay sensitivity function, $q(w)$, is given by

$$q(w) := \sum_{i=1}^K \frac{\lambda_i}{\Lambda} q_i(w).$$

The mean cross-selling time for the combined segment is estimated by

$$\frac{1}{\mu^{cs}} = \sum_{i=1}^K \frac{\lambda_i q_i}{\sum_{j=1}^K \lambda_j q_j} \cdot \frac{1}{\mu_i^{cs}}.$$

This is a reasonably precise estimate assuming that the waiting times are small. Moreover, the comparison result in Proposition 4 below holds when one uses a more precise estimate that takes into account the waiting times. The combined willingness-to-pay distribution F is computed indirectly as follows. Let $F(p, w)$ be equal to the probability that the willingness-to-pay of a customer that agreed to listen to the cross-selling offer after a waiting time of w time units is less than or equal to p . Then, $q(w)$ and $\bar{F}(p, w)$ satisfy the following intuitive relation

$$q(w)\bar{F}(p, w) = \sum_{i=1}^K \frac{\lambda_i}{\Lambda} q_i(w)\bar{F}_i(p, w),$$

from which we can solve for $F(p, w)$.

The deterministic relaxation for the combined segment is now easy to solve by specializing the results of §3 to a single segment with characteristics $q(w)$ and $F(p, w)$. Specifically, it is again optimal to set $w = 0$, which together with (10) gives the following objective:

$$\bar{\Pi}^a(\Lambda, p) := -cR + \Lambda q(p\bar{F}(p, 0) - c/\mu^{cs}) \vee 0, \tag{13}$$

where the superscript “ a ” is meant to associate this expression to model (a), and

$$q := \sum_{i=1}^K \frac{\lambda_i}{\Lambda} q_i \quad \text{and} \quad \bar{F}(p, 0) := \sum_{i=1}^K \frac{\lambda_i q_i}{\sum_{j=1}^K \lambda_j q_j} \bar{F}_i(p, 0). \tag{14}$$

As shown in §4, one can study this deterministic formulation by separately optimizing over the price p , and then considering the resulting staffing and cross-selling problem at that price.

The pricing decision. The optimal price that the call center should use in this deterministic relaxation is given by the solution to the following problem:

$$\max_{p \in \bar{\mathcal{P}}} p\bar{F}(p, 0), \tag{15}$$

which we denote by \bar{p}^a , and let $r^a = \bar{p}^a \bar{F}(\bar{p}^a, 0)$ and $\bar{\mathcal{P}} = \mathcal{P}_1 \cap \mathcal{P}_2 \cap \dots \cap \mathcal{P}_K$. Note that despite our assumptions regarding the unimodality of $p_i \bar{F}_i(p_i, 0)$, $p\bar{F}(p, 0)$ need not be unimodal itself. However, one can always find its optimizer through a single-parameter search (assuming that the set $\bar{\mathcal{P}}$ is not empty).

The staffing and cross-selling decisions. Plugging \bar{p}^a into (13) and using the results of §3, the solution of the deterministic relaxation can be divided into two cases:

Case i. If $r^a \geq c/\mu^{cs}$: the call center cross-sells to all customers and staffs with $R_{\max} := R(1 + \bar{z}^a)$ servers, where $\bar{z}^a = \Lambda q/(R\mu^{cs})$.

Case ii. If $r^a < c/\mu^{cs}$: the call center will *not* cross-sell to any customer and staff with R servers. Using (13) and (14), the resulting profit rate in the deterministic relaxation is given by

$$\bar{\Pi}^a(\Lambda) := \begin{cases} -cR + \sum_{i=1}^K \lambda_i q_i (\bar{p}^a \bar{F}_i(\bar{p}^a, 0) - c/\mu_i^{cs}) & \text{if } r^a > c/\mu^{cs}, \\ -cR & \text{otherwise,} \end{cases} \quad (16)$$

which is again an upper bound for the optimal profit rate of the stochastic call center system.

As in §3, the natural implementation of the above policies in case i would be to cross-sell as long as the queue is below an appropriate threshold that serves to limit excessive delays. In case ii, the system may still elect to cross-sell, but only if either the queue is very small or there are a sufficient number of agents that are idle. Moreover, in that case the staffing level should be inflated to $R + x\sqrt{R}$ for an appropriate constant x to provide moderate delays. The asymptotic analysis of §3 does apply to the single-segment model when the solution of the deterministic relaxation falls into case i, but it does not cover case ii, where the system exercises negligible cross-selling. That case was studied in detail in Armony and Gurvich (2006) and will not be further reviewed here.

The Effect of Customer Segmentation. The key differences between the two systems, with and without segmentation, are best illustrated through their respective deterministic relaxations, which are simple and accurate, in the sense that they capture the structure of the underlying optimal policies and their resulting performance asymptotically.

1. *Cross-selling (all-or-none versus selected types).* For both models, the call center will do significant cross-selling only if the expected revenue from doing so exceeds the capacity cost involved in that activity. With no segmentation capability in place, the system will either choose to cross-sell to all of its callers if $r^a \geq c/\mu^{cs}$, or to none. In the first case, this may involve cross-selling to customer segments to which it is strictly unprofitable to do so, whereas in the second case, it involves forgoing profitable cross-selling opportunities that cannot be singled out from the larger pool of callers (the latter follows from Assumption 1). Using customer segmentation, the system will only cross-sell to types $i = 1, \dots, \bar{k}$ for which $r_i(\bar{p}_i) \geq c/\mu_i^{cs}$, i.e., for which cross-selling is profitable. Finally, we note that although Assumption 1 guarantees that the call center will always choose to cross-sell to some subset of the customer types, if these can be segmented out, it does not guarantee that it is profitable to do so in a system with no segmentation capability.

2. *Staffing.* The model with no segmentation will either staff with $R_{\max} = R(1 + \bar{z}^a)$ or $R + x\sqrt{R}$ servers, depending on whether it will cross-sell or not. In contrast, the model

with segmentation will staff with $R(1 + \bar{z})$ servers; $\bar{z} < \bar{z}^a$, unless it is profitable to cross-sell to all customer types.

3. *Uniform versus customized pricing and profitability.* Most structural differences between the two systems originate from the pricing policies adopted by the call center in each case, and the corresponding expected revenue that they will generate per customer that agrees to enter the cross-selling phase. As explained earlier, the system that segments its customers will customize its prices, \bar{p}_i , for each type according to (11), whereas the system with no segmentation will use a uniform price, \bar{p}^a , defined through (15). An immediate consequence of the above is that

$$r^a = \sum_{i=1}^K \frac{\lambda_i q_i}{\sum_{j=1}^K \lambda_j q_j} \bar{p}^a \bar{F}_i(\bar{p}^a, 0) \leq \sum_{i=1}^K \frac{\lambda_i q_i}{\sum_{j=1}^K \lambda_j q_j} \bar{p}_i \bar{F}_i(\bar{p}_i, 0).$$

Premultiplying by $\sum_{j=1}^K \lambda_j q_j$ and subtracting out the corresponding capacity cost, we find that

$$\begin{aligned} \left(\sum_{j=1}^K \lambda_j q_j \right) (r^a - c/\mu^{cs}) &\leq \sum_{i=1}^K \lambda_i q_i (r_i(\bar{p}_i) - c/\mu_i^{cs}) \\ &\leq \sum_{i=1}^K \lambda_i q_i (r_i(\bar{p}_i) - c/\mu_i^{cs})^+. \end{aligned}$$

The right-hand side (RHS) of the above expression is equal to the profit contribution due to cross-selling in the system with segmentation, which is clearly nonnegative. This allows us to strengthen this inequality to the following:

$$\left(\sum_{j=1}^K \lambda_j q_j \right) (r^a - c/\mu^{cs})^+ \leq \sum_{i=1}^K \lambda_i q_i (r_i(\bar{p}_i) - c/\mu_i^{cs})^+, \quad (17)$$

where, in turn, the left-hand side (LHS) of (17) is the profit contribution due to cross-selling in the system with no segmentation. The above inequality is strict as long as there exists a type i for which $\bar{p}^a \bar{F}_i(\bar{p}^a, 0) < \bar{p}_i \bar{F}_i(\bar{p}_i)$, which by the definition of \bar{p}_i and the unimodality of $p\bar{F}_i(p, 0)$, reduces to

$$\exists i \in \{1, \dots, K\} \quad \text{for which } \bar{p}_i \neq \bar{p}^a, \quad (18)$$

or equivalently, to

$$\exists i, j \in \{1, \dots, K\} \quad \text{such that } \bar{p}_i \neq \bar{p}_j. \quad (19)$$

Unless customer types have trivial differences with respect to their willingness to pay, conditions (18) or (19) are likely to be satisfied, in which case the ability to segment the customer pool would lead to significant profit gains. For example, if the willingness-to-pay distributions for the various types were exponential with parameters b_i , then the above conditions would require that at least two of these types had different parameters $b_i \neq b_j$. If the distributions were logistic with scale parameters b_i (these are commonly used in the literature in modelling different customer segments), then again (18) would require that the parameters of at least two segments are different. A simple extension of our previous results yields the following characterization of the potential value of customer segmentation in the underlying stochastic call center systems.

PROPOSITION 4. Under Assumption 1, if (18) (or equivalently (19)) holds, then for all Λ , $\bar{\Pi}(\Lambda) - \bar{\Pi}^a(\Lambda) = \delta\Lambda$, where δ is the difference of the RHS and LHS of (17) normalized by Λ . Moreover, if we let Λ grow large, keeping λ_i/Λ constant for all i , then

$$\Pi^*(\Lambda) - \Pi^{*,a}(\Lambda) = \delta\Lambda + o(\Lambda),$$

where $\Pi^*(\Lambda)$, $\Pi^{*,a}(\Lambda)$ are the optimal expected profit rates for the underlying stochastic systems with and without segmentation, respectively.

The above proposition together with the results of Theorems 1 and 3 suggest that the staffing and cross-selling policies proposed in this paper would realize most of the profit differential that can be attributed to customer segmentation. Operationally, the latter also leads to more efficient capacity utilization because call centers that do not segment their callers, but try to cross-sell to them, end up pursuing too many customer prospects that are unlikely to lead to a sale. Our stylized yet insightful analysis can be used to assess the magnitude of this potential benefit, which is useful in deciding the value proposition of an investment in technology and agent training that would be needed to support a sophisticated customer segmentation and cross-selling strategy.

6. Numerical Results

Our results are organized in three categories. The first offers a representative numerical illustration of the accuracy of our asymptotic analysis. The second examines the quality of the proposed policies, and in particular shows the sensitivity of the system performance to changes in staffing and threshold levels that are used in the cross-selling decisions. The last one gives some examples of the potential value of using customer segmentation in such a call center. For simplicity, we assume throughout this section that $\mu_i^{cs} = \mu^{cs}$ for all i .

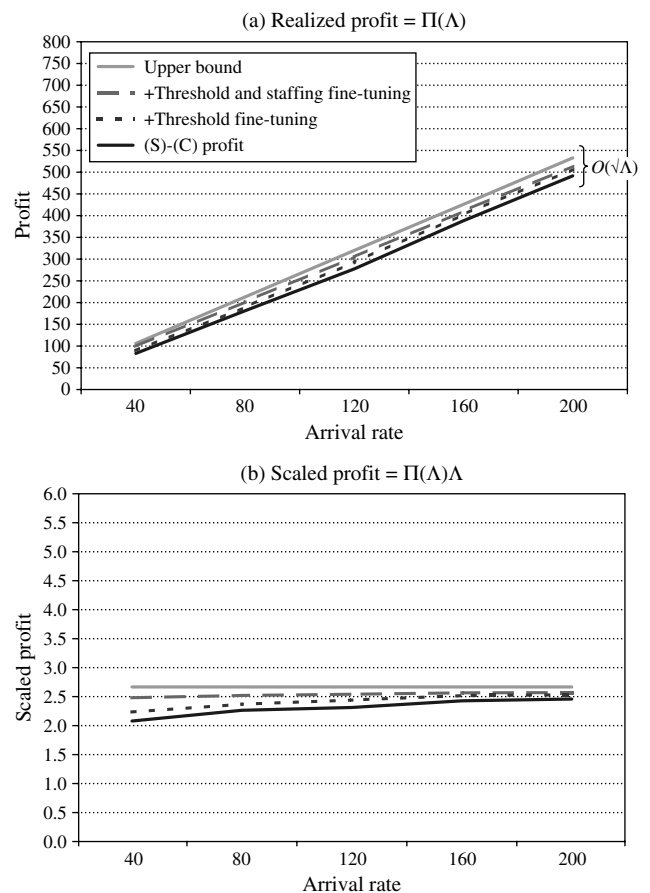
The Accuracy of Large-Scale Asymptotics. We illustrate the accuracy of the proposed (S)-(C) heuristic by experimenting on a system with four customer classes. The service rates are $\mu^s = 1$ and $\mu^{cs} = 2$; one may regard all subsequent parameters as normalized with respect to μ^s . The arrival rates are $\lambda_1 = \lambda_2 = \frac{1}{3}\Lambda$ and $\lambda_3 = \lambda_4 = \frac{1}{6}\Lambda$, whereas the aggregate arrival rate, Λ , will be varied over a range of values in our experiment. The product prices are exogenously given and result in expected revenues per type- i customer who goes through cross-selling, given by $r_1 = 7$, $r_2 = 5$, and $r_3 = r_4 = 0.4$, regardless of the realized waiting time.⁵ For simplicity, we assume that the customers' willingness-to-listen functions are common across types and given by the linear function $q_i(w) = [1 - 0.1w]^+$. The staffing cost is normalized to $c = 1$, and for concreteness we consider the constrained formulation with an upper bound for the waiting time equal to $1/6$; if the natural time

units are minutes, then this upper bound is 10 seconds. Under this choice of parameters, we have that $\bar{z} = \frac{1}{3} > 0$ and $\bar{k} = 2$, i.e., the center will cross-sell to types 1 and 2 only. These values of \bar{z} and \bar{k} and the above set of revenue and cost parameters give $\bar{\Pi}(\Lambda) = (2.67)\Lambda$ as an upper bound on the system's profit rate.

We have simulated the system behavior under three variants of the policy (S)-(C) for Λ ranging from 40 to 200. The first variant is a direct translation of the solution of the deterministic relaxation, with a threshold $\eta_2 = \lceil \frac{1}{6}\Lambda \rceil$ chosen according to the recommendation in §3.2; recall that type 2 is the least profitable type to which the system cross-sells. (For simplicity, we set $\eta_1 = \infty$, i.e., the system would always cross-sell to type 1 customers.) The other two policy variants had η_2 and the staffing level N further optimized via exhaustive simulation, i.e., by performing a search over all possible values of N . The simulation code was written in c++. Each sample path contained 800,000 customer arrivals from which we formed time averages of the queue length and of the fraction of customers of each type that were cross-sold to. The length of each simulated path ensured that our estimates were close to the actual steady-state behavior.

First, we note from Figure 2(a) that the absolute deviation between the profits achieved through the three

Figure 2. Performance of (S)-(C).



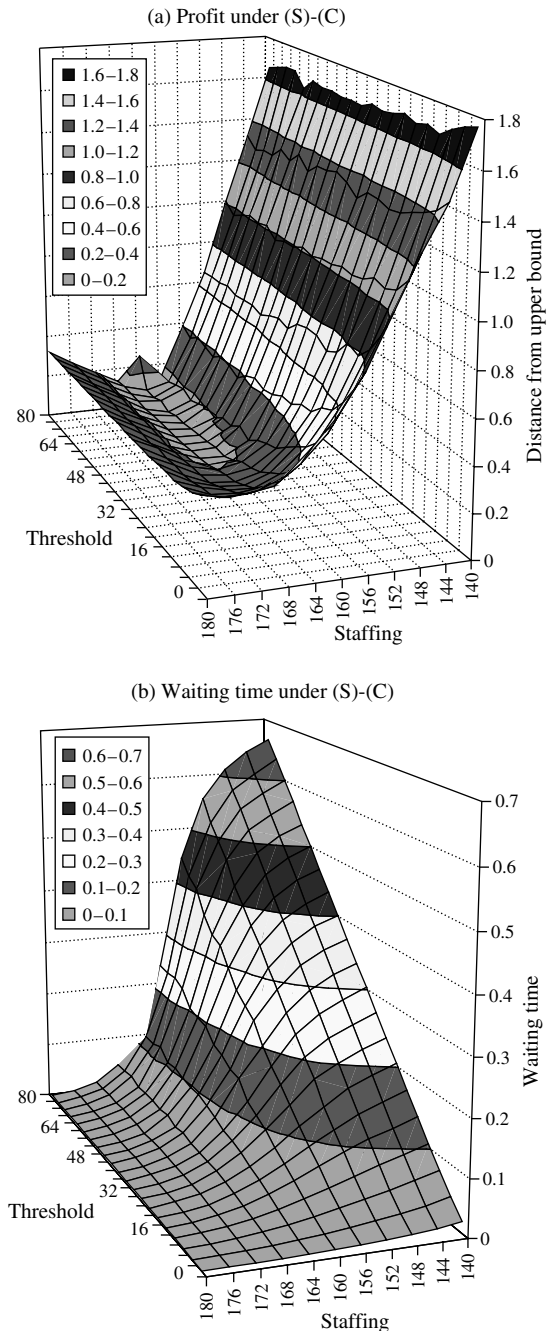
candidate policies as well as their difference against the deterministic upper bound increases with the scale of the system, as measured by the aggregate call volume Λ . However, Figure 2(b) illustrates that if normalized by Λ , which is the multiplicative factor by which the above quantities are growing, then the respective difference decays to zero. In fact, this decay is of order $1/\sqrt{\Lambda}$. The above findings are representative of many examples that we tested. Second, we observe that as the size of the system increases, most of the profit gains from fine-tuning the cross-selling threshold parameter and staffing level can be attributed to the former. This is practically appealing because it makes the model more robust to forecasting errors, because adjustments can be made online. The next set of results that we present studies this issue in more detail, and also reviews the waiting time constraint qualification.

Performance Sensitivity with Respect to the Cross-Selling Threshold and the Staffing Level. Figure 3 offers a more detailed look at the effect of these two parameters to the center’s profitability and the steady-state expected waiting time experienced by its callers for the system examined above for $\Lambda = 120$. The parameters extracted from the deterministic relaxation are $\bar{z} = 1/3$ and $\bar{k} = 2$, which would translate to a nominal staffing of $N = 160$ servers, and a nominal threshold of $\eta_2 = \Lambda \bar{W}^\Lambda = 20$; i.e., the center would stop cross-selling to type 2 customers when there are more than 20 customers in queue. Specifically, Figure 3(a) shows the distance between the realized profit and its upper bound for various values of η_2 and N . Figure 3(b) depicts the expected waiting time for each of these parameter combinations; the respective constraint requires that this falls below 0.167.

It is worth noting that the center’s profitability is fairly insensitive to the staffing level around its nominal value of 160 servers because the effect of the latter can be compensated for by appropriately adjusting the cross-selling threshold. As expected, the waiting time is decreasing in the staffing level and increasing in the value of the cross-selling threshold; i.e., more servers reduce the overall load, whereas higher thresholds imply that the system is willing to tolerate longer waiting times. In fact, as expected from an informal application of Little’s law, the expected waiting time increases almost linearly as a function of the threshold. The effect of the threshold on the profit is less significant, which is consistent with our asymptotic results that showed that (S)-(C) with practically any threshold level performs very close to the upper bound in large systems. Taken together, the above comments suggest that call centers of reasonably large size can use the nominal staffing level extracted through the deterministic analysis, and subsequently select the cross-selling threshold to achieve constraint qualification and improve profits.

The Value of Market Segmentation. We conclude this section through a set of numerical experiments that

Figure 3. Performance as function of staffing and threshold levels.



strive to illustrate the potential value of market segmentation. The analysis here is crude in the sense that it is limited to the deterministic relaxation. The asymptotic performance guarantees and the numerical results presented above suggest that the profit gap between the respective deterministic relaxations will persist in the stochastic systems as well. To facilitate the presentation of our results, we will mostly focus on a two-type system, for which $\mu^s = 1$, $\mu^{cs} = 2$, $c = 1$, $\Lambda = 100$, and $\lambda_1 = \lambda_2 = 0.5\Lambda$. The waiting cost d or the waiting time upper bound \bar{W} do not play any role

in the deterministic analysis, and hence there is no need to specify them.

It remains to specify the customer choice behavior. As in the previous examples, we assume that the delay preferences of both types are the same with $q_i(w) = [1 - 0.1w]^+$. Type- i customers are assumed to have an exponentially distributed willingness to pay with parameter b_i for which $\bar{F}_i(p_i) = e^{-b_i p_i}$, $i = 1, 2$. We assume that prices can obtain values on the bounded interval $[0, 20]$ in each case. For the system that segments its customers, the optimal prices are given by $\bar{p}_i = 1/b_i \wedge 20$ (where $x \wedge y = \min\{x, y\}$), for which $r_i(\bar{p}_i) = 1/b_i \wedge 20 e^{-(20b_i \wedge 1)}$. Note that the optimal price $1/b_i$ in the absence of the price bound of \$20 is equal to the average of the distribution F_i , and that $r_i(\bar{p}_i)$ is linear in $1/b_i$ as long as $b_i \geq 0.05$. The solution to the deterministic relaxation will cross-sell to type- i provided that $r_i(\bar{p}_i) \geq c/\mu^{cs}$, which in this model translates to $b_i \leq 0.74$ ($=2/e$), and that $1/b_i \geq 1.36$. The optimal price for the system that cannot segment the two customer types does not admit a closed-form solution, and is computed numerically using (14) and (15).

To test specific numerical system instances, we have generated 250 independent realizations of the pair (b_1, b_2) by drawing each of the b_i s independently from a uniform distribution on $[0, 2]$. For each realization of (b_1, b_2) , we solved the deterministic relaxations with and without segmentation. This involved computing the optimal prices, deciding to which types to cross-sell, if any, calculating the corresponding staffing level, and finally the profit rate. Figure 4 displays the relative increase in profits, $(\bar{\Pi}(\Lambda) - \bar{\Pi}^a(\Lambda))/\bar{\Pi}^a(\Lambda)$, versus the maximum of the average willingness to pay among the two types, given by $\max(1/b_1, 1/b_2)$. The average profit increase through segmentation in this two-class experiment was around 24%. We have repeated this experiment several times, and in all of the experiments the average profit increase was above 20%. Figure 4 is rather intuitive. There will be no profit gap between the two systems if $b_1 = b_2$ or if the b_i s are different, but are such that no system decides to

cross-sell to any customer. In settings where at least one type has a very large average willingness to pay, both systems will be very profitable in their cross-selling activities, and the relative difference in profit will be small (the RHS of the figure). In settings where both parameters $1/b_i$ are small, then again the profit differential will be small because cross-selling is barely compensating for the cost of capacity. The difference between the two systems is more pronounced when $1/b_1$ and $1/b_2$ are of moderate size, in which case the relative added value from (a) price customization and (b) selective cross-selling (i.e., the capability to cross-sell to only one of the two types) is significant. For example, 20% of the 250 instances that we generated are such that the system with segmentation will choose to only cross-sell to one type, whereas the system with no segmentation capability will not cross-sell at all.

Finally, as the number of customer types and the availability of information on potential segmentation increases, the overall profit contribution due to segmentation becomes more substantial. In a set of experiments that we ran with four customer types, the average relative profit increase was 40% (up from 24% for the system with two types). Also, as the number of types was increased, we observed more instances where the cross-selling recommendations of the two systems would differ significantly.

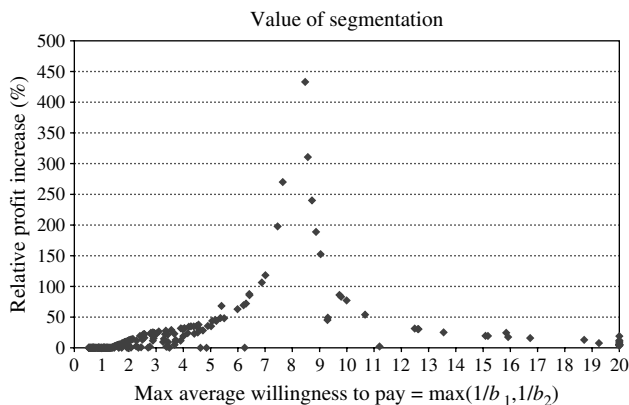
7. Concluding Remarks

To summarize, this paper proposes a tractable deterministic relaxation for studying the various control problems that arise in call center systems with cross-selling capability, paying particular attention to the effect of customer segmentation on the structure of the staffing, cross-selling, and routing policies that the system may choose to adopt. The policies that are generated through this analysis are simple to implement, intuitive, and achieve near-optimal performance.

Our analysis can be extended in several directions to better model the operational complexity of modern call center systems, as well as that of customer behavior. In the former, this may include systems that have multiple pools of agents with different processing capabilities, as well as more complicated service requirements, that may need a sequence of steps to be handled by the same or different agents. With respect to the latter, one could allow the customer’s decision of whether to listen to the cross-selling offer to include information from the initial phase of service experienced by the customer, such as his service time, whether his initial request was successfully resolved, etc. Another extension would be to allow for customers to abandon the queue if their waiting time is excessive. All of the above generalizations increase the complexity of the underlying system substantially, but can be addressed using our approximate analysis with little additional effort.

Finally, an interesting extension would examine the staffing and control decisions in the face of nonstationary arrival patterns or parameter estimation and forecasting

Figure 4. Profit comparison of systems with and without customer segmentation.



errors. Our asymptotic optimality results in this paper apply only to the stationary case with known arrival rates. For this setting, our asymptotic analysis and experience with numerical examples show that only the smallest threshold, η_k , has an important effect on the system performance. Still, the control (C)—with its multiple thresholds—was designed with more general settings in mind. Indeed, it seems plausible that in settings with nonstationarity and estimation errors, these larger threshold will play an important role by providing the system with a significant level of adaptability and robustness.

8. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at <http://or.journal.informs.org/>.

Endnotes

1. A recent study by McKinsey & Co. (Eichfeld et al. 2006) suggests that bank call centers can generate revenues that are equivalent to 10% of the revenue generated through the retail branch channels.
2. In a recent study, a Purdue University research group (Anton 2005) has estimated that call centers may attempt to cross-sell to as many as 60% of all its callers.
3. The first step involves the identification of appropriate attributes along which to segment the customer pool. The accuracy of the estimation of the customer-type characteristics will be greatly improved if the center can keep track of data on customers that refused to listen to the cross-selling offer, and on those that listened but did not buy. Finally, there is a trade-off between the number of customer segments and the accuracy of this estimation procedure, which may result in coarse segmentation, as opposed to segmenting down to the level of each customer.
4. For example, if the problem of original interest has $\Lambda' = 100$ and $\bar{W}' = 20$ seconds, then \hat{W} is selected so that $\bar{W}' = \hat{W}/\sqrt{\Lambda'}$, which in this case would give $\hat{W} = 200$ seconds. One should then study an asymptotic version of (2) as Λ grows large and \bar{W} is scaled according to $200/\sqrt{\Lambda}$; note that the original formulation is recovered for $\Lambda = 100$.
5. This is equivalent to assuming that in this case the willingness to pay is independent of the realized waiting time.

References

- Akşin, O. Z., P. T. Harker. 1999. To sell or not to sell: Determining the trade-offs between service and sales in retail banking phone centers. *J. Service Res.* 2(1) 19–33.

- Anton, J. 2005. Best practices in cross-selling and up-selling. <http://www.benchmarkportal.com>.
- Armony, M. 2005. Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems* 51(3–4) 287–329.
- Armony, M., I. Gurvich. 2006. When promotions meet operations: Cross-selling and its effect on call-center performance. Working paper, New York University and Columbia University, New York.
- Armony, M., C. Maglaras. 2004a. Contact centers with a call-back option and real-time delay information. *Oper. Res.* 52(4) 527–545.
- Armony, M., C. Maglaras. 2004b. On customer contact centers with a call-back option: Customer decisions, routing rules and system design. *Oper. Res.* 52(2) 271–292.
- Bassamboo, A., J. M. Harrison, A. Zeevi. 2006. Design and control of a large call center: Asymptotic analysis of an LP-based method. *Oper. Res.* 54(3) 419–435.
- Borst, S., A. Mandelbaum, M. Reiman. 2004. Dimensioning large call centers. *Oper. Res.* 52(1) 17–34.
- Byers, R. E., R. So. 2007a. A mathematical model for evaluating cross-sales policies in telephone service centers. *Manufacturing Service Oper. Management* 9(1) 1–8.
- Byers, R. E., R. So. 2007b. The value of information-based cross-sales policies in telephone service centers. Working paper, University of California, Irvine.
- Eichfeld, A., T. D. Morse, K. W. Scott. 2006. Using call centers to boost revenue. *McKinsey Quart.* (May).
- Erlang, A. K. 1948. On the rational determination of the number of circuits. E. Brockmeyer, H. L. Halstrom, A. Jensen, eds. *The Life and Works of A. K. Erlang*. The Copenhagen Telephone Company, Copenhagen, 216–221.
- Feldman, Z., A. Mandelbaum, W. A. Massey, W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* 54(2) 324–338.
- Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* 5(2) 79–141.
- Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* 4(3) 208–227.
- Güneş, E., O. Z. Akşin. 2004. Value creation in service delivery: Relating job designs, incentives and operational performance. *Manufacturing Service Oper. Management* 6(4) 338–357.
- Gurvich, I., M. Armony, A. Mandelbaum. 2008. Service level differentiation in call centers with fully flexible servers. *Management Sci.* 54(2) 279–294.
- Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29(3) 567–588.
- Maglaras, C., A. Zeevi. 2005. Pricing and design of differentiated services: Approximate analysis and structural insights. *Oper. Res.* 53(2) 242–262.
- Mandelbaum, A., S. Zeltyn. 2009. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Oper. Res.* Forthcoming.
- Örmeci, E. L., O. Z. Akşin. 2007. Revenue management through dynamic cross-selling in call centers. Working paper, Koc University, Istanbul, Turkey.
- Talluri, K. T., G. J. van Ryzin. 2004. *The Theory and Practice of Revenue Management*. Kluwer Academic Publishers, Boston.