Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Customer Acquisition, Retention, and Service Access Quality: Optimal Advertising, Capacity Level, and Capacity Allocation

http://orcid.org/0000-0002-2010-6983Philipp Afèche, Mojtaba Araghi, Opher Baron

To cite this article:
http://orcid.org/0000-0002-2010-6983Philipp Afèche, Mojtaba Araghi, Opher Baron (2017) Customer Acquisition, Retention, and Service Access Quality: Optimal Advertising, Capacity Level, and Capacity Allocation. Manufacturing & Service Operations Management

Published online in Articles in Advance 25 Sep 2017

. https://doi.org/10.1287/msom.2017.0635

http://pubsonline.informs.org/journal/msom/

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Articles in Advance, pp. 1–18
ISSN 1523-4614 (print), ISSN 1526-5498 (online)

Customer Acquisition, Retention, and Service Access Quality: Optimal Advertising, Capacity Level, and Capacity Allocation

Philipp Afèche,[a] Mojtaba Araghi,[b] Opher Baron[a]

[a] Rotman School of Management, University of Toronto, Toronto, Ontario M5S 3E6, Canada; [b] Lazaridis School of Business and Economics, Wilfrid Laurier University, Waterloo, Ontario N2L 3C5, Canada

Contact: afeche@rotman.utoronto.ca, http://orcid.org/0000-0002-2010-6983 (PA); maraghi@wlu.ca (MA); opher.baron@rotman.utoronto.ca (OB)

Received: November 17, 2012
Revised: June 10, 2015; September 23, 2016; December 20, 2016
Accepted: January 3, 2017
Published Online in Articles in Advance: September 25, 2017

https://doi.org/10.1287/msom.2017.0635

Abstract. *Problem definition*: We provide guidelines on three fundamental decisions of customer relationship management (CRM) and capacity management for profit-maximizing service firms that serve heterogeneous repeat customers, whose acquisition, retention, and behavior depend on their service access quality to bottleneck capacity: how much to spend on customer acquisition, how much capacity to deploy, and how to allocate capacity and tailor service access quality levels to different customer types. *Academic/practical relevance*: These decisions require a clear understanding of the connections between customers' behavior and value, their service access quality, and the capacity allocation. However, existing models ignore these connections. *Methodology*: We develop and analyze a novel fluid model that accounts for these connections. Simulation results suggest that the fluid-optimal policy also yields nearly optimal performance for large stochastic queueing systems with abandonment. *Results*: First, we derive new customer value metrics that extend the standard ones by accounting for the effects of the capacity allocation, the resulting service access qualities, and customer behavior: a customer's lifetime value; her $V\mu$ index, where V is her one-time service value and μ her service rate; and her policy-dependent value, which reflects the $V\mu$ indices of other served types. Second, we link these metrics to the profit-maximizing policy and to new capacity management prescriptions, notably, optimality conditions for rationing capacity and for identifying which customers to deny service. Further, unlike standard index policies, the optimal policy prioritizes customers based not on their $V\mu$ indices, but on policy- and type-dependent functions of these indices. *Managerial implications*: First, our study highlights the importance of basing decisions on more complete metrics that link customer value to the service access quality; marketing-focused policies that ignore these links may reduce profits significantly. Second, the proposed metrics provide guidelines for valuing customers in practice. Third, our decision guidelines help managers design more profitable policies that effectively integrate CRM and capacity management considerations.

Supplemental Material: The online appendix is available at https://doi.org/10.1287/msom.2017.0635.

Keywords: abandonment • advertising • call centers • capacity management • congestion • customer relationship management • fluid models • marketing–operations interface • priorities • promotions • service quality • staffing • queueing systems

1. Introduction

Many service firms serve heterogeneous repeat customers whose acquisition, retention, and behavior during their lifetime in the customer base depend on their service quality. A key dimension of service quality is access to bottleneck capacity, such as call centers, field operations, or fulfillment operations. We study three interrelated strategic decisions for a profit-maximizing firm in this situation: how much to spend on new customer acquisition, how much capacity to deploy, and how to allocate this capacity and tailor service access quality levels to different customer types.

This problem was motivated by several questions that emerged in our conversations with managers of a credit card company on how to manage customer access to its call center. What is the value of a call and the lifetime value of a customer? What are the effects of the service access quality on these metrics? How should the optimal decisions consider these effects, and what are the consequences of ignoring them? These questions highlight the importance of considering the effects of service access quality on the overall customer relationship, not only on individual calls, consistent with the key premise of customer relationship management (CRM). The importance of a CRM approach to call center capacity management is supported by customer surveys and other researchers. As Anton et al. (2004) estimate, firms that use call centers conduct 80% of interactions with their customers through this channel, and 92% of customers base their opinion of a company on their call center service experiences. Moreover, these experiences can

have a dramatic impact on customer satisfaction and retention. Long waiting times are cited by 67% of customers as a major cause of frustration, and a poor call center experience is cited by 40% of customers as the sole reason for terminating their relationship with a business (Genesys Telecommunications Labs 2007). According to Akşin et al. (2007, p. 682), "firms would benefit from a better understanding of the relationship between customers' service experiences and their repeat purchase behavior, loyalty to the firm, and overall demand growth in order to make better decisions about call center operations." The importance of understanding this relationship for call centers has recently also been recognized in the marketing literature (see Sun and Li 2011).

This paper provides a foundation for building such understanding, based on a new fluid model that integrates CRM and service operations management. Whereas the call center context motivated this research, our model emphasizes general features that apply to any bottleneck capacity of a service firm. For instance, at a high level, Amazon.com faces a similar problem in allocating its fulfillment capacity among new customers, Prime customers, and non-Prime customers.

We model homogeneous new customers and heterogeneous base (repeat) customer types that differ in their service request rates, service-dependent and -independent profits and costs, and reactions to their service access quality, measured by their service probability. The firm uses advertising to control new customer arrivals, and capacity level and allocation to control customers' service probabilities and, in turn, their customer base transitions, from conversion to base customers of different types, to switching among types, to defection. The novelty of this model is to link the makeup and value of the customer base to both the capacity allocation, unlike prior CRM models, and the service access quality of past interactions, unlike prior capacity management models.

This paper makes the following contributions to CRM and service capacity management:

1. *Customer value metrics that depend on the capacity allocation policy*. We identify novel metrics that, unlike standard ones, link the value of a customer to the capacity allocation policy and the resulting service probabilities and customer base transitions of all types: (i) The customer lifetime value (CLV) of base customers. (ii) The $V\mu$ index of a customer type, where V is her one-time service value (OTV) and μ her service rate. A type's $V\mu$ index depends on the CLVs and quantifies the value generated by allocating one unit of capacity to serving that type, including her instant service-dependent profit plus the expected future *service-independent* profits of the types that she may switch to. (iii) The policy-dependent value of a customer type, which reflects not only her own $V\mu$ index, but also the $V\mu$ indices of the served types that she may switch to.

2. *Guidelines for service capacity management*. We contribute (i) new analytical prescriptions for and (ii) important implications that follow from the profit-maximizing advertising, capacity, and capacity allocation policies.

(i) The optimal capacity allocation policy in our model features two key differences to standard index policies: First, the $V\mu$ indices consider the effect of service on customers' future requests and financial impact. Second, since the value of serving a type also depends on the $V\mu$ indices of other types, customers' optimal priority ranking does *not* correspond to that of their $V\mu$ indices. Furthermore, we find that under mild conditions it is optimal to ration capacity, whereby the firm serves only new and lucrative base customers but denies service to unprofitable base customers. We derive optimality conditions for capacity rationing, first for the case where base customers' service quality only affects their retention, then for cases where they also respond to service quality by switching their type or by spreading negative word of mouth (WOM).

(ii) Our results have a number of important implications: First, the customer attributes may have subtle effects on the optimal policy. The flexibility of our model allows managers to account for these effects. Second, in settings with repeat customers, the capacity allocation policy plays a previously ignored key role in controlling the customer base composition through differentiated service levels. Third, marketing-focused policies that ignore the effect of service probabilities on the CLV may yield suboptimal decisions and reduce profits significantly. Finally, simulation results show that the optimal policy in the fluid model also yields nearly optimal performance in a corresponding stochastic many-server queueing model with abandonment. This suggests that our fluid model approach may prove effective in tackling further problems of joint CRM and capacity management for stochastic service systems, such as the credit card call center that motivated this research.

In Section 2 we review the related literature. In Section 3 we present the model and problem formulation. In Section 4 we develop the main results for the case without base customer switching. In Section 5 we discuss how these results generalize under base customer type switching and under word of mouth about service quality. In Section 6 we discuss the implications of our results. In Section 7 we offer concluding remarks. All proofs and certain derivations are in the online supplement.

2. Literature Review

This paper bridges research streams on advertising, CRM, service capacity management to serve demand

that is independent of past service quality, and operations management to serve demand that depends on past service quality.

2.1. Advertising

The vast majority of papers on advertising, unlike ours, ignore supply constraints. Feichtinger et al. (1994) offer an extensive review. Papers that do consider supply constraints study different settings. For example, in Sethi and Zhang (1995) demand is independent of past service quality, whereas in Olsen and Parker (2008) customers are homogeneous; both studies consider systems with inventory.

2.2. CRM

Models of CRM and CLV are of central concern in marketing. For reviews see Rust and Chung (2006) and Reinartz and Venkatesan (2008) on CRM, and Gupta et al. (2006) on CLV. Studies that propose and empirically demonstrate the value of CLV-based frameworks for customer selection and marketing resource allocation include Rust et al. (2004) and Venkatesan and Kumar (2004). Based on the CLV components, CRM initiatives can be classified as focusing on customer acquisition, growth, and/or retention. Studies that focus on the relationship between acquisition and retention spending include Blattberg and Deighton (1996), Reinartz et al. (2005), Musalem and Joshi (2009), Pfeifer and Ovchinnikov (2011) and Ovchinnikov et al. (2014). Studies that focus on the effects of marketing actions on customer growth include Bitran and Mondschein (1996), Lewis (2005), Li et al. (2005), Rust and Verhoef (2005), and Günes et al. (2010). Papers that focus on explaining or predicting customer retention and churn include Verhoef (2003), Braun and Schweidel (2011), and Ascarza and Hardie (2013). Studies that link service quality, customer satisfaction, retention, and other CLV components include Anderson and Sullivan (1993), Rust et al. (1995), Zeithaml et al. (1996), Bolton (1998), Ho et al. (2006) and Aflaki and Popescu (2014).

Compared to the CRM literature, the key distinction of our model is that we explicitly link customer acquisition and retention to the capacity-allocation-dependent service quality. Studies that optimize some notion of service quality (Ho et al. 2006 and Aflaki and Popescu 2014) ignore capacity constraints (and customer acquisition). The papers that consider a capacity constraint (Pfeifer and Ovchinnikov 2011 and Ovchinnikov et al. 2014) study a firm's optimal spending on customer acquisition and retention of two base customer types ("low" and "high") that do not switch. However, their models ignore the link between capacity allocation and service quality: all base customers are served in each period, whereas their retention depends on the firm's spending. These models are therefore not equipped to study the design of optimal capacity allocation policies and, unlike our model, cannot yield optimal capacity rationing in the absence of uncertainty.

2.3. Service Capacity Management to Serve Demand That Is Independent of Past Service Quality

There is a vast literature on service capacity management. Capacity allocation studies consider operational tools including admission, priority and routing controls, marketing controls such as cross selling and pricing, or a combination of both. Call centers are among the most extensively studied service systems. Gans et al. (2003), Akşin et al. (2007), and Green et al. (2007) survey the call center literature. Hassin and Haviv (2003) and Hassin (2016) survey the literature on managing queueing systems with rational customers. These research streams focus on service access quality, i.e., waiting time and/or abandonment, assuming a fixed and perfect service delivery quality. However, there is growing interest in considering service delivery failures (e.g., de Véricourt and Zhou 2005) or tradeoffs between the service delivery quality and waiting time (e.g., Anand et al. 2011).

In contrast to this paper, these research streams model a firm's customer base as independent of the capacity allocation policy and the service quality of past interactions. That is, they model the arrivals of potential customer requests as exogenous but allow the outcomes of these requests (such as balking, retrial, abandonment, or spending amount) to depend on the service quality. Because these models do not keep track of repeat customers, their service-level prescriptions reflect only transaction-based metrics such as waiting and abandonment costs. In contrast, the prescriptions in our model also consider the effect of service levels on customers' future demand and their CLV.

2.4. Operations Management to Serve Demand That Depends on Past Service Quality

A number of papers consider the link between demand and past service quality with a different focus from ours. Schwartz (1966) appears to be the first to consider how demand depends on past inventory availability. Gans (2002) and Bitran et al. (2008b) consider a general notion of service quality in the absence of capacity constraints, the former for oligopoly suppliers, the latter for a monopoly. Hall and Porteus (2000), Liu et al. (2007), Gaur and Park (2007), and Olsen and Parker (2008) study equilibrium capacity/inventory control strategies and market shares of firms that compete for customers who switch among them in reaction to poor service. These papers, unlike ours, consider homogeneous customers and a single service level for each firm. Sun and Li (2011) empirically estimate how retention depends on customers' allocation to onshore versus offshore call centers, and on their waiting and service

time. Their results underscore the value of modeling the link between these service quality metrics, retention, and CLV. Farzan (2013) considers a first-in-first-out (FIFO) queue with repeat purchases that depend on past service quality, but quality is independent of the capacity allocation, unlike in our model. Adelman and Mersereau (2013) study the dynamic capacity allocation problem of a supplier with a fixed set of heterogeneous customers whose demands depend on their past fill rates. In their model, past service quality affects customers' profitability, not their retention.

3. Model and Problem Formulation

The credit card company that motivated this research uses call centers as the main customer contact channel. The firm also uses other channels such as the web, email, and Short Message Service (SMS), but these are much less costly and less suited for requests that require interaction with customer service representatives.

The focus of our model, one of the firm's call centers, is both sales- and service-oriented. Potential new customers call in response to advertised credit card offers. Existing cardholders call with service requests, e.g., to increase their credit limit, report a lost or stolen card, and so on.

Marketing and operations decisions are made each month according to the following procedure. The marketing department sets the monthly advertising spending level and the volume of credit card offers to be mailed out, based on some measures of customer profitability. The operations department is not involved in these decisions, and they ignore operational factors, specifically, service level considerations and staffing cost. Once the monthly advertising level is determined, credit card offers are mailed evenly over the course of that month.

Following the advertising decisions, the operations department forecasts the monthly call volume based on a number of factors, including the number of offers to be mailed to new customers and historical data on their demand response, and the number and calling pattern of base customers.

Based on the monthly call volume forecast, the operations department determines the staffing level for that month to meet certain predetermined service-level targets.

This decision procedure raises the following issues: (1) there is no clarity about the value of a call, the value of a customer, and how these measures depend on the call center's service quality; (2) advertising and staffing decisions either ignore service-level considerations or make arbitrary assumptions about service-level targets; (3) it is not clear whether to prioritize certain customers and, if so, on what basis; and (4) marketing and operations do not adequately coordinate their decisions.

These issues are pertinent not only for this company, but also for other firms facing the challenge of how to manage costly capacity in serving heterogeneous customers. To address these issues, in Section 3.1 we develop a deterministic fluid model that captures the strategic marketing and operations trade-offs. In Section 3.2 we discuss the model assumptions. In Section 6.4 we show that this model also closely approximates large-scale stochastic queueing systems such as the credit card company's call center.

3.1. The Model

Consider a firm that serves new and base (i.e., existing) customers with some service capacity. Following their first interaction with the firm, new customers may turn into any of m base customer types. Base customers make up the firm's customer base and repeatedly interact with the firm. We index new customers by $i = 0$ and base customers by $i \in \{1, 2, \ldots, m\}$. We say type i customers when $i \in \{0, 1, 2, \ldots, m\}$ and type i base customers when $i \in \{1, 2, \ldots, m\}$. Service requests arrive as detailed below. Let μ_i denote the service rate for type i requests. Table 1 summarizes the notation.

We consider the system in steady state under three stationary controls. The advertising policy controls the new customer arrival rate λ_0 as detailed below. The capacity policy controls the capacity N, defined as the total processing time available per unit time, at a unit cost C per unit time (in a call center N corresponds to the number of servers). The capacity allocation policy controls the service probabilities: let the decision variable q_i denote the steady-state type i service probability and $\mathbf{q} := (q_1, \ldots, q_m)$ the $1 \times m$ vector of base customer service probabilities.

New customers arrive to the system according to a deterministic process with constant rate λ_0, which depends on the firm's advertising spending. We assume

Table 1. Summary of Notation

	Decision variables
N	Capacity
λ_0	Arrival rate of new customers
q_i	Service probability of type i customers
	System parameters
μ_i	Service rate of type i customers
r_i	Arrival rate per type i base customer
$\theta_{ij}(q_i)$	Probability that a type i customer, served with probability q_i, switches to type j customer
γ_i	Service-independent departure rate per type i base customer
	Economic parameters
p_i	Profit per served request of type i customers
c_i	Cost per denied request of type i customers
R_i	Service-independent profit rate per type i base customer
C	Capacity cost rate
	Steady-state performance measures
x_i	Average number of type i base customers
Π	Profit rate

that the firm cannot target advertising to acquire specific customer types. Let $S(\lambda_0)$ be the advertising spending rate per unit time as a function of the corresponding new customer arrival rate. Advertising spending has diminishing returns (Simon and Arndt 1980), so $S(\lambda_0)$ is strictly increasing and strictly convex in λ_0. For analytical convenience we assume that S is twice continuously differentiable and $S'(0) = 0$.

The customer base evolution depends as follows on the customer flows and the service probabilities. Customers' service requests may be served or denied; requests that are denied are lost. The service probability q_i is the fraction of type i requests that are served. Customer transitions into, within, and out of the customer base depend as follows on their service probabilities. Let $\theta_{ij}(q_i)$ be the probability that, following her service request, a customer of type $i \in \{0, \ldots, m\}$ who is served with probability q_i switches to a base customer of type $j \in \{1, \ldots, m\}$. We assume that

$$\theta_{ij}(q_i) = q_i\bar{\theta}_{ij} + (1-q_i)\underline{\theta}_{ij}, \quad i = 0, 1, 2, \ldots, m;\ j = 1, 2, \ldots, m. \tag{1}$$

The parameters $\bar{\theta}_{ij}$ and $\underline{\theta}_{ij}$ denote the conditional ij-switching probabilities given that type i has received or has been denied service, respectively, where $\bar{\theta}_{ij} > \underline{\theta}_{ij}$ or $\bar{\theta}_{ij} < \underline{\theta}_{ij}$ depending on the attributes of types i and j. New customers do not join the customer base if their service request is denied, that is, $\underline{\theta}_{0j} = 0$ for $j \in \{1, \ldots, m\}$. Therefore, a new customer converts to a type j base customer with probability $\theta_{0j}(q_0) = q_0\bar{\theta}_{0j}$, and new customers join the customer base with rate $\lambda_0 q_0 \sum_{j=1}^m \bar{\theta}_{0j}$.

A type i base customer generates service requests with rate r_i. Given service probability q_i, after a service request such a customer remains in the customer base with probability $\sum_{j=1}^m \theta_{ij}(q_i)$, which we assume to be increasing in q_i, that is, $\sum_{j=1}^m(\bar{\theta}_{ij} - \underline{\theta}_{ij}) > 0$ from (1). A type i customer therefore leaves for service-dependent reasons with rate $r_i(1 - \sum_{j=1}^m \theta_{ij}(q_i))$. Base customers may also leave the company for service-independent reasons, such as relocation, switching to a competitor, or death. Let $\gamma_i > 0$ be the service-independent departure rate of a type i customer.

Let x_i denote the steady-state number of type i base customers or simply the type i customer base. In steady state, the flows into and out of this customer base must balance. Using (1) we have

$$\lambda_0 q_0 \bar{\theta}_{0i} + \sum_{j=1}^m x_j r_j (q_j\bar{\theta}_{ji} + (1-q_j)\underline{\theta}_{ji}) = x_i(\gamma_i + r_i) \quad i = 1, 2, \ldots, m, \tag{2}$$

where the left-hand side accounts for the inflow rates due to (new and base) customers of type $j \neq i$ who switch to, and those of type i who remain as, type i after their service request; the right-hand side sums the outflow rates due to service-independent attrition and service requests. Figure 1 shows the customer flows for the case where base customers do not switch type, that is, $\theta_{ij}(q_i) \equiv 0$ for $i \neq j \in \{1, \ldots, m\}$; we focus on this case in Section 4.

To write (2) in matrix form, define the $1 \times m$ vectors $\mathbf{x} := (x_1, x_2, \ldots, x_m)$ and $\bar{\boldsymbol{\theta}}_0 := (\bar{\theta}_{01}, \bar{\theta}_{02}, \ldots, \bar{\theta}_{0m})$ and the $m \times m$ matrices $\bar{\boldsymbol{\Theta}} := \{\bar{\theta}_{ij}\}$ and $\underline{\boldsymbol{\Theta}} := \{\underline{\theta}_{ij}\}$ for $i, j \in \{1, \ldots, m\}$, $\mathbf{D}_\gamma := \mathrm{diag}(\gamma_1, \gamma_2, \ldots, \gamma_m)$, $\mathbf{D}_\mathbf{r} := \mathrm{diag}(r_1, r_2, \ldots, r_m)$, and $\mathbf{D}_\mathbf{q} := \mathrm{diag}(q_1, q_2, \ldots, q_m)$. Then we have from (2) that

$$\lambda_0 q_0 \bar{\boldsymbol{\theta}}_0 + \mathbf{x}\mathbf{D}_r(\underline{\boldsymbol{\Theta}} + \mathbf{D}_\mathbf{q}(\bar{\boldsymbol{\Theta}} - \underline{\boldsymbol{\Theta}})) = \mathbf{x}(\mathbf{D}_\gamma + \mathbf{D}_\mathbf{r}). \tag{3}$$

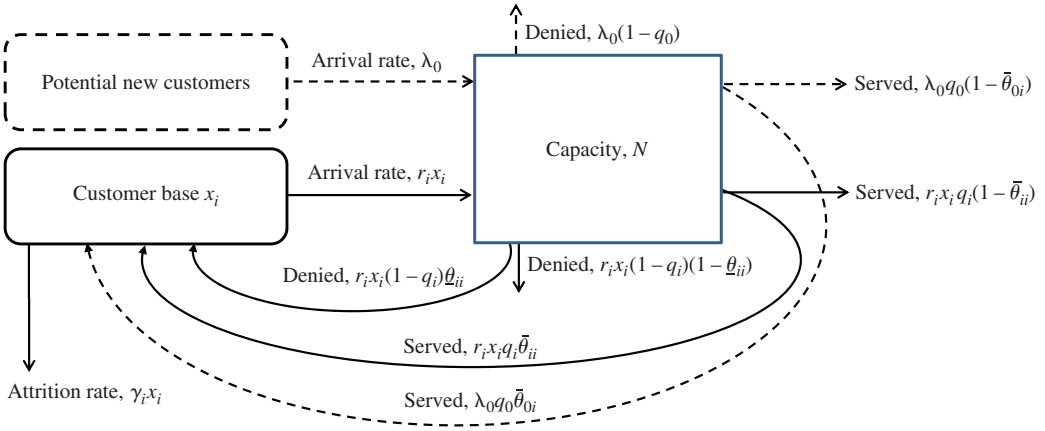
Define the $m \times m$ matrix

$$\mathbf{T}(\mathbf{q}) := (\mathbf{D}_\gamma + \mathbf{D}_\mathbf{r}(\mathbf{I} - \underline{\boldsymbol{\Theta}} - \mathbf{D}_\mathbf{q}(\bar{\boldsymbol{\Theta}} - \underline{\boldsymbol{\Theta}})))^{-1}, \tag{4}$$

where $\mathbf{I}$ is the $m \times m$ identity matrix. It follows from (3) and (4) that

$$\mathbf{x}(\mathbf{q}) = \lambda_0 q_0 \bar{\boldsymbol{\theta}}_0 \mathbf{T}(\mathbf{q}). \tag{5}$$

Figure 1. (Color online) Flows of New (Type 0) Customers (Dashed Lines) and Type i Base Customers (Solid Lines) for the Case Where Base Customers Do Not Switch Type



Then $T_{ij}(\mathbf{q})$ is the average time that a type i base customer spends as type j before leaving the customer base, as a function of the base customer service probabilities $\mathbf{q}$, where $T_{ij}(\mathbf{q}) < \infty$ since $\gamma_j > 0$ for $j \in \{1, \ldots, m\}$. If base customers do not switch type, (4) yields $T_{ij}(\mathbf{q}) = 0$ for $i \neq j$ and

$$T_{ii}(\mathbf{q}) = \frac{1}{\gamma_i + r_i(1 - q_i\bar{\theta}_{ii} - (1 - q_i)\underline{\theta}_{ii})} \tag{6}$$

is the mean lifetime as a type i base customer; the denominator is her departure rate, and $\bar{\theta}_{ii}$ and $\underline{\theta}_{ii}$ are her loyalty probabilities given that she receives or is denied service, respectively. Then by (5) and (6),

$$x_i = \frac{\lambda_0 q_0 \bar{\theta}_{0i}}{\gamma_i + r_i(1 - q_i\bar{\theta}_{ii} - (1 - q_i)\underline{\theta}_{ii})}. \tag{7}$$

By Little's law, the type i customer base equals throughput multiplied by sojourn time.

Let Π denote the firm's steady-state profit rate. The profit has service-dependent and service-independent components. On average a type i request yields a profit of p_i if served and a cost of c_i if denied. Therefore, the service-dependent profit rate of a type i base customer equals $r_i(p_i q_i - c_i(1 - q_i))$. Let $R_i \geq 0$ denote the average service-independent profit rate of a type i base customer. For credit card companies, R_i corresponds to call-center-independent profits such as subscription fees and interest payments from cardholders and transaction fees from merchants.

The firm solves the following profit-maximization problem by choosing the new customer arrival rate λ_0, the capacity N, and the new and base customer service probabilities q_0 and $\mathbf{q}$:

$$\begin{aligned} \underset{q_0, \mathbf{q} \geq \mathbf{0}, N \geq 0, \lambda_0 \geq 0}{\text{maximize}} \quad \Pi = {} & \lambda_0(p_0 q_0 - c_0(1 - q_0)) \\ & + \sum_{i=1}^{m} x_i(R_i + r_i[p_i q_i - c_i(1 - q_i)]) \\ & - CN - S(\lambda_0) \end{aligned} \tag{8}$$

subject to

$$x_i = \lambda_0 q_0 \sum_{j=1}^{m} \bar{\theta}_{0j} T_{ji}(\mathbf{q}), \quad i = 1, 2, \ldots, m, \tag{9}$$

$$q_i \leq 1, \quad i = 0, 1, 2, \ldots, m, \tag{10}$$

$$\frac{\lambda_0 q_0}{\mu_0} + \sum_{i=1}^{m} \frac{x_i r_i q_i}{\mu_i} \leq N. \tag{11}$$

In the profit rate (8), the first product is the new customer profit rate, the sum is over the base customer profit rates, the third term is the capacity cost rate, and the last is the advertising cost rate. The customer base equations (9) correspond to (5). In the capacity constraint (11), the left-hand side expresses the total processing time required to achieve the desired service probabilities.

3.2. Discussion of Model Assumptions

We discuss our assumptions on service quality, service capacity, and service demand, focusing on the credit card company that motivated this work. We note that our model can be tailored to a range of firms that serve heterogeneous repeat customers with costly capacity.

3.2.1. Service Quality. We model the service access quality as controllable and the service delivery quality as fixed.

Controllable service access quality. We model the service probabilities q_i as the measures of service access quality. Our focus on controlling the service access quality, taking the service delivery quality as fixed, is motivated by the call center of the credit card company introduced above: whereas it was clear that the quality of each customer encounter must meet strict service delivery standards, it was less clear what service access quality to offer to different customer types.

One could also reinterpret our model by assuming that all requests are served, but with discretionary task completion (Hopp et al. 2007): that is, service rates depend on the capacity allocation, so q_i is type i's service delivery quality that varies with its service rate.

Markovian customer response to service access quality. We assume that a base customer type's response to the outcome of her most recent request is independent of her service history. This "recency effect" is commonly assumed in models that link demand to past service levels (e.g., Hall and Porteus 2000, Ho et al. 2006, Liu et al. 2007) though some (e.g., Aflaki and Popescu 2014) consider a customer's service history. However, by allowing for an arbitrary number of types and a variety of switching behaviors, our model can capture history-dependent customer responses.

Fixed service delivery quality. The fixed parameters $\bar{\theta}_{ij}$, $\underline{\theta}_{ij}$, r_i, γ_i, p_i, and R_i serve as aggregate measures of customers' responses to the service delivery quality, the firm, and its products and services relative to competitors. The credit card company call center managers did not control these parameters and so treated them as given. Optimization over these parameters (e.g., to study the effects of training or product improvements on loyalty) is feasible in our model but outside the scope of this paper. We illustrate the sensitivity of our results to some of these parameters in Section 6.1.

3.2.2. Service Capacity. *Focus on bottleneck capacity*. Our model focuses on a single service capacity as the bottleneck for quality, because, for the credit card company, the call center is both the costliest and the most important customer contact channel in terms of customer acquisition and retention. Therefore, the company's key trade-off between the value and cost of service quality focuses on call center operations. In contrast, no such trade-offs arise for other attributes of the company's product and service quality, such as the

features of its credit cards, the invoice accuracy, and so on. Indeed, the company keeps these quality attributes relatively fixed, as discussed previously.

Premium versus low-cost channel. One can view our single-channel model as capturing a premium high-cost channel explicitly and a self-service low-cost channel such as the web implicitly and approximately: the average cost of serving a customer through the web is an order of magnitude cheaper than serving her via call centers. In this interpretation the profit of serving a type i request through the web corresponds to $-c_i$. Our model can be extended to multiple channels by considering multiple capacity pools with different financial parameters, service rates, interaction rates, and so on.

3.2.3. Service Demand. *Independence of base customer demand from advertising*. To focus on the effect of service access quality on customer behavior, we do not model the effect of advertising on base customers. However, our model implicitly captures these advertising effects through the request rate r_i, the profits R_i and p_i, the probabilities $\bar{\theta}_{ij}$ and $\underline{\theta}_{ij}$, and the service-independent defection rate γ_i.

Stationary demand. We assume that the new customer arrival rate is a concave, stationary, and deterministic function of the advertising spending rate. Under these assumptions the optimal advertising policy is to spend continuously at an even rate and yields a constant demand rate (Sasieni 1971, Feinberg 2001). Indeed, although the credit card company sets the advertising level at the beginning of each month, it mails out the offers steadily over the month, which yields a steady demand response. To focus on strategic-level guidelines, our model ignores the typical predictable intraday and intraweek arrival rate fluctuations, to focus on the first-order relationship between advertising spending and demand response levels. Nevertheless, for operations that can flexibly adjust the capacity over time (this holds for the credit card company call center and increasingly for other companies, because of flexible workforce contracts and outsourcing providers), our results can also be used at the tactical level to adjust capacity to both predictable and unpredictable short-term arrival rate fluctuations, so as to maintain the optimal service levels (see Sections 4.2.1 and 4.2.2).

4. Main Results: The Case Without Base Customer Switching

For simplicity we present the main results for the case without base customer switching. As shown in Section 5.1, these results extend to the case with base customer switching. In Section 4.1 we reformulate problem (8)–(11) in terms of customers' $V\mu$ *indices*, which are novel service-probability-dependent customer value metrics, and the capacity allocation. In Section 4.2 we characterize the optimal policy.

4.1. Service Quality, $V\mu$ Indices, Capacity Allocation, and Customer Value

In Section 4.1.1 we derive a customer type's $V\mu$ index from the service-probability-dependent metrics of base customer lifetime value. In Section 4.1.2 we reformulate problem (8)–(11), and in Section 4.1.3 we define the value of a new customer, in terms of these $V\mu$ indices and the capacity allocation policy.

4.1.1. Service-Probability-Dependent Customer Lifetime Value and the $V\mu$ Index. Let $L_i(\mathbf{q})$ denote the mean type i *base customer lifetime value* (*CLV*) as a function of the base customer service probabilities $\mathbf{q}$. It is given by

$$L_i(\mathbf{q}) := T_{ii}(\mathbf{q})(R_i + r_i(p_i q_i - c_i(1-q_i))) = \frac{R_i + r_i(p_i q_i - c_i(1-q_i))}{\gamma_i + r_i(1 - q_i\bar{\theta}_{ii} - (1-q_i)\underline{\theta}_{ii})}, \tag{12}$$

the product of her mean customer base sojourn time by her mean profit rate. Clearly, without base customer switching, the type i CLV only depends on her own service probability q_i.

Let V_0 denote the mean *one-time service value* (*OTV*) of a new customer, i.e., the value of serving a new customer's current request, but none of her future requests:

$$V_0 := p_0 + c_0 + \sum_{i=1}^{m} \bar{\theta}_{0i} L_i(\mathbf{0}). \tag{13}$$

Serving a new customer yields an instant profit of p_0, plus a future service-independent profit stream; with probability $\bar{\theta}_{0i}$ the new customer turns into a type i base customer with CLV $L_i(\mathbf{0})$ (given that she is not served). Not serving a new customer yields a loss of c_0; the difference yields (13).

Similarly, let V_i denote the mean OTV of a type i base customer:

$$V_i := p_i + c_i + (\bar{\theta}_{ii} - \underline{\theta}_{ii}) L_i(\mathbf{0}), \quad i = 1, 2, \ldots, m. \tag{14}$$

Serving a type i customer's current request but none of her future requests yields $p_i + \bar{\theta}_{ii} L_i(\mathbf{0})$, and serving none of her requests yields $-c_i + \underline{\theta}_{ii} L_i(\mathbf{0})$, so the difference yields (14). From (12) and (14), the base customer OTV and CLV satisfy the following intuitive relationship (see Online Appendix A):

$$V_i = \frac{L_i(\mathbf{e}_i) - L_i(\mathbf{0})}{r_i T_{ii}(\mathbf{e}_i)}, \quad i = 1, 2, \ldots, m, \tag{15}$$

where the policy $\mathbf{q} = \mathbf{e}_i$ serves all type i base customer requests ($q_i = 1$) but no other types ($q_j = 0$ for $j \neq i$), and $r_i T_{ii}(\mathbf{e}_i)$ is the resulting mean lifetime number of type i requests.

We define a type's $V\mu$ index as the product of her OTV by her service rate; it quantifies the value generated by allocating one unit of capacity to serving that

type, including her instant service-dependent profit plus the expected future service-independent profits of the types that she may switch to. For $i \geq 1$ we assume that $V_i\mu_i \geq V_{i+1}\mu_{i+1}$ without loss of generality and $V_i\mu_i > V_{i+1}\mu_{i+1}$ for simplicity. (Cases with $V_i\mu_i = V_{i+1}\mu_{i+1}$ add cumbersome detail but no insight to the analysis.)

4.1.2. Profit Maximization in Terms of the $V\mu$ Indices and the Capacity Allocation. Let N_i be the capacity, i.e., the total processing time per unit time, that is allocated to and consumed by type i customers. By Little's law the capacity allocated to type i equals throughput multiplied by service time, so the service probabilities map as follows to the capacity allocation:

$$N_0 := \frac{\lambda_0 q_0}{\mu_0}, \tag{16}$$

$$N_i := \frac{x_i r_i q_i}{\mu_i}, \quad i = 1, 2, \ldots, m. \tag{17}$$

Define the $1 \times (m+1)$ vector $\mathbf{N} := (N_0, N_1, \ldots, N_m)$. From (3), (4), (16), and (17), the customer base depends as follows on the capacity allocation (see Online Appendix A):

$$x_i = (N_0\mu_0\bar{\theta}_{0i} + N_i\mu_i(\bar{\theta}_{ii} - \underline{\theta}_{ii}))T_{ii}(\mathbf{0}), \quad i = 1, 2, \ldots, m. \tag{18}$$

By (8), (12)–(14), and (16)–(18), the total customer value satisfies (see Online Appendix A)

$$\lambda_0(p_0q_0 - c_0(1-q_0)) + \sum_{i=1}^{m} x_i(R_i + r_i[p_iq_i - c_i(1-q_i)]) = \sum_{i=0}^{m} N_iV_i\mu_i - \lambda_0 c_0, \tag{19}$$

where $N_iV_i\mu_i$ expresses the value generated by allocating N_i units of capacity to type i customers.

Write s_0 for the mean service time of a new customer request. Then

$$s_0 := \frac{1}{\mu_0} \tag{20}$$

and $\lambda_0 s_0$ is the offered load of new customers. Let s_i be the expected total processing time of all type i base customer requests that may be generated as a result of serving a new customer. Then

$$s_i := \bar{\theta}_{0i}T_{ii}(\mathbf{e}_i)\frac{r_i}{\mu_i}, \tag{21}$$

where $\bar{\theta}_{0i}$ is the probability that a new customer who is served joins as a type i base customer, $T_{ii}(\mathbf{e}_i)$ is a type i customer's expected lifetime if all her requests are served, and r_i/μ_i is her expected total processing requirement per unit time. Therefore, $\lambda_0 q_0 s_i$ is the offered load of type i base customers if new customers have arrival rate λ_0 and service probability q_0.

Let $\Pi(\mathbf{N}, N, \lambda_0)$ denote the profit rate as a function of the capacity allocation vector $\mathbf{N}$, the total capacity N, and the new customer arrival rate λ_0. The problem (8)–(11) is equivalent to

$$\underset{\mathbf{N}\geq\mathbf{0},\, N\geq 0,\, \lambda_0\geq 0}{\text{maximize}}\ \Pi(\mathbf{N}, N, \lambda_0) = \sum_{i=0}^{m} N_iV_i\mu_i - \lambda_0 c_0 - CN - S(\lambda_0) \tag{22}$$

subject to

$$N_0 \leq \lambda_0 s_0, \tag{23}$$

$$N_i \leq N_0\mu_0 s_i, \quad i = 1, 2, \ldots, m, \tag{24}$$

$$\sum_{i=0}^{m} N_i \leq N. \tag{25}$$

The profit (22) follows from (8) and (19). By (6), (16)–(18), (20), and (21), the capacity allocation constraints (23) and (24) are equivalent to the service probability constraints (10) and ensure that the capacity consumed by each type does not exceed its offered load. Finally, (25) corresponds to (11).

4.1.3. The Maximum Value of a New Customer Depends on All $V\mu$ Indices. As a preliminary to the optimal policy, we characterize the maximum value of a new customer. This value depends on the capacity allocation policy through the service probabilities of base customers.

Let $\bar{s}_i$ be the expected total processing time of all requests generated by a new customer under the policy that serves her first request and her subsequent ones if and only if she turns into one of the i highest-value base customer types (in terms of their $V\mu$ indices):

$$\bar{s}_i := \sum_{j=0}^{i} s_j, \quad i = 0, \ldots, m. \tag{26}$$

Let $\bar{V}_i$ denote the expected total value per processing time of a new customer under this policy:

$$\bar{V}_i = \frac{p_0 + c_0 + \sum_{j=1}^{i}\bar{\theta}_{0j}L_j(\mathbf{e}_j) + \sum_{j=i+1}^{m}\bar{\theta}_{0j}L_j(\mathbf{0})}{\bar{s}_i}, \quad i = 0, \ldots, m. \tag{27}$$

The numerator sums the immediate profit of serving a new customer plus her expected future profit as a base customer; with probability $\bar{\theta}_{0j}$ she switches to type j and her CLV is $L_j(\mathbf{e}_j)$ or $L_j(\mathbf{0})$, depending on whether type j is served or not, respectively. Equivalently, from (13), (15), (21), and (27), $\bar{V}_i$ equals the load-weighted convex combination of the $V\mu$ indices of types that are served:

$$\bar{V}_i := \frac{\sum_{j=0}^{i} s_jV_j\mu_j}{\bar{s}_i}, \quad i = 0, \ldots, m, \tag{28}$$

where $s_j/\bar{s}_i$ is type j's share of the total processing time generated by a new customer.

Remark 1. Importantly, the value of serving a new customer, $\bar{V}_i$, depends on the service policy, that is, not only on her own $V\mu$ index, but also on the $V\mu$ indices of base customers that are served. As a result, new customers' optimal priority ranking does not correspond to the ranking of their $V\mu$ index, in contrast to standard index policies in the literature. These key implications derive from two distinctive features of our model: the presence of customer base transitions and their dependence on the service quality. Under base customer switching, these implications extend to the values of and the priority ranking among base customers, as we show in Section 5.1.

Lemma 1 characterizes the service policy that yields the maximum new customer value, $\bar{V}_k$. This metric and the $V\mu$ indices drive the optimal policy under fixed advertising in Sections 4.2.1 and 4.2.2.

Lemma 1. *Consider a system without base customer switching. Define*

$$\begin{aligned} k &:= 0 \quad \text{if } \bar{V}_0 > \bar{V}_1 \quad \text{and} \\ k &:= \max\{1 \le i \le m : \bar{V}_{i-1} \le \bar{V}_i\} \quad \text{if } \bar{V}_0 \le \bar{V}_1. \end{aligned} \tag{29}$$

Then $\bar{V}_k$ is the maximum value of a new customer per processing time:

$$\bar{V}_k = \max_{0 \le i \le m} \bar{V}_i, \tag{30}$$

$$\bar{V}_0 < \cdots < \bar{V}_{k-1} \le \bar{V}_k > \bar{V}_{k+1} > \cdots > \bar{V}_m, \tag{31}$$

$$V_i\mu_i \ge \bar{V}_i \quad \text{for } i \le k, \quad \text{and} \quad \bar{V}_i > V_i\mu_i \quad \text{for } i > k. \tag{32}$$

When the advertising level is a decision rather than fixed, serving every new customer is optimal, so the *net* value of a new customer equals $\bar{V}_i$ minus the service denial cost c_0 (per processing time):

$$\tilde{V}_i := \bar{V}_i - \frac{c_0}{\bar{s}_i}, \quad i = 0, \ldots, m. \tag{33}$$

Lemma 2 characterizes the service policy that yields the maximum new customer *net* value, $\tilde{V}_{k^*}$. This metric and the $V\mu$ indices drive the jointly optimal policy, under optimal advertising, in Section 4.2.3.

Lemma 2. *Consider a system without base customer switching. Define*

$$\begin{aligned} k^* &:= 0 \quad \text{if } \tilde{V}_0 > \tilde{V}_1 \quad \text{and} \\ k^* &:= \max\{1 \le i \le m : \tilde{V}_{i-1} \le \tilde{V}_i\} \quad \text{if } \tilde{V}_0 \le \tilde{V}_1. \end{aligned} \tag{34}$$

Then $k^* \ge k$ and $\tilde{V}_{k^*}$ is the maximum net value of a new customer per processing time:

$$\tilde{V}_{k^*} = \max_{0 \le i \le m} \tilde{V}_i, \tag{35}$$

$$\tilde{V}_0 < \cdots < \tilde{V}_{k^*-1} \le \tilde{V}_{k^*} > \tilde{V}_{k^*+1} > \cdots > \tilde{V}_m, \tag{36}$$

$$V_i\mu_i \ge \tilde{V}_i \quad \text{for } i \le k^*, \quad \text{and} \quad \tilde{V}_i > V_i\mu_i \quad \text{for } i > k^*. \tag{37}$$

4.2. Optimal Policy: Capacity Allocation, Capacity Level, and Advertising

We present in Section 4.2.1 the optimal capacity allocation for fixed capacity and advertising, in Section 4.2.2 the optimal capacity allocation and level for fixed advertising, and in Section 4.2.3 the jointly optimal policy.

4.2.1. Optimal Capacity Allocation for Fixed Capacity and Advertising Levels.

Consider the problem of optimizing the capacity allocation for fixed capacity and new customer arrival rate. This problem may arise due to hiring lead times, time lags between advertising and demand response, unplanned demand bursts, or poor marketing–operations coordination.

The total offered load if all new customers are served is $\lambda_0\bar{s}_m$. We say capacity is "*rationed*" if $N < \lambda_0\bar{s}_m$. The key property of a capacity allocation policy is the *priority ranking* that determines which customer types are served when capacity is rationed. Proposition 1 establishes the optimal priority ranking and the corresponding capacity allocation.

Proposition 1. *Consider a system without base customer switching. Fix the new customer arrival rate λ_0 and the capacity N. Determine the index k from Lemma 1.*

1. *It is optimal to prioritize base customers of type $i \le k$ over new customers, new customers over base customers of type $i > k$, and base customers in decreasing order of their $V_i\mu_i$ index.*
2. *The optimal capacity allocation satisfies*

$$N_i^* = \begin{cases} \min(\lambda_0\bar{s}_k, N)\dfrac{s_i}{\bar{s}_k}, & i \le k, \\ \min(\lambda_0 s_i, (N - \lambda_0\bar{s}_{i-1})^+), & i > k. \end{cases} \tag{38}$$

3. *The optimal profit rate is concave in the capacity N and satisfies*

$$\begin{aligned} \Pi^*(N, \lambda_0) := {} & \min(\lambda_0\bar{s}_k, N)\bar{V}_k \\ & + \sum_{i=k+1}^{m} \min(\lambda_0 s_i, (N - \lambda_0\bar{s}_{i-1})^+)V_i\mu_i \\ & - \lambda_0 c_0 - CN - S(\lambda_0). \end{aligned} \tag{39}$$

The greedy policy that prioritizes all types according to their $V\mu$ index is *not* optimal in general, as noted in Remark 1; it is only optimal if new customers have the largest $V\mu$ index. Rather, the optimal policy allocates the capacity to maximize the value generated per processing time. By Lemma 1, this requires prioritizing all base customers of type $i \le k$ ahead of new customers, where k is determined so that new customers' policy-dependent value, $\bar{V}_k$, is smaller than the $V\mu$ indices of base customers with higher priority and larger than the indices of those with lower priority. Then, if $N < \lambda_0\bar{s}_k$, it is optimal to turn away enough new customers to ensure service for all base customers of type $i \le k$, and to serve none of type $i > k$. However, if $N \ge \lambda_0\bar{s}_k$, it is

optimal to serve all requests of type $i \le k$, and those of type $i > k$ in decreasing order of their $V\mu$ indices. In Section 6.4 we discuss how to implement this policy in, and contrast it with standard policies for, stochastic queueing systems with abandonment.

The profit rate (39) reflects the value generated under the optimal capacity allocation: for $N < \lambda_0 \bar{s}_k$ each unit of capacity has a value of $\bar{V}_k$, and for $N > \lambda_0 \bar{s}_k$ the value of each additional capacity unit is given by the largest $V\mu$ index among types with unserved requests.

4.2.2. Optimal Capacity Allocation and Capacity Level for a Fixed Advertising Level.

We turn to the problem of optimizing the capacity allocation and the capacity level for fixed advertising. This problem arises because the advertising policy is often a strategic decision that affects the more tactical operational decisions, and also in response to unplanned bursts of arrivals. The optimal capacity, denoted by N^*, is the largest capacity level with nonnegative marginal profit. The marginal profit of a unit of capacity equals its value under the optimal capacity allocation, minus its cost C. Proposition 2 follows from the optimal profit rate (39) in Proposition 1.

Proposition 2. *Consider a system without base customer switching. Fix the new customer arrival rate λ_0 and determine the index k from Lemma 1. Under the optimal capacity allocation policy it is profitable to operate if and only if $\tilde{V}_k > C$ in which case the following holds*:

1. *It is optimal to serve new customers, all base customers of type $i \le k$, and base customers of type $j > k$ if and only if $V_j \mu_j \ge C$.*
2. *The optimal capacity level is*

$$N^* = \lambda_0 \left(\bar{s}_k + \sum_{i=k+1}^{m} s_i 1_{\{V_i \mu_i \ge C\}} \right), \tag{40}$$

and rationing capacity is optimal if and only if $C > V_m \mu_m$.

3. *The optimal profit rate is*

$$\Pi^*(\lambda_0) := \lambda_0 \left(\bar{s}_k (\tilde{V}_k - C) + \sum_{i=k+1}^{m} s_i (V_i \mu_i - C)^+ \right) - S(\lambda_0). \tag{41}$$

An important implication of Proposition 2 is that *rationing capacity is optimal* under practically plausible conditions. Mathematically, this holds if and only if $\tilde{V}_k > C > V_m \mu_m$; that is, the capacity cost is smaller than the maximum value of a new customer per processing time, but larger than the smallest $V\mu$ index of base customers. As we show in Section 6.1, these conditions are easily met, which suggests that they may commonly hold in practice. Under these conditions, the optimal policy achieves two goals: (1) It serves all "high-value" base customers (of type $i \le k$) and all new customers because they generate these base customers. Note that, taken on their own, these requests of new customers may not be profitable; that is, their $V\mu$ index may be lower than the capacity cost (i.e., $V_0 \mu_0 < C < \bar{V}_k$). (2) It serves "low-value" base customers (of type $i > k$) if and only if their $V\mu$ index exceeds the capacity cost. Importantly, the *optimal policy deliberately denies service to unprofitable base customers* who were acquired along with high-value customers.

4.2.3. Jointly Optimal Capacity Allocation, Capacity Level, and Advertising Level.

Let λ_0^* denote the optimal new customer arrival rate, that is, under the optimal advertising level. Proposition 3 summarizes the solution to the profit-maximization problem (22)–(25).

Proposition 3. *Consider a system without base customer switching. Determine the index k^* from Lemma 2. Under the jointly optimal advertising, capacity, and capacity allocation policies, it is profitable to operate if and only if $\tilde{V}_{k^*} > C$ in which case the following holds*:

1. *It is optimal to serve new customers, all base customers of type $i \le k^*$, and base customers of type $j > k^*$ if and only if $V_j \mu_j \ge C$.*
2. *The optimal capacity level is*

$$N^* = \lambda_0^* \left(\bar{s}_{k^*} + \sum_{i=k^*+1}^{m} s_i 1_{\{V_i \mu_i \ge C\}} \right), \tag{42}$$

and rationing capacity is optimal if and only if $C > V_m \mu_m$.

3. *The optimal new customer arrival rate satisfies $\lambda_0^* > 0$ and*

$$\bar{s}_{k^*} (\tilde{V}_{k^*} - C) + \sum_{i=k^*+1}^{m} s_i (V_i \mu_i - C)^+ = S'(\lambda_0^*), \tag{43}$$

and the optimal profit rate satisfies

$$\Pi^* = \lambda_0^* \left(\bar{s}_{k^*} (\tilde{V}_{k^*} - C) + \sum_{i=k^*+1}^{m} s_i (V_i \mu_i - C)^+ \right) - S(\lambda_0^*). \tag{44}$$

The profitability condition and the optimal capacity level and allocation policy in Proposition 3 parallel their counterparts in Proposition 2, adjusted for the service denial cost effect explained in Section 4.1.3. In particular, Proposition 3 suggests that rationing capacity is sensible in practice; it is optimal if and only if $\tilde{V}_{k^*} > C > V_m \mu_m$. These conditions are easily met as shown in Section 6.1.

The optimal advertising spending specified in (43) balances the maximum value of a new customer with the marginal advertising cost of acquiring this customer. Like the optimal capacity level, the optimal advertising depends on the optimal capacity allocation policy and the resulting CLV.

5. Extensions

We discuss how the optimality conditions for capacity rationing generalize under base customer type switching in Section 5.1 and under word of mouth about service quality in Section 5.2.

5.1. Switching Among Base Customer Types

We extend the main results of Section 4 under base customer switching (see Online Appendix A for details).

5.1.1. Service-Probability-Dependent Customer Lifetime Value and the $V\mu$ Index. The formula (12) for the mean type i CLV generalizes to

$$L_i(\mathbf{q}) := \sum_{j=1}^{m} T_{ij}(\mathbf{q})(R_j + r_j(p_j q_j - c_j(1-q_j))). \quad (45)$$

Each summand in (45) measures the type j contribution to this CLV, which equals the product of the average time $T_{ij}(\mathbf{q})$ (defined in (4)) that a type i base customer spends as type j before leaving the customer base, multiplied by a type j customer's average profit rate per unit time.

The formulas (13) and (14) for the customer OTVs generalize to

$$V_i := p_i + c_i + \sum_{j=1}^{m} (\bar{\theta}_{ij} - \underline{\theta}_{ij}) L_j(\mathbf{0}), \quad i = 0, 1, 2, \ldots, m, \quad (46)$$

where $(\bar{\theta}_{ij} - \underline{\theta}_{ij})L_j(\mathbf{0})$ for $i \neq j$ measures the effect of serving a type i customer on her future profits as type j. For a new customer ($i = 0$), (46) agrees with (13) since $\underline{\theta}_{0j} = 0$ for all j.

The relationship (15) between base customer OTV and CLV continues to hold: by (4) and (45),

$$V_i = \frac{L_i(\mathbf{e}_i) - L_i(\mathbf{0})}{r_i T_{ii}(\mathbf{e}_i)}, \quad i = 1, 2, \ldots, m.$$

5.1.2. Profit Maximization in Terms of the $V\mu$ Indices and the Capacity Allocation. The expression (18) for the customer base as a function of the capacity allocation generalizes to

$$x_i = \sum_{j=1}^{m} \left(N_0 \mu_0 \bar{\theta}_{0j} + \sum_{k=1}^{m} N_k \mu_k (\bar{\theta}_{kj} - \underline{\theta}_{kj}) \right) T_{ji}(\mathbf{0}), \quad i = 1, 2, \ldots, m, \quad (47)$$

which follows from (3), (4), (16), and (17). The expression (19) for the total customer value in terms of the capacity allocation and $V\mu$ indices continues to hold: (16) and (17) and (45)–(47) imply

$$\lambda_0(p_0 q_0 - c_0(1-q_0)) + \sum_{i=1}^{m} x_i (R_i + r_i[p_i q_i - c_i(1-q_i)]) = \sum_{i=0}^{m} N_i V_i \mu_i - \lambda_0 c_0. \quad (48)$$

Finally, we generalize (21) and (24) to quantify the offered load of each base customer type in terms of the capacity allocated to other types. Let s_{ji} denote the expected total processing time of all type i base customer requests that may be generated as a result of serving a type j customer, under the policy that serves all requests of type i but none of other types (i.e., $\mathbf{q} = \mathbf{e}_i$). Then

$$s_{ji} = \sum_{k=1}^{m} (\bar{\theta}_{jk} - \underline{\theta}_{jk}) T_{ki}(\mathbf{e}_i) \frac{r_i}{\mu_i}, \quad i = 1, 2, \ldots, m;\ j = 0, 1, 2, \ldots, m. \quad (49)$$

Note that, without base customer switching, s_{0i} defined in (49) agrees with s_i defined in (21).

It follows from (48) and (49) that the profit-maximization problem (22)–(25) for the case without base customer switching continues to hold if base customers switch type, except that the capacity allocation constraints for base customers generalize from $N_i \leq N_0 \mu_0 s_i$ in (24) to

$$N_i \leq N_0 \mu_0 s_{0i} + \sum_{j \neq i \geq 1}^{m} N_j \mu_j s_{ji}, \quad i = 1, 2, \ldots, m. \quad (50)$$

The right-hand side of (50) expresses the offered load of type i base customers as a function of the capacity allocated to all other types.

5.1.3. Optimality Conditions for Capacity Rationing. Because the profit function (22) in the maximization problem (22)–(25) has the same structure under base customer switching, the main results of Section 4 generalize to this case: It is optimal to serve new customers and a subset of base customer types and to deny service to any remaining types by rationing capacity. However, as noted in Remark 1, under base customer switching the value of a base customer type is no longer limited to its own $V\mu$ index: serving her affects other types' offered loads as shown in (50), so their capacity allocations and their values are interdependent. Therefore, the optimal priority ranking among base customers no longer corresponds to that of their $V\mu$ indices. As a result, determining the optimal capacity allocation policy requires solving the full problem.

Nevertheless, the optimality conditions for capacity rationing in Proposition 3, $\tilde{V}_{k^*} > C > V_m \mu_m$, generalize naturally under base customer switching, by noting that $\tilde{V}_{k^*}$ is the maximum over all type subsets of the ratio of *total* customer value generated to *total* processing time required, and $V_m \mu_m$ is the minimum over all types of the ratio of *marginal* customer value generated to *marginal* processing time required. To formalize these measures under base customer switching, let $\mathcal{I} = \{1, 2, \ldots, m\}$, let $\mathcal{C}$ be an arbitrary subset of $\mathcal{I}$ and define $\mathcal{C}_{-i} := \mathcal{C} \backslash \{i\}$. Under the policy that serves new customers and all base customer types in $\mathcal{C}$ but no other types, the capacity allocation constraints (23) for new customers and (50) for base customer types in $\mathcal{C}$ are binding. Given this capacity allocation policy, let $\bar{N}_i(\mathcal{C})$ for $i \in \mathcal{I}$ denote the ratio of the type i customer

throughput $N_i\mu_i$ to new customer throughput $N_0\mu_0$. Then we have $\bar{N}_i(\mathscr{C}) = 0$ for $i \notin \mathscr{C}$, and from (50)

$$\bar{N}_i(\mathscr{C}) = s_{0i}\mu_i + \sum_{j\in\mathscr{C}_{-i}} \bar{N}_j(\mathscr{C})s_{ji}\mu_i, \quad i \in \mathscr{C}. \tag{51}$$

Let $\tilde{V}(\mathscr{C})$ be the ratio of total customer value to total processing time for policy $\mathscr{C}$. By (48) and (51)

$$\tilde{V}(\mathscr{C}) := \frac{V_0 - c_0 + \sum_{i\in\mathscr{C}}(s_{0i} + \sum_{j\in\mathscr{C}_{-i}} \bar{N}_j(\mathscr{C})s_{ji})V_i\mu_i}{1/\mu_0 + \sum_{i\in\mathscr{C}}(s_{0i} + \sum_{j\in\mathscr{C}_{-i}} \bar{N}_j(\mathscr{C})s_{ji})}. \tag{52}$$

Let $\Delta V(\mathscr{F}, \mathscr{F}_{-i})$ denote the ratio of marginal value generated to marginal processing time required when changing the policy from serving all types except i, to serving all types. By (48) and (51)

$$\begin{aligned}\Delta V(\mathscr{F}, \mathscr{F}_{-i}) &:= \left[\sum_{j\in\mathscr{F}_{-i}} V_j\mu_j \sum_{k\in\mathscr{F}_{-j}} (\bar{N}_k(\mathscr{F}) - \bar{N}_k(\mathscr{F}_{-i}))s_{kj} + V_i\mu_i\left(s_{0i} + \sum_{k\in\mathscr{F}_{-i}} \bar{N}_k(\mathscr{F})s_{ki}\right)\right] \\ &\quad \cdot \left[\sum_{j\in\mathscr{F}_{-i}} \sum_{k\in\mathscr{F}_{-j}} (\bar{N}_k(\mathscr{F}) - \bar{N}_k(\mathscr{F}_{-i}))s_{kj} + \left(s_{0i} + \sum_{k\in\mathscr{F}_{-i}} \bar{N}_k(\mathscr{F})s_{ki}\right)\right]^{-1}. \end{aligned} \tag{53}$$

Note that, without base customer switching, (52) for $\mathscr{C} = \{1, 2, \ldots, i\}$ agrees with (33), that is, $\tilde{V}(\mathscr{C}) = \tilde{V}_i$, and (53) implies $\Delta V(\mathscr{F}, \mathscr{F}_{-i}) = V_i\mu_i$. Under base customer switching, capacity rationing is optimal if and only if $\max_{\mathscr{C}\subset\mathscr{F}} \tilde{V}(\mathscr{C}) > C > \min_{i\in\mathscr{F}} \Delta V(\mathscr{F}, \mathscr{F}_{-i})$: the first inequality must hold for profitable operation, and the second must hold for profits to increase by denying service to at least one type.

5.2. Word of Mouth About Service Quality

The policy of rationing capacity and denying service to some base customers may adversely effect the firm's ability to attract new customers. We show that negative WOM from base customers about their service quality reduces, but does not eliminate, the capacity cost range in which such a policy is optimal. For simplicity we focus on the case of homogeneous base customers ($m = 1$), but this intuitive result continues to hold in the case of heterogeneous base customers.

We consider λ_0 to be the *maximum* new customer arrival rate that is generated by spending $S(\lambda_0)$ on advertising. To model negative WOM about service quality, we assume that the *effective* new customer arrival rate, λ_0^e, decreases in the base customer service denial rate, $x_1r_1(1 - q_1)$:

$$\lambda_0^e := \lambda_0 - \delta x_1 r_1(1 - q_1) = \lambda_0 - \delta(x_1r_1 - N_1\mu_1). \tag{54}$$

The parameter $\delta \geq 0$ captures the WOM intensity. The second equality in (54) follows from (17).

We denote the effective new customer arrival rate by $\lambda_0^e(\mathbf{N}, \lambda_0)$, as it also depends on the capacity allocation $\mathbf{N}$ as discussed below. Let $\Pi^w(\mathbf{N}, N, \lambda_0)$ be the profit function with WOM.

WOM has two effects on problem (22)–(25). First, in the profit function (22) the maximum new customer service denial cost changes from $\lambda_0 c_0$ to $\lambda_0^e(\mathbf{N}, \lambda_0)c_0$. For $m = 1$ we get

$$\Pi^w(\mathbf{N}, N, \lambda_0) = N_0V_0\mu_0 + N_1V_1\mu_1 - \lambda_0^e(\mathbf{N}, \lambda_0)c_0 - CN - S(\lambda_0). \tag{55}$$

Second, the capacity allocation constraint for new customers (23) changes from $N_0 \leq \lambda_0 s_0$ to

$$N_0 \leq \lambda_0^e(\mathbf{N}, \lambda_0)s_0. \tag{56}$$

To consider how $\lambda_0^e(\mathbf{N}, \lambda_0)$ depends on the capacity allocation, let $a := \bar{\theta}_{01}r_1T_{11}(0)$. Then we have from (54), together with (6), (18), (20), and (21), that

$$\lambda_0^e(\mathbf{N}, \lambda_0) = \lambda_0 - \delta a\left(\frac{N_0}{s_0} - \frac{N_1}{s_1}\right). \tag{57}$$

Increasing N_0 increases the number of new customers served at a rate $1/s_0$, which raises the base customer service denial rate by a/s_0. However, increasing N_1 increases the number of base customers served at a rate $1/s_1$, which reduces their service denial rate by a/s_1. That is, a measures the increase (decrease) in base customer service denials per new (base) customer served. If all base customers are served ($N_1/s_1 = N_0/s_0$), there is no negative WOM, and $\lambda_0^e(\mathbf{N}, \lambda_0) = \lambda_0$ in (57).

In summary, the problem with WOM is to maximize the profit (55) subject to the constraints (24), (25), (56), and (57). Let λ_0^{w*} denote the optimal new customer arrival rate and N^{w*} the optimal capacity level with WOM. Proposition 4 shows that WOM reduces, but does not eliminate, the capacity cost range in which rationing capacity is optimal (see Online Appendix A for details).

Proposition 4. *Consider a system with homogeneous base customers ($m = 1$). WOM affects the optimal policy and reduces the optimal profit if and only if $\tilde{V}_0 > C > V_1\mu_1$. Let*

$$\bar{V}_1^w(\delta) := \frac{s_0(\delta a/(1 + \delta a))\tilde{V}_0 + s_1V_1\mu_1}{s_0(\delta a/(1 + \delta a) + s_1}. \tag{58}$$

1. *If $\tilde{V}_0 > \bar{V}_1^w(\delta) > C > V_1\mu_1$, then with WOM it is optimal to serve all customers, $N^{w*} = \lambda_0^{w*}\bar{s}_1$, but without WOM it is optimal to serve only new customers, $N^* = \lambda_0^* s_0$. WOM yields a lower advertising level, $\lambda_0^{w*} < \lambda_0^*$, but a higher or lower capacity level than without WOM.*

2. *If* $\tilde{V}_0 > C > \bar{V}_1^w(\delta) > V_1\mu_1$, *then with or without WOM, it is strictly optimal to serve only new customers, but WOM reduces advertising and capacity levels*: $\lambda_0^{w*} < \lambda_0^*$ *and* $N^{w*} = \lambda_0^{w*}s_0/(1+\delta a) < N^* = \lambda_0^* s_0$. *In the limit as* $\delta \leftarrow \infty$, *WOM makes it unprofitable to operate.*

The metric $\bar{V}_1^w(\delta)$ captures, for a system that prioritizes new customers, the average value per processing time from serving additional base customer requests (with value $V_1\mu_1$) and all the new customer requests (with value $\tilde{V}_0$) that are generated as a result. With WOM it is therefore optimal to serve base customers if and only if $\bar{V}_1^w(\delta) > C$. The ratio $\delta a/(1+\delta a)$ in $\bar{V}_1^w(\delta)$ measures the WOM-driven increase in the effective new customer arrival rate per increase in base customer throughput. Without WOM ($\delta = 0$) this ratio is clearly 0, so that $\bar{V}_1^w(0) = V_1\mu_1$, as in the basic model. With extreme WOM ($\delta \to \infty$) this ratio is one; that is, each additional base customer served generates an additional new customer. In this limiting case WOM *imposes* a policy that serves all customers, so $\bar{V}_1^w(\infty) = \tilde{V}_1 < C$; this policy is unprofitable by Part 2 of Proposition 4.

6. Implications

In Section 6.1 we illustrate how the optimal service policy depends on customer attributes. In Section 6.2 we highlight the key role of the optimal capacity allocation policy in controlling the customer base. In Section 6.3 we show that ignoring the effect of service probabilities on the CLV may significantly reduce performance. Finally, in Section 6.4 we discuss the application of our deterministic fluid model to approximate and optimize large-scale stochastic queueing systems with abandonment.

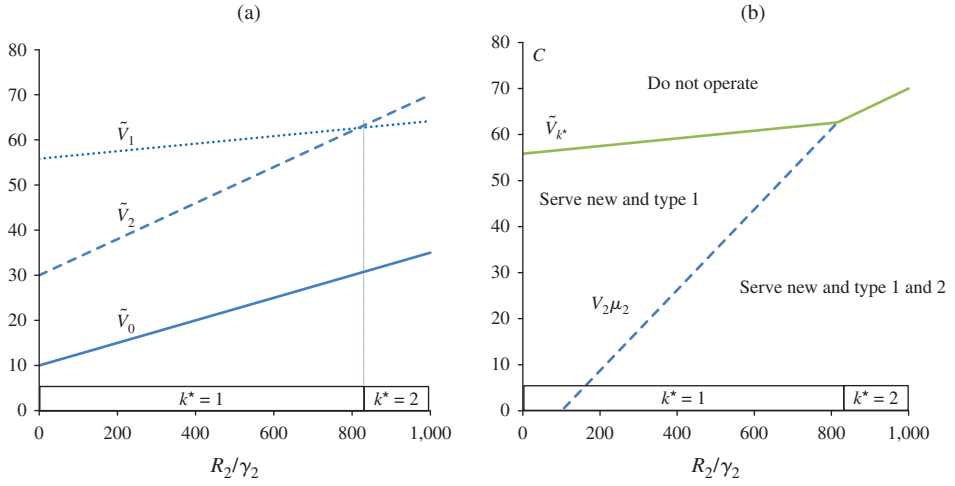
6.1. The Effects of Customer Attributes on the Optimal Service Policy

We present two examples of how the optimal policy of Proposition 3 depends on customer attributes. Throughout we assume $p_0 = -10$ and $c_0 = 0$ for new customers, two base customer types with $p_i = -10$, $c_i = 10$, $\bar{\theta}_{0i} = 0.2$, $\bar{\theta}_{ii} = 1$, and $r_i/\gamma_i = 10$ for $i = 1, 2$, and $\mu_i = 1$ for $i \in \{0, 1, 2\}$.

Example 1. *Effect of service-independent profit on optimal service policy.*

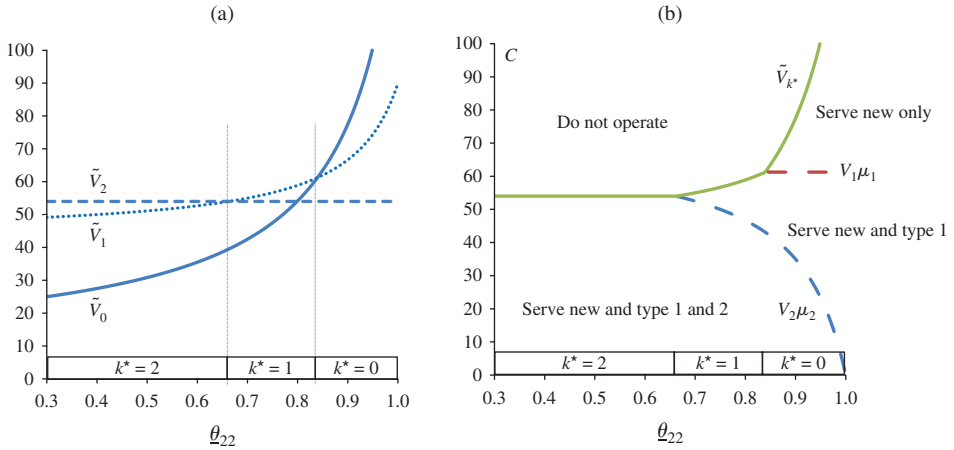
Consider the case where the service-independent profit of type 1 exceeds that of type 2: We fix $R_1/\gamma_1 =$ 1,000 and vary $R_2/\gamma_2 \in [0, 1{,}000]$. We assume equally loyal base customers: $\underline{\theta}_{11} = \underline{\theta}_{22} = 0.3$. Leaving the capacity cost aside, Figure 2(a) shows how the service-independent type 2 profit R_2/γ_2 effects a new customer's total value per processing time under three policies: serving no base customers ($\tilde{V}_0$), serving only those of type 1 ($\tilde{V}_1$), or serving both types ($\tilde{V}_2$), respectively. The policy that maximizes this new customer value denies service to type 2 customers (so $k^* = 1$) if their service-independent profit is below a threshold ($R_2/\gamma_2 < 820$), but serves them otherwise ($k^* = 2$ for $R_2/\gamma_2 \geq 820$). By Proposition 3 the optimal service policy depends not only on the maximum new customer value $\tilde{V}_{k^*}$, but also on the $V\mu$ index of lower ranked types $i > k^*$, and on the capacity cost C. Figure 2(b) identifies the optimal policy depending on the capacity cost C and the service-independent type 2 profit R_2/γ_2: by Proposition 3, it is not profitable to operate if the capacity cost exceeds $\tilde{V}_{k^*}$; it is optimal to deny service to type 2 customers if $k^* = 1$ and $\tilde{V}_1 > C > V_2\mu_2$, which holds if their service-independent profit is

Figure 2. (Color online) Example 1: Effect of Service-Independent Profit on Optimal Service Policy



Note. Service-independent type 1 profit: $R_1/\gamma_1 = 1{,}000$. (a) New customer value per processing time as a function of the service policy and the service-independent type 2 profit R_2/γ_2. (b) Optimal policy as a function of the capacity cost C and R_2/γ_2.

Figure 3. (Color online) Example 2: Effect of Loyalty on Optimal Service Policy



Note. Type 1 loyalty probability: $\underline{\theta}_{11} = 0.3$. (a) New customer value per processing time as a function of the service policy and the type 2 loyalty probability $\underline{\theta}_{22}$. (b) Optimal policy as a function of the capacity cost C and $\underline{\theta}_{22}$.

sufficiently low ($R_2/\gamma_2 < 820$), but to serve all customers otherwise.

Example 2. *Effect of loyalty on optimal service policy.*

Consider the case where type 1 customers are less loyal than type 2 customers after service denial: We fix $\underline{\theta}_{11} = 0.3$ and vary $\underline{\theta}_{22} \in [0.3, 1.0]$. We assume equally lucrative base customers: $R_1 = R_2 = 800$. As shown in Figure 3(a), leaving the capacity cost aside, the policy that maximizes a new customer's total value per processing time serves all requests only if type 2 customers' loyalty is below a threshold ($k^* = 2$ for $\underline{\theta}_{22} \le 0.66$), but denies service to type 2 if their loyalty is moderate ($k^* = 1$ for $0.66 < \underline{\theta}_{22} \le 0.83$) and to all base customers if type 2 are very loyal upon service denial ($k^* = 0$ for $\underline{\theta}_{22} > 0.83$). Figure 3(b) indicates the optimal service policy depending on the capacity cost C and the type 2 loyalty probability $\underline{\theta}_{22}$: By Proposition 3, determining the optimal service policy requires comparing the capacity cost C with $V_2\mu_2$ if $k^* = 1$, but with both $V_1\mu_1$ and $V_2\mu_2$ if $k^* = 0$. These comparisons yield the four policies in Figure 3(b).

6.2. Using the Capacity Allocation Policy to Control the Customer Base

Our analysis implies that the capacity allocation policy plays a key role in controlling both the composition and the size of the customer base: by targeting different service levels to different customer types, the allocation policy can influence both the switching rates among, and the retention rates of, base customer types. Here we focus on the retention effects in the absence of base customer switching. From (7) the ratio of the customer bases of two types, say 1 and 2, satisfies

$$\frac{x_1}{x_2} = \frac{\lambda_0 q_0 \bar{\theta}_{01}/(\gamma_1 + r_1(1 - q_1\bar{\theta}_{11} - (1-q_1)\underline{\theta}_{11}))}{\lambda_0 q_0 \bar{\theta}_{02}/(\gamma_2 + r_2(1 - q_2\bar{\theta}_{22} - (1-q_2)\underline{\theta}_{22}))}, \quad (59)$$

where $\lambda_0 q_0 \bar{\theta}_{0j}$ is the type j joining rate and $1/(\gamma_j + r_j(1 - q_j\bar{\theta}_{jj} - (1-q_j)\underline{\theta}_{jj}))$ is the type j lifetime.

To illustrate how the optimal policy can effect this ratio we revisit Examples 1 and 2 of Section 6.1, assuming that both types have equal service-independent attrition rates ($\gamma_1 = \gamma_2$). Recall that both types have equal probabilities for joining ($\bar{\theta}_{01} = \bar{\theta}_{02} = 0.2$) and remaining ($\bar{\theta}_{11} = \bar{\theta}_{22} = 1$) in the customer base after being served, and equal normalized service request rates ($r_1/\gamma_1 = r_2/\gamma_2 = 10$). It follows from (59) that, if giving equal service to both types is optimal (so $q_1^* = q_2^*$), then their customer bases are equal; that is, $x_1^* = x_2^*$. However, if serving only type 1 is optimal (so $q_1^* = 1$, $q_2^* = 0$), then the customer base of type 1 is larger than that of type 2: by (59) we have

$$\frac{x_1^*}{x_2^*} = \frac{\gamma_2 + r_2(1 - \underline{\theta}_{22})}{\gamma_1} = 1 + 10(1 - \underline{\theta}_{22}), \quad (60)$$

where the fraction represents the ratio of type 2 to type 1 departure rates, and the second equality holds since $\gamma_1 = \gamma_2$ and $r_2/\gamma_2 = 10$. The type 2 customer base is smaller because its members leave at a larger rate because of service denial, at rate $r_2(1 - \underline{\theta}_{22})$ where r_2 is their service request rate and $\underline{\theta}_{22}$ is their loyalty probability given service denial. In Example 1, this probability is fixed at $\underline{\theta}_{22} = 0.3$, so from (60) we have $x_1^* = 8x_2^*$ whenever it is optimal to deny service to type 2. However, in Example 2 the type 2 loyalty probability $\underline{\theta}_{22}$ varies, so the customer base ratio x_1^*/x_2^* depends not only on the optimal policy but also on this loyalty probability. To illustrate this point, consider in Figure 3(b) how the optimal policy at a capacity cost of $C = 50$ varies with $\underline{\theta}_{22}$. For $\underline{\theta}_{22} \le 0.75$ it is optimal to serve all customers so that $x_1^* = x_2^*$. For larger values of $\underline{\theta}_{22}$ the optimal policy is to deny service to type 2 customers, but by (60) the customer base ratio x_1^*/x_2^* decreases

in $\underline{\theta}_{22}$: the higher the loyalty of type 2 customers after service denial, the larger their base relative to that of type 1. In the limit as $\underline{\theta}_{22} \to 1$, type 2 customers are so loyal that they leave only for service-independent reasons, and we again have $x_1^* = x_2^*$.

6.3. Ignoring the Effect of Service Probabilities on the CLV Hurts Performance

Standard approaches in the marketing literature ignore the effect of service probabilities on the CLV. We show that doing so may significantly reduce performance. We focus on Example 1 in Section 6.1 with $R_2/\gamma_2 = 250$, so $\tilde{V}_1 > \tilde{V}_2 > V_2\mu_2$ and $k^* = 1$ by Figure 2. By Proposition 3 the optimal policy is to deny service to type 2 customers if $\tilde{V}_1 > C > V_2\mu_2$ but to serve all customers if $V_2\mu_2 \geq C$. We contrast this policy with two alternative policies, *marketing driven* and *uncoordinated*.

In the *marketing-driven* policy, the marketing department optimizes the advertising and capacity levels, *assuming* that all requests must be served. Let λ_0^M and N^M denote, respectively, the optimal new customer arrival rate and capacity level under this policy. The new customer arrival rate λ_0^M balances the marginal cost of acquiring a new customer with the profit when serving all her requests, that is, $\lambda_0^M > 0$ satisfies

$$\bar{s}_2(\tilde{V}_2 - C) = S'(\lambda_0^M) \quad \text{if } \tilde{V}_2 > C, \tag{61}$$

and $\lambda_0^M = 0$ otherwise. The capacity level $N^M = \lambda_0^M(s_0 + s_1 + s_2)$ because all requests are served.

In the *uncoordinated* policy, the marketing department also optimizes the advertising level, assuming that all requests will be served, but the operations department optimizes the capacity allocation policy and capacity level (in line with Proposition 2), given the new customer arrival rate set by marketing. Let λ_0^U and N^U denote, respectively, the optimal new customer arrival rate and capacity level under this policy. Then $\lambda_0^U = \lambda_0^M$, because both policies determine the new customer arrival rate by (61). By Proposition 2 the optimal capacity satisfies $N^U = \lambda_0^U(s_0 + s_1)$ if $\tilde{V}_1 > C > V_2\mu_2$, and $N^U = \lambda_0^U(s_0 + s_1 + s_2)$ if $V_2\mu_2 \geq C$ (note that $\tilde{V}_1 = \bar{V}_1$ since $c_0 = 0$).

Figure 4 compares the capacity and the new customer arrival rate for the optimal, marketing-driven, and uncoordinated policies, depending on the capacity cost. For $C > V_2\mu_2$, both the marketing-driven and the uncoordinated policies reduce the arrival rate, and even cause the system to shut down for $C \geq \tilde{V}_2$, because they *impose* a suboptimal service level that reduces the new customer CLV (since $\tilde{V}_2 < \tilde{V}_1$). For $C > V_2\mu_2$, the uncoordinated policy also yields a lower-than-optimal capacity, whereas the capacity under the marketing-driven policy may be lower or higher than optimal, because of two countervailing effects: the new customer arrival rate is lower than optimal (i.e., $\lambda_0^M < \lambda_0^*$), but all requests are served, rather than only those of new customers, as is optimal.

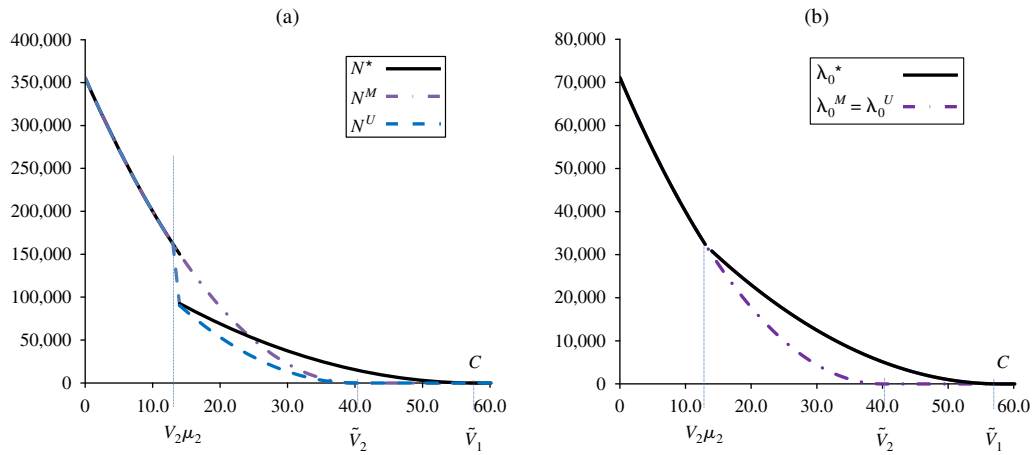
These policies may result in significant profit losses. At a cost of $C = 25$, the loss relative to the optimal profit is 56% for the marketing-driven and 15% for the uncoordinated policy. These losses increase in the capacity cost, up to the extreme case of suboptimal shutdown for $C \in [\tilde{V}_2, \tilde{V}_1)$.

This discussion underscores the importance of accounting for the *optimal* service level in evaluating the CLV, particularly when this metric drives substantial resource allocation decisions. Any policy that imposes arbitrary service levels (e.g., FIFO for all customer types with an industry-standard service access quality) would be similarly suboptimal.

6.4. Implementing the Fluid-Optimal Policy in a Large Stochastic System

We discuss the application of the deterministic fluid model to optimize large-scale stochastic queueing

Figure 4. (Color online) Ignoring the Effect of Service Probabilities on the CLV Hurts Performance



Notes. (a) Capacity and (b) new customer arrival rate as functions of the capacity cost, for the optimal (∗), marketing driven (M), and uncoordinated (U) policies. (Parameters of Example 1, with $R_2/\gamma_2 = 250$.)

systems with abandonment, such as the credit card call center that motivated this research.

6.4.1. The Primitives of the Stochastic Queueing Model. Consider a service facility such as an inbound call center operating as a stochastic queueing system with N identical parallel servers. New customer arrivals follow a Poisson process with rate λ_0. Type i base customers' call interarrival times, service times, and service-independent sojourn times are independent draws from exponential distributions with means $1/r_i$, $1/\mu_i$, and $1/\gamma_i$, respectively. The financial parameters (R_i, p_i, and c_i) and the switching probabilities $\theta_{ij}(q_i)$ are as in the deterministic model.

Unlike in the deterministic model, because of queueing delays and customer impatience, the stochastic system may fail to serve all requests even if it is underutilized. We model type i customers' impatience by independent and exponentially distributed abandonment times with mean τ_i. Abandonments are equivalent to service denials (and denied requests are lost, as in the fluid model). In this model customer base transitions depend on waiting times only through the resulting service probabilities q_i: conditional on the outcome of receiving or being denied service, a customer's transition is independent of her waiting time. This assumption is consistent with the notions of "critical incidents" and "end effects" in the service literature. The prevalent assumption is that customers think about terminating their relationship with providers only when some critical incident occurs (Keaveney 1995, Gremler 2004). In our setting the abandonment is the natural critical incident: this is how customers signal that their waiting times are too long. Bitran et al. (2008a) (see Section 4.2.1 and references therein) suggest the presence of an "end effect," whereby the outcome of a service encounter may dominate the memory of the preceding waiting experience.

6.4.2. Implementing the Fluid-Optimal Policy in the Stochastic Queueing Model. Because of the state dependence and feedback in customer flows, the relationships between the capacity allocation policy and the service probabilities seem to be analytically intractable in the stochastic model, unlike their counterparts (16) and (17) in the fluid model. It is therefore difficult to optimize the stochastic system directly. However, simulation results summarized in Section 6.4.3 suggest that the following natural implementation of the fluid-optimal capacity allocation policy yields nearly optimal performance for a stochastic system with sufficiently many servers: operate a head-of-the-line priority policy that gives customer types strict (nonpreemptive) priorities according to the ranking specified in part 1 of Proposition 1. (In the case with base customer switching, the fluid-optimal priority ranking is determined by solving the problem numerically.) This implementation assumes flexible servers, so the firm can pool capacity across types. This is common in practice, including in the credit card company discussed above. Another practice is to dedicate capacity to each type.

The fluid-optimal policy of Proposition 1 features two key differences from standard index policies for systems with abandonment, such as the $c\mu/\theta$ rule (Atar et al. 2010) and $c\mu$ type policies (Tezcan and Dai 2010). (1) The $V\mu$ indices consider the effect of service on customers' *future* requests and financial impact. (2) The values $\bar{V}_k$ and $\tilde{V}_k$ that determine the priority of one type, new customers, also depend on the $V\mu$ indices of *other* types. These differences reflect that, unlike standard models, ours captures customer base transitions that link future requests to past service quality.

6.4.3. Summary of Simulation Results. We briefly summarize results from simulations (see Online Appendix B for details) that evaluate the performance of the fluid-optimal policy in the stochastic queueing model described above. Focusing for simplicity on the case of homogeneous base customers, we consider a number of customer parameter combinations, some where it is never optimal and others where it may be optimal to deny service to base customers, according to Proposition 3. For each parameter combination we vary the server cost C in a range that yields enough servers, 100 or more, for the fluid model results to be applicable. We compare (1) the fluid-optimal new customer arrival rate, capacity level, and priority ranking with their counterparts from simulation-based optimization and (2) the simulation-based profit under the fluid-optimal prescriptions with the simulation-based optimal profit.

The main finding from over 350 simulation experiments is that on average the fluid-optimal prescriptions yield a relative profit loss below 1%. We observe worse profit performance (losses around 6%) only at capacity costs with jumps in the fluid-optimal service and capacity levels. (The fluid-optimal new customer arrival rate and capacity typically deviate more from simulation-optimal levels, whereas the fluid-optimal priority ranking is typically optimal for the stochastic system.)

7. Concluding Remarks

We study the profit-maximizing advertising, capacity, and capacity allocation policies for a service firm with heterogeneous repeat customers whose acquisition, retention, and behavior during their lifetime in the customer base depend on their service probability. We develop our results in the context of a new fluid model that integrates CRM and capacity management. This model links the makeup and value of the customer base both to the capacity allocation, unlike prior CRM models, and to the service access quality of past interactions, unlike prior capacity management models.

We make several contributions to CRM and service capacity management: We derive new metrics that link the value of a customer to the capacity allocation policy and the resulting service probabilities and customer base transitions of all types: the CLV of base customers, the $V\mu$ index, and the policy-dependent value of a customer type. We provide new capacity management prescriptions that hinge on these metrics, notably, optimality conditions for rationing capacity and for identifying which customers to deny service. These results have important implications: firms need to understand how customer attributes affect the optimal policy; with repeat business the capacity allocation policy plays a previously ignored key role in controlling the customer base; marketing-focused policies that ignore the effect of service probabilities on the CLV may reduce profits significantly; and the fluid model approach may also prove effective for CRM and capacity management for stochastic systems such as the credit card call center that motivated this work.

Our results prescribe a "bang-bang" structure for the optimal capacity allocation: customers get either perfect service or none at all. This may seem unrealistic in some cases. However, it is easy to modify these prescriptions to ensure a minimum, strictly positive, service probability even for the least profitable types. The corresponding modified fluid-optimal capacity allocation can be determined by adding minimum-capacity constraints to the profit-maximization problem. Implementing this modified allocation in a stochastic system would increase the fluid-optimal number of servers and thereby reduce the delays and the abandonment rates of the lower-priority classes.

Our results also point to the interplay between the targeting levels of the advertising and capacity allocation policies. On one hand, as shown in Section 6.3, imposing or assuming suboptimal nontargeted service levels reduces the value of an acquired customer and yields lower than optimal advertising spending. On the other hand, our results suggest that, if the firm could better target advertising in order to selectively acquire more profitable customer types (our model assumes it cannot), then the optimal service policy would target high service levels to more, or possibly all, customer types.

Our study focuses on a stationary environment that yields constant optimal arrival rates. In nonstationary environments, for example, if the new customers' demand response to advertising changes over time, the optimal service probabilities may be time dependent. Establishing the structure of the optimal policies becomes more complicated as a result of the interplay between time-dependent service probabilities and customer value metrics. However, if the capacity level can be adjusted on the time scale of demand fluctuations, restricting attention to stationary service policies while dynamically adjusting the advertising and capacity levels not only simplifies the analysis (the customer value metrics presented in this paper remain valid) but may also be practically appealing and nearly optimal. Araghi (2014) studies a special case where the firm follows a periodic advertising policy and the new customer arrival rate decays exponentially between advertising pulses.

We focus for simplicity on controlling the customer base through differentiated service levels via the capacity allocation policy. However, our framework can be extended to allow for targeted advertising to base customers. For example, in the case of the credit card company, base customers' credit card spending, which corresponds to their service-independent profit R_i, may depend both on their service access quality and on advertisements. In this case the switching probabilities θ_{ij} would be functions of both the service probability and the advertising dollars targeted to type i.

Finally, our model parameters can be estimated based on data that can be tracked. Such estimates would help quantify the effects of service access quality attributes on CLV, a key requirement for the practical implementation of capacity management policies that reflect CRM principles.

Acknowledgments

The authors wish to thank the review team for insightful and detailed comments that helped improve the paper significantly. Support for this research was provided by grants from the Natural Science and Engineering Research Council of Canada.

References

Adelman D, Mersereau AJ (2013) Dynamic capacity allocation to customers who remember past service. *Management Sci.* 59(3): 592–612.

Aflaki S, Popescu I (2014) Managing retention in service relationships. *Management Sci.* 60(2):415–433.

Akşin OZ, Armony M, Mehrotra V (2007) The modern call-center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* 16(6):665–688.

Anand KS, Paç MF, Veeraraghavan S (2011) Quality–speed conundrum: trade-offs in customer-intensive services. *Management Sci.* 57(1):40–56.

Anderson EW, Sullivan MW (1993) The antecedents and consequences of customer satisfaction for firms. *Marketing Sci.* 12(2): 125–143.

Anton J, Setting T, Gunderson C (2004) Offshore company call centers: A concern to U.S. consumers. Technical report, Purdue University Center for Customer-Driven Quality, Lafayette, IN.

Araghi M (2014) Problems in service operations with heterogeneous customers. Ph.D. thesis, University of Toronto, Toronto, Canada.

Ascarza E, Hardie BGS (2013) A joint model of usage and churn in contractual settings. *Marketing Sci.* 32(4):570–590.

Atar R, Giat C, Shimkin N (2010) The $c\mu/\theta$ rule for many-server queues with abandonment. *Oper. Res.* 58(5):1427–1439.

Bitran GR, Mondschein S (1996) Mailing decisions in the catalog sales industry. *Management Sci.* 42(9):1364–1381.

Bitran GR, Ferrero J, Rocha e Oliveira P (2008a) Managing customer experiences: perspectives on the temporal aspects of service encounters. *Manufacturing Service Oper. Management* 10(1):61–83.

Bitran GR, Rocha e Oliveira P, Schilkrut A (2008b) Managing customer relationships through price and service quality. Working paper, IESE, Madrid.

Blattberg RC, Deighton J (1996) Manage marketing by the customer equity test. *Harvard Bus. Rev.* 74(4):136–145.

Bolton RN (1998) A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing Sci.* 17(1):45–65.

Braun M, Schweidel DA (2011) Modeling customer lifetimes with multiple causes of churn. *Marketing Sci.* 30(5):881–902.

de Véricourt F, Zhou Y-P (2005) Managing response time in a call-routing problem with service failure. *Oper. Res.* 53(6):968–981.

Farzan A (2013) Quality and capacity decisions in service processes. Ph.D. thesis, University of Washington, Seattle.

Feichtinger G, Hartl RF, Sethi SP (1994) Dynamic optimal control models in advertising: Recent developments. *Management Sci.* 40(2):195–226.

Feinberg FM (2001) On continuous-time optimal advertising under S-shaped response. *Management Sci.* 47(11):1476–1487.

Gans N (2002) Customer loyalty and supplier quality competition. *Management Sci.* 48(2):207–221.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review and research prospects. *Manufacturing Service Oper. Management* 5(2):79–141.

Gaur V, Park Y (2007) Asymmetric consumer learning and inventory competition. *Management Sci.* 53(2):227–240.

Genesys Telecommunications Labs (2007) *Genesys Global Consumer Survey* (Genesys Telecommunications Lab, Daly City, CA).

Green LV, Kolesar PJ, Whitt W (2007) Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper. Management* 16(1):13–39.

Gremler (2004) The critical incident technique in service research. *J. Service Res.* 7(1):65–89.

Günes ED, Akşin OZ, Örmeci EL, Özden SH (2010) Modeling customer reactions to sales attempts: If cross-selling backfires. *J. Service Res.* 13(2):168–183.

Gupta S, Hanssens D, Hardie B, Kahn W, Kumar V, Lin N, Ravishanker N, et al. (2006) Modeling customer lifetime value. *J. Service Res.* 9(2):139–155.

Hall J, Porteus E (2000) Customer service competition in capacitated systems. *Manufacturing Service Oper. Management* 2(2):144–165.

Hassin R (2016) *Rational Queueing* (Taylor & Francis, Boca Raton, FL).

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

Ho TH, Park YH, Zhou YP (2006) Incorporating satisfaction into customer value analysis: Optimal investment in lifetime value. *Marketing Sci.* 25(3):260–277.

Hopp WJ, Iravani SMR, Yuen GY (2007) Operations systems with discretionary task completion. *Management Sci.* 53(1):61–77.

Keaveney MS (1995) Customer switching behavior in service industries: An exploratory study. *J. Marketing* 59(2):71–82.

Lewis MA (2005) Dynamic programming approach to customer relationship pricing. *Management Sci.* 51(6):986–994.

Li S, Sun B, Wilcox RT (2005) Cross-selling sequentially ordered products: An application to consumer banking. *J. Marketing Res.* 42(2):233–239.

Liu L, Shang W, Wu S (2007) Dynamic competitive newsvendors with service-sensitive demands. *Manufacturing Service Oper. Management* 9(1):84–93.

Musalem A, Joshi YV (2009) How much should you invest in each customer relationship? A competitive strategic approach. *Marketing Sci.* 28(3):555–565.

Olsen TL, Parker RP (2008) Inventory management under market size dynamics. *Management Sci.* 54(10):1805–1821.

Ovchinnikov A, Boulu-Reshef B, Pfeifer PE (2014) Balancing acquisition and retention spending for firms with limited capacity. *Management Sci.* 60(8):2002–2019.

Pfeifer PE, Ovchinnikov A (2011) A note on willingness to spend and customer lifetime value for firms with limited capacity. *J. Interactive Marketing* 25(3):178–189.

Reinartz W, Venkatesan R (2008) Decision models for customer relationship management (CRM). Wierenga B, ed. *Handbook of Marketing Decision Models* (Springer Science, Berlin), 291–326.

Reinartz W, Thomas JS, Kumar V (2005) Balancing acquisition and retention resources to maximize customer profitability. *J. Marketing* 69(1):63–79.

Rust RT, Chung TS (2006) Marketing models of service and relationships. *Marketing Sci.* 25(6):560–580.

Rust RT, Verhoef PC (2005) Optimizing the marketing interventions mix in intermediate-term CRM. *Marketing Sci.* 24(3):477–489.

Rust RT, Lemon K, Zeithaml V (2004) Return on marketing: Using customer equity to focus marketing strategy. *J. Marketing* 68(1): 109–127.

Rust RT, Zahorik AJ, Keinigham TL (1995) Return on quality (ROQ): Making service quality financially accountable. *J. Marketing* 59(2):58–70.

Sasieni MW (1971) Optimal advertising expenditure. *Management Sci.* 18(4):64–72.

Schwartz BL (1966) A new approach to stockout penalties. *Management Sci.* 12(12):B538–B544.

Sethi SP, Zhang Q (1995) Multilevel hierarchical decision making in stochastic marketing-production systems. *SIAM J. Control Optim.* 33(2):528–553.

Simon JL, Arndt J (1980) The shape of the advertising response function. *J. Adv. Res.* 20(4):11–28.

Sun B, Li S (2011) Learning and acting on customer information: A simulation-based demonstration on service allocations with offshore centers. *J. Marketing Res.* 48(1):72–86.

Tezcan T, Dai J (2010) Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Oper. Res.* 58(1):94–110.

Venkatesan R, Kumar V (2004) A customer lifetime value framework for customer selection and resource allocation strategy. *J. Marketing* 68(4):106–125.

Verhoef PC (2003) Understanding the effect of CRM efforts on customer retention and customer share development. *J. Marketing* 67(4):30–45.

Zeithaml VA, Berry LL, Parasuraman A (1996) The behavioral consequences of service quality. *J. Marketing* 60(4):31–46.