IESE Business School
University of Navarra

Working Paper
WP-750
April, 2008

MANAGING CUSTOMER RELATIONSHIPS THROUGH PRICE AND SERVICE QUALITY

Gabriel R. Bitran[1]

Paulo Rocha e Oliveira[2]

Ariel Schilkrut[3]

[1] Professor of Management, MIT Sloan School of Management

[2] Professor of Marketing, IESE

[3] Chief Operating Office, Scopix

IESE Business School – University of Navarra
Av. Pearson, 21 – 08034 Barcelona, Spain. Phone: (+34) 93 253 42 00 Fax: (+34) 93 253 43 43
Camino del Cerro del Águila, 3 (Ctra. de Castilla, km 5,180) – 28023 Madrid, Spain. Phone: (+34) 91 357 08 09 Fax: (+34) 91 357 29 13

Managing Customer Relationships Through Price and Service Quality

Gabriel R. Bitran • Paulo Rocha e Oliveira • Ariel Schilkrut

School of Management, Massachusetts Institute of Technology, Cambridge MA 02142

IESE Business School, Universidad de Navarra, Barcelona, Spain

Scopix, San Diego, CA 92130

gbitran@mit.edu • paulo@iese.edu • Ariel.Schilkrut@sloan.mit.edu

---

This paper examines the ways in which a service provider's policies on pricing and service level affect the size of its customer base and profitability. The analysis begins with the development of a customer behavior model that uses customer satisfaction and depth of relationship as mediators of the impact of price and service level on profitability. Based on this model of customer behavior, the system is analyzed as a queueing network from which the properties of the aggregate population's behavior are derived. The analysis reveals the counterintuitive result that a policy that involves a decrease in prices or an increase in service level may lead to a smaller customer base. However, this policy may also lead to higher profits. The novelty of this result lies in the explanation of the phenomenon that when the customer base decreases due to a change in prices or service quality, companies may experience gains in profit that result not from a decrease in costs associated with serving fewer customers but from an increase in revenues resulting from the indirect effects of the lower prices or higher level of service on customer behavior. The application of optimization techniques to the model developed in this paper yields optimality conditions through which managers can assess the long-term profitability of their pricing and service-level policies.

---

1 Introduction

This paper examines the impact of pricing and service quality on the size of the customer base and profitability. The setting in which the analysis takes place is a subscription-based, capacity-constrained service. The focus is on understanding the interdependence of the pricing policy and service level and their impact on customers' potential to generate revenue and customer behavior in terms of usage of the service. The key to the analysis is the development of a model where customers choose the depth of their relationship with the company based on their level of satisfaction. Deeper relationships increase the strain already faced by a capacity-constrained service-delivery system. If customers are satisfied and choose to pursue deeper relationships, the company will have to either lower its service quality or make investments to improve capacity. This paper provides a mathematical model that sheds light on the underlying dynamics governing such service-delivery systems, providing useful insights into the optimality of price- and quality-based managerial decisions.

A recent study of a cellular phone company conducted by Bain & Company, Inc. revealed instances when, within the same segment, customers with high usage levels were more likely to churn than customers with lower usage levels. Bittencourt and Sellmeister Bueno (2003) report a similiar finding in the financial services industry. This is a surprising result, particularly in light of the homogeneity of preferences across the customers examined. The predominant paradigm is that customers who don't use a service very often are the ones most likely to defect. Customer satisfaction has been shown to have a positive impact on both usage of services (e.g.: Heskett et al. 1994, Bolton and Lemon 1999) and customer retention (e.g.: Jones and Sasser 1995). Gourville and Soman (2002) make the causal relationship more explicit when they show that the probability that customers will cancel their membership to a service is inversely proportional to how often they use the service. Empirical support for this result comes from industries as diverse as health clubs (DellaVigna and Malmendie 2001) and cable television (Lemon, White and Winer 2002).

The apparent incongruence between the managerial observations brought to our attention and the academic predictions described above suggest that the relationship between service quality, customer retention, service usage and profitability could be more complex than previously supposed. Indeed, one of the objectives of this paper is to show that results such as those observed by Bain and by Bittercourt and Sellmeister Bueno (2003) are consistent with rational customer behavior if we incorporate into our model the dynamics of the customer's

choice of depth of relationship.

The frequency with which customers choose to interact with companies is often considered to be a reliable predictor of their lifetime value, a fact well known to the many managers in the catalog industry who have been successfully using the RFM (recency, frequency, and monetary value) framework for years. This follows from the intuition that customers who are satisfied with a service are likely to use it more often. Customers who like their cellular phones are more likely to use them in place of their regular phones, and those who like pay-per-view movies will use this service more often than they will rent movies. However, as customers use these services more often, they are increasingly more likely to erode the firm's profits for at least two reasons. First, because the intrinsic variability of service-delivery systems will lead to a higher number of service failures. Second, because the increased use of a service facility can lead to either a decrease in supplier responsiveness and service quality or higher costs in the form of further investments to prevent such failures.

Service quality is repeatedly cited (e.g., Rust et al. 1995, Bolton and Lemon 1999) to be a key determinant of switching behavior. An increase in quality or a decrease in price will make services more valuable to customers, but the effects of these policies on the long-term financial performance of a firm are not easily determined. A decrease in price or increase in service level may actually lead to a decrease in the size of the customer base. This can happen because these policy changes may yield the expected result of increased usage, and this puts a higher load on the system. Consequently, customers may experience more service failures, which lead to lower levels of satisfaction, resulting in a decrease in the size of the customer base in the long run. In this way, a managerial action which objectively gives more value to customers may ultimately drive some of them away.

This result is important to managers seeking to maximize their customer base in pursuit of higher profits, but when examining how changes in price and service quality can impact the size of the customer base, it is important to note that a decrease in the number of customers does not always leads to lower profits. The number of customers that a company has can actually be a remarkably poor indicator of the value of the customer base. Financial analysts often valued dotcoms based on their number of customers, and the market showed its disapproval of this metric when the prices of such companies crashed (Gupta, Lehman, and Stewart 2004). There is a complex relationship between pricing, service quality, and profitability, as the negative impact that changes (e.g., higher levels of quality) have on profit may have a revenue-based component as well as a cost-based component.

This paper connects the operational decisions of pricing and service quality with the behavior of customers reacting to these policies. In order to study this problem, §2 reviews the relevant literature from Marketing as well as Operations. Next, §3.1 develops the individual customer behavior model that is consistent with the relevant results from the literature and serves as the building block for the behavior of the customer base. Then, §3.2 shows how aggregating several customers behaving according to this model leads to a population behavior model whose steady-state behavior (analyzed in §4) can account for the phenomena observed by the companies described above. The analysis continues in §5, where the pricing and service-quality decisions are analyzed through a nonlinear program whose objective function is profit maximization. Finally, §6 discusses the managerial and academic relevance of these results and provides some directions for further research.

2 Literature Review

The range of issues addressed in this paper requires an interdisciplinary approach. Two streams of research in the pricing literature are particularly relevant. First, the literature from Operations Management and Queueing Theory informs our understanding of the impact of pricing and service level on system load. Second, the study of two-part pricing structures (mostly from Economics and Marketing) helps define the types of policies most suitable for subscription services. This paper also draws on an already interdisciplinary stream of research which studies the relationship between service quality, customer satisfaction, customer loyalty, and profitability.

Traditional Operations Management and Operations Research literature has focused on optimizing the firm's internal processes, making relatively simple assumptions about customer behavior and the cost and impact of service quality (Bitran, Ferrer and Oliveira 2008). Within these fields, there is a significant body of literature that analyzes queueing systems where the arrival rate depends on pricing and waiting time. The use of pricing as a mechanism for regulating the size of queues was first studied by Naor (1969), who introduced the notion of levying tolls to prevent customers from joining the queue during times of heavy congestion. He showed that social optima can be achieved through tolls or administrative constraints on the waiting space. Several papers generalized this model (e.g., Yechiali 1971, Knudsen 1972, Edelson and Hildebrand 1975, Lipmann and Stidham 1977, Mendelson and Yechiali 1981) by studying pricing decisions under fixed capacity. More recent extensions

include Van Mieghem's (2000) addition of the managerial control of scheduling and Chen and Frank's (2001) model where managers observe queue length and dynamically adjust their prices accordingly. Mendelson (1985) and Mendelson and Whang (1990) focused on problems related to optimal pricing and capacity allocation. Dewan and Mendelson (1990) extended these results to include customers with heterogeneous value functions. As in these papers, the analysis in the present paper is based on studying the impact of policies on the steady-state behavior of a queueing system. There have been significant subsequent advances concerning the existence of solutions, their stability, and the effect of small disturbances on equilibrium for these queueing systems (Stidham 1992, Friedman and Landsberg 1993 and 1996, Rump and Stidham 1998).

The present paper diverges from the traditional queueing literature by rejecting the assumption that each additional job submitted to the facility increases the social (gross) value. This assumption contends that at every period, either each customer uses the facility only once, or if she accesses the service more than once, the value of each interaction is not related to the number of jobs already submitted. Furthermore, the traditional queueing models also assume that customers use the same estimate of waiting time when deciding whether or not to use the service. In contrast, the present paper assumes that the value of the relationship to the customer depends explicitly on that customer's level of usage, the aggregation of which results in the expected total value. Finally, this paper takes into account the impact of past experiences on customer satisfaction and customer behavior, consistent with studies such as Bolton's (1998).

Unlike the models cited in the preceding paragraphs, the queueing system developed in this paper is controlled through service quality and a two-part pricing structure (also called dual pricing systems or two-part tariffs) which consists of a subscription fee that customers must pay in order to have access to the service and a usage fee that must be paid each time the service is used. This pricing structure was chosen for two reasons. First, because it can be interpreted as a generalization of subscription-only pricing or usage-only pricing by setting one of the price parameters to zero. Second, because it is a very commonly-used tool for price discrimination in practice (Tirole 1988), particularly in telecommunications and financial services, two of the industries that provided the main motivation for the present study. The analysis of two-part pricing structures in usage-based services dates back at least to Oi's (1971) pioneering work in the context of Disneyland tickets. The recent increase in the use of two-part pricing brought about by the Internet, telecommunications, and paid

television has renewed academic and managerial interest in various facets of the subject. Danaher (2002) built on the work of Mahajan et al. (1982) in order to find the prices that maximize the adoption rate of a new cellular phone service. Essegaier et al. (2002) developed a game-theoretic model to analyze the mediating effect of capacity constraints on the firm's optimal pricing strategy when consumers consistently behave as either "light" or "heavy" users. The problem studied in this paper requires an innovative approach for at least two reasons: first, because the existing literature on two-part pricing does not address the long-term effects on the size of the customer base and profitability; second, because the pricing and service-quality decisions must be considered simultaneously, along with their impact on customer behavior (e.g., a light user may become a heavy user if service quality improves).

The quality revolution in manufacturing spilled over to the service industry in the late 1980s and early 1990s and brought about a large number of managerial papers and books advocating the virtues of quality-oriented companies (e.g., Reichheld and Sasser, Jr. 1990, Reichheld and Teal 1996). The service-profit chain (Heskett et al. 1994), a framework connecting operational investments to profitability through service quality, took a prominent place in managerial circles. It did not take very long before this unprecedented high emphasis on quality came under scrutiny. Service firms often do not experience the economies of scale and corresponding cost reductions that were brought about by the implementation of quality programs in manufacturing firms. Consequently, many service companies faced the disastrous consequences of implementing financially unsound quality programs (e.g., Hill 1993, Wiesendanger 1993). Quality is a costly investment that must be linked to profitability.

The first step in linking quality to profitability involves linking quality to behavioral intentions (repurchasing intentions, in particular). There is extensive Marketing literature on this area (e.g., Rust and Zahorik 1993, Boulding et al. 1993). Bolton's (1998) analysis is particularly relevant to the present paper, as it studies the effect of customer satisfaction on loyalty, using data from a cellular communications firm. Her analysis reveals that the effect of a bad experience is smaller for customers who have been with the company longer, a result which was further supported by Rust et al. (1999).

Hall and Porteus (2000) and Gans (2002) made key contributions in modeling and understanding the connection between quality and customer loyalty in capacity-constrained services. In Hall's and Porteus' paper, customers switch between service providers based on their past service experience. Gans builds on this work by relating a firm's selection of service level to the duration of a customer's relationship with that firm. He considers the

setting of the service level to be a strategic decision and assumes that the firm can provide the targeted service level by adjusting its operational parameters (e.g., a call center can add or remove service representatives or a retailer can adjust the inventory policy). The model presented in this paper adds to this stream of research in many ways, three of which are particularly important. First, the updating mechanism allows the customers' current level of satisfaction with the service provider (as measured by their estimate of the service level) to be a function of the number of past experiences, previous estimates, and satisfaction from the last interaction. This set of assumptions is consistent with the empirical work of Bolton (1998) and Rust et al. (1999). Second, customers choose the depth of their relationship (operationalized through the rate of interactions) based on their level of satisfaction, allowing for the quantification of relationship depth in a way that is not possible in a model where interactions can only occur at predetermined points. This is an essential capability if we want to understand the way in which depth of relationship mediates the impact of customer satisfaction on profitability. Third, the present model allows for the simultaneous optimization of two-part pricing and service quality. This is an important managerial contribution, as price and service-quality level can usually be controlled by managers and two-part tariffs are the norm in a number of relevant industries.

Rust et al.'s (1995) Return on Quality (ROQ) framework and Kamakura et al.'s (2002) empirical implementation of the service-profit chain have established the final link between quality and profitability. The ROQ model provided a fundamental building block by explicitly quantifying the operational costs and increases in revenue associated with quality. Kamakura et al.(2002) took a similar approach as they incorporated quality-related costs into their operationalization of the service-profit chain. The present paper contributes to this stream of research by explicitly considering the effect of an additional phenomenon impacting the financial accountability problem: the negative impact that an investment in quality can have on profitability can actually come from the revenue side, not just the cost side.

3 Model Description

3.1 Individual Customer Behavior

Customers interact with the company repeatedly over time and are capable of initiating a service encounter whenever they choose to do so (the terms "service encounter" and "service

interaction" are used interchangeably in this paper). The service quality experienced by the customer during the k'th interaction is denoted by w_k, which depends on the company's internal service-quality level (denoted by W, which can be interpreted, for example, as the average waiting time) as well as a random factor discussed in the paragraphs below. After each interaction, customers update their expectation of the company's service quality through the recursive relationship

$$\tilde{w}_k = \alpha_k w_k + (1 - \alpha_k)\, \tilde{w}_{k-1} \tag{1}$$

where $\alpha_k \in (0, 1)$ is the weight assigned to the last experience. In this way, the customer's $\tilde{w}_k$ can be connected to the number of previous transactions. Two important special cases are $\alpha_k = \frac{1}{k}$, where the service estimate is the average of all previous service experiences, and $\alpha_k = \alpha$, which corresponds to exponential smoothing. This type of Bayesian updating mechanism where customers combine their last experience with a summary statistic of previous experiences is standard in the marketing literature and has been empirically validated in a variety of settings (e.g., Bolton 1998). Furthermore, the α_k parameter in the specific formulation presented in Eq (1) takes into account the fact that the customers' service-quality estimates can depend not only on their previous estimate and the actual level of service experienced in the last interaction, but also on the number of previous interactions, in accordance with the findings of Bolton (1998) and Rust et al (1999) discussed in §2.

Customer satisfaction is represented by the customer's estimate of the firm's true service level and is denoted by x, a random variable whose distribution is given by $F\left(x|\tilde{w}_k\right) = \tilde{F}_k$. This assumption has its limitations, but these do not affect the generality of the analytical approach or the results of this paper. The consumers' expectations, $F\left(x|\tilde{w}_k\right)$,are based on a prior $\tilde{w}_0$ (which is independent of W) and on a vector of observations $(w_1, w_2, ..., w_k)$ which are generated by the stochastic process governed by reality, or $F\left(\right)$. There is no value of W for which the consumer will never defect because the customer's estimate of $F\left(\right)$ will always be based on the observed realizations of this stochastic process, not on W itself. The methodology used in this paper can be extended to accommodate estimates and perceptions of higher moments of perceived service quality, allowing for more complex operational definitions of $\tilde{F}_k$. In spite of its real-world appeal, this additional complexity leads to the same results and insights, and therefore the simpler definition of customer satisfaction is preferred for the sake of clarity of exposition (the details of this analysis are

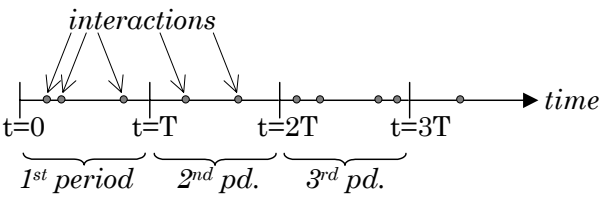

Figure 1: Example of service interactions over time

available from the authors).

3.1.1 Individual Customer Dynamics

The dynamics of the customer-company interactions are summarized in Figure 2. Note that this figure explicitly depicts how the firm's decision variables—p_u (the usage fee), p_s (the subscription fee), and W (the average waiting time)—are used to control the customer interface.

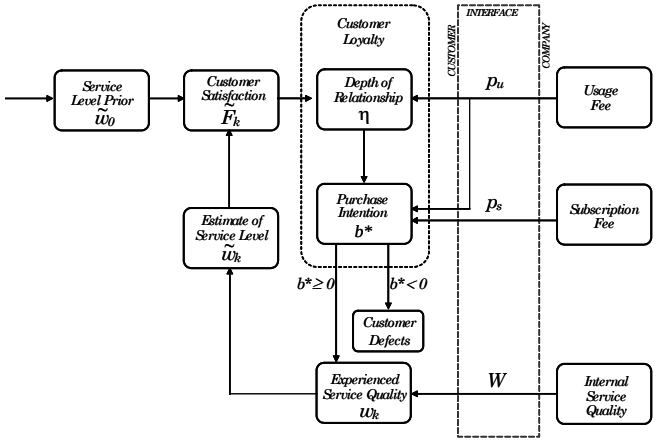

Figure 2: Dynamics of Customer-Company interactions

At the beginning of each period, active customers (i.e., those who have not defected) pay a subscription fee p_s to renew their subscription to the service. Periods are defined to be of length T (see Figure 1), which can represent days, weeks, or months depending on the specific application. Customers who choose to renew the subscription agree to enter a service contract whereby they will pay usage fee p_u every time they use the service. This is one of the most frequently used pricing structures in the telecommunications industry (cf.

Danaher 2002). Furthermore, this two-part pricing includes the case where customers pay a fixed subscription fee and have unlimited usage ($p_u = 0$) as well as the case where there is no subscription fee and customers only pay when they use the service ($p_s = 0$).

Customers use the service at the rate η, which (as depicted in Figure 2) is a function of their level of satisfaction and the usage fee (p_u) but not the subscription fee (p_s). In contrast to the models of Hall and Porteus (2000) and Gans (2002), there can be more than one interaction per period. This is how the present model captures the customer's choice of depth of relationship. Figure 1 depicts a situation where there were three interactions in the first period and two in the second period.

The utility customers derive from each service interaction has a fixed component as well as a random component. The fixed component is given by $(v(\eta) - \eta p_u)$, where $v(\eta)$ is the intrinsic benefit of receiving the service at rate η for one period. The random component, denoted by $c(w)$, is a function of the experienced service quality, w. Since capacity-constrained service-delivery systems are commonly modeled as queues, quality is assumed (without loss of generality) to be measured in terms of the waiting time, and therefore the customer utility function is decreasing in w. In this case, the function $c(w)$ can be interpreted as the cost of waiting.

3.1.2 Subscription and Defection

Customers will choose to subscribe (or renew their subscription) whenever the expected benefits outweigh the expected costs—more precisely, whenever

$$(v(\eta) - \eta p_u) - \eta(\bar{c}(\tilde{w})) - p_s > 0, \tag{2}$$

where

$$\bar{c}(\tilde{w}) = \int_0^\infty c(x) dF(x|\tilde{w}) \tag{3}$$

is the amount of dissatisfaction due to waiting that the customers expect to experience given their current estimate of the service level. Inequality (2) must also hold for new customers who must decide whether or not to subscribe. Each potential customer arrives with an expectation of service quality $\tilde{w}_0$. Expectations are updated after each service encounter, and customers defect immediately after a service encounter if their current estimate of $\tilde{w}$ does not satisfy inequality (2). In other words, a potential customer only becomes a new

customer if inequality (2) is satisfied when $\tilde{w} = \tilde{w}_0$.

Even though new arrivals into the system depend strictly on an exogenous arrival rate and are independent of p_u and p_s, they are implicitly endogenous in that the probability that they will become actual customers also depends on the fees, since it is possible for a customer to defect before paying the subscription fee and undergoing the first interaction (note in Figure 2 that it is possible for a new customer to go directly to the"customer defects" stage without actually experiencing the service).

There are very few service situations where the actual mean service level W is perfectly observable by customers. Typically, for a new customer, W would affect $\tilde{w}_0$ only indirectly, through word-of-mouth. However, the effect of word-of-mouth is often biased and can be overwhelmed by explicit service promises (advertising, personal selling, contracts, other communications), implicit service promises (tangibles, prices), and the customer's past experience with other firms in the same industry as well as firms in other industries (Zeithaml, Berry and Parasuraman 1993). With this in mind, we believe that an initial approach to the problem where $\tilde{w}_0$ is independent of W is legitimate, but that a real-world estimation of the arrival rate should take many more factors into account.

The analysis in the sections that follow will make use of the function $b\left(\eta; p_s, p_u, \tilde{F}(\cdot)\right)$, defined as

$$b\left(\eta; p_s, p_u, \tilde{F}(\cdot)\right) = (v(\eta) - \eta p_u) - \eta\left(\bar{c}(\tilde{w})\right) - p_s. \tag{4}$$

This function represents the customer's expected net utility per unit of time, where the unit is the period of length T. The way the two-part tariff is incorporated into the consumer's decision model is consistent with recent behavioral findings that consumers devise a summary statistic based on the cost per unit of a two-part tariff and compare it with their expected usage level (Redden and Hoch 2006). The customer will choose the usage rate η^* that maximizes $b\left(\eta; p_s, p_u, \tilde{F}(\cdot)\right)$ and will defect whenever

$$b\left(\eta^*; p_s, p_u, \tilde{F}(\cdot)\right) = b^* < b_{\min}. \tag{5}$$

The threshold $b_{\min}$ represents the value a customer expects to receive from the competition minus any applicable switching costs and can be set to 0 (as in (2)) without loss of generality. Indeed, (5) can be made equivalent to the commonly-used logit and probit model by making

particular assumptions about the functional form of $b(\cdot)$ and the distribution of the error terms of the probabilistic variables. This formulation of the defection decision is consistent with the conclusions Gupta and Zeithaml (2006) reached after examining the published literature connecting customer metrics to financial results: there is a strong correlation between customer satisfaction (captured here by $\tilde{F}(\cdot)$) and customer retention. The threshold, $b_{\min}$, which represents the expected utility of switching to the competition, it is assumed to be constant for a number of reasons. First, because customer expectations of a given firm's service quality are shaped by several different factors (described in our justification of why $\tilde{w}_0$ is independent of W) according to the results of Zeithaml, Berry and Parasuraman (1993). Second, because in most real-world applications our actual service level, W, is not perfectly observable by the competitors, so there is no a priori reason to assume that they would adapt their service levels optimally and rationally based on our decision of W. Finally, even if competitors choose to adapt their service levels, these decisions can often not be made instantaneously. The delay in the competitors' change of capacity would then be followed by a delay in the time until the customers' perceptions of quality changes, making the constant threshold a reasonable assumption given the time frame in which these decision are made by actual firms.

3.1.3 Depth of Relationship

The next propositions present monotonicity properties of the customer's optimal choice of depth of relationship and the corresponding level of utility the customer derives from interacting with the firm. Proposition 1 connects the customer's optimal depth of relationship with the firm's decision variables, and Proposition 2 connects the customer's expected net utility per unit time with the firm's decision variables.

Proposition 1 *(a) The set of usage rates that maximize the expected utility per period is nonincreasing in p_u.*

(b) The set of optimal usage rates is also nonincreasing in the expected waiting cost $\bar{c}$.

(c) If the cost function c is increasing, then the set of optimal usage rates is nonincreasing in $\tilde{w}$.

Proof. See Appendix A.1. ■

Proposition 1 establishes the monotonicity of the customer's usage of the service as a function of the usage fee, the expected interaction cost, and the service-level estimate. The

behavior of the expected net utility with respect to these parameters is characterized in Proposition 2.

Proposition 2 *(a) The expected net utility per period, $b^* = \max_{\eta \geq 0} b(\eta)$, is nonincreasing in the expected cost $\bar{c}$. If $\tilde{F}(x|\tilde{w})$ is decreasing in $\tilde{w}$ and $c(\cdot)$ is nondecreasing, then b^* is also nonincreasing in $\tilde{w}$.*

(b) The expected net utility per period is nonincreasing in both the expected value of the usage fee p_u and the periodic membership fee p_s.

Proof. See Appendix A.1. ■

This monotonicity property allows the company to use the customer's estimate of service level as a concrete and manageable measure of customer utility. The properties of Propositions 1 and 2 will hold for any specific form of the function $v(\eta)$. Different types of customer behavior can be modeled by varying the cost and the value functions. For example, a step-increasing value function will result in a step-decreasing depth of relationship. When $v(\eta)$ is concave, the maximization problem has a unique solution (η^*), and the result of Proposition 1 applies to the unique optimal usage rate η^*.

3.2 Dynamics of Aggregate Customer Behavior

The evolution of the relationship between customers and the firm is modeled as a Markov process, where each state is defined by the number of previous interactions and the customer's current level of satisfaction (operationalized through the estimation of the average waiting time). Customers remain with the company as long as $b^* > b_{\min}$ (as in (5)). The expected net utility is monotonic in the current level of satisfaction, as shown in Proposition 2. Thus, requiring the customers' utility to be above the threshold is equivalent to requiring the estimate $\tilde{w}$ to be below $\tilde{w}_{\max}$, where $\tilde{w}_{\max}$ is implicitly defined by $b^*(\tilde{w}_{\max}) = b_{\min}$.

In order to define a finite set of states, the interval $[0, \tilde{w}_{\max}]$ is partitioned into a set I of S disjoint subintervals, where $I = \{I_1, ..., I_S\} = \{[0, u_1), [l_2, u_2), ..., [l_S, u_S]\}$, and $u_i = l_{i+1},\ i = 1, ..., S-1$, and $u_S = \tilde{w}_{\max}$. Given the monotonicity properties of the net utility function, this partition is equivalent to partitioning the interval $[b^*(0), b_{\min}]$ into S subintervals. The states are defined so that a customer who has interacted with the firm k times and whose level of satisfaction $(\tilde{w}_k)$ falls in the interval I_i is in state (i, k). A customer in state (i, k) who accesses the service for the $(k+1)^{th}$ time experiences service quality w_{k+1}, updates the

level of satisfaction according to (1), and moves to state $(\tilde{i}, k+1)$, where $\tilde{i}$ is defined so that $\tilde{w}_{k+1} \in I_{\tilde{i}}$. Figure 3 provides a representation of this model, where customers move downward through the network while they are with the firm.

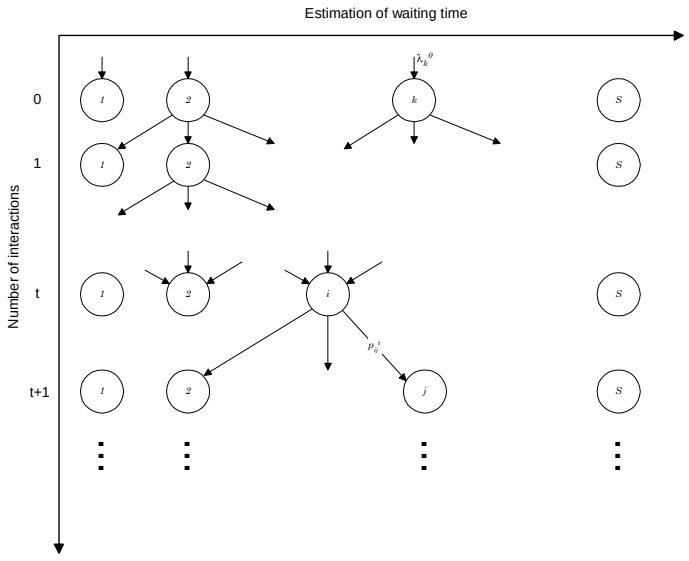

Figure 3: Evolution of customers

The transition probabilities in this network are a function of the true distribution of waiting time (which depends on W, the service-quality decision variable) and the customer's update mechanism. If the true distribution of waiting time is $F(x)$ and the customer's estimate of the mean of the service quality is updated according to Equation (1), then

$$\Pr\left(\tilde{w}_{k+1} \leq x | \tilde{w}_k = y\right) = \Pr\left(w_k \leq \frac{x - (1-\alpha_k)\, y}{\alpha_k}\right) = F\left(\frac{x - (1-\alpha_k)\, y}{\alpha_k}\right).$$

Given that a customer is in state $(i, \cdot)$, $\tilde{w}$ is assumed to be uniformly distributed in $[l_i, u_i)$, i.e., $f_{\tilde{w}_k}\left(y | \tilde{w}_k \in I_i\right) = \frac{1}{\Delta_i}$, where $\Delta_i = u_i - l_i$. More precisely,

$$\begin{aligned}\Pr\left(\tilde{w}_{k+1} \leq x | \tilde{w}_k \in I_i\right) &= \int_{l_i}^{u_i} \Pr\left(w_{k+1} \leq x | w_k = y\right) f_{\tilde{w}_k}\left(y | \tilde{w}_k \in I_i\right) dy \\ &= \tfrac{1}{\Delta_i} \int_{l_i}^{u_i} F\left(\tfrac{x-(1-\alpha_k)y}{\alpha_k}\right) dy.\end{aligned}$$

The transition probability $p_{ij}^k = \Pr\left\{\tilde{w}_{k+1} \in I_j | \tilde{w}_k \in I_i\right\}$ (moving from state (i,k) to $(j, k+$

1)) is then given by

$$p_{ij}^{k} = \Pr\left(l_j \leq \tilde{w}_{k+1} \leq u_j | \tilde{w}_k \in I_i\right) = \Pr\left(\tilde{w}_{k+1} \leq u_j | \tilde{w}_k \in I_i\right) - \Pr\left(\tilde{w}_{k+1} \leq l_j | \tilde{w}_k \in I_i\right)$$
$$= \frac{1}{\Delta_i}\left[\int_{l_i}^{u_i} F\left(\frac{u_j-(1-\alpha_k)y}{\alpha_k}\right) dy - \int_{l_i}^{u_i} F\left(\frac{l_j-(1-\alpha_k)y}{\alpha_k}\right) dy\right].$$

Note that the assumption that the estimates are uniformly distributed is asymptotically exact as $\Delta_i \to 0$. From the Lebesgue density theorem, it follows that

$$\lim_{\Delta_i \to 0} \frac{1}{\Delta_i}\left[\int_{l_i}^{l_i+\Delta_i} F\left(\frac{u_j-(1-\alpha_k)\,y}{\alpha_k}\right) dy - \int_{l_i}^{l_i+\Delta_i} F\left(\frac{l_j-(1-\alpha_k)\,y}{\alpha_k}\right) dy\right]$$
$$= F\left(\frac{u_j-(1-\alpha_k)\,l_i}{\alpha_k}\right) - F\left(\frac{l_j-(1-\alpha_k)\,l_i}{\alpha_k}\right).$$

A scaling factor p_D can be introduced into the model to account for customer defections that are caused by exogenous factors (e.g., moving, dying) that are assumed to remain constant over time (c.f. Schmittlein, Morrison and Colombo 1987). If this factor is introduced, all transition probabilities are multiplied by $(1-p_D)$. In what follows, it is assumed that $p_D > 0$ unless otherwise noted. Note, however, that the model will still converge if $p_D = 0$, as the necessary condition for convergence, $\Pr(\tilde{w} > w_{\max}) > 0$, is true under the assumption that customers will not tolerate an arbitrarily low service quality.

4 System Behavior in Steady State

This section will analyze the customer base as a migration process and explain the counterintuitive result that decreasing price or increasing service can result in a decrease in the number of customers in the system. This analysis will answer questions concerning the effect of the service level and pricing on the expected duration of customer relationships, the total number of customers, and the level of demand experienced by the service facility by modeling the customer base as an open migration process. The company will then be analyzed as a network of infinite server queues. The states correspond to stations in the network, and the time between interactions corresponds to service time.

The arrival process of new customers to state $(i,0)$ at the top of the network is assumed to be a homogeneous Poisson process with rate λ_i^0. For customers in states $(i,\cdot)$, the mean time between interactions is $\frac{1}{\eta_i}$, where $\eta_i = \frac{1}{\Delta_i}\int_{l_i}^{u_i} \eta^*(\tilde{w})\, d\tilde{w}$. Since $\eta^*(\tilde{w})$ is a monotonic

function of $\tilde{w}$, η_i will also be monotonic. If the time between interactions for customers in state (j,k) has a general distribution with mean $\frac{1}{\eta_j}$, then the result of Proposition 3 follows from queueing network theory.

Proposition 3 *(Arrival rate to each state) In equilibrium, the number of customers at each state (j,k) is independent and has a Poisson distribution with parameter $\frac{\lambda_j^k}{\eta_j}$, where λ_j^k is the arrival rate to state (j,k) and is the solution to the following system:*

$$\lambda_j^k = \sum_{i:I_i \in \hat{I}} p_{ij}^{k-1} \lambda_i^{k-1} \quad \textit{for all } j : I_j \in \hat{I} \textit{ and for } k = 1, 2, ... \tag{6}$$

where $\hat{I} = \{I_i \in I : [\tilde{w} \in I_i] \rightarrow [b^*(\tilde{w}) \geq b_{\min}]\}$.

Letting $\boldsymbol{\lambda}_k = \left(\lambda_i^k : s_i \in \bar{U}_{b_{\min}}\right)$ be the vector whose elements are the arrival rates to the states corresponding to customers of age k, and letting $\mathbf{P}_k = \left[p_{ij}^k\right]$ be the matrix of transition probabilities, (6) can be written as:

$$\begin{aligned} \boldsymbol{\lambda}_k &= \boldsymbol{\lambda}_{k-1}\mathbf{P}_{k-1} \\ &= \boldsymbol{\lambda}_0\mathbf{P}_0\mathbf{P}_1 \cdots \mathbf{P}_{k-1}.^1 \end{aligned} \tag{7}$$

The demand for service will be the sum of the demand of customers at every state, so the total arrival rate to the service facility is

$$\lambda = \sum_{k=0}^{\infty} \left(\boldsymbol{\lambda}_k \mathbf{e}\right), \tag{8}$$

where $\mathbf{e}$ is the unit column vector. N, the expected total number of customers in the system, is given by

$$N = \sum_{k=0}^{\infty} \sum_{s_j \in \bar{U}_{b_{\min}}} \frac{\lambda_j^k}{\eta_j}. \tag{9}$$

This system has an infinite number of states. However, the total arrival rate to the service facility and the expected total number of customers in the system is always finite, as established by the proposition below.

Proposition 4 *(Finiteness of the total arrival rate and the expected number of customers)*

(a) The total rate of arrival to the service facility is finite.

(b) The expected total number of customers in the system is finite.

Proof. See Appendix A.1. ■

After every interaction, some customers leave the system, and the rate of arrival to the next level of states decreases. If customers stay long enough in the system and become insensitive to actual realizations of service quality, the rate of arrival to the service facility and the expected number of customers in the system could become unbounded. This is of no concern, as it would only happen when $p_D = 0$ and $\Pr(\tilde{w} > w_{\max}) = 0$, which in turn requires $w_{\max}$ to be unreasonably high and a_k to be decreasing and tending to 0—a highly unlikely situation in a realistic scenario.

Another question to explore is how the service level affects the aggregate demand as well as the number of customers in the system. In order to answer this question, we must explore the behavior of the arrival rate of customers who request service at the service facility as a function of the level of service chosen by the firm. The following proposition assumes that $F(x|W)$ is nonincreasing in the average waiting time W in order to derive an important monotonicity property.

Proposition 5 *The total arrival rate to the service facility is nonincreasing in W.*

Proof. See Appendix A.1. ■

An alternative approach yielding the same results is to make the more general assumption that the firm selects a distribution of service levels from a family of distributions $\{F_i(x), i \geq 0\}$, where the index i can be either discrete or continuous. In this case, it is also necessary to assume that the distributions can be stochastically ordered so that either $F_i(x) \geq_{st} F_j(x)$ or $F_i(x) \leq_{st} F_j(x)$, i.e., the firm selects the service level from a stochastically ordered set.

Figure 4 presents results from a numerical example (see Appendix A.2 for details) which illustrates the effect of the service level on the the firm's customer base. Here one can observe the counterintuitive effect that as the quality of service decreases, the number of customers does not always decrease (it may in fact increase). This effect, which can be predicted by the preceding analysis, can be intuitively explained as follows. As the service quality decreases, the probability that a customer will decide to leave the company increases, and thus that customer's total number of interactions decreases. At the same time, the intervals between

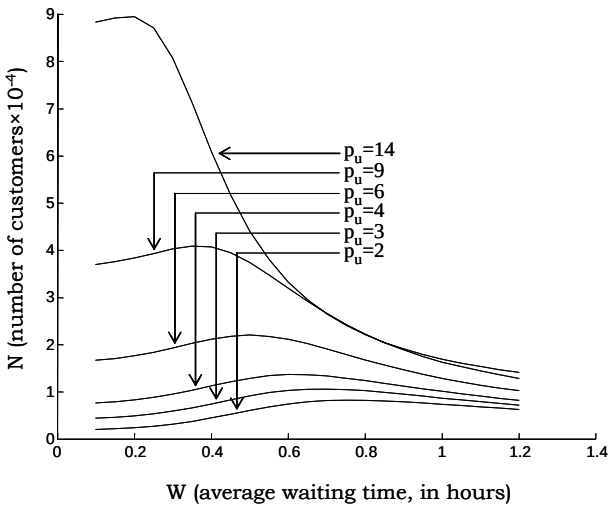

Figure 4: Effect of the service level on the total number of customers for different usage fees (p_u).

that customer's consecutive interactions increase. The expected length of stay is defined by the relationship

E(Length of Stay)=E(Total Number of Interactions)×(Average Interval Between Interactions),

where E() denotes the mathematical expectation. The net effect on the average length of stay will depend on the relative magnitude of these two terms. According to Little's Law, the expected number of customers in the system will be equal to the product of the arrival rate of potential customers and the average length of stay of each customer. Therefore, the total expected number of customers will ultimately depend on the relative sizes of those same opposing elements: the expected number of interactions (which goes down when quality goes down) and the average length of the intervals between them (which goes up when quality goes down). As service quality decreases, the increased interval between interactions can be high enough to offset the decrease in the total number of interactions and produce a greater total "Length of Stay," which leads to a greater number of customers in steady state; hence, we obtain the counterintuitive result that the size of the customer base can actually increase when service quality goes down.

Alternatively, consider the partial derivative of N with respect to W, $\frac{\partial N}{\partial W} = \sum_k \sum_j \frac{\partial\left(\frac{\lambda_j^k}{\eta_j}\right)}{\partial W}$.

Note that η_j does not depend on W. Since η_j is defined as the average usage rate for customers in state (j,k) (i.e., for customers with a given estimate of service level), η_j depends only on p_u and p_s. It follows that

$$\frac{\partial N}{\partial W} = \sum_k \sum_j \frac{1}{\eta_j} \frac{\partial \lambda_j^k}{\partial W}.$$

Letting $J^+ \doteq \left\{(j,k) : \frac{\partial \lambda_j^k}{\partial W} \geq 0\right\}$ and $J^- \doteq \left\{(j,k) : \frac{\partial \lambda_j^k}{\partial W} < 0\right\}$, it can be affirmed that N is increasing in W whenever

$$\sum_{(j,k)\in J^+} \frac{1}{\eta_j} \frac{\partial \lambda_j^k}{\partial W} > \sum_{(j,k)\in J^-} \sum_k \frac{1}{\eta_j} \frac{\partial \lambda_j^k}{\partial W}.$$

When the usage rate is constant, the total number of customers in the system is also nonincreasing in W. This property can be verified by observing that the total number of customers, N, is bounded from above by

$$\bar{N} \doteq \frac{1}{\min\left[\eta_j\right]} \sum_k \sum_j \lambda_j^k \tag{10}$$

and from below by

$$\underline{N} \doteq \frac{1}{\max\left[\eta_j\right]} \sum_k \sum_j \lambda_j^k, \tag{11}$$

with both bounds being nonincreasing in W.

The remainder of this section will examine the effects of the firm's pricing policies on the arrivals to the service facility (service requests) and on the number of customers in the system. It is important to note that the transition rates between states corresponding to customers who stay in the system are functions of the service level provided by the firm and the customers' estimation procedure. These rates are affected by neither the usage fee p_u nor the subscription fee p_s. However, pricing affects each customer's decision of whether or not to stay with the company. Intuitively, one might expect that the higher the price, the higher the customer's demand for service quality. In the present model this translates into a contraction of the subspace of service-quality estimates that would be sufficiently high for the customer to stay with the firm.

Recall that customers stay with the company as long as their expected net utility per

period is above a threshold $b_{\min}$. For fixed values of p_u and p_s, this is equivalent to requiring the customer's estimate to be below a critical value $w_{\max}$ (this result follows immediately from the monotonicity property in part (a) of Proposition 2). Since the net utility per period is also monotonic in p_u and p_s (part (b) of Proposition 2), it follows that $w_{\max} \doteq w : b^*(w, p_u, p_s) = b_{\min}$ is decreasing in p_u and in p_s. That is, in response to higher prices, customers will demand higher levels of service. This observation leads to the next result, which is analogous to Proposition 5, with the difference that the decision variables are now p_u and p_s rather than W.

Proposition 6 *The total arrival rate to the service facility is nonincreasing in both the average usage fee p_u and the subscription fee p_s.*

Proof. See Appendix A.1. ■

In the case of expected number of customers in the system as a function of the usage fee p_u, there are, as was the case with service quality, two opposing effects—recall that the number of customers depends on average length of stay, which is the product of the number of interactions and the average interval between them. Based on the monotonicity property of the usage rate (Proposition 1), an increase in the usage fee will produce an increase in the time between customer interactions. On the other hand, an increase in price will make the customer's criteria for staying in the company more stringent, which will tend to reduce the number of interactions. Thus, the net result of an increase in price on the size of the customer base is not necessarily monotonic (as is illustrated by the numerical example in Figure 5) and will depend on the relative importance of those two effects (number of interactions and average interval between them). This dynamic is analogous to the one in Figure 4, and the bounds for N obtained in (10) and (11) are also nonincreasing in p_u. Note that in both Figures 4 and 5, the number of customers increases for sufficiently small values of the control variable (in response to what customers might objectively consider undesirable changes in price and service quality) before it decreases as managers might expect, so careful managerial selection of the pricing and service-quality levels is particularly important to yield the desired results.

The effect of the subscription fee on the rate of usage is simpler to analyze. An increase in p_s will not influence the rate of usage. This result is formalized in the proposition below.

Proposition 7 *The expected total number of customers in the system is nonincreasing in the periodic membership fee p_s.*

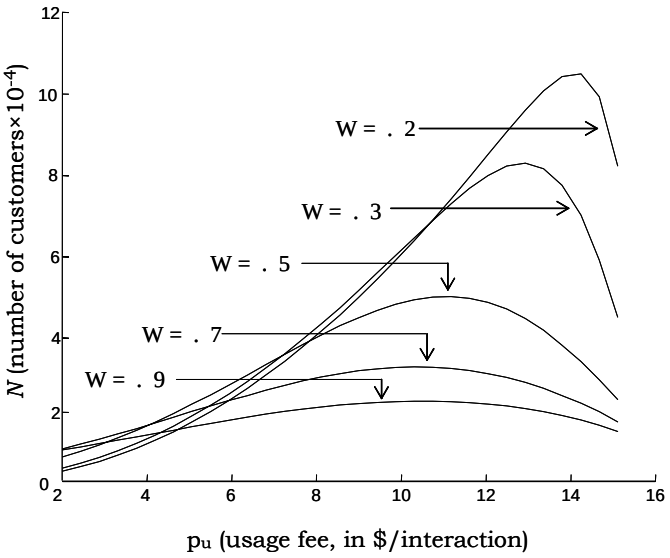

Figure 5: Effect of the usage fee on the total number of customers for different levels of service (W, average waiting time, in hours).

Proof. See Appendix A.1. ■

5 Profit Optimization

This section examines how price and service quality affect profitability. The previous section has shown that the size of the customer base in steady state is non-monotonic in the service quality and usage fee. As mentioned earlier, this result raises questions concerning the adequacy of maximizing the size of the customer base or using the size of the customer base as a proxy for profitability in the long run, since fewer customers sometimes yield higher profits. This section provides a rigorous approach to the issue of profit maximization.

The function to be maximized is the rate at which profit is generated in steady state. The revenue per period, denoted R, is given by

$$R = \lambda p_u + N p_s.$$

Let $C(\lambda, W)$ be the cost per period of providing service level W to a set of customers arriving at rate λ. $C(\lambda, W)$ is assumed to be increasing in λ and decreasing and convex in W. The

profit optimization problem can then be formulated as the following nonlinear program:

$$\max \Pi = \lambda p_u + N p_s - C(\lambda, W)$$

$$s.t.$$

$$\lambda = g(p_u, p_s, W), \qquad N = h(p_u, p_s, W), \qquad W \geq 0.$$

The restrictions on λ and N correspond to the implicit functions for the arrival rate and the size of the customer base, derived in (8) and (9) respectively.

Let

$$\mathcal{L} = \Pi - \gamma_1 [g(p_u, p_s, W) - \lambda] - \gamma_2 [h(p_u, p_s, W) - N],$$

where γ_1 and γ_2 are Lagrange multipliers. Using the notation where y_x is the partial derivative of y with respect to x, the necessary optimality conditions are given by the following set of equations:

$$\begin{array}{lll} \lambda - \gamma_1 g_{p_u} - \gamma_2 h_{p_u} = 0, & N - \gamma_1 g_{p_s} - \gamma_2 h_{p_s} = 0, & -C_W - \gamma_1 g_W - \gamma_2 h_W = 0, \\ p_u - C_\lambda + \gamma_1 = 0, & p_s + \gamma_2 = 0, & \\ g(p_u, p_s, W) - \lambda = 0, & h(p_u, p_s, W) - N = 0, & \text{and } W \geq 0. \end{array}$$

Straightforward algebraic manipulation yields

$$p_u \frac{\partial \lambda}{\partial W} + p_s \frac{\partial N}{\partial W} = \frac{\partial C}{\partial W} + \frac{\partial C}{\partial \lambda} \frac{\partial \lambda}{\partial W}, \tag{12}$$

$$\lambda + p_u \frac{\partial \lambda}{\partial p_u} + p_s \frac{\partial N}{\partial p_u} = \frac{\partial C}{\partial \lambda} \frac{\partial \lambda}{\partial p_u}, \tag{13}$$

$$\text{and } N + p_u \frac{\partial \lambda}{\partial p_s} + p_s \frac{\partial N}{\partial p_s} = \frac{\partial C}{\partial \lambda} \frac{\partial \lambda}{\partial p_s}. \tag{14}$$

These equations are analogous to equating marginal revenues with marginal costs for each of the three decision variables: service quality W in (12), usage fee p_u in (13), and subscription fee p_s in (14).

The LHS of (12) reveals that the marginal revenues due to improving service quality can be decomposed into two parts. The first part corresponds to the effect of service quality on the arrival rate into the service facility (service requests). The second part corresponds to

the effect of service quality on the number of customers in the system. Proposition 5 asserts that the arrival rate is nonincreasing in W, but the size of the customer base can be either increasing or decreasing, as illustrated in Figure 4. The impact of an increase in service quality on profitability will depend on the relative size of these effects (changes in the arrival rate and the number of customers) in light of the prices. Note that the cost side has two terms. This is because increasing the service quality has a direct effect on costs (providing good service is assumed to be more costly than providing bad service) as well as an indirect effect (service quality has an effect on how many customers will demand the service).

In Equation (13) the marginal revenue gained from increasing the usage fee is decomposed into three parts. The first is the direct effect of a price increase on the revenue. This is the additional revenue gained assuming that the number of customers demanding service will remain constant in spite of the price difference. The second and third parts correspond to the indirect effects, which are analogous to those explained in the previous paragraph. The first indirect effect $\left(\frac{\partial \lambda}{\partial p_u}\right)$ is the simplest to understand, since the result of Proposition 6 guarantees that the arrival rate is nonincreasing in p_u. The last term, on the other hand, cannot be understood as intuitively. As depicted in Figure 5, increasing the usage fee can have a positive or a negative impact on the size of the customer base. As in the case of the waiting time, an increase in p_u may also result in an increase or a decrease in profitability.

Finally, Equation (14) shows that the marginal revenue gained from increasing the subscription fee is decomposed into three parts. The first is the direct effect of a price increase on the revenue due to the number of customers that will be paying the subscription fee in steady state (i.e., the size of the customer base). The second and third terms on the LHS correspond to the indirect effects. Propositions 6 and 7 ensure that both $\frac{\partial \lambda}{\partial p_s}$ and $\frac{\partial N}{\partial p_s}$ are negative, making this the most intuitive of the optimality conditions, as the dynamics involved are the same ones found in traditional pricing problems.

6 Discussion

This paper provides an important building block for understanding the underlying structure of the dynamics governing the impact of price and service quality on customer satisfaction and profitability. Changes in price and service quality that provide more value to customers sometimes result in fewer rather than more customers in the long run. Another puzzling phenomenon has been that in some cases where the size of the customer base actually has

increased as expected, managers have been observing profits go down. This paper provides an explanation for these phenomena and recommends actions that will enable managers to act optimally under these circumstances. The key factor driving the results is the incorporation of a variable that captures each customer's depth of relationship, chosen according to each individual's perception of service quality.

In a wider context, the analysis addresses the issue of customer interface design from the perspective of a long-term relationship between a company and its clients. Service quality plays a key role in long-term relationships: it acts as one of the main drivers of customer satisfaction, which in turn determines loyalty and hence the length of the relationship itself. Companies must carefully manage their service-quality levels in order to differentiate their service offerings. Fee structures also play a determining role in long-term relationships, as pricing exerts a large influence on the customers' frequency of interaction and propensity to defect.

The focus of this paper is on the impact of service quality and fee structures on customer behavior and the resulting effect on the long-term value of customer relationships to the firm, in order to provide insights into the design of customer interfaces (choosing fees and service quality) to optimize profitability. In particular, the results explain the interdependence of the pricing structure, service level, and long-term behavior of customers and quantify the effect of these collective factors on the expected utilization of the service facility, the size of the customer base, and the revenues.

The most intriguing result of this paper relates to the fact that the number of clients that a company will have does not necessarily increase as service quality increases. In order to explain this phenomenon, the analysis in §4 reveals that the number of clients at any given point in time depends on the rate at which new customers are acquired and the average length of their relationship with the company. The length of the relationship, in turn, is determined by the total number of interactions multiplied by the average time between interactions. If the service level falls below a certain threshold, customers leave. As the level of service decreases, the number of interactions decreases because the probability that the customer will cross the threshold will increase. However, the length of time between interactions will also increase because customers will not use the service as often as they would if the service level were higher. The increase in the average length of time between interactions can offset and even surpass the decrease in the total number of interactions. Companies that earn revenues each time the customer uses their service are worse off in this

situation (they would profit better from making decisions which maximize the number of interactions). Companies that charge a subscription fee, on the other hand, can be better off: they decrease their operational costs due to the reduced number of interactions while simultaneously increasing their customer base. This result is rather unintuitive and serves to show the value of mathematical analysis in devising and implementing strategic policies.

The models developed in this paper can help managers design the way in which the firm will interact with its customers. These models can be used to evaluate the consequences of decisions—such as changes in service quality, fee structures, and product quality—in terms of customer retention, revenues, and costs. These are high-level decisions that have important implications regarding the type of customers that companies attract and retain, which in turn determines the company's sources of revenue. The result that lower levels of quality may lead to a larger customer base and higher profitability is of great interest to managers deciding how to position their services in light of their competitors' offers and the expectations of their target market.

The findings of this paper open the way for a number of new research opportunities. One possible extension is to incorporate the impact of marketing expenditure on the arrival rate of new customers. It would be interesting to investigate the case where the impact is made constantly over time as well as the effect of a single campaign on the steady state of the system. This paper can also be extended to the case where the firm interacts with a heterogeneous customer base. If the customer base can be divided into discrete segments that can be targeted separately, the firm's control problem reduces to simple replications of the one solved in this paper. However, the problem becomes more complex as the representation of heterogeneity and the addressability of individual customers become more difficult. Developing this paper's model into a decision-support tool for controlling the customer interface, compatible with various models of customer heterogeneity, is a rich research topic worth pursuing. Finally, generalizing the results of this paper to the case where customers interact with the firm through multiple channels is the subject of current investigation by the authors.

A Appendices

A.1 Proofs

Propositions 1

Proof. The proofs are trivial under the assumptions that $v(\eta)$ is concave and $\overline{c}\,(\tilde{w})$, the expectation of the function c (given by (3)), is increasing in $\tilde{w}$. These assumptions are quite robust for most applications. For the general case, assuming b is smooth we have $\frac{\partial^2 b}{\partial \eta \partial p_u} \leq 0$. It then follows from Topkis (1978) that b has decreasing differences in (η, p_u) and therefore $b^*(\eta)$ is nonincreasing in p_u. This proves part (a). For part (b), we note that $\frac{\partial^2 b}{\partial \eta \partial c} \leq 0$ and the result follows from the same argument used above. Finally, for part (c), note that if $\frac{\partial \bar{c}}{\partial \tilde{w}} > 0$ then b has decreasing differences in $(\eta, \tilde{w})$, since $\frac{\partial^2 b}{\partial \eta \partial \tilde{w}} = \frac{\partial^2 b}{\partial \eta \partial \bar{c}} \frac{\partial \bar{c}}{\partial \tilde{w}} \leq 0$. ■

Proposition 2

Proof. The proof for this proposition is based on the envelope theorem. For part (a), first note that $\frac{\partial b^*}{\partial \bar{c}} = \frac{\partial b(\bar{c},\eta)}{\partial \bar{c}} = -\eta$,where $\eta \in \boldsymbol{\eta}^*(\bar{c})$. By definition, $\eta \geq 0$, and therefore, $\frac{\partial b^*}{\partial \bar{c}} \leq 0$. If c is increasing, $\frac{\partial \bar{c}}{\partial \tilde{w}} \geq 0$, implying that $\frac{\partial b(\tilde{w},\eta)}{\partial \tilde{w}} = \frac{\partial b(\bar{c},\eta)}{\partial \bar{c}} \frac{\partial \bar{c}}{\partial \tilde{w}} \leq 0$. The argument when c is decreasing is symmetric, concluding the proof of part (a). For part (b), note that $\frac{\partial b^*}{\partial p_u} = \frac{\partial b}{\partial p_u} = -\eta$. Since $\eta \geq 0$, it follows that $\frac{\partial b^*}{\partial p_u} \leq 0$. For the membership fee, p_s, we simply note that $\frac{\partial b^*}{\partial p_s} = \frac{\partial b}{\partial p_s} = -1 < 0$. ■

Proposition 3

Proof. Follows immediately from Kelly (1979). ■

Proposition 4

Proof. For part (a), to see that $\boldsymbol{\lambda}$ is always finite we can check that $\lim_{t\to\infty} \frac{\boldsymbol{\lambda}_{k+1}\mathbf{e}}{\boldsymbol{\lambda}_k \mathbf{e}} < 1$ (D'Alembert's ratio test). First, note that

$$\frac{\boldsymbol{\lambda}_{k+1}\mathbf{e}}{\boldsymbol{\lambda}_k \mathbf{e}} = \frac{\boldsymbol{\lambda}_0 \mathbf{P}_0 \mathbf{P}_1 \cdots \mathbf{P}_{k-1} \mathbf{P}_k \mathbf{e}}{\boldsymbol{\lambda}_0 \mathbf{P}_0 \mathbf{P}_1 \cdots \mathbf{P}_{k-1} \mathbf{e}} = \frac{\boldsymbol{\lambda}_k \mathbf{P}_k \mathbf{e}}{\boldsymbol{\lambda}_k \mathbf{e}}.$$

Next, note that the sum of every row of $\mathbf{P}_k$ is less than 1 as long as there exists a level of service that will cause at least one customer to defect (an assumption which is trivially satisfied in practice if customers will not tolerate an arbitrarily low level of service). Therefore, every element of $\mathbf{P}_k \mathbf{e}$ is less than 1 and $\frac{\boldsymbol{\lambda}_k \mathbf{P}_k \mathbf{e}}{\boldsymbol{\lambda}_k \mathbf{e}} = \frac{\boldsymbol{\lambda}_{k+1}\mathbf{e}}{\boldsymbol{\lambda}_k \mathbf{e}} < 1$. For part (b), assuming that customers require a positive expected utility and that $v(0) = 0$, $f \geq 0$ implies that at every state the access rate η_j is positive ($\eta_j = 0$ implies a nonpositive utility). Then, the expected number

of customers in the system, $N = \sum_k \sum_j \frac{\lambda_j^k}{\eta_j}$, is bounded from above by $\frac{1}{\min_j\{\eta_j\}} \sum_k \sum_j \lambda_j^k$. Given that $\frac{1}{\min_j\{\eta_j\}} > 0$ and that $\sum_k \sum_j \lambda_j^k$ is finite (part (a)), we conclude that N is also finite. ■

Proposition 5

Proof. If customers base their decisions on the first moment of the distribution of waiting time, then the probability that a customer in state (i,k) will leave the system is:

$$\begin{aligned}\left(1 - \sum_j p_{ij}^k\right) &= \Pr\left(w_{k+1} \geq w_{\max} | w_k \in I_i\right) \\ &= \left(1 - \frac{1}{\Delta_i} \int_{l_i}^{u_i} F\left(\frac{w_{\max} - (1-\alpha)\, y}{\alpha}\right) dy\right).\end{aligned}$$

For any $W_1 < W_2$,

$$F\left(\frac{w_{\max} - (1-\alpha)\, y}{\alpha} | W_1\right) \geq F\left(\frac{w_{\max} - (1-\alpha)\, y}{\alpha} | W_2\right).$$

Thus,

$$\int_{l_i}^{u_i} F\left(\frac{w_{\max} - (1-\alpha)\, y}{\alpha} | W_1\right) dy \geq \int_{l_i}^{u_i} F\left(\frac{w_{\max} - (1-\alpha)\, y}{\alpha} | W_2\right) dy$$

for $i = 1, ..., S$, and for every $k \geq 0$, the sum of every row of $\mathbf{P}_k(W_1)$ is greater or equal to the sum of every row of $\mathbf{P}_k(W_2)$. It then follows from (7) and (8) that $\lambda(W_1) \geq \lambda(W_2)$. ■

Proposition 6

Proof. The probability that a customer in state s_i^t will leave the system is given by:

$$\begin{aligned}\left(1 - \sum_{j=1}^{S} p_{ij}^t\right) &= \Pr\left(w_{t+1} \geq w_{\max} | w_t \in I_i\right) \\ &= \left(1 - \frac{1}{\Delta_i} \int_{l_i}^{u_i} F\left(\frac{w_{\max} - (1-\alpha)\, y}{\alpha}\right) dy\right).\end{aligned}$$

Let $w_{\max}(p_u, p_s) \doteq w : b^*(w, p_u, p_s) = b_{\min}$. For any $p_u \leq p_{u2}$ and any p_s, $w_{\max}(p_{u1}, p_s) \geq w_{\max}(p_{u2}, p_s)$, and thus $F\left(\frac{w_{\max}(p_{u1},p_s)-(1-\alpha)y}{\alpha}\right) \geq F\left(\frac{w_{\max}(p_{u2},p_s)-(1-\alpha)y}{\alpha}\right)$, and for $i = 1, ..., S$,

$$\int_{l_i}^{u_i} F\left(\frac{w_{\max}(p_{u1}, p_s) - (1-\alpha)\, y}{\alpha}\right) dy \geq \int_{l_i}^{u_i} F\left(\frac{w_{\max}(p_{u1}, p_s) - (1-\alpha)\, y}{\alpha}\right) dy.$$

Then, for every $t \geq 0$, the sum of every row of $\mathbf{P}_t(p_{u1})$ is greater or equal to the sum of every row of $\mathbf{P}_t(p_{u2})$, and from (7) and (8) it follows that $\lambda(p_{u1}) \geq \lambda(p_{u2})$. Analogously, for any $p_{s1} \leq p_{s2}$ and any p_u, $w_{\max}(p_u, p_{s1}) \geq w_{\max}(p_u, p_{s2})$ and the result follows. ■

Proposition 7

Proof. $\frac{\partial N}{\partial p_s} = \sum_{k=0}^{\infty} \sum_j \frac{\partial\left(\frac{\lambda_j^k}{\eta_j}\right)}{\partial p_s}$, where $\frac{\partial\left(\frac{\lambda_j^k}{\eta_j}\right)}{\partial p_s} = \frac{\frac{\partial \lambda_j^k}{\partial p_s}\eta_j - \frac{\partial \eta_j}{\partial p_s}\lambda_j^k}{\left(\eta_j\right)^2}$. Since η_j does not depend on p_s, $\frac{\partial N}{\partial p_s} = \sum_{k=0}^{\infty} \sum_j^S \frac{\frac{\partial \lambda_j^k}{\partial p_s}}{\eta_j}$. Given that $\lambda_j^k = \sum_{s_i \in \bar{U}_{b_{\min}}(p_s)} p_{ij}^{k-1} \lambda_i^{k-1}$, p_{ij} does not depend on p_s, and $\bar{U}_{b_{\min}}$ is decreasing in p_s, it follows that for every (j,k), λ_j^k is also decreasing in p_s. Since for every j, $\eta_j \geq 0$, the result follows. ■

A.2 Parameters for simulations

In Figure 4 the customers' utility functions are given by $v(\eta) = k_1\sqrt{\eta}$, the waiting cost is given by $c(x) = k_2 x^2$, and the customers' estimates of waiting time are exponentially distributed with mean $\tilde{w}$. Thus, $\bar{c}(\tilde{w}) = \int_0^{\infty} k_2 x^2 \frac{1}{\bar{w}} e^{-\frac{1}{\bar{w}}x} dx = 2k_2\bar{w}^2$ and $\eta^* = \frac{k_1}{2(p+2k_2\bar{w}^2)^2}$. The value of the parameters is given by $W = 0.5$, $\alpha = 0.6$, $\beta = 0$, $p_D = 0.5 \times 10^{-3}$, $k_1 = 12$, $k_2 = 2$, $b_{\min} = 0$, and $p_s = 2$.

References

Bitran, Gabriel, J. C. Ferrer and P. R. Oliveira. 2008. Managing Customer Experiences: Perspectives on the Temporal Aspects of Service Encounters. *Manufacturing & Service Operations Management,* **10** (1) 61-83

Bittencourt, L. G., J. Sellmeister Bueno. 2003. The challenges of implementing CRM in the financial services industry. S.M. thesis, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA.

Bolton, R. N. 1998. A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing Science* **17**(1) 45-65.

Bolton, R.N. and Lemon, K. N. 1999. A Dynamic Model of Customers' Usage of Services: Usage as an Antecedent and Consequence of Satisfaction. *Journal of Marketing Research*, **36** (2) 171-186.

Boulding, W., A. Kalra, R. Staelin, V. A. Zeithaml. 1993. A dynamic process of service quality. *Journal of Marketing Research* **30**(February) 7-27.

Chen, H., M. Z. Frank. 2001. State dependent pricing with a queue. *IIE Trans.* **33** 847-860.

Danaher, P. J. 2002. Optimal pricing of new subscription services: Analysis of a market experiment. *Marketing Science* **21**(2) 119-138.

DellaVigna, S. and U. Malmendier. 2001. Self-Control in the Market: Evidence from the Health Club Industry. Working paper (Harvard University).

Dewan, S., H. Mendelson. 1990. User delay cost and internal pricing for a service facility. *Management Sci.* **36**(12) 1502-1517.

Edelson, N. M., D. K. Hildebrand. 1975. Congestion tolls for Poisson queueing processes. *Econometrica* **43**(1) 81-92.

Essegaier, S., S. Gupta, Z. J. Zhang. 2002. Pricing access services. *Marketing Science* **21**(2) 139-159.

Friedman, E., A. S. Landsberg. 1993. Short-run dynamics of multi-class queues. *Oper.Res. Lett.* **14** 221-229.

Friedman, E., A. S. Landsberg. 1996. Long-run dynamics of queues: Stability and chaos. *Oper. Res. Lett.* **18** 185-191.

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

Gourville, J. and D. Soman. 2002. Pricing and the Psychology of Consumption. *Harvard Business Review* **80**(9) 90-96.

Gupta, S. Lehman, D. and Stuart, J. 2004. Valuing costumers. *Journal of Marketing Research* **41**(February) 7–18.

Hall, J., E. Porteus. 2000. Customer service competition in capacitated systems. *Manufacturing and Services Operations Management* **2**(2) 144-165.

Gupta, S. and V. Zeithaml. 2006. Customer Metrics and Their Impact on Financial Performance. Working paper, Columbia

University, New York.

Heskett, J. L., T. O. Jones, G. W. Loveman, W. E. Sasser, Jr., L. Schlesinger. 1994. Putting the service-profit chain to work. *Harvard Business Review* **72**(2) 164-174.

Hill, R. C. 1993. When the going gets rough: A Baldridge award winner on the line. *Academy of Management Executives* **7**(August) 75-79.

Hughes, Arthur. 2005. Strategic Database Marketing, 3rd ed. New York: McGraw-Hill.

Jones, T.O. and Sasser, W.E. Jr. 1995. Why satisfied customers defect. *Harvard Business Review* **73**(6) 88-99.

Kamakura, W. A., V. Mittal, F. Rosa, J. A. Mazzon. 2002. Assessing the service-profit chain. *Marketing Science* **21**(3) 294-317.

Kelly, F. P. 1979. *Reversibility and Stochastic Networks.* Wiley, New York.

Knudsen, N. C. 1972. Individual and social optimization in a multiserver queue with a general cost-benefit structure. *Econometrica* **40**(3) 515-528.

Lemon, K. N., White, T.B. and R. S. Winer. 2002. Dynamic Customer Relationship Management: Incorporating Future Considerations into the Service Retention Decision. *Journal of Marketing* **66**(January) 1-14.

Lipmann, S. A., S. Stidham. 1977. Individual versus social optimization in exponential congestion systems. *Oper. Res.* **25**(2) 233-247.

Mahajan, V., P. E. Green, S. M. Goldberg. 1982. A conjoint model for measuring self- and cross-price/demand relationships. *Journal of Marketing Research* **29** 334-342.

Mendelson, H. 1985. Pricing computer services: Queueing effects. *Communications of the ACM* **28**(3) 312-321.

Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Oper. Res.* **38** 870-883.

Mendelson, H., U. Yechiali. 1981. Controlling the M/G/1 queue by conditional acceptance of customers. *Eur. J. Oper. Res.* **7** 77-85.

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

Oi, W. Y. 1971. A Disneyland dilemma: Two-part tariffs for a Mickey Mouse monopoly. *Quart. J. Econom.* **85** 77-96.

Reichheld, F. F., W. E. Sasser, Jr. 1990. Zero defections: Quality comes to service. *Harvard Business Review* (Sep-Oct) 105-111.

Reichheld, F. F., T. Teal. 1996. *The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Value.* Harvard Business School Press, Boston, MA.

Redden, J.P. and S. J. Hoch. 2006. The psychology of two-part tariffs. *Wharton School Working Paper*, April.

Rump, C., S. Stidham. 1998. Stability and chaos in input pricing for a service facility with adaptive customer response to congestion. *Management Sci.* **44**(2) 246-261.

Rust, R. T., A. J. Zahorik. 1993. Customer satisfaction, customer retention, and market share. *Journal of Retailing* **69**(2) 193-215.

Rust, R. T., A. J. Zahorik, T. L. Keiningham. 1995. Return on Quality (ROQ): Making service quality financially accountable. *Journal of Marketing* **59**(April) 58-70.

Rust, R. T., J.I. Inman, J. Jia, J. and A. J. Zahorik. 1999. What You Don't Know About Customer-Perceived Quality: The Role of Customer Expectation Distributions. *Marketing Science* **18**(1) 77-92.

Schmittlein, David C., Donald G. Morrison, and Richard Colombo. 1987. Counting Your Customers: Who They Are and What Will They Do Next? *Management Science*, **33** (January), 1-24.

Stidham, S. 1992. Pricing and capacity decisions for a service facility: Stability and multiple local optima. *Management Sci.* **38**(8) 1121-1139.

Topkis, D. M. 1978. Minimizing a submodular function on a lattice. *Oper. Res.* **26**(2) 305-331.

Van Mieghem, J. A. 2000. Price and service discrimination in queueing systems: Incentive compatibility of Gcμ scheduling. *Management Sci.* **46**(9) 1249-1267.

Wiesendanger, B. 1993. Deming's luster dims at Florida Power and Light. *Journal of Business Strategy* **14**(Sep-Oct), 60-61.

Yechialy, U. 1971. On optimal balking rules and toll charges in the G/M/1 queue process. *Oper. Res.* **19** 348-370.

Zeithaml, V., L. Berry, and A. Parasuraman. 1993. The Nature and Determinants of Customer Expectations of Service. *Journal of the Academy of Marketing Science* **21**(1) 1-12.