Information Systems Research
Vol. 22, No. 4, December 2011, pp. 721–738
ISSN 1047-7047 | EISSN 1526-5536 | 11 | 2204 | 0721

http://dx.doi.org/10.1287/isre.1100.0283

Pricing Digital Goods: Discontinuous Costs and Shared Infrastructure

Ke-Wei Huang
Department of Information Systems, National University of Singapore, Singapore 117543,
huangkw@comp.nus.edu.sg

Arun Sundararajan
Leonard N. Stern School of Business, New York University, New York, New York 10012,
asundara@stern.nyu.edu

In this paper, we analyze a model of usage pricing for digital products with discontinuous supply functions. This model characterizes a number of information technology-based products and services for which variable increases in demand are fulfilled by the addition of *blocks* of computing or network infrastructure. Such goods are often modeled as information goods with zero variable costs; in fact, the actual cost structure resembles a mixture of zero marginal costs and positive periodic fixed costs. This paper discusses the properties of a general solution for the optimal nonlinear pricing of such digital goods. We show that the discontinuous cost structure can be accrued as a virtual constant variable cost. This paper applies the general solution to solve two related extensions by first investigating the optimal technology capacity planning when the cost function is both discontinuous and declining over time, and then characterizing the optimal costing for the discontinuous supply when it is shared by several business profit centers. Our findings suggest that the widely adopted full cost recovery policies are typically suboptimal.

Key words: pricing digital goods; nonlinear pricing; infrastructure cost; IT chargeback
History: Paulo Goes, Senior Editor; Tunay Tunca, Associate Editor. This paper was received on April 9, 2007, and was with the authors $16\frac{1}{2}$ months for 3 revisions. Published online in *Articles in Advance* March 17, 2010.

1. Introduction

This paper presents a study of usage pricing for a digital product whose supply function is discontinuous. This unique cost structure is the main focus of the paper and is a characteristic of a number of digital products for which variable increases in demand are fulfilled by the addition of *blocks* of computing or network infrastructure. Each additional block has a fixed cost, enabling a seller to fulfill a fixed, large level of demand at zero marginal cost, often leading to these products being viewed as *information goods*. At the same time, this discontinuous infrastructure cost is declining over time and is often shared among several information technology (IT) services. Contrary to conventional wisdom, indicating that infrastructure costs may not affect pricing, our paper shows that ignoring the discontinuous cost structure is detrimental to the monopoly profit.

Our study is most easily motivated by some examples. Given a fixed amount of infrastructure, the marginal cost of providing an additional unit of Internet service is typically zero. However, as the total number of users served by the Internet service provider (ISP) increases, the ISP needs to add blocks of infrastructure, including modems to their modem pool, and equipment and bandwidth to the networks that connect their users to the Internet backbone. Each additional block of infrastructure provides the ISP with the ability to fulfill an additional (fixed) amount of demand for Internet service with negligible marginal cost. The cost structures for providers of streaming media and online trading are similar: costs of adding infrastructure are incurred in discontinuous blocks, each of which enables the provider to stream a fixed amount of additional content or to execute a fixed number of additional trades per unit of time, with no additional marginal cost.[1]

Examples of discontinuous costs are likely to increase as *on-demand computing* models are more widely adopted. An on-demand IT service provider typically offers corporate buyers access to hosted software via the Internet, outsourced computing resources, and infrastructure management with a

[1] Other examples include local or cell phone service providers, video-on-demand providers, Internet caching services, shared grid computing services, shared data storage, and rendering server farms for digital animation. There are also examples of IT-enabled "nondigital" goods that share this cost structure—for instance, the provision of call center services (where each agent facilitates fulfilling a fixed number of additional calls), or the addition of new flights to an existing route of an airline.

usage-based payment structure. The vision of on-demand computing is often referred to as "utility computing," the process by which companies pay as they go for hardware and software as if paying for water, gas, or electricity. The cost structure for such IT service providers is not only discontinuous but also rapidly declining and shared among multiple products. For example, Amazon.com provides a collection of more than 10 Web services based on a shared powerful computing infrastructure. The goal of this study is to investigate how to appropriately incorporate the modern IT infrastructure costs into the pay-per-use pricing (usage pricing) structure. Specifically, major questions exist when pricing on-demand services, such as whether vendors should completely ignore the discontinuous infrastructure cost and how on-demand vendors should split the cost of a shared infrastructure among multiple Web services.

This discontinuous cost structure differs from the standard models in an important way: it models costs that are somewhere in between variable and fixed. Traditional variable costs are modeled as being incurred in the short run, varying continuously with the level of demand that a seller fulfills, and playing a critical role in pricing. Traditional fixed costs, while discontinuous, are modeled as being long-run costs that are incurred very infrequently, the magnitude of which does not directly affect a seller's short-run choice of pricing. In each of the examples described above, neither of these traditional cost components captures the actual supply of digital goods. In contrast, the relevant costs appear to be similar to fixed costs of digital goods in that they increase the *capacity* of the seller across periods but similar to variable costs in that they are incurred in the short run and vary in (possibly large) steps based on the demand that a seller faces in each period.

Our analysis proceeds as follows. We construct a model following standard assumptions in the nonlinear pricing literature in economics. We start by defining the "constrained demand problem": a standard nonlinear pricing problem with an additional capacity constraint (Spulber 1993a). Next, we show that the optimal pricing schedule with discontinuous costs coincides with the solution to a specific instance of the constrained demand problem. This reduces the former to a problem of identifying the optimal number of *blocks* of supply (e.g., optimal size of the server farm) using a simple rule, which is analogous to equating the average cost of the marginal revenue to that of the marginal block. With this general solution concept, this paper considers two extensions: rapidly declining and shared discontinuous supply costs. We show in a stylized, two-period model that pricing may not be affected by anticipated cost declines up to a point; this result highlights one implication of the discontinuous nature of the cost function. In the second case, we demonstrate how the infrastructure cost is accrued to multiple products based on the usage rate. This paper also shows that it would be typically suboptimal to use the widely adopted total cost recovery method as the transfer pricing mechanism.

The remainder of this paper is organized as follows. Section 2 provides the literature review. Section 3 describes the model setup. Section 4 develops the optimal pricing of the constrained demand problem and the pricing problem with discontinuous costs. Section 5 investigates the other two cost features. Concluding remarks are given in §6.

2. Literature Review

Our research is built on the nonlinear pricing literature within economics. This line of research starts from the seminal work of Mussa and Rosen (1978), who investigated second-degree price discrimination in a versioning context. Maskin and Riley (1984) considered a generalized model with nonlinear pricing over quantities to explore, among other things, whether quantity discounts are optimal. They also provided the theoretical equivalence between versioning and nonlinear pricing models. Based on these findings, researchers have applied similar models to study optimal taxation, incentive contracts, and dynamic price discrimination (Laffont and Martimort 2001). For a complete review of this literature, Varian (1989) has provided an early comprehensive summary of the baseline model, Wilson (1993) has provided a thorough treatment of nonlinear pricing with detailed explanations and examples, and Stole (2007) has provided a recent survey focusing on advances in multidimensional and competitive nonlinear pricing models. Laffont and Martimort (2001) have provided a masterly textbook exposition of the modern theory.

This paper is most similar to nonlinear pricing models with extensions related to the cost function or demand-capacity constraint. Our research question is similar to that discussed in the seminal work of Oren et al. (1985) who study optimal nonlinear pricing with capacity costs, particularly in the context of electricity pricing, using a general two-dimensional (consumption time and usage) nonlinear pricing model. To tackle the technical difficulties of multidimensional nonlinear pricing, the assumption on the cost structure follows conventional constant variable cost assumptions.[2] Several researchers (Braid 1989, 1996; Fischer and Serra 2003; Calzada 2007) have extended

[2] The definition of Oren et al. (1985) regarding the capacity cost is different from ours. Contrary to our setup, Oren et al. (1985) did not consider one constraint on the aggregated demand. The capacity cost was defined on the individual level rather than the aggregate level.

the work of Oren et al. (1985) to investigate peak-load pricing for electricity or telephony services in a two-dimensional setup. This stream of research centers around the substitutability between the consumption during peak and off-peak hours, whereas our emphasis is the discontinuous cost structure.

Theoretically, our model is most similar to that of Spulber (1993a) in which a monopoly nonlinear pricing model with a capacity constraint is analyzed. A minor difference is that Spulber (1993a) assumed finite consumer types while this paper assumes continuous consumer types. Spulber (1992; 1993a, b) showed that optimal nonlinear pricing schemes for the monopolist can be derived as the standard solution plus a virtual constant cost term, the shadow price associated with capacity constraint, a finding that is consistent with our results in the "constrained demand problem." This finding is also the building block of our main analysis of three unique IT cost structures: discontinuous, declining, and shared supply cost. Later, Monteiro and Page (1996) provided a technical proof of the existence of the solution in a nonlinear pricing model with a general cost function. The proposed solution is technically more complicated, but is intuitively similar to that found in Spulber (1993a). Thomas (2001) discussed the optimal degradation and exclusion properties in the same setup. Two other papers in the literature discussed related issues. Cornelli (1996) considered a nonlinear pricing problem in which the buyers pay first and then the monopolist can cancel the transaction if the total revenue is less than an exogenously given fixed cost. Essegaier et al. (2002) investigated the competitive two-part tariff pricing plans for access service industries, for example, the online service, telecommunication or fitness club industries, by a stylized model setup with a capacity constraint. Our work adds to the existing literature by examining three unique cost features of digital goods—discontinuous, rapidly declining, and shared among multiple products, all of which have not been explicitly discussed in the literature to date.

Our paper also contributes to a growing literature on applying the nonlinear pricing model to the pricing of digital goods. Jain and Kannan (2002) studied search-based and subscription fee pricing for online information services (IS), and Sundararajan (2004a) has shown that fixed-fee pricing can increase the profits earned with pure nonlinear pricing because of the administrative cost from a large number of price discrimination plans. Sundararajan (2004b) analyzed optimal pricing and technological deterrence levels with digital piracy, and Hitt and Chen (2005) adopted a similar pricing model to investigate customized bundling with fixed fees. Masuda and Whang (2006) compared the optimal three-part tariff with nonlinear pricing plans in the telecommunication industry. Chen and Seshadri (2007) have shown that a versioning strategy is optimal when the buyers have heterogeneous outside options. These papers assume that digital goods are "information goods" with zero marginal costs, an aspect that distinguishes their analyses from ours. This paper contributes by providing a simple method to incorporate the discontinuous infrastructure costs into existing studies, while at the same time validating the robustness of existing findings under a discontinuous cost structure.

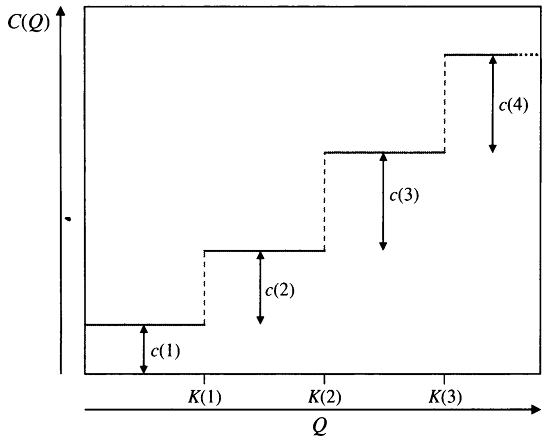
Another relevant stream of literature is the pricing of IS with congestion costs. Mendelson (1985) highlighted the differences between the optimal short- and long-run transfer pricing of IT department services with congestion externalities. Subsequently, Mendelson and Whang (1990) demonstrated that priority pricing for resources can be expressed as a base price plus a priority surcharge. Dewan and Mendelson (1990) analyzed congestion pricing with general delay costs. Konana et al. (2000) have shown that dynamic priority pricing, including a congestion premium for accessing a real-time database, outperforms a variety of standard priority rules. Nadiminti et al. (2002) discussed asymmetric information about user preferences and, among other things, highlighted the optimality of volume discounted (nonlinear) pricing. Afeche and Mendelson (2004) provided a generalized model allowing delay costs and consumer value to be interdependent. To the best of our research, the discontinuous cost structure has not been explored thus far in this stream of literature.[3]

3. Model Overview

Consider the case of a monopolist that sells a digital product that may be used by customers in continuously varying quantities. The cost function of the monopolist is described by a pair of functions $c(i)$ and $K(i)$, where $K(i)$ is the total demand that i units of infrastructure enables the monopolist to fulfill, and

[3] By abstracting away from delay costs (which form the basis for unknown customer heterogeneity in most of this literature), we can use a more general specification of heterogeneity in consumer value, which is appealing, as delay costs are not the central basis for choice in many of our examples. In a sense, one might approximately account for queuing effects in our model if one thinks of a seller with a target (exogenous) quality level that it aims to achieve and incorporates the effect of congestion by adjusting the effective additional demand that each *unit* of supply can achieve while still maintaining this quality level. Of course, explicitly modeling queuing effects could be attractive in terms of expanding the model's generality; however, it will lead to a substantially more complex model. Moreover, this would shift the focus of the model away from those unique (and underresearched) aspects of pricing IT-based products that our model aims to highlight.

Figure 1 An Illustration of the Shape of the Cost Function $C(Q)$



$c(i)$ is the incremental cost of the ith unit of infrastructure.[4] Therefore the cost of supplying a total quantity Q is specified by the function

$$C(Q) = \sum_{i=1}^{n(Q)} c(i), \tag{3.1}$$

where $n(Q)$ is the minimum units of infrastructure that provide the ability to fulfill demand of at least Q: $n(Q) = \min\{j: K(j) \geq Q\}$. To examine the average cost of each unit of infrastructure, we define $k(i) \equiv K(i) - K(i-1)$. The cost function illustrated in Figure 1 is therefore a step function with discontinuous increases at a specific integer value of Q, with potentially different jumps at each of these values.

We assume that the average cost of the monopolist, $c(i)/k(i)$, is nondecreasing in i, or loosely speaking, costs are *discontinuously convex*. This convexity can arise in many different ways. For example, the complexity of managing a server farm in which each server adds roughly the same capacity increases with the number of servers; this would correspond to an increasing level of $c(i)$, for constant $k(i)$. Analogously, the increase in the effective processing capacity of a grid of computers reduces as one adds more nodes (each of which costs about the same); this would correspond to a constant $c(i)$ and a decreasing level of $k(i)$. The latter argument might hold for any digital product whose supply is based on an IT system that resembles a multiserver queue: the incremental arrival rate $k(i)$ that can be handled by the addition of an extra server at a constant cost $c(i)$, while keeping service levels constant declines as the number of servers increases.

Customers are heterogeneous, indexed by their type, $\theta \in [0, 1]$. The preference of a customer of type θ is represented by the function

$$w(q, \theta, p) = U(q, \theta) - p, \tag{3.2}$$

where q is the usage and p is the total price paid by the customer. Our formulation of preferences follows the standard nonlinear pricing model (Maskin and Riley 1984), in which $U(q, \theta)$ is referred to as the customer's utility function, and has the following properties, for each $\theta \in [0, 1]$, $\forall q$:

(1) Increasing and concave value: $U(0, \theta) = 0$; $U_1(q, \theta) \geq 0$, $U_{11}(q, \theta) < 0$.

(2) Higher customer types get higher utility: $U(q, 0) = 0$ and $U_2(q, \theta) > 0$.

(3) These increases in utility with each type are diminishing: $U_{22}(q, \theta) \leq 0$.

(4) Spence-Mirrlees single-crossing condition: $U_{12}(q, \theta) > 0$.

(5) Nonincreasing absolute risk aversion: $\partial/(\partial\theta) \cdot [(-U_{11}(q, \theta))/(U_1(q, \theta))] \leq 0$.

Throughout the paper, numbered subscripts of functions represent derivatives with respect to the corresponding variable. Assumptions (1), (2), (4), and (5) are standard in nonlinear pricing models. A detailed discussion of the implications of these assumptions can be found, for instance, in §2.1 of Sundararajan (2004a). While Assumption (3) is made for mathematical reasons, it is a reasonable description of preferences. Intuitively, these assumptions imply that $U(q, \theta)$ is increasing and concave in both arguments. For example, the value of Amazon's Web services, Salesforce.com's on-demand CRM, or EMC's data storage farm to their respective clients is typically increasing and concave in the usage of those services. Assumption 2 implies that clients may have different valuations per unit usage, which is captured by θ. Alternatively, θ can be interpreted as other sources of heterogeneity. For example, the clients of on-demand computing are heterogeneous in their priority on IT cost reduction, IT capabilities, concerns over IT reliability or security, preferences to keep core business data in-house, or the uptime of IT applications. A typical functional form of (3.2) is $U(q, \theta) = \theta q - q^2$ (when $q \leq \theta$), which is solved in §4.3 as an example.

The sequence and information structure of the game follows the standard setup. The monopolist does not observe the type of any customer but knows $F(\theta)$, the

[4] In general, the seller deploys a fixed level of infrastructure, represented by the vector $K = (k1, k2, \ldots, k_n)$. The components of infrastructure could include hardware, software licenses, disk storage, customer support infrastructure, administration and maintenance staff, and so on. For ease of exposition, we treat K as a variable rather than a vector because here we only need the constrained demand and the associated fixed cost for that chunk of demand capacity. The seller also guarantees a fixed level of quality of service to each of its customers, which restricts the aggregate level of demand that it can fulfill at a choice of infrastructure K to a maximum of $Q(K)$.

probability distribution of types in the customer population,[5] which is assumed to be absolutely continuous (and thereby has a density function $f(\theta)$, which is nonzero and finite), and assumed to have a nonincreasing inverse hazard rate $H(\theta)$, where $H(\theta) \equiv (1-F(\theta))/(f(\theta))$. The monopolist first decides the optimal number of blocks given the discontinuous cost structure. Next, the monopolist announces a pricing schedule (i.e., $p(q)$) that assigns a specific total payment for each level of usage. Because the monopolist cannot explicitly distinguish between customer types prior to contracting, the entire schedule must be available to all customers. Lastly, each customer self-selects the value of q that maximizes his or her utility.

A standard approach used to solve this problem is to employ the revelation principle: this ensures that the monopolist can restrict his or her attention to direct mechanisms, under which one specific quantity price pair is designed for each customer type, where it is rational and optimal for the customer to choose the quantity price pair that was designed for his or her type. The pricing schedule is, therefore, represented by a menu of quantity price pairs $(q(t), p(t))$, where $t \in [0,1]$, which satisfies the following two standard constraints:

[IC]: $U(q(\theta),\theta)-p(\theta) \ge U(q(t),\theta)-p(t), \quad \forall\theta, \forall t.$

[IR]: $U(q(\theta),\theta)-p(\theta) \ge 0, \quad \forall\theta.$

[IC] ensures that each buyer will self-select the $(q(t), p(t))$ designed for him or her while [IR] ensures that all buyers receive a nonnegative surplus. For an incentive-compatible schedule $(q(t), p(t))$, the cumulative quantity demanded by all customers will be $\int_{\underline{\theta}}^{1} q(\theta)f(\theta)d\theta$, which means that the monopolist's profit-maximization problem is given by

$$\max_{p(\theta),\,q(\theta)} \int_{\underline{\theta}}^{1} p(\theta)f(\theta)\,d\theta - C\Big(\int_{\underline{\theta}}^{1} q(\theta)f(\theta)\,d\theta\Big) \quad (3.3)$$

$$\text{subject to } [\text{IC}], [\text{IR}], p(\theta) \ge 0, \quad \text{and} \quad q(\theta) \ge 0, \quad \forall\theta \in [0,1], \quad (3.4)$$

where $\underline{\theta}$ is the consumer who feels indifferent between buying or not. In contrast, most existing studies, with the exception of Spulber (1993a) analyze the following objective function:

$$\int_{\underline{\theta}}^{1}[p(\theta) - c \times q(\theta)]f(\theta)\,d\theta,$$

where c is the constant variable cost typically assumed to be zero for digital goods in the literature, leading to the idea that cost does not matter at all when pricing information goods.

4. Nonlinear Pricing with Discontinuous Costs

This section explains the solution of the proposed problem. The sequence of analysis leading to the section's main result is summarized in Table 1. Because of the technical complexity of this problem, formal theorems and proofs are deferred to the appendix. Only intuitions will be discussed in this section.

4.1. Nonlinear Pricing with Demand Constraints

This section considers the nonlinear pricing problem when the monopolist incurs no fixed or variable costs but faces a constraint, denoted by K, on its capacity. While this subproblem may seem tangential to the main problem, it is important because the optimal pricing schedule in the main problem depends critically on the solution of this subproblem. The monopolist cannot increase K, and incurs no costs for supplying quantity $Q \le K$. The pricing problem is therefore

$$\max_{q(\cdot),\,p(\cdot)} \int_{\underline{\theta}}^{1} p(\theta)f(\theta)\,d\theta \quad (4.1)$$

$$\text{subject to (3.4) and } \int_{\underline{\theta}}^{1} q(\theta)f(\theta)\,d\theta \le k. \quad (4.2)$$

This problem is referred to as the "constrained demand pricing problem." A similar model with a finite number of buyers is solved in Spulber (1993a). The optimal solution is shown to be equivalent to the solution of a standard nonlinear problem with a constant variable cost λ, which is the Lagrange multiplier associated with the demand capacity constraint (4.2). Formally, (4.1) and (4.2) can be rewritten as

$$\max_{q(\cdot),\,p(\cdot)} \int_{\underline{\theta}}^{1} [p(\theta) - \lambda \times q(\theta)]f(\theta)\,d\theta \qquad \text{subject to (3.4).} \quad (4.3)$$

This result also implies that with constant variable cost c and demand constraint K, the solution can be derived by setting the constant variable cost at $c+\lambda$. All standard results in the nonlinear pricing literature have a counterpart in the current setup. For example,

Table 1 A Brief Summary of the Sequence of Analysis

(A) Lemma 1 formulates a constrained demand pricing problem in a form that makes it tractable.

(B) Lemma 2 provides a characterization of the solution to this problem. With a constraint on demand K, the optimal pricing schedule is $(p^c(\theta,K), q^c(\theta,K))$.

(C) Theorem 1 shows that if the optimal solution to the main problem involves the monopolist incurring its first n^* "units" of cost, then the pricing schedule must be $p^*(\theta) = p^c(\theta, K(n^*))$, $q^*(\theta) = q^c(\theta, K(n^*))$.

(D) Theorem 2 characterizes the optimal choice of n^*, which leads immediately to the optimal pricing schedule based on Theorem 1 and Lemma 2.

[5] The interpretation of $F(\theta)$ is slightly different from the standard model (Spulber 1993a, b) because of the fact that capacity constraint is on the integral of all quantities. One interpretation is that there are many buyers and the value of θ is unknown to the seller. However, our model's result is consistent with the findings of Spulber (1993a, b).

there is no distortion of the consumption for the highest type, while lower types are inefficiently underserved by the monopolist (Maskin and Riley 1984).

The other economic interpretation of λ is the shadow value of the demand capacity constraint; that is, the "marginal revenue function" of the monopolist in terms of K. λ, therefore, measures the marginal increase in revenue with one additional unit of demand that the monopolist is able to fulfill, after the monopolist adjusts its pricing function in response to the relaxation of the constraint. This seemingly simple fact is critical to our main solution with discontinuous costs.

We denote the optimal solutions in terms of K by $q^C(\theta, K)$ and $p^C(\theta, K)$. The results of the comparative statics analysis are summarized as follows.

Proposition 1. *When the capacity constraint is binding,*

(i) $q_2^C(\theta, K) > 0$: *relaxing the demand constraint induces an increase in total consumption.*

(ii) $d[p^C(\theta, K)/q^C(\theta, K)]/dK < 0$: *relaxing the demand constraint decreases the unit price for all customers, and the impact on total price can be shown to be ambiguous.*

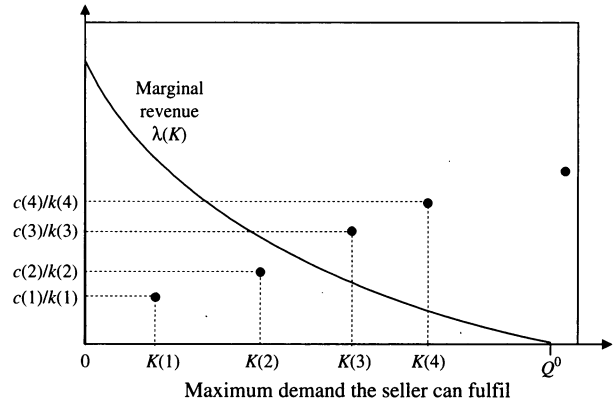
(iii) $\underline{\theta}_1(K) < 0$: *relaxing the demand constraint increases the fraction of participating customers.*

The result, $q_2^C(\theta, K) > 0$, is independently interesting from a managerial perspective. An alternative response would be for the monopolist to focus on the high-end market by using zero marginal cost pricing while shutting out more low-end customers to satisfy the capacity constraint, for example, by no longer offering contracts designed for such customers. This seems intuitively consistent with getting as much value as possible from one's allowed total demand. This proposition shows that to stop serving low-end customers is never a profit-maximization strategy. The optimal strategy is to charge a higher price to all customers, perhaps by raising unit prices in all contracts rather than offering fewer contracts, so that the usage of all customers is reduced and the demand capacity will be satisfied accordingly. This is because, although low-end consumers are less valuable to the monopolist, they may have relatively high marginal willingness to pay when they have very low consumption compared to that of high-end consumers with high consumption. Given part (i) of Proposition 1, parts (ii) and (iii) become intuitive—when the demand constraint is relaxed, the monopolist can sell more products in total. In response, the monopolist will lower the unit price, leading to a larger customer base and higher usage by all users.

4.2. Optimal Pricing with Discontinuous Costs

In the appendix, Theorem 1 shows that the solution of the main problem with discontinuous costs must occur at the discontinuous points. Under the assumption that $c(i)/k(i)$ is nondecreasing (discontinuously convex), Theorem 2 provides a simple way of identifying this optimal number, by comparing its incremental revenue to its incremental average cost. We do not see general results when $c(i)/k(i)$ is nonmonotonic or rapidly decreasing in i, the reasons for which will be discussed toward the end of the section.

Figure 2 An illustration of Part (i) of Theorem 2



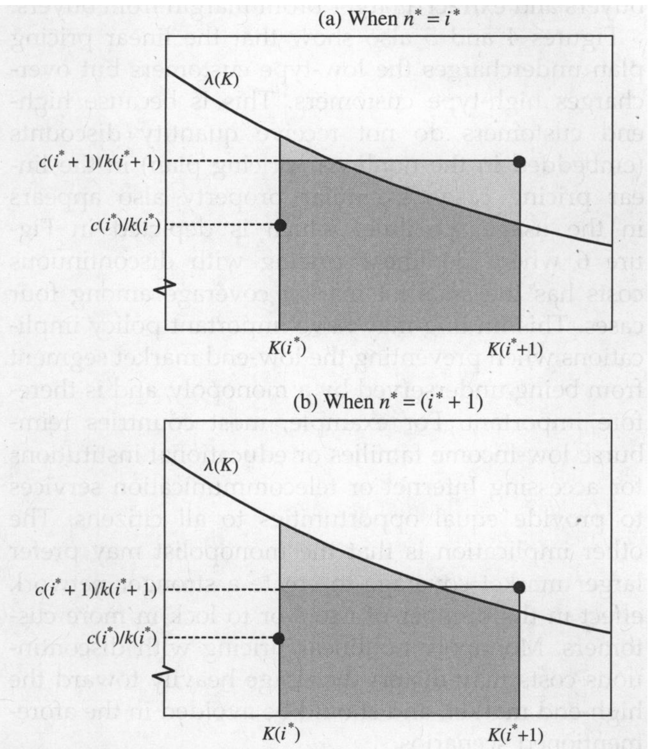
Notes. The solid downward sloping curve is the marginal revenue λ of capacity K, while the series of upward sloping points are successive values of the average cost of capacity $c(i)/k(i)$. As illustrated, $i^* = 2$, since $\lambda(K(2)) > c(2)/k(2)$ and $\lambda(K(3)) < c(3)/k(3)$.

The main idea of Theorem 2 is illustrated in Figure 2. Marginal revenue decreases as the monopolist's ability to fulfill demand increases, and the corresponding average cost, $c(i)/k(i)$, of fulfilling each incremental block of demand increases with i. Since the monopolist gets no further revenue from increasing its ability to fulfill demand beyond Q^0, the revenue-maximization level of demand, $\lambda(K)$ is zero for $K > Q^0$. The proof that $\lambda(K)$ is continuous and decreasing in K is provided in the Online Appendix B.[6] Therefore, there are always two successive values of i such that the marginal revenue exceeds the average cost at the former, and the average cost exceeds the marginal revenue at the latter. The optimal solution must be either i^* or $i^* + 1$ because the monopolist can add or reduce blocks to strictly improve profits whenever $i < i^*$ (or $i > i^* + 1$).

To pick the optimal solution from i^* and $i^* + 1$, it is necessary to compare the profits from these two cases. The area under the marginal revenue curve over each unit of cost represents the actual additional revenue that the monopolist can get by incurring this unit of cost. This incremental revenue stems from the optimal changes in its pricing schedule that reflect the monopolist's ability to fulfill $k(i^* + 1)$ additional units of demand. Figure 3 illustrates this result further.

[6] An electronic companion to this paper is available as part of the online version that can be found at http://isr.journal.informs.org/.

Figure 3 Illustrations of the Result of Part (ii) of Theorem 2



Note. The area under the $\lambda(K)$ curve between $K(i^*)$ and $K(i^*+1)$ is the incremental revenue (horizontal stripes); the area in the rectangle between $K(i^*)$ and $K(i^*+1)$ and under the $c(i^*+1) = k(i^*+1)$ line is the incremental cost $c(i^*+1)$. Panel (a) illustrates a scenario under which incremental cost exceeds incremental revenue, in which case $n^* = i^*$, while panel (b) illustrates the opposite, in which case $n^* = i^*+1$.

This solution procedure can be generalized as a continuously convex cost function. First, we can observe that the optimal solution still coincides one instance of the *constrained demand pricing problem.* Therefore we can depict the marginal revenue curve, $\lambda(K)$, as in Figure 2. The main difference is that the supply curve is a continuous marginal cost curve, rather than those average cost points in Figure 2. As a consequence, the optimal capacity can be determined by directly equating the marginal revenue and marginal cost.

4.3. A Numerical Example: Linear and Nonlinear Pricing

Our baseline model provides the theoretical benchmark of other simplified pricing plans, such as linear pricing and two- and three-part tariffs, particularly mobile phone pricing plans. The proposed solution concept can be applied to these simplified pricing plans as well because identifying the optimal number of blocks depends only on the cost function and the marginal revenue curve, as discussed in §4.2, but cannot be applied to the functional form of pricing plans. The current section illustrates this argument by comparing linear and nonlinear pricing solutions. Formally, a simplified pricing strategy is equivalent to an additional constraint in the baseline model. For example, if the monopolist is restricted to linear pricing, the maximization problem becomes the objective function in (3.3) plus an additional constraint $p(\theta) = b \times q(\theta)$, where b is a positive constant. It can be verified that all solution procedures in §4.1 still apply, and as in (4.3), the linear pricing problem can be transformed to

$$\max_{q(\cdot),\, b} \int_{\underline{\theta}}^{1} [b \times q(\theta) - \lambda \times q(\theta)] f(\theta)\, d\theta \quad \text{subject to (3.4)}, \qquad (4.4)$$

which suggests that the discontinuous cost structure can be accrued as a virtual variable cost λ also in the linear pricing case.

To shed more light on our results, we conduct a numerical analysis to contrast the results of linear and nonlinear pricing with and without discontinuous costs. We assume that customer types follow the uniform distribution between 0 and 1, and the utility function is quadratic: formally, $F(\theta) = \theta$ and

$$U(q,\theta) = \begin{cases} \theta q - \frac{1}{2}q^2, & \text{when } q \le \theta; \\ \theta, & \text{else.} \end{cases}$$

Closed-form solutions are reported in Table C.1 in the Online Appendix C. The optimal pricing functions when $\lambda = 0$ and 0.5 are depicted in Figure 4. Unit price and usage plans are depicted in Figures 5 and 6, respectively.

There are two distinctive features in our model: discontinuous costs and nonlinear pricing. To examine the impact of the discontinuous cost, we compare the results when $\lambda = 0$ and 0.5 in either the linear or nonlinear pricing cases. When $\lambda = 0$, the solution is "zero marginal cost pricing." Figures 4 and 5 demonstrate

Figure 4 Numerical Solutions of $p(q)$ in the Linear and Nonlinear Pricing Cases When $\lambda = 0$ or 0.5

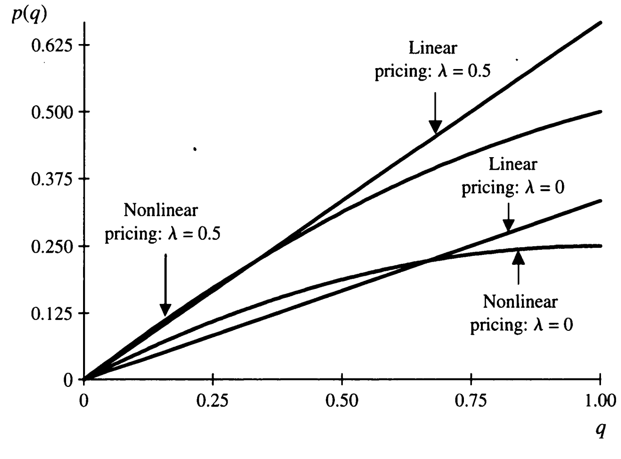
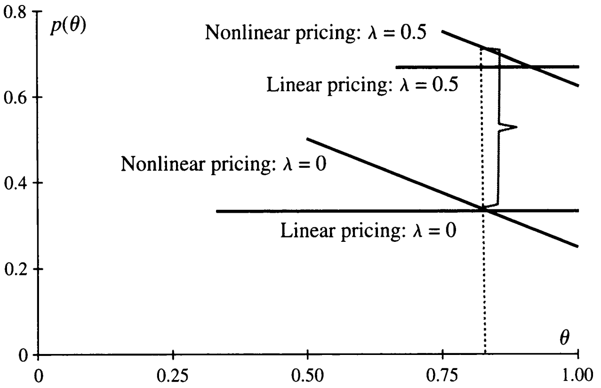


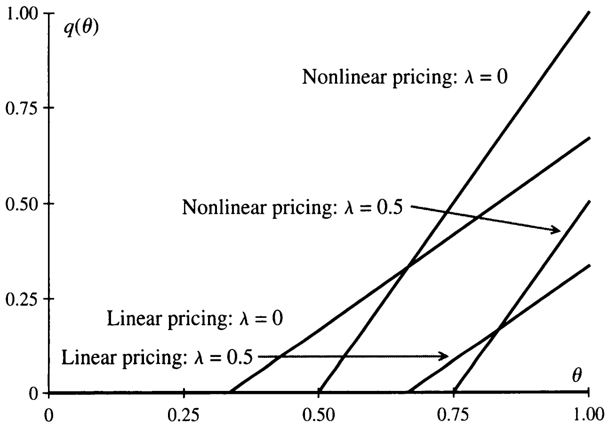
Figure 5 **Numerical Solutions of $p(\theta)/q(\theta)$ in the Linear and Nonlinear Pricing Cases When $\lambda = 0$ or 0.5**



that the pricing is too low for *all* buyers when the monopolist treats the discontinuous cost as a conventional fixed cost. In the Online Appendix C, we also present the results of a numerical analysis to investigate the profit loss from ignoring the discontinuous cost by varying the block size and average cost of blocks. We find that the profit loss could be as high as 70% under reasonable parameter settings (Table C-2 in Online Appendix C). The amount of profit loss depends critically on the actual optimal capacity that equates the marginal average cost and marginal revenue. The smaller the optimal capacity, the larger the profit loss will be. Theoretically, the profit loss could be close to 100% in examples in which the optimal number of blocks is one, whereas the revenue-maximization level is a large number of blocks.

Comparing linear and nonlinear pricing results, we find that in both cases, the monopolist should raise the unit price by a constant. In the Online Appendix C, we show that in the nonlinear pricing case, this constant should be $3\lambda/4$, compared with only $2\lambda/3$ in the linear pricing case. In other words, nonlinear pricing can transfer a larger portion of the virtual variable cost to buyers because a nonlinear pricing plan can more effectively price discriminate buyers and extract a larger profit margin from buyers.

Figure 6 **Numerical Solutions of $q(\theta)$ in the Linear and Nonlinear Pricing Cases When $\lambda = 0$ or 0.5**



Figures 4 and 5 also show that the linear pricing plan undercharges the low-type customers but overcharges high-type customers. This is because high-end customers do not receive quantity discounts (embedded in the nonlinear pricing plan) in the linear pricing case. A similar property also appears in the usage schedule, which is depicted in Figure 6 where nonlinear pricing with discontinuous costs has the smallest market coverage among four cases. This finding may have important policy implications when preventing the low-end market segment from being underserved by a monopoly, and is therefore important. For example, most countries reimburse low-income families or educational institutions for accessing Internet or telecommunication services to provide equal opportunities to all citizens. The other implication is that the monopolist may prefer larger market coverage to create a stronger network effect in the number of users or to lock in more customers. Monopoly nonlinear pricing with discontinuous costs may distort the usage heavily toward the high-end market, and should be avoided in the aforementioned scenarios.

5. Discontinuous, Declining, and Shared Costs

Our solution procedure proposed in §4 demonstrates a nice separability between the optimal pricing and IT capacity planning problems, particularly in identifying the optimal number of blocks of supply, for digital goods with discontinuous costs. This is largely because of the fact that the solution procedure in §4.2 depends only on the marginal revenue curve and the discontinuous cost function, but not on the exact pricing strategies. Section 5 extends the baseline model to the cases in which the supply cost is not only discontinuous but also declining over time or shared among multiple business units.

5.1. Declining Costs and Evolving Demand

This section extends the analysis to multiple periods: the monopolist can generate revenue in an infinite number of periods while, for simplicity, the IT capacity procurement is only allowed in the first two periods. The discontinuous cost is assumed to decline in the second period. The goal of this section is to investigate the optimal IT capacity planning when the cost is both discontinuous and declining in a stylized setting.

To isolate the impact of the discontinuous costs, we assume that customers are short lived in each period with identical and independent distribution of types. Because consumers are short lived, contract renegotiation issues and dynamic pricing issues are ruled out by this assumption, which greatly simplifies the following analysis. The main trade-off faced by the monopolist is clear cut: capacity cost saving by delaying installation to the second period versus revenue loss in the first period because of smaller installed capacity. This section shows that the discontinuous cost provides additional incentives for the monopoly not to defer capacity installation.

We denote the solution of revenue function of the single-period demand-constrained problem solved in §4 by $R(K(n))$, which is a function of the number of blocks. Formally, the single-period problem can be expressed as

$$\max_n R(K(n)) - C(K(n)).$$

Let the number of blocks installed in each period be n_1 and n_2, respectively. The objective function is given by

$$\begin{aligned}&\underbrace{R(K(n_1)) - C(K(n_1))}_{\text{First-period profit}}\\&+\underbrace{\delta[\alpha'_R R(K(n_1+n_2)) - \alpha_c C(K(n_1+n_2)) + \alpha_c C(K(n_1))]}_{\text{Second-period profit}}\\&+\underbrace{\sum_{t=2}^{\infty}\delta^t(\alpha'_R)R(K(n_1+n_2))}_{\text{Future profit}}.\end{aligned}$$

The first two terms are the profits from the first period. δ is the discount factor and is between 0 and 1. The terms in the bracket are the profits earned in the second period. Since n_2 is defined as the incremental number of blocks purchased in the second period, the total capacity in the second period is $n_1 + n_2$. Therefore the capital expenditure in the second period is $\alpha_C[C(K(n_1+n_2)) - C(K(n_1))]$, where $1 > \alpha_C > 0$ is defined as the degree of the declining cost in the second period. In the second and third terms, $\alpha'_R > 0$ models the market demand expansion or contraction in future periods.[7] For example, for most of the on-demand computing services, demand increases and the hardware costs decline over time because of its infancy. For fixed-line telephony services, dial-up ISP, and integrated services digital network ISP, demand and costs both decline over time. For ease of exposition, the objective function can be rewritten as

$$\begin{aligned}R(K(n_1)) - C(K(n_1)) + \delta[\alpha_R R(K(n_1+n_2))\\ - \alpha_C C(K(n_1+n_2)) + \alpha_C C(K(n_1))],\end{aligned}\qquad(5.1)$$

where $\alpha_R > 0$ models the second period present value of all future revenues. The main finding is summarized in the following proposition.

Proposition 2. (i) *When $\delta \geq 1/\alpha_R - 1/\alpha_C$, the monopolist will not defer capacity installation ($n_2^* = 0$).*

(ii) *When $\delta < 1/\alpha_R - 1/\alpha_C$, the monopolist will also choose $n_2^* = 0$ when*

$$\begin{aligned}R(K(n^*)) - R(K(n^*-1))\\ > (1-\delta\alpha_C)[C(K(n^*)) - C(K(n^*-1))],\end{aligned}\qquad(5.2)$$

where $n^* \equiv \max\{n\colon \arg\max_n \alpha_R \times R(K(n)) - \alpha_c \times C(K(n))\}$.

(iii) *Otherwise, the monopolist will defer capacity installation, and therefore $n_2^* > 0$.*

The inequality $\delta \geq (1/\alpha_R) - (1/\alpha_C)$ illustrates the main trade-off in this section. When the current revenue is more important (α_R is smaller) or the cost saving is small (α_C is larger), this right-hand side (RHS) becomes larger and the monopolist tends to defer capacity installation to the second period. In economics, the discount factor on the left-hand side (LHS) has two components: real interest rate and risk premium. The second component is critical to IT capacity planning: firms with high uncertainty about future demand and costs will bear a larger risk premium and will therefore have a smaller discount factor, δ. Hence, firms operating with new technologies, such as WiMax mobile computing or on-demand ERP, have smaller discount factors and should defer their capacity expansion to the second period after the uncertainty is resolved.

In parts (i) and (ii) of Propositions 2, $n_2^* = 0$, but the explanations behind the two cases differ. Parts (i) and (iii) of Propositions 2 result from the main trade-off of delaying capacity installation but do not result from the discontinuous cost. In contrast, part (ii) of Proposition 2 results mainly from the discontinuous cost. When the cost function is continuous, this regime disappears. The intuition is that when the cost is discontinuous, the monopoly cannot defer less than one block, making it more costly for the monopoly to defer capacity installation.

The intuition behind (5.2) is as follows: the LHS is the revenue lost in the first period and the RHS is the cost saving in the second period. The definition of n^* represents the largest number of blocks that can maximize the term in the bracket. Equation (5.2)

[7] The multiplicative ratio change of revenue is derived under the following two assumptions: (1) $F(\theta)$ remains the same and (2) the total number of buyers increases over time. In second-degree price discrimination models, the total number of customers is typically normalized to 0 because it does not affect the solutions. Only $F(\theta)$ affects solutions. Hence the total revenue is proportional to the total number of customers as long as $F(\theta)$ remains the same. These two assumptions fit the context in which the market demand expansion is not skewed toward high- or low-type adopters. In other words, these two assumptions fit most IT services in the mature stage of the product life cycle in which the total market size expands or contracts proportionally in both high- and low-end market segments.

implies that the larger the block size is, the less likely the monopoly will defer capacity installation. This is because $[R(K(n^*)) - R(K(n^*-1))]/[C(K(n^*)) - C(K(n^*-1))]$ is the discontinuously marginal revenue divided by marginal cost. Since we have shown that the marginal revenue is decreasing while the marginal cost is discontinuously increasing based on our assumptions, this ratio is decreasing in K and n^*. As a consequence, the larger the block size, the larger this ratio will be, and the easier it will be for (5.2) to hold.

Next, with Proposition 1, we can analyze the dynamics of the monopolist's optimal pricing plans. Let the benchmarking case be the pricing solutions when $n_2^* = 0$. In this case, the IT capacity and associated optimal nonlinear pricing remain the same in all periods. When $n_2^* > 0$, we show in the appendix that the capacity in the first period is smaller, while the capacity in the second period is greater than that in the benchmarking case. Proposition 1 provides the directional impacts of the installed capacity on endogenous variables. Because we know the relative IT capacity in each period, we can order endogenous variables in three cases in Table 2.

Table 2 suggests that when the monopolist defers capacity installation to the second period, the unit price will be higher in the first period and smaller in the second period, compared with the benchmarking case. At the same time, usage and market coverage become smaller in the first period and increase in the second period. This result provides one explanation for the decreasing average revenue per user over time in several IT service industries, such as telecommunication, mobile, and ISP industries. The rationale provided by our model is that, when facing rapidly declining infrastructure costs, IT service providers may defer the installation of capacity. At the same time, they will implement nonlinear pricing plans with higher average prices to inhibit the usage of users (Period 1 in Table 2). Over time, as infrastructure costs decline, the total installed capacity increases and the capacity constraint becomes less of a critical issue, resulting in a lower average price for each type of user (Period 2 in Table 2).

Table 2 Contrasting the Results When a Monopoly Defers Capacity Installation with the Case Without Deferring Capacity Installation

Notations	Definition of variables	Period 1 when $n_2^* > 0$	Period 2 when $n_2^* > 0$	All periods when $n_2^* = 0$
n_t^*	Installed blocks of IT capacity	1	3	2
$1-\underline{\theta}$	Market coverage	1	3	2
$q(\theta)$	Usage plan	1	3	2
$p(\theta)/q(\theta)$	Unit price	3	1	2

Note. 3, highest value; 2, second-highest value; 1, lowest value.

5.2. Infrastructure Costs and IT Chargeback

The current section examines IT infrastructure costing when the monopolist provides multiple IT services by a shared IT infrastructure. For example, Google and Yahoo! both provide a number of services, including e-mail, chat, spreadsheets, calendars, and social networking based on a powerful shared IT infrastructure. Applications based on a shared infrastructure have been increasing because of the advances in grid computing and Web service technologies. This application is important also because, traditionally, IT infrastructure costs were treated as fixed overhead costs. These kinds of costs are either inappropriately absorbed by IT departments or charged out equally to all business units regardless of individual usage of the infrastructure. This section will show that such indiscriminant cost-allocation schemes typically lead to a suboptimal profit overall.

Suppose the monopolist offers m products where $m > 1$. For simplicity, we assume that the demand for each of these products is independent. The assumption is clearly restrictive, but our purpose is to investigate optimal infrastructure costing, so ignoring complicated interactions among the demands of multiple products, such as bundling effects or intra firm conflicts, helps to isolate the impacts of the underlying discontinuous cost structure and may help to produce transparent insights. We continue to maintain the assumptions made in §4.

The revenue function for product i is denoted by $R_i(Q_i)$, which is constructed by an optimal pricing schedule under the demand capacity constraint at Q_i, a problem solved in §4.1. We also allow the provision of each of the services to entail the use of varying "amounts" of the infrastructure (denoted by u_i). The units of the infrastructure could be interpreted as CPU hours, storage device inputs/outputs, lines printed, and server connection hours, depending on which metric best fits the target application. When business unit i provides Q_i units of service i, it costs $u_i \times Q_i$ units of the infrastructure.

As in practice, we assume that each product's pricing is determined by independent business units (profit centers). The monopolist is responsible for IT capacity planning and chargeback mechanism design. All product pricing decisions are delegated to the business units. Formally, the sequence of the game is as follows: in the first period, the monopolist chooses a number of blocks of computing resources, n, and sets an IT chargeback (transfer price) at t per unit usage of infrastructure. The monopolist maximizes total profit from all business units. The chosen n and t are common knowledge to all business units. In the second period, m business units simultaneously announce pricing plans, $p_i(q_i)$, which indirectly determines Q_i in each market segment. Formally,

$$\text{First period: } \max_{n,t} \sum_{i=1}^{m} R_i(Q_i) - C(K(n))$$

$$\text{subject to } \sum_{i=1}^{m} \mu_i \times Q_i = K(n), \tag{5.3}$$

$$\text{Second period: } \max_{Q_i} R_i(Q_i) - t \times u_i \times Q_i, \quad \forall i. \tag{5.4}$$

This problem abstracts three important business decisions: (1) IT capacity planning (n), (2) IT chargeback (t), and (3) the pricing of products based on a shared infrastructure ($p_i(q_i)$). Next, we define the "full cost recovery" as the transfer pricing policy in which the total transfer price equals the total infrastructure cost. Under the assumption that the cost structure is strictly discontinuously convex, we present the following proposition.

Proposition 3. (i) *The equilibrium IT chargeback price is* λ: $t^* = \lambda$, *where* λ *is the Lagrange multiplier associated with the demand capacity constraint at* $K(n^*)$. *The optimal number of blocks* (n^*) *is chosen by comparing* $\lambda(K)$ *with the discontinuous average cost.*

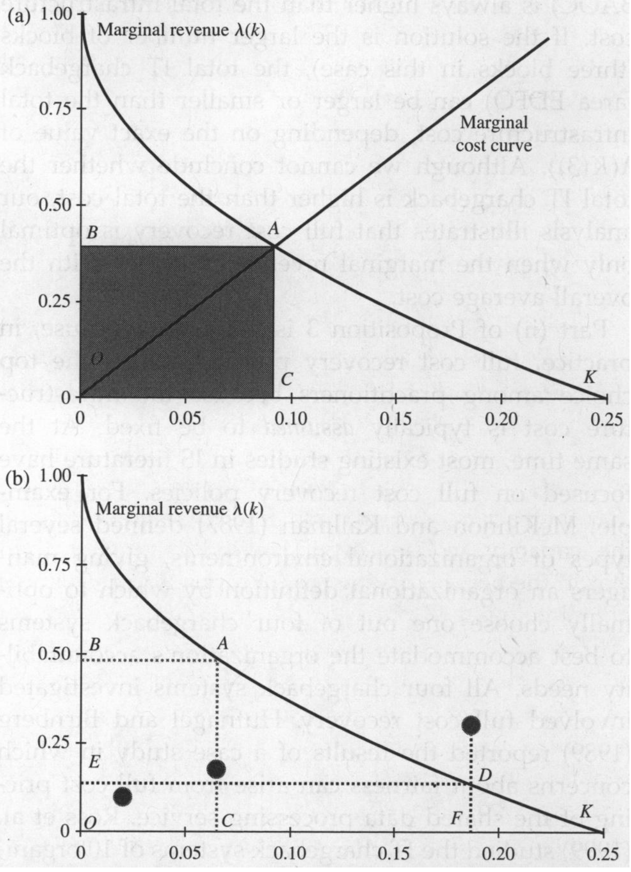
(ii) *When the block size is sufficiently small or the cost function is continuously convex, full cost recovery is never an optimal transfer pricing policy because the total transfer price is larger than the total cost.*

$$\underbrace{\sum_{i=1}^{m} t^* \times u_i \times Q_i}_{\textit{Total IT chargeback}} > \underbrace{C(K(n)).}_{\textit{Total infrastructure cost}} \tag{5.5}$$

(iii) *When the block size is large, the IT chargeback can be smaller than the total infrastructure cost. Only under rare coincidences can the total IT transfer price equal total infrastructure cost.*

The intuition behind part (i) of Proposition 3 is that, if the monopolist decides the product pricing for all business units, the optimal pricing plans are determined by costing the shared infrastructure by a virtual variable cost λ, the Lagrange multiplier associated with the capacity constraint.[8] As a result, when the transfer pricing is set at λ, the decentralized solution will be equivalent to the centralized solution. As a result, the monopolist cannot improve its profit by using other transfer pricing plans, including nonlinear transfer pricing plans. The details are described in the appendix.

Figure 7 Illustrations of Proposition 3



Note. Panel (a) illustrates a scenario under which incremental cost exceeds incremental revenue, in which case $n^* = i^*$, while panel (b) illustrates the opposite, in which case $n^* = i^* + 1$.

Next, part (ii) of Proposition 3 concludes that full cost recovery rules cannot be optimal when the block size is smaller than a threshold. The reason becomes clear when the cost function is continuously convex, which is illustrated in Figure 7(a). In §4, we show that an optimal λ is equal to the marginal cost, which is greater than the overall average cost because of the convexity of the cost function. As a result, the transfer price, λ, is higher than the overall average cost, which implies that the total transfer pricing (area BAOC in Figure 7(a)) must be greater than the total infrastructure cost (area AOC in Figure 7(a)).

When the infrastructure cost is discontinuous, this problem becomes more complicated. When the block size is small, the marginal revenue is still close to the marginal average cost and the claims in the preceding paragraph still apply. When the block size becomes larger, Figure 7(b) illustrates the optimal solution, which is either two or three blocks. If the solution is the smaller number of blocks (two blocks in this case), the marginal revenue is still larger than the

[8] The intuition is similar to our baseline model: the marginal average cost must equal the marginal revenue of the per unit computing power. In the multiproduct case, the capacity constraint is associated with a Lagrange multiplier, which is also the marginal revenue of each unit of the computing infrastructure. Because the demand of each product depends only on that product's price, the optimal solution must produce the same marginal revenue from all products. Otherwise, the monopolist can sell more units of the product that has higher marginal revenue to improve its overall profit.

last block's average cost: total IT chargeback (area BAOC) is always higher than the total infrastructure cost. If the solution is the larger number of blocks (three blocks in this case), the total IT chargeback (area EDFO) can be larger or smaller than the total infrastructure cost, depending on the exact value of $\lambda(K(3))$. Although we cannot conclude whether the total IT chargeback is higher than the total cost, our analysis illustrates that full cost recovery is optimal only when the marginal revenue coincides with the overall average cost.

Part (ii) of Proposition 3 is important because, in practice, full cost recovery policies remain the top choice among practitioners because the infrastructure cost is typically *assumed* to be fixed. At the same time, most existing studies in IS literature have focused on full cost recovery policies. For example, McKinnon and Kallman (1987) defined several types of organizational environments, giving managers an organizational definition by which to optimally choose one out of four chargeback systems to best accommodate the organization's accountability needs. All four chargeback systems investigated involved full cost recovery. Hufnagel and Birnberg (1989) reported the results of a case study in which concerns about fairness can arise from full cost pricing of the shared data processing service. Ross et al. (1999) studied the IT chargeback systems of 10 organizations, and found these 10 systems to be quite different in many aspects, except that all were cost-recovery methods. Our study shows that full cost recovery policies are typically suboptimal when the infrastructure cost is discontinuously convex. The optimal transfer pricing is simple: it is the marginal revenue at the capacity limit of the infrastructure.

6. Concluding Remarks

A number of IT-based products and services are modeled as information goods that have large fixed costs of production, but no variable costs of production or distribution. It is widely recognized that pricing for information goods differs significantly from those which are optimal for goods with positive variable costs. However, we posit that IT-based products and services—Internet services, telephony, online trading, on-demand software, digital music, streamed video on demand, and grid computing—are not really information goods. Variable increases in demand are fulfilled by the addition of *blocks* of computing or network infrastructure, and their actual cost structure resembles a mixture of positive periodic fixed costs and zero marginal costs.

We have discussed a general approach for the appropriate costing of such discontinuous costs for digital goods. The optimal nonlinear pricing schedule with discontinuous supply functions coincides with the solution to a standard nonlinear pricing problem with a virtual constant variable cost, which is the Lagrange multiplier associated with the constraint from the limit of IT capacity. This virtual variable cost is also the marginal revenue of the IT capacity. Additionally, we examined two other cost features common in IT services. In the first case, we show that the discontinuous cost structure may reduce the incentive of the monopolist to delay IT capacity procurement to take advantage of the declining hardware costs. In the second case, we analyze the optimal costing for the shared IT infrastructure. Our model shows a surprising result: the full cost recovery IT chargeback system is typically suboptimal.

For each of the three properties of costs, there exist unanswered questions for future research. For our baseline model with discontinuous costs, we do not consider the competition effects. For example, it is likely that Google is able to compete more effectively because of its technological prowess in implementing a shared infrastructure, which delivers tremendous power at a cost-performance ratio far lower than that of its competitors. An extension of our model to infrastructure-based competition would add to our understanding of how pricing power is influenced by technological capability in these increasingly common shared infrastructure environments.

Second, our model can be generalized one step further: as long as the overall objective function is discontinuously convex, our solution procedure applies well when the monopolist generates advertising revenue from the total usage, as is the case with Google and many other e-commerce sites. Formally, we can solve the following pricing problem:

$$\max_{p(\theta),\, q(\theta)} \int_{\underline{\theta}}^{1} p(\theta) f(\theta)\, d\theta + A\Big(\int_{\underline{\theta}}^{1} q(\theta) f(\theta)\, d\theta\Big)$$
$$\text{subject to (3.4).} \tag{6.1}$$

Similar to our baseline case, this optimization problem can be further transformed to

$$\max_{p(\theta),\, q(\theta)} \int_{\underline{\theta}}^{1} [p(\theta) + \lambda \times q(\theta)] f(\theta)\, d\theta$$
$$\text{subject to (3.4),} \tag{6.2}$$

where the advertising income is incorporated into the objective function as a per unit virtual revenue, λ. The *virtual advertising income* may induce the monopolist to price products lower than zero, making profits only from advertising income, a result consistent with many real-world applications, including Web mail services, instant messengers, peer-to-peer applications (i.e., Kazaa), social networking sites (i.e., Facebook, Friendster, or Twitter), YouTube, informational service websites (i.e., TripAdvisor.com), and specialized search engines.

Finally, to isolate the impacts of cost structures on pricing, we impose several simplifying assumptions in §5. In the rapidly declining cost case, we assume that customers are short lived. The dynamic price discrimination with IT capacity planning is a practically important research question. Our conjecture is that, when customers are long lived, buyers will defer their purchase decisions because they expect the product prices to drop, which puts pressure on the monopoly's first-period pricing and induces the monopoly to increase its IT capacity in the early stages. A related research direction is to generalize the present model to investigate optimal subscription pricing plans over different lengths of time. A complete solution is beyond the scope of this paper because of the complexity of dynamic pricing models. In the shared infrastructure case, we assume that demands are independent in different markets. When bundling is not allowed and there are additional costs or benefits from selling two products together, we can still use the marginal benefit of us the shared IT infrastructure to determine the optimal scale of IT capacity. When a bundling strategy is allowed and the monopoly offers several pure bundles, our model could be extended to this case because the bundled products can be considered as one product (one demand function) and our results still apply. When the monopoly uses a mixed bundling strategy, the pricing problem becomes more complicated and deserves further investigation because increasingly more IT products and features are sold as a large bundle (Bakos and Brynolfsson 1999, 2000).

7. Electronic Companion

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

Acknowledgments

The authors thank the associate editor, two anonyomous reviewers, Nicholas Economides, Ravi Mantena, Ivan Png, Ying-Ju Chen, and especially Roy Radner for helpful discussions, as well as seminar participants at New York University and CIST 2007 for feedback on an earlier draft of this paper. The authors take all responsibility for errors and omissions. This research is funded by NET Institute Summer Research Grant in 2006.

Appendix. Proofs

We proceed the proofs by the following standard procedure used in the nonlinear pricing literature. In Lemma 1, we first transform the optimization problems, (3.3) and (3.4), in §3 into a tractable form. The solution to the transformed optimization problem is derived in Lemma 2. Given these two lemmas, the proofs of the main theorems become straightforward.

Proof of Theorem 1

Lemma 1. *An equivalent formulation of the pricing problem in* (4.1)–(4.2) *is*

$$\max_{q(\cdot)} \int_{\underline{\theta}}^{1} [U(q(\theta),\theta) - U_2(q(\theta),\theta)H(\theta)]f(\theta)\,d(\theta) \quad (A.1)$$

$$\text{subject to } q(\theta) \ge 0 \quad \forall\,\theta \in [\underline{\theta},1], \quad (A.2)$$

$$q_1(\theta) > 0 \quad \forall\,\theta \in [\underline{\theta},1], \quad (A.3)$$

$$\int_{\underline{\theta}}^{1} q(\theta)f(\theta)\,d\theta \le K. \quad (A.4)$$

Proof. The transformation of the objective function into the virtual profit function is standard in the literature (see Sundararajan 2004a, b; Salanie 2005). First, the (IC) condition implies that customers will self-select the contract designed for him or her in the equilibrium. In other words, $\theta = \arg\max_t U(q(t),\theta) - p(t)$. The first-order condition of maximizing $U(q(t),\theta) - p(t)$ yields

$$U_1(q(\theta),\theta)q_1(\theta) - p_1(\theta) = 0. \quad (A.5)$$

The second-order condition leads to (see Salanie 2005, p. 30 for details)

$$U_{12}(q(\theta),\theta)q_1(\theta) \ge 0.$$

Because we have assumed the *single-crossing condition*, $U_{12}(q(\theta),\theta) \ge 0$, this inequality holds when $q_1(\theta) > 0$, which is (A.3). Next, we define the consumer surplus function by

$$s(\theta) \equiv U(q(\theta),\theta) - p(\theta). \quad (A.6)$$

Fully differentiating this term, it follows that

$$\frac{ds(\theta)}{d\theta} = U_1(q(\theta),\theta)q_1(\theta) - p_1(\theta) + U_2(q(\theta),\theta) = U_2(q(\theta),\theta),$$

where the second equality results from the necessary condition (A.5). As a result,

$$s(\theta) = s(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} U_2(q(t),t)\,dt, \quad (A.7)$$

where $\underline{\theta}$ is the consumer who feels indifferent between buying or not. Given this expression, we can transform the objective function as follows:

$$\begin{aligned}\int_{\underline{\theta}}^{1} p(\theta)f(\theta)\,d\theta &= \int_{\underline{\theta}}^{1}\Big[U(q(\theta),\theta) - s(\theta)\Big]f(\theta)\,d\theta\\ &= \int_{\underline{\theta}}^{1}\Big[U(q(\theta),\theta) - s(\underline{\theta})\\ &\qquad - \int_{\underline{\theta}}^{\theta} U_2(q(t),t)\,dt\Big]f(\theta)\,d\theta.\end{aligned}$$

The last term in the bracket can be further simplified by changing the order of integration as follows:

$$\begin{aligned}\int_{\underline{\theta}}^{1}\left[\int_{\underline{\theta}}^{\theta} U_2(q(t),t)\,dt\right]f(\theta)d\theta &= \int_{\underline{\theta}}^{1}\left[\int_{t}^{1} U_2(q(t),t)f(\theta)d\theta\right]dt\\ &= \int_{\underline{\theta}}^{1} U_2(q(t),t)\left[\int_{t}^{1} f(\theta)d\theta\right]dt\\ &= \int_{\underline{\theta}}^{1} U_2(q(t),t)[1-F(t)]\,dt\\ &= \int_{\underline{\theta}}^{1} U_2(q(\theta),\theta)[1-F(\theta)]\,d\theta.\end{aligned}$$

Substituting the last term back into the objective function yields

$$\int_{\underline{\theta}}^{1}\{U(q(\theta),\theta)-s(\underline{\theta})-U_2(q(\theta),\theta)H(\theta)\}f(\theta)\,d\theta,$$
$$\text{where}\quad H(\theta)=[1-F(\theta)]/f(\theta).$$

This term can be simplified to the final objective function because the lowest type of buyer receives zero surplus, $s(\underline{\theta})=0$. If this is not the case, the monopolist can raise the price (equivalent to lower $s(\theta)$) to all buyers by a small positive constant and earn more profit. Finally, the (IR) constraint is satisfied because $s(\theta)=s(\underline{\theta})+\int_{\underline{\theta}}^{\theta}U_2(q(t),t)\,dt=\int_{\underline{\theta}}^{\theta}U_2(q(t),t)\,dt\geq 0$, since $U_2(q(t),t)>0$ by Assumption 2. □

Next, in Lemma 2, we derive the solution to the constrained demand problem. Given any constant $\lambda>0$, we define $q(\theta,\lambda)=\max\{0,q\}$, where q is the solution of

$$U_1(q,\theta)-U_{12}(q,\theta)H(\theta)=\lambda. \tag{A.8}$$

Assuming $U_{122}(q,\theta)<0$, we can show in Lemma 2 that (A.8) is the first-order condition for the optimal quantity schedule. (A.8) is also the key equation for deriving the optimal usage schedule. This expression is exactly the same as that in the standard nonlinear pricing models in which λ is the constant variable cost. Next, we define $q^0(\theta)$ and $p^0(\theta)$ as the solution when $\lambda=0$:

$$q^0(\theta)=q(\theta,0),$$
$$p^0(\theta)=U(q^0(\theta),\theta)-\int_{\underline{\theta}}^{\theta}U_2(q^0(t),t)\,dt.$$

$(q^0(\theta),p^0(\theta))$ is referred to as the "revenue-maximization" pricing schedule, since it maximizes the monopolist's revenues with a zero marginal cost (equivalently, without any capacity constraint). $\underline{\theta}$ is an endogenous variable that determines the market coverage. Given a demand capacity constraint K, we denote the pricing schedule that solves the constrained demand pricing problem by $(q^C(\theta,K), p^C(\theta,K))$, and the lowest adopting type by $\underline{\theta}(K)$. Superscript C is used for the quantity and price schedules in terms of q and K. To minimize our future use of inline integrals, we also define the revenue-maximizing level of total demand Q^0 as $Q^0\equiv\int_0^1 q^0(\theta)f(\theta)d\theta$.

Lemma 2. (i) *If $K\geq Q^0$, then the demand constraint is nonbinding, and the monopolist chooses $q^0(\theta)$ and $p^0(\theta)$ to maximize profits in the absence of a capacity constraint:*

$$q^C(\theta,K)=q^0(\theta)\ \textit{for each}\ \theta\in[\underline{\theta}(K),1]. \tag{A.9}$$

$$\underline{\theta}(K)\ \textit{is the solution of}\ U(q^0(\theta),\theta)-U_2(q^0(\theta),\theta)H(\theta)=0. \tag{A.10}$$

(ii) *If the constraint is binding that is, if $K<Q^0$, then the monopolist chooses the pricing schedule that would be chosen if it incurred linear variable costs equal to the marginal revenue $\lambda(K)$:*

$$q^C(\theta,K)=q(\theta,\lambda(K))\ \textit{for each}\ \theta\in[\underline{\theta}(K),1]. \tag{A.11}$$

$$\begin{aligned}\underline{\theta}(K)\ \textit{is the solution of}\ &U(q^C(\theta,K),\theta)-U_2(q^C(\theta,K),\theta)H(\theta)\\ &=\lambda(K)\cdot q^C(\theta,K).\end{aligned} \tag{A.12}$$

In each case, the corresponding total price for customer type θ is

$$p^C(\theta,K)=U(q^C(\theta,K),\theta)-\int_{t=\underline{\theta}(K)}^{\theta}U_2(q^C(t,K),t)\,dt. \tag{A.13}$$

Moreover, for a given K, the contract $(q^C(\theta,K),p^C(\theta,K))$ and the lowest adopting customer type $\underline{\theta}(K)$ are uniquely specified by (A.11)–(A.13).

Proof. (i) When the constraint is not binding, this problem becomes a standard nonlinear pricing with zero marginal cost. The necessary condition is simply the derivative of the integrand in (A.1) for each type θ w.r.t. $q(\theta)$. As a result,

$$U_1(q(\theta),\theta)-U_{12}(q(\theta),\theta)H(\theta)=0,$$

which shows that $q^C(\theta,K)=q^0(\theta)$. Because $U(q^0(\theta),\theta)-U_2(q^0(\theta),\theta)H(\theta)$ is the virtual profit term,[9] the monopolist will serve all customers with positive profit, and thus $\underline{\theta}(K)$ is determined by equating the virtual profit term to zero.

(ii) Because of the additional demand capacity constraint, the above solution cannot be applied immediately. Mathematically, our optimization problem with a demand capacity constraint is documented as a standard isoperimetric problem, which confirms that we can treat this dynamic

[9] The *virtual profit for each consumer type* term is not specific to our problem and is standard in the nonlinear pricing models (e.g., Laffont and Martimort 2001, Sundararajan 2004a). The intuition behind this term is as follows: in a second-degree price discrimination model, when the seller increases $q(\theta)$, the usage schedule offered to a specific type θ, there are two effects: (1) the direct profit from that customer changes, and (2) indirectly, the monopoly has to offer more usage to higher types so that those high types will not switch to the lower type's contract. The virtual profit term includes both effects. Formally, the first term in the objective function is the utility of a buyer and represents the direct profit from each type. The second term is the rent left for buyers because of information asymmetry (also called information rent). These two terms together is referred as *virtual profit for θ*.

programming problem as a static optimization problem using the Lagrange multiplier approach:

$$\begin{aligned}\max_{\underline{\theta},q(\cdot)} L &= \int_{\theta=\underline{\theta}}^{1} [U(q(\theta),\theta) - U_2(q(\theta),\theta)H(\theta)]f(\theta)\,d\theta \\ &\quad + \lambda\left[K - \int_{\underline{\theta}}^{1} q(\theta)f(\theta)\,d\theta\right] \\ &= \int_{\theta=\underline{\theta}}^{1} [U(q(\theta),\theta) - U_2(q(\theta),\theta)H(\theta) - \lambda q(\theta)]f(\theta)\,d\theta + \lambda K\end{aligned}$$

subject to (A.2) and (A.3).

Maximizing the integrand pointwise with respect to $q(\theta)$, we can derive the necessary condition (A.8). The necessary condition (A.12) for $\underline{\theta}$ can be found in any textbook on dynamic programming (see, for example, Seierstad and Sydsæster, 1987 p. 39, Equation (41b)). (A.13) results from the total price equals to the consumer surplus minus the information rent offered to each type, as shown in the proof of Lemma 1. The optimal usage plan can be shown to satisfy $q_1^C(\theta, K) > 0$ with the assumption $U_{122} \leq 0$, as in Sundararajan (2004a). Without assuming $U_{122} \leq 0$, this problem is still tractable but involves a complicated ironing procedure (Maskin and Riley 1984). The uniqueness is more technically involved and is provided in Online Appendix B. Readers who are not interested in the technical details can skip this part without losing information. □

Theorem 1. *Let $q^*(\theta)$ and $p^*(\theta)$ be the optimal pricing schedule when the monopolist's cost function is as defined in* (3.1), *and let n^* be the corresponding optimal number of units of cost incurred by the monopolist. Then, either*

$$q^*(\theta) = q^0(\theta) \text{ and } p^*(\theta) = p^0(\theta), \tag{A.14}$$

for each θ; that is, the monopolist chooses the revenue-maximization pricing schedule, or

$$q^*(\theta) = q^C(\theta, K(n^*)) \text{ and } p^*(\theta) = p^C(\theta, K(n^*)), \tag{A.15}$$

for each θ; that is, the optimal pricing schedule is identical to the constrained demand pricing schedule with an upper bound $K(n^*)$ on demand.

Proof. Given any optimal solution $K(n^*) > Q^0$, the capacity constraint is not binding since the revenue-maximization pricing solves the problem by definition. When $K(n^*) \leq Q^0$, the demand constraint will be binding ($Q^* = K(n^*)$). Since there is no variable cost of using each block of capacity, the solution must occur at the capacity limit of n^* blocks. Therefore the solution is the one presented in Lemma 2. □

Proof of Theorem 2. Given Theorem 1, the monopolist's problem is reduced to identifying the optimal value of n^*, the number of discontinuous units of cost. A direct way of doing this is to solve Equations (A.15) for each feasible value of n, compare the corresponding profits, and choose the best one. Theorem 2 specifies how to identify n^* in a more efficient and intuitive way.

Theorem 2. (i) *There exists a unique number of units of infrastructure $i^* \geq 0$ such that*

$$\lambda(K(i^*)) \geq \frac{c(i^*)}{k(i^*)}; \tag{A.16}$$

$$\lambda(K(i^*+1)) < \frac{c(i^*+1)}{k(i^*+1)}. \tag{A.17}$$

(ii) *The value of i^* in* (A.16) *and* (A.17) *defines the optimal number of units of cost that a monopolist should incur, and consequently, the optimal pricing schedule with discontinuous costs.*

(a) *If* $\int_{K(i^*)}^{K(i^*+1)} \lambda(x)\,dx < c(i^*+1)$, *then* $n^* = i^*$.

(b) *If* $\int_{K(i^*)}^{K(i^*+1)} \lambda(x)\,dx \geq c(i^*+1)$, *then* $n^* = (i^*+1)$,

where the function $\lambda(x)$ is defined in Lemma 2.

Proof. The proof has been discussed in detail in the main text. □

*Proof of Proposition 1. Part (i) Applying the implicit function theorem on (A.8) yields

$$q_2^C(\theta, K) = \frac{dq^C(\theta,\lambda)}{d\lambda} = \frac{-1}{-[U_{11} - U_{112}H(\theta)]} < 0, \quad \forall \theta,$$

where $U_{11} < 0$ and $U_{112}H(\theta) > 0$, by assumptions.

Part (ii) We first derive the result of the total price and next the result of the unit price. We can show that the sign of the total price is ambiguous. After differentiating (A.13), we see that

$$\begin{aligned}\frac{dP^C}{dK} &= \left[U_1(q^C(\theta,K),\theta)\frac{dq^C(\theta,K)}{dK}\right. \\ &\qquad \left. - \int_{t=\underline{\theta}(K)}^{\theta} U_{12}(q^C(t,K),t)\frac{dq^C(t,K)}{dK}\,dt\right] \\ &\quad + U_2(q^C(\underline{\theta},K),\,\underline{\theta})\frac{d\underline{\theta}}{dK}.\end{aligned}$$

The last term is zero because $q^C(\underline{\theta}, K) = 0$, which is shown in Online Appendix B. Hence the sign of LHS depends on terms in the bracket. When $\theta \to \underline{\theta}$, this term is positive because the second term goes to zero while the first term is still positive. In other words, when K increases, the total payment for lower types are higher.

We show that the sign is ambiguous, in general, by showing that $dP^C/dK|_{\theta=1}$ can be negative.

$$\left.\frac{dP^C}{dK}\right|_{\theta=1} = \lambda \cdot \frac{dq^C(1,K)}{dK} - \int_{t=\underline{\theta}(K)}^{1} U_{12}(q^C(t,K),t)\frac{dq^C(t,K)}{dK}\,dt.$$

In this equation, we have replaced $U_1(q^C(\theta, K), \theta)$ by λ, which is derived from (A.8) because $H(1) = 0$. As a result, when $\lambda = 0$, the first term becomes 0 and $dP^C/dK|_{\theta=1}$ is negative because $U_{12} > 0$ and $dq^C(t, K)/dK \geq 0$. $dP^C/dK|_{\theta=1}$ can be positive because when K goes to 0, the second term goes to 0 as $\underline{\theta}(K)$ goes to 1 while the first term is still positive, and hence $dP^C/dK|_{\theta=1}$ is positive.

As for the unit price, it follows that

$$\frac{d(P^C(\theta)/q^C(\theta))}{dK} = \frac{q^C \cdot dP^C/dK - P^C \cdot dq^C/dK}{q^C(\theta)^2}.$$

The sign of the LHS depends on the numerator of the RHS, which goes to 0 when θ goes to $\underline{\theta}(K)$ since q and P^C are both zero at $\underline{\theta}(K)$. If the numerator is decreasing in θ, then RHS is negative for all θ. The last part of the proof is to show this claim. The numerator can be written as

$$P^C\frac{dq^C}{dK}\left[\frac{q^C\cdot dP^C/dK}{P^C\cdot dq^C/dK}-1\right]=P^C\frac{dq^C}{dK}\left[\frac{q^C}{P^C}\frac{dp^C}{dq^C}-1\right]$$
$$=P^C\frac{dq^C}{dK}\frac{q^C}{P^C}\left[\frac{dp^C}{dq^C}-\frac{p^C}{q^C}\right].$$

Note that $P^C((dq^C/dK)(q^C/P^C))>0$ because all three terms are positive. If $p^C(q^C)$ is a concave function, then $dp^C/dq^C-p^C/q^C$ is negative (i.e., marginal price is always less than the average price for all types). For most of the distributional assumptions, the optimized $p^C(q^C)$ is indeed a concave function (e.g., beta distribution with quadratic utility function). Moreover, if $p^C(q^C)$ is convex, that means the unit price is increasing, which will provide opportunities for buyers to arbitrage. As a result, $p^C(q^C)$ is typically assumed to be concave in the literature.

Part (iii) By $dq^C(\theta,\lambda)/d\lambda<0$ and $q(\underline{\theta})=0$, we can conclude that $d\underline{\theta}/d\lambda>0$, which is equivalent to $d\underline{\theta}/dK<0$.

Proof of Proposition 2. For ease of exposition, we define the single-period optimal number of blocks in terms of a revenue-to-cost ratio, x, as follows:

$$n(x)=\max\left\{n\colon \arg\max_n x\cdot R(K(n))-C(K(n))\right\}. \quad (A.18)$$

The max$\{\cdot\}$ function is used to pick the largest solution when there are multiple solutions. By this definition, $n(1)$ is the single-period baseline solution in §4.2. Also bear in mind that the value of $n(x)$ is an integer. Because of the discontinuous cost, $n(x)$is a left-continuous step function with a shape similar to the discontinuous cost function. The complete solution of §5.1 depends on three critical values of $n(x)$: $n(\alpha_R/\alpha_C)$, $n(1+\delta\alpha_R)$, and $n(1/(1+\delta\alpha_C))$.

Following the standard procedure, we solve the profit-maximization problem backward. There are two cases in the second period: $n_2^*=0$ or $n_2^*>0$. When $n_2^*>0$, in the second period the monopolist will choose an infrastructure level that balances the second-period marginal revenue and *marginal average cost* because there is no benefit to defer installation in the last period of this game. Mathematically, n_2^* is chosen by maximizing the following two terms:

$$\max_{n_2}\delta[\alpha_R R(K(n_1+n_2))-\alpha_C C(K(n_1+n_2))]$$
$$=\alpha_C\times\delta\times\max_{n_2}\left[\frac{\alpha_R}{\alpha_C}R(K(n_1+n_2))-C(K(n_1+n_2))\right]. \quad (A.19)$$

By definition, the solution is $n_1^*+n_2^*=n(\alpha_R/\alpha_C)$. The important implication is that as long as the first-period capacity, n_1^*, is strictly smaller than $n(\alpha_R/\alpha_C)$, the monopoly will adjust n_2^* to make up the difference in the second period. We define $n(\alpha_R/\alpha_C)$ as n^*, as in the main text.

Returning to the first-period's decision, there are also two possibilities: the monopolist expects the second-period capacity to be either $n_2^*=0$ or $n_2^*=n(\alpha_R/\alpha_C)-n_1^*$. In the first case, (5.1) can be simplified to

$$\max_{n_1}R(K(n_1))-C(K(n_1))+\delta\alpha_R R(K(n_1))$$
$$=\max_{n_1}[(1+\delta\alpha_R)R(K(n_1))-C(K(n_1))]. \quad (A.20)$$

The optimal solution of this problem is by definition $n_1^*=n(1+\delta\alpha_R)$.

In the second case, when $n_2^*>0$, since $n_1^*+n_2^*=n(\alpha_R/\alpha_C)$, n_1^* does not affect the sum of these two terms in (5.1). Therefore we only need to consider the other three terms in (5.1),

$$\max_{n_1}R(K(n_1))-C(K(n_1))+\alpha_C\delta C(K(n_1))$$
$$=(1-\alpha_C\delta)\max_{n_1}\left[\frac{1}{(1-\alpha_C\delta)}R(K(n_1))-C(K(n_1))\right]. \quad (A.21)$$

The solution of this problem is $n_1^*=n(1/(1-\alpha_C\delta))$ by definition.

The complete solution of this proposition depends on the ordering of $n(\alpha_R/\alpha_C)$, $n(1+\delta\alpha_R)$, and $n(1/(1+\delta\alpha_C))$. It can be verified that

(1) When

$$\delta\geq\frac{1}{\alpha_R}+\frac{1}{\alpha_C},\ \frac{1}{1-\delta\alpha_C}\geq 1+\delta\alpha_R\geq\frac{\alpha_R}{\alpha_C}.$$

(2) When

$$\delta<\frac{1}{\alpha_R}+\frac{1}{\alpha_C},\ \frac{1}{1-\delta\alpha_C}<1+\delta\alpha_R<\frac{\alpha_R}{\alpha_C}.$$

Because of the discontinuous cost structure, $n(x)$ is a stepwise function that is weakly increasing in x. As a result, we can conclude that when $\delta\geq 1/\alpha_R+1/\alpha_C$,

$$n\left(\frac{1}{1-\delta\alpha_C}\right)\geq n(1+\delta\alpha_R)\geq n\left(\frac{\alpha_R}{\alpha_C}\right),$$

which implies that $n_2^*=0$ and completes the proof of Proposition 2(i).

When $\delta<(1/\alpha_R)+(1/\alpha_C)$, we cannot reach the same conclusion. This is because $n(1/(1+\delta\alpha_C))$ can be equal to $n(\alpha_R/\alpha_C)$ because of the discontinuous cost structure. Our next step is to identify the condition under which $n_1^*=n(1/(1-\alpha_C\delta))<n(\alpha_R/\alpha_C)$. We know that (A.21) is globally concave by assumption. That is, (A.21) is first increasing and then decreasing in n_1 (e.g., the objective function is unimodal). It is sufficient to compare the profit at n^* and n^*-1 to decide whether n_1^* is strictly smaller than $n(\alpha_R/\alpha_C)$. Formally,

$$R(K(n^*))-C(K(n^*))+\alpha_C\delta C(K(n^*))$$
$$<R(K(n^*-1))-C(K(n^*-1))+\alpha_C\delta C(K(n^*-1)),$$

which is equivalent to

$$R(K(n^*))-R(K(n^*-1))$$
$$<(1-\delta\alpha_C)[C(K(n^*))-C(K(n^*-1))]. \quad (A.22)$$

When (A.22) holds, the monopolist's profit is higher when delaying at least one block to the second period, which implies that $n_1^*<n(\alpha_R/\alpha_C)$ and $n_2^*>0$. This result establishes Proposition 2(iii). When (A.22) does not hold, then $n_1^*\geq n(\alpha_R/\alpha_C)$ and $n_2^*=0$. This result establishes Proposition 2(ii). □

Proof of Proposition 3.

Proof of (i). We start the proof by deriving the centralized solution when the monopolist also chooses Q_i for all business units. Formally,

$$\max_{n,\,Q_i} \sum_{i=1}^{m} R_i(Q_i) - C(K(n)) \quad \text{subject to } \sum_{i=1}^{m} u_i \times Q_i = K(n).$$

This is a simple generalization of our model in §5. We start by solving the demand-constrained problem at K. The objective function is given by

$$\max_{n,\,Q_i} \sum_{i=1}^{m} R_i(Q_i) + \lambda\left(K - \sum_{i=1}^{m} u_i \times Q_i\right),$$

which is equivalent to solving each business unit's problem individually. The optimization problem for business unit i is given by

$$\max_{Q_i} R_i(Q_i) - \lambda \times u_i \times Q_i,$$

which shows that when a business unit sets the transfer pricing t at λ, the decentralized solution in Proposition 3(i) is the same as that in the centralized case. Because the profit of the centralized case is weakly better than the decentralized case, $t^* = \lambda$.

Following the same procedure in §4.2, we can determine the optimal number of blocks (n) by comparing the average cost of each block with $\lambda(K)$.

Proof of (ii). When the monopolist chooses one of the two solutions (i^* and i^*+1) in (A.16) and (A.17), if the optimal solution is i^*, then

$$t^* = \lambda(K(i^*)) \geq \frac{c(i^*)}{k(i^*)} > \frac{C(K(i^*))}{K(i^*)},$$

where the second inequality results from the discontinuously convex cost. In other words, we can conclude that t^* is greater than the average cost, and hence the total transfer pricing is greater than total infrastructure cost. If the optimal solution is i^*+1, then

$$t^* = \lambda(K(i^*+1)) < \frac{c(i^*+1)}{k(i^*+1)}.$$

By the continuously convexity, we can conclude that

$$\frac{c(i^*+1)}{k(i^*+1)} > \frac{C(K(i^*+1))}{K(i^*+1)}.$$

In general, $\lambda(K(i^*+1))$ can be larger or smaller than the average cost. When the block size approaches zero, the marginal revenue can be arbitrarily close to the last block's average cost. As a result,

$$\frac{C(K(i^*+1))}{K(i^*+1)} < \lambda(K(i^*+1)) < \frac{c(i^*+1)}{k(i^*+1)}.$$

We show that the marginal revenue can be smaller than the average cost by an example. Suppose $\lambda(K) = 1 - K$, $k(1) = k(2) = 0.5$, $c(1) = 0.10$, $c(2) = 0.12$. We only need to consider $i^* = 1$ or 2. The revenues of two cases are $R(k(1)) = 0.375$ and $R(k(2)) = 0.5$, respectively. The profit of two cases are $0.375 - 0.1 < 0.5 - 0.1 - 0.12$, and the optimal solution is $i^* = 2$. The marginal revenue and transfer pricing at $i^* = 2$ is 0, which is clearly smaller than the average cost, 0.22.

References

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

Bakos, Y., E. Brynjolfsson. 1999. Bundling information goods: Pricing, profits, and efficiency. *Management Sci.* **45**(12) 1613–1630.

Bakos, Y., E. Brynjolfsson. 2000. Bundling and competition on the Internet. *Marketing Sci.* **19**(1) 63–82.

Braid, R. M. 1989. Uniform versus peak-load pricing of a bottleneck with elastic demand. *J. Urban Econom.* **26**(3) 320–327.

Braid, R. M. 1996. Peak-load pricing of a transportation route with an unpriced substitute. *J. Urban Econom.* **40**(2) 179–197.

Calzada, J. 2007. Capacity-based versus time-based access charges in telecommunications. *J. Regulatory Econom.* **32**(2) 153–172.

Chen, Y. J., S. Seshadri. 2007. Product development and pricing strategy for information goods under heterogeneous outside opportunities. *Inform. Systems Res.* **18**(2) 150–172.

Cornelli, F. 1996. Optimal selling procedures with fixed costs. *J. Econom. Theory* **71**(1) 1–30.

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

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

Fischer, R., P. Serra. 2003. Energy prices in the presence of plant indivisibilities. *Energy Econom.* **25**(4) 303–314.

Hitt, L. M., P. Y. Chen. 2005. Bundling with customer self-selection: A simple approach to bundling low-marginal-cost goods. *Management Sci.* **51**(10) 1481–1493.

Hufnagel, E. M., J. G. Birnberg. 1989. Perceived chargeback system fairness in decentralized organizations: An examination of the issues. *MIS Quart.* **13**(4) 415–429.

Jain, S., P. K. Kannan. 2002. Pricing of information products on online servers: Issues, models, and analysis. *Management Sci.* **48**(9) 1123–1142.

Konana, P., A. Gupta, A. Whinston. 2000. Integrating user preferences and real-time workload in electronic commerce. *Inform. Systems Res.* **11**(2) 177–196.

Laffont, J. J., D. Martimort. 2001. *The Theory of Incentives: The Principal-Agent Model.* Princeton University Press, Princeton, NJ.

Maskin, E., J. Riley. 1984. Monopoly with incomplete information. *RAND J. Econom.* **15**(2) 171–196.

Masuda, Y., S. Whang. 2006. On the optimality of fixed-up-to tariff for telecommunications service. *Inform. Systems Res.* **17**(3) 247–253.

McKinnon, W. P., E. Kallman. 1987. Mapping chargeback systems to organizational environments. *MIS Quart.* **11**(5) 5–20.

Mendelson, H. 1985. Pricing computer services: Queuing effects. *Comm. ACM* **28**(3) 312–321.

Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the m/m/1 queue. *Oper. Res.* **38**(5) 870–883.

Monteiro, P. K., F. H. Page. 1996. Non-linear pricing with a general cost function. *Econom. Lett.* **52**(3) 287–291.

Mussa, M., S. Rosen. 1978. Monopoly and product quality. *J. Econom. Theory* **18**(2) 301–317.

Nadiminti, R., T. Mukhopadhyay, C. Kriebel. 2002. Research report: Intrafirm resource allocation with asymmetric information and negative externalities. *Inform. Systems Res.* **13**(4) 428–434.

Oren, S., S. Smith, R. Wilson. 1985. Capacity pricing. *Econometrica* **53**(3) 545–567.

Ross, J. W., M. R. Vitale, C. M. Beath. 1999. The untapped potential of IT chargeback. *MIS Quart.* **23**(2) 215–237.

Salanie, B. 2005. *The Economics of Contracts: A Primer*, 2nd ed. MIT Press, Cambridge, MA.

Seierstad, A., K. Sydsæster. 1987. *Optimal Control with Economic Applications*. North-Holland, Amsterdam.

Spulber, D. 1992. Optimal nonlinear pricing and contingent contracts. *Internat. Econom. Rev.* **33**(4) 747–772.

Spulber, D. 1993a. Monopoly pricing of capacity usage under asymmetric information. *J. Indust. Econom.* **41**(3) 241–257.

Spulber, D. 1993b. Monopoly pricing. *J. Econom. Theory* **59**(1) 222–234.

Stole, L. 2007. Price discrimination and competition. M. Armstrong, R. H. Porter, eds. *Handbook of Industrial Organization*, Vol. 3. North Holland, Amsterdam, 2221–2299.

Sundararajan, A. 2004a. Nonlinear pricing of information goods. *Management Sci.* **50**(12) 1660–1673.

Sundararajan, A. 2004b. Managing digital piracy: Pricing and protection. *Inform. Systems Res.* **15**(3) 287–308.

Thomas, L. 2001. Cost functions in non-linear pricing. *Econom. Lett.* **72**(1) 53–59.

Varian, H. 1989. Price discrimination. R. Schmalensee, R. Willig, eds. *Handbook of Industial Organization*, Vol. 1. North Holland, Amsterdam, 597–654.

Wilson, R. 1993. *Nonlinear Pricing*. Oxford University Press, Oxford, UK.