NET Institute*

www.NETinst.org

Working Paper #06-11

September 2006

**Pricing Digital Goods: Discontinuous Costs and Shared Infrastructure**

Ke-Wei Huang and Arun Sundararajan
Leonard N. Stern School of Business, New York University

# Pricing Digital Goods: Discontinuous Costs and Shared Infrastructure[1]

**Ke-Wei Huang and Arun Sundararajan**

Leonard N. Stern School of Business, New York University
44 West 4th Street, New York, NY 10012
*khuang0@stern.nyu.edu, asundara@stern.nyu.edu*

September 30, 2006

**Abstract:** We develop and analyze a model of pricing for digital products with discontinuous supply functions. This 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. Examples include internet service, telephony, online trading, on-demand software, digital music, streamed video-on-demand and grid computing. These goods are often modeled as information goods with variable costs of zero, although their actual cost structure features a mixture of positive periodic fixed costs, and zero marginal costs. The pricing of such goods is further complicated by the fact that rapid advances in semiconductor and networking technology lead to sustained rapid declines in the cost of new infrastructure over time. Furthermore, this infrastructure is often shared across multiple goods and services in distinct markets.

The main contribution of this paper is a general solution for the optimal nonlinear pricing of such digital goods and services. We show that this can be formulated as a finite series of more conventional constrained pricing problems. We then establish that the optimal nonlinear pricing schedule with discontinuous supply functions coincides with the solution to one specific constrained problem, reduce the former to a problem of identifying the optimal number of "blocks" of demand that the seller will fulfil under their optimal pricing schedule, and show how to identify this optimal number using a simple and intuitive rule (which is analogous to "balancing" the marginal revenue with average "marginal cost"). We discuss the extent to which using "information-goods" pricing schedules rather than those that are optimal reduce profits for sellers of digital goods. A first extension includes the rapidly declining infrastructure costs associated with Moore's Law to provide insight into the relationship between the magnitude of cost declines, infrastructure planning and pricing strategy. A second extension examines multi-market pricing of a set of digital goods and services whose supply is fulfilled by a shared infrastructure.

Our paper provides a new pricing model which is widely applicable to IT, network and electronic commerce products. It also makes an independent contribution to the theory of second-degree price discrimination, by providing the first solution of monopoly screening when costs are discontinuous, and when costs incurred can only be associated with the total demand fulfilled, rather than demand from individual customers.

Keywords: digital goods, price discrimination, nonlinear pricing, screening, discontinuous costs, shared infrastructure, Moore's Law

JEL Codes: D41, D82, L1

[1]We thank seminar participants at New York University for their feedback. Financial support from the NET Institute (http://www.netinst.org/) is gratefully acknowledged.

## 1. Introduction

This paper studies nonlinear pricing for a digital product whose supply function is discontinuous and varies significantly over time. This is 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 costs a fixed amount, and enables a seller to fulfill a fixed additional level of demand at zero marginal cost. The latter feature often leads to these products being modeled as "information goods". We present the first general formulation of the monopoly nonlinear pricing problem with discontinuous costs of this kind, and analyze how its solution differs from the problem of pricing zero-marginal-cost information goods. We examine how this optimal pricing policy is sensitive to (1) *multi-period* pricing when rapid technological progress reduces these costs over time, (2) *multi-market* pricing for differentiated digital goods whose supply depends on a shared IT infrastructure. Our research will thus extend a standard model of second-degree price discrimination to accommodate key (and previously unexplored) aspects of cost and supply that are unique to IT and network-based digital products.

Our study is most easily motivated by some examples. Given a fixed level of infrastructure, the marginal cost of providing an additional unit of Internet service is typically zero. However, as the total quantity of service supplied by the ISP increases, the ISP need to add blocks of infrastructure: modems to their modem pool, and equipment/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. An application service provider who offers access to hosted software incurs costs towards installing new servers and buying Internet bandwidth; these costs are incurred in blocks, each instance of a "fixed" cost enables the fulfilling of a fixed amount of

additional demand at no marginal cost. The cost structure for a provider of streaming media, or of online trading is similar – costs towards 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 time, at no additional marginal cost.[2].

In each of these examples, the seller faces consumers who demand variable (and individually varying) levels of usage, the seller's cost function is non-decreasing in the total demand it fulfils, with periodic discontinuous increases as additional blocks are added, and the seller incurs low or no variable costs for fulfilling additional demand between these "jumps". This is the cost structure that underlies the supply function in our model. Judging by the examples described above, it appears to be applicable to an increasing number of digital goods and services. Furthermore, it differs from the standard models of supply used in price-discrimination models in three specific ways. First, it models costs that are somewhere in between variable and fixed costs. Traditional variable costs are modeled as being incurred in the short-run, varying continuously with the level of demand a seller fulfils, and often having a significant impact on pricing policy. Traditional fixed costs, while discontinuous, are modeled as being long-run costs that are incurred very infrequently — for instance, the costs associated with building a factory — and whose magnitude does not directly affect a seller's short-run choice of pricing. In each of the examples we describe above, neither of these traditional components of costs capture the actual supply of the good; in contrast, the relevant costs appear to be similar to fixed costs in that they increase the "capacity" of the seller across periods, but similar to variable costs in

[2]Other examples might include local phone or cell phone service providers, video-on-demand for a digital cable provider, Internet caching services, collocation, shared grid computing, shared data storage, and rendering server farms for digital animation; this set of examples is likely to increase as "on-demand computing" and "apps-on-tap" models are more widely adopted.

There are examples of IT-enabled "non-digital" goods that share this cost structure as well – 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.

that they are incurred in the short-run, and vary in (possibly large) steps based on the demand a seller faces in each period.

Our research can also be distinguished from the existing pricing literature in other important ways. Standard models of nonlinear pricing associate a variable cost function with the demand of *each* customer, rather than with the *total* demand fulfilled. This makes no difference when variable costs are continuous and linear (or zero, for that matter). However, the distinction is important when a specific cost increase cannot generically be associated with each unit demand increase – an implicit assumption that underlies the specification of any standard cost function. Rather, when a collective increase in demand of a group of customers results in periodic jumps in the cost of supply, as is the case with the digital products described in the examples above, this changes the formulation of the problem of designing optimal second-degree price discrimination in a significant way. Furthermore, technological progress changes the cost function of sellers of these goods rapidly over time, and it is not clear how this kind of trend, which is unique to IT infrastructure (and empirically rendered in what is widely referred to as Moore's Law), will affect the design of pricing. These IT infrastructures are of increasing power and ubiquity, leading sellers to base the supply of different digital goods on common large-scale infrastructures of this kind. The Internet backbone (Economides, 2005) is an ubiquitous example of such a shared infrastructure, but there are company specific ones as well. For example, Google's search, mail, news and maps each rely on the same massive on-demand computing infrastructure, while providing different products in distinct markets.

Our analysis proceeds as follows. Starting with a single period problem in section 3, we first derive the optimal nonlinear pricing schedule for a monopolist who can fulfill demand up to a pre-specified level of demand at zero marginal cost. We show that the solution to

this problem (which we term the *constrained demand* problem) coincides with the solution to an unconstrained problem with positive linear marginal costs, a problem well studied in the literature. The demand constraint of this problem generates a function which measures the marginal revenue from increasing the seller's ability to fulfill demand. Next, we show that the optimal nonlinear pricing schedule with discontinuous costs coincides with the solution to a specific instance of the constrained demand problem, and reduce the former to a problem of identifying the optimal number $n^*$ of "blocks" of demand (which may be of differing sizes) that the seller will fulfill under their optimal pricing schedule. We then show how to identify $n^*$ using a simple and intuitive rule (which is analogous to "balancing" the marginal revenue with average "marginal cost") that identifies the appropriate level of average cost to operate at. The pricing schedule that solves the corresponding constrained demand problem is optimal for a seller who faces discontinuous costs. We discuss some managerial implications of this result, provide an example that highlights how pricing and profits can vary significantly relative to those suggested by a model with zero marginal costs, and describe how changes in some key model parameters affect the optimal pricing schedule.

In section 4, we extend this analysis in two ways. First, we examine infrastructure costs that remain discontinuous but decline over time. We show that when a seller faces a finite horizon of product viability, pricing may not be affected by anticipated cost declines up to a point; this result highlights one implication of the "capacity" nature of the cost function of the digital products we model. However, if the anticipated drop costs is significant enough, it results in a decline in average price, but often increases the total price paid by each participating buyer. Next, we discuss the extension in which the seller provides multiple products based on a common infrastructure. The solution to this pricing problem also coincides with that

of an unconstrained problem with linear marginal costs. Furthermore, our results indicate that the shadow linear marginal cost of each product is proportional to its rate of utilization of the underlying infrastructure. In other words, the more a product utilizes the underlying infrastructure, the higher its optimal price should be.

Our paper adds to a growing literature on the optimal pricing of digital goods. This body of research has shown, among other things, that large-scale pure bundling can increase a monopolist's profits (Bakos and Brynjolfsson, 1999) so long as the value each customer places on different goods in the bundle does not vary too much (Geng, Stinchcombe and Whinston, 2005), that product versioning may not be an optimal pricing strategy (Bhargava and Choudhary, 2001), and that fixed-fee pricing can increase the profits from pure second-degree price discrimination (Sundararajan, 2004). Each of these papers assumes that digital goods are "information goods" with zero marginal costs, an aspect that distinguishes their analyses from ours, and also suggests a significant direction for future research. Our model differs significantly from those studied in the literature on nonlinear capacity pricing or "peak-load" pricing (for instance, in Oren, Smith and Wilson, 1985, Wilson 1993), since their focus is on capacity planning to account for short-run variability in demand, and pricing that optimally controls this variation.

Some of the literature on pricing of information systems with queuing effects addresses issues related to ours. For example, Mendelson (1985) highlights the differences between optimal short-run and long-run transfer pricing an IT-based system whose supply is subject to capacity constraints In particular, when capacity can be varied in the long-run, he shows that the optimal adjustment of pricing to account for queuing externalities depends only on relative utilization, and is independent of the specific queuing characteristics of the system. Many subsequent

papers have extended Mendelson's central results on queuing effects; these include Mendelson and Whang (1990) who demonstrate that priority pricing for a constrained resource can be expressed in a simple form: a base price for the lowest class plus an increasing priority surcharge. Dewan and Mendelson's (1990) analysis of congestion pricing with general delay costs establishes a relationship between expected delay costs and the marginal cost of capacity, a connection our analysis explores indirectly as well. Konana, Gupta and Whinston (2000) show that dynamic priority pricing that charges a congestion premium for accessing real-time database outperforms a variety of standard priority rules; this result is established both analytically, and in a more detailed simulation that takes the operational intricacies of real-time databases into account. Westland (1992) recognizes the presence of both positive and negative demand externalities. More recently, Nadiminti, Mukhopadhyay, and Kriebel (2002) extend Mendelson (1985) to admit asymmetric information about user preferences, and among other things, highlight the optimality of volume discounted (nonlinear) pricing. The model of Afeche and Mendelson (2004) allows delay costs and consumer value to be interdependent (rather than additive): their result describing the deviation of the revenue maximizing solution with the one that maximizes welfare is related to the contrast we draw between revenue maximizing (under the "information goods" assumption) and the actual profit maximizing pricing policy[3].

Some of these papers model uniform pricing mechanism while others model second degree price discrimination based on priority (or quality) – in contrast, our baseline results model second degree price discrimination based on *quantity* for a demand-constrained monopolist – while both latter sets of models have related mechanism design problems, our constraint on demand

[3]Other notable papers in this literature include Ha's (1998) analysis of joint production between customers and servers, So and Song (1998) who examine the effect of delivery time guarantees on capacity choices, Stidham's (1992) extension of Dewan and Mendelson (1990), and the model of pricing systems of distributed congestible resources (specifically, the Internet) of Gupta et al. (1997).

has not been explored thus far, and has a different impact on the optimization problem. By abstracting away from delay costs (which form the basis for unknown customer heterogeneity in most of this literature), we can admit a more general specification of heterogeneity in consumer value, which is appealing, since delay costs are not the central basis for choice in many of our examples. In a sense, one might approximately account for queuing effects into our model if one thinks of a seller who has a target (exogenous) quality level that it aims to achieve, and incorporates the effect of congestion by adjusting the effective additional demand $k(i)$ that each "unit" of supply can achieve while still maintaining this quality level. The fact that both $k(i)$ and $c(i)$ are independently defined for each unit makes this possible. 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 under-researched) aspects of pricing IT-based products that our model aims to highlight.

## 2. Overview of model

A monopolist 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 seller to fulfill, and $c(i)$ is the variable cost of the $i^{th}$ unit of infrastructure[4]. Therefore, the cost of supplying

[4]In general, the seller deploys a fixed level of infrastructure, represented by the vector $K = (k_1, k_2, ..., k_n)$. 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 scalar rather than a vector since here we only need the constrained demand and the associated fixed cost for that chunk of demand capacity.

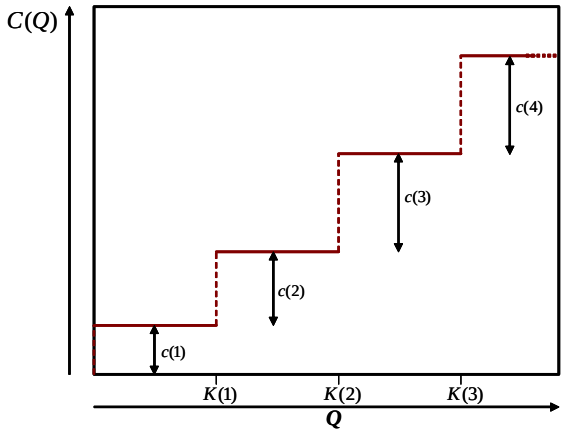


Figure 2.1: Illustrates the shape of the cost function $C(Q)$.

a total quantity $Q$ is specified by the function

$$C(Q) = \sum_{i=1}^{n(Q)} c(i), \tag{2.1}$$

where $n(Q)$ is the minimum units of infrastructure that provide the ability to fulfill demand[5] of at least $Q$, and thus $n(Q) = \min\{j : K(j) \geq Q\}$. To examine the average cost of each unit of infrastructure, we define: $k(i) \equiv K(i) - K(i-1)$ and we set $c(0)/k(0) \equiv 0$. The latter assumption is for future analytical brevity, so that we do not have to analyze the seller using just one unit of supply as a special case. The cost function, illustrated in Figure 2.1, is therefore a step function with discontinuous increases at specific integer value of $Q$, and with potentially different jumps at each of these values[6].

We assume that the average cost of the seller $c(i)/k(i)$ is non-decreasing in $i$, or loosely

[5]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 $Q(K)$.

[6]Often, there is a fixed cost $F$ associated with setting up the ability to fulfill demand; equivalently, $c(1)$ may be substantially higher than succeeding values of $c(i)$. A generalization of our model that incorporates this kind of setup cost is straightforward.

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 or distribution is based on an IT system that resembles a multi-server 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 preferences of a customer of type $\theta$ are represented by the function

$$w(q, \theta, p) = U(q, \theta) - p, \tag{2.2}$$

where $q$ is the quantity of the product used and $p$ is the total price paid by the customer. Our formulation of preferences follows the standard general model of nonlinear pricing (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 $\theta \in [0, 1]$:

1. Increasing and concave value: $U(0, \theta) = 0; U_1(q, \theta) \geq 0, U_{11}(q, \theta) < 0$ for all $q$.
2. Higher customer types get higher utility: $U(q, 0) = 0$ and $U_2(q, \theta) > 0$ for all $q > 0$.
3. These increases in utility with type are diminishing: $U_{22}(q, \theta) \leq 0$
4. Spence-Mirrlees single-crossing condition: $U_{12}(q, \theta) > 0$ for all $q$.

5. Non-increasing absolute risk-aversion: $\frac{\partial}{\partial \theta}\left[\frac{-U_{11}(q,\theta)}{U_1(q,\theta)}\right] \leq 0$.

Through the paper, numbered subscripts of functions represent derivatives with respect to the corresponding variable. Assumptions (1), (2), (4) and (5) are standard in models of second-degree price discrimination. A discussion of their implications can be found, for instance, in section 2.1 of Sundararajan (2004). While assumption (3) is made for mathematical reasons, it seems like a reasonable description of preferences.

The sequence and information structure of the model is as follows. The seller does not observe the type of any customer, but knows $F(\theta)$, the probability distribution of types in the customer population[7], which is assumed to be absolutely continuous (and thereby has a density function $f(\theta)$ which is non-zero and finite), and to have a non-increasing inverse hazard rate $H(\theta)$, where $H(\theta) \equiv \frac{1-F(\theta)}{f(\theta)}$. The seller prices its product by announcing a pricing schedule that assigns a specific total payment for each level of usage $q$. Since the seller cannot explicitly distinguish between customer types prior to contracting, the entire schedule (menu of quantity-price pairs) must be available to all customers. The revelation principle ensures that the seller can restrict its attention to direct mechanisms, under which one specific quantity-price pair is designed for each customer type, and 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 two standard constraints:

[IC]: For each $\theta$, $U(q(\theta),\theta) - p(\theta) \geq U(q(t),\theta) - p(t)$, for all $t \in [0,1]$.

[7]The interpretation of $F(\theta)$ is slightly different from the ordinary screening model due to the capacity constraint. One interpretation is that there is no uncertainty of $\theta$ but the values of $\theta$ is unknown to the seller. The other (less rigorous) interpretation is that $\theta$ is random but the seller faces a very large number of buyers and thus the total demand is almost always the same as that in the first case. As a consequence, our model can still describe the market in the second case.

[IR]: For each $\theta$, $U(q(\theta), \theta) - p(\theta) \geq 0$.

When the menu of quantity-price pairs satisfies (IC) and (IR), every customer of type $\theta$ will choose the pair $q(\theta), p(\theta)$. A schedule satisfying these constraints is simply referred to as *incentive-compatible*. An incentive-compatible schedule is said to be *optimal* if it yields profits that are at least as high as any other incentive-compatible schedule.

The sequence of events is as follows: the seller designs and announces the schedule $q(t), p(t)$, customers make their purchasing choices, and each party receives its payoff. For an incentive-compatible schedule $q(t), p(t)$, the cumulative quantity demanded by all customers will be $\int_{\theta=0}^{1} q(\theta) f(\theta) d\theta$, which means that the seller's profit is

$$\int_0^1 p(\theta) f(\theta) d\theta - C\left(\int_0^1 q(\theta) f(\theta) d\theta\right), \tag{2.3}$$

and the seller consequently aims to design the incentive-compatible schedule that maximizes (2.3) subject to [IC] and [IR].

## 3. Pricing for a single period

This section presents the optimal nonlinear pricing schedule of a seller who incurs discontinuous costs, and who prices for a single-period. The sequence of analysis leading to the section's main result is summarized in Table 3.1.

### 3.1. A preliminary result: Pricing with demand constraints

This section derives the structure of the optimal pricing schedule when the seller incurs no fixed or variable costs but faces a constraint on its capacity to fulfill demand. While this sub-

| |
|---|
| (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 seller incurring its first $n^*$ "units" of cost, then its 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 Proposition 1 and Lemma 2. |

Table 3.1: Brief summary of the sequence of analysis in this section

problem may seem tangential to the main problem described thus far, it is important because we subsequently show that the design of the optimal pricing schedule (and its comparative statics) for both the static and dynamic versions of the main problem can be characterized in terms of the results of this subsection[8].

The sub-problem in this section assumes that the seller has a fixed upper bound on the demand it can fulfill, (equivalently, its *capacity*), denoted by $K$. The seller cannot increase $K$, and incurs no costs for supplying any quantity $Q \leq K$. The pricing problem solved in this section is therefore:

$$\max_{q(.),p(.)} \quad \int_0^1 p(\theta) f(\theta) d\theta \tag{3.1}$$

$$\text{subject to} \quad \text{[IC], [IR], and}$$

$$q(\theta) \geq 0 \ \forall \ \theta \in [0,1] \tag{3.2}$$

$$\textstyle\int_0^1 q(\theta) f(\theta) d\theta \leq K. \tag{3.3}$$

[8]This subsection may be somewhat mathematically detailed for some readers. Structuring the paper so that some of these details were relegated to the appendix. led to exposition problems and lack of transparency in subsequent sections, since many key intermediate functions necessary for a clear exposition of the paper's main results are defined in this subsection.

This problem is referred to as the *constrained demand pricing problem.*

**Lemma 1.** *An equivalent formulation of the pricing problem in (3.1)-(3.3) is:*

$$\max_{\underline{\theta},q(.)} \quad \int_{\theta=\underline{\theta}}^{1} [U(q(\theta),\theta) - U_2(q(\theta),\theta)H(\theta)]\, f(\theta)d\theta \tag{3.4}$$

$$\textit{subject to} \quad q(\theta) \geq 0 \;\forall\; \theta \in [\underline{\theta}, 1] \tag{3.5}$$

$$q_1(\theta) > 0 \;\forall\; \theta \in [\underline{\theta}, 1] \tag{3.6}$$

$$\textstyle\int_{\underline{\theta}}^{1} q(\theta) f(\theta) d\theta \leq K \tag{3.7}$$

Unless otherwise specified, all proofs are relegated to Appendix A. Lemma 1 transforms the constrained demand pricing problem in two ways. The first is a standard transformation of the objective function (3.1) into the *virtual profit* function (3.4) that internalizes the incentive compatibility constraints while adding an additional set of monotonicity constraint (3.6). Additionally, we reformulate this transformed problem so that the seller directly chooses the fraction of customer types $(\underline{\theta}, 1]$ who will purchase positive quantities, and chooses the appropriate quantity each of these customer types is induced to purchase. This reformulation is useful because its solution provides a direct way of examining how market coverage (that is, the subset $[\underline{\theta}, 1]$ of customers who purchase) varies with changes in $K$.

The solution to the constrained demand problem is derived in Lemma 2, which uses the expression defined in equation (3.8). Given any constant $\lambda > 0$, define $q(\theta, \lambda)$ as follows:

$$q(\theta,\lambda) = \max\{0, q\}, \text{ where } q \text{ is the solution of } U_1(q,\theta) - U_{12}(q,\theta)H(\theta) = \lambda, \tag{3.8}$$

This equation is simply the F.O.C. of the integrand of (3.4) with a Lagrange multiplier $\lambda$

associated with (3.7). Intuitively, if the seller had no demand constraint, and instead incurred linear variable costs at the rate $\lambda$, then $q(\theta, \lambda)$ is the optimal quantity that the seller would induce customers of type $\theta$ to consume. Next, define $q^0(\theta)$ and $p^0(\theta)$, the solution when $\lambda = 0$, as follows:

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

Together, $q^0(\theta)$ and $p^0(\theta)$ are referred to as the *revenue maximizing* pricing schedule, since it maximizes the seller's revenues with a zero marginal cost (equivalently, without any constraint). The *marginal revenue function* (which is the shadow value of the demand constraint, more on this following Lemma 2) defined as:

$$\lambda(K) \text{ is the solution of } \left[\int_{\theta=0}^{1} q(\theta, \lambda) f(\theta) d\theta\right] = K. \tag{3.9}$$

Given a demand constraint $K$, denote the pricing schedule that solves the constrained demand pricing problem as $q^C(\theta, K), p^C(\theta, K)$, and the lowest adopting type as $\underline{\theta}(K)$. Here, we use the superscript $C$ for the quality and price schedules in terms of $\theta$ and $K$. To minimize our future use of in-line integrals, 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.** *(a) If $K \geq Q^0$, then the demand constraint is non-binding, and therefore the seller chooses $q^0(\theta)$ and $p^0(\theta)$ to maximize its profits in the absence of a capacity constraint:*

$$q^C(\theta, K) = q^0(\theta) \text{ for each } \theta \in [\underline{\theta}(K), 1] \tag{3.10}$$

$$\underline{\theta}(K) \text{ is the solution of } U(q^0(\theta), \theta) - U_2(q^0(\theta), \theta) H(\theta) = 0 \tag{3.11}$$

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

$$q^C(\theta, K) = q(\theta, \lambda(K)) \text{ for each } \theta \in [\underline{\theta}(K), 1] \tag{3.12}$$

$$\underline{\theta}(K) \text{ is the solution of } U(q^C(\theta, K), \theta) - U_2(q^C(\theta, K), \theta)H(\theta) = \lambda(K) \cdot q^C(\theta, K). \tag{3.13}$$

*In each case, the corresponding total price for customer type $\theta$ is*

$$p^C(\theta, K) = U(q^C(\theta, K), \theta) - \int_{t=\underline{\theta}(K)}^{\theta} U_2(q^C(t, K), t)dt. \tag{3.14}$$

*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 (3.12)-(3.14)*

The definition of $\lambda(K)$ as the marginal revenue function follows from Lemma 2: it is the value of the Lagrangian multiplier of the demand constraint when the upper bound on demand $K$. It therefore measures the marginal increase in revenue with demand: that is, the marginal increase in the value of (3.1) at its maximizing values of $p(.)$ and $q(.)$ for a marginal increase in the demand the seller is able to fulfill, after the seller adjusts its pricing function in response to the relaxation of the constraint

A corollary of this lemma indicates how pricing, demand and fraction of customers who purchase vary with changes in $K$.

**Corollary 1.** *When $K < Q^0$, we have the following results of comparative statics analysis:*
*(a) $q_2^C(\theta, K) > 0$, $d\left[p^C(\theta, K)/q^C(\theta, K)\right]/dK < 0$: relaxing the demand constraint induces an increase in total consumption (and decrease in average price) for all customers.*

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

Corollary 1 indicates that any binding constraint on the seller's ability to fulfill demand always reduces *both* the fraction of customers who purchase a positive quantity, and the quantity purchased by each of these customers. Moreover, as $K$ increases and the seller is less constrained, the fraction of adopting customers and the total usage induced from each of these customers both increase. This result is quite intuitive: when the demand constraint is relaxed, the seller can sell more products, which will lead to lower average product prices. Since all types of consumers will consume more ($q_2^C(\theta, K) > 0$), it is straightforward that more consumers may be served when $K$ is larger. The result, $q_2^C(\theta, K) > 0$, is independently interesting from a managerial perspective, for the following reason: an alternative response would be for the seller to focus ONLY on the high-end of the market by inducing as much consumption as possible from a fraction of high-valuation customers while shutting out customers with lower willingness to pay. This seems intuitively consistent with getting as much value as possible from one's allowed total demand. The corollary shows that while intuitively appealing, this is never a profit-maximizing strategy. The reason is that although low-end consumers are less valuable, they may still have relatively higher marginal willingness-to-pay when they have very low consumption compared with that of high-end consumers with high consumption.

### 3.2. Optimal pricing with discontinuous costs

In this section, we return to the main problem described in Section 2, and present two of the paper's main results. First, Theorem 1 relates the solution of the main problem to the result of Lemma 2, thereby reducing the seller's problem to one of simply identifying the optimal number of units of supply. Next, under the assumption that $c(i)/k(i)$ is non-decreasing, Theorem

2 provides a simple way of identifying this optimal number, by comparing their incremental revenue to their incremental cost. The optimal pricing schedule follows immediately from Lemma 2 and Theorem 1. We do not have general results when $c(i)/k(i)$ is non-monotonic or decreasing in $i$, and the reasons will be discussed towards the end of the section.

Our first theorem relates the solution of the main problem to the solution of the constrained demand pricing problem solved in Section 3.1

**Theorem 1.** *Let $q^*(\theta)$ and $p^*(\theta)$ be the optimal pricing schedule when the seller's cost function is as defined in (2.1), and let $n^*$ be the corresponding optimal number of units of cost incurred by the seller. Then, either:*

$$q^*(\theta) = q^0(\theta) \text{ and } p^*(\theta) = p^0(\theta), \tag{3.15}$$

*for each $\theta$, that is, the seller chooses the revenue-maximizing pricing schedule, or*

$$q^*(\theta) = q^C(\theta, K(n^*)) \text{ and } p^*(\theta) = p^C(\theta, K(n^*)), \tag{3.16}$$

*for each $\theta$, 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 $K(n^*) > Q^0(\theta)$, revenue-maximizing pricing solves the problem. Otherwise, $(q^0(\theta), p^0(\theta))$ is not feasible and cannot be the solution. When $K(n^*) \leq Q^0(\theta)$, the demand constraint will be binding ($Q^* = K(n^*)$) at some $n^*$ and therefore the solution is the one presented in Lemma 2. ■

Given the result of theorem 1, the seller's problem has now been reduced to identifying the

optimal value of $n^*$, the number of discontinuous units of cost the seller should optimally incur. A direct way of doing this is to solve equations (3.16) for each feasible value of $n$, compare the corresponding profits, and choose the best one. Our next result specifies how to identify $n^*$ in a more efficient and intuitive way.

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

$$\lambda(K(i^*)) \geq \frac{c(i^*)}{k(i^*)}; \tag{3.17}$$

$$\lambda(K(i^*+1)) < \frac{c(i^*+1)}{k(i^*+1)}. \tag{3.18}$$

*(b) The value of $i^*$ in (3.17-3.18) defines the optimal number of units of cost a seller should incur, and consequently, the optimal pricing schedule with discontinuous costs.*

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

*(ii) 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.*

The result of Theorem 2(a) is illustrated in Figure 3.1. Marginal revenue decreases as the seller'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 seller gets no further revenue from increasing its ability to fulfill demand beyond $Q^0$, the revenue-maximizing level of demand, $\lambda(K)$ is zero for $K > Q^0$. Therefore, there are always two successive values of $i$ such that the marginal revenue exceeds the average cost at the former, and the the average cost exceeds the marginal revenue at the latter.

The result of Theorem 2(b) is more subtle. The area under the marginal revenue curve over each unit of cost represents the actual additional revenue the seller can get by incurring this

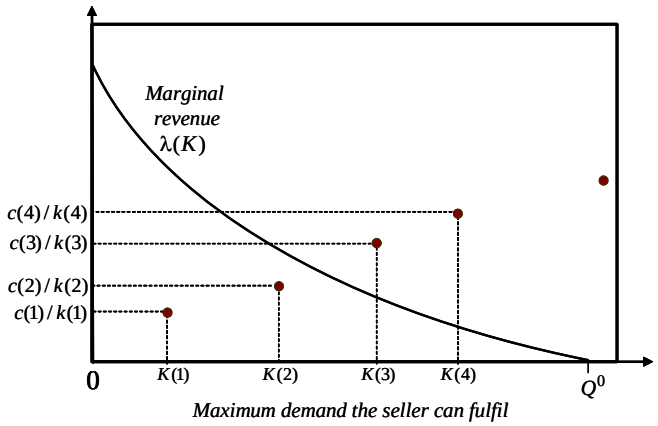


Figure 3.1: Illustrates part (a) of Proposition 2. The solid downward sloping curve is the marginal revenue $\lambda$ of capacity $K$, while the series of upward sloping points are successive values of the average cost of capacity $c(i)/k(i)$. As illustrated, $i^* = 2$, since $\lambda(K(2)) > c(2)/k(2)$ and $\lambda(K(3)) < c(3)/k(3)$.

unit of cost. This incremental revenue stems from the optimal changes in its pricing schedule that reflect the seller's ability to fulfill $k(i+1)$ additional units of demand, and the resulting changes in realized demand as discussed in Section 3.1, this is because the function $\lambda(K)$ is the value of the multiplier of the demand constraint in an optimization problem whose objective function is the seller's revenue. Figure 3.2 illustrates this result further.

Theorem 2 has a number of implications. Its result provides a complete solution to a new nonlinear pricing problem, one that describes the monopoly screening problem faced specifically by sellers of a wide variety of information-technology based products and services. Additionally, this solution shows that when faced with this non-standard IT-specific problem, a seller can use an appropriately modified version of their standard pricing techniques, rather than having to understand and apply an entirely new theoretical formulation.

The specific prescription of our result – pricing based (approximately) on the average value of one's last discontinuous unit of cost – is relatively straightforward and intuitive. However,

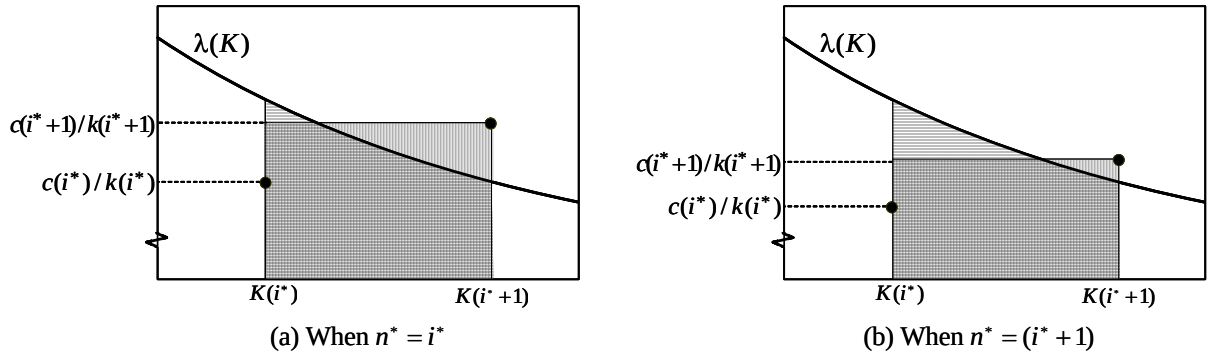


Figure 3.2: Illustrates the result of Proposition 2(b). The area under the $\lambda(K)$ curve between $K(i^*)$ and $K(i^*+1)$ is the incremental revenue (horizontal stripes); the area in the rectangle between $K(i^*)$ and $K(i^*+1)$ and under the $c(i^*+1)/k(i^*+1)$ line is the incremental cost $c(i^*+1)$. Figure 3(a) illustrates a scenario under which incremental cost exceeds incremental revenue, in which case $n^* = i^*$, while Figure 3(b) illustrates the opposite, in which case $n^* = (i^*+1)$.

it can also change pricing and profitability in a significant way. For example, if there are discontinuously diseconomies of scale (e.g., resulting from an OS platform with poor scalability), the average cost of the last block will be much higher than overall average cost or zero. The seller is liable to sell excessively unless it recognizes that its pricing should be based on the average cost of the *last* unit of cost incurred.

If a seller bases its pricing on the near-zero marginal costs observed between successive increases in capacity (that is, if it treats its product like an *information good*), it will choose the revenue maximizing pricing schedule. This can affect profits significantly. To draw out some managerial implications of this result, we present a simple example in which utility is quadratic and the customer type distribution follows the beta distribution[9] with parameters $a = 1$ and $b > 0$, that is $U(q, \theta) = \theta q - \frac{1}{2}q^2$ and $F(\theta) = 1 - [1-\theta]^b$.

Given a supply function $c(i), K(i)$, the optimal menu of quantity-price pairs, along with their

[9]The beta distribution describes a family of curves that are unique in that they are nonzero only on the interval (0 1). The shape of the beta distribution is quite variable depending on the values of the parameters. Uniform distribution ($a = b = 1$) is a special case of beta distribution. The general pdf of beta distribution is $\frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$, where $B(a,b)$ is the beta function.

associated profit and welfare expressions are available upon request. These yield the optimal $n^*$ according to Theorems 1 and 2. The actual pricing function (that is, price as a function of quantity) takes the following form: a two-part tariff with a volume discount.

$$p(q) = \frac{1}{1+b}\left[b\lambda(K(n^*)) + q(1-\frac{q}{2})\right]. \tag{3.19}$$

Note that all of the dependence of pricing on the supply function $c(i), K(i)$ is contained in $n^*$. This is a consequence of the sequence we have chosen for our theoretical formulation, and is advantageous when applying it to specific examples, since allows one to examine properties of the pricing schedule without too much algebraic complexity. In this specific case, (3.19) indicates that when the seller treats its good as an information good, then $\lambda(K(n^*) = 0$. In contrast, when the seller takes its discontinuous costs into account, $\lambda(K(n^*)) > 0$, and prices at all levels of usage should be higher (by a fixed amount in this example, though this is not true in general). For higher levels of $c(i)$, the number of units $n^*$ will be lower, $\lambda(K(n^*))$ will be higher, and zero marginal cost pricing will be farther from the optimal level.

Since the cost of the infrastructure associated with delivering digital products and services (servers, networking equipment, ...) declines fairly rapidly over time, it is natural to investigate how the seller's optimal pricing schedule changes with this kind of sustained decline. Comparative statics analysis of the base model has two weaknesses. First, it ignores the durability of the infrastructure associated with delivering digital goods: costs incurred in fulfilling demand in one period are likely to fulfill demand in subsequent periods. Second, by collapsing a dynamic problem into a comparative statics calculation, it does not capture the effects that anticipated future cost declines might have on current choices (a seller may deploy less infrastructure in anticipation of future cost declines, for instance). Consequently, we investigate the effect of

declining costs on pricing further in the first extension to our model in section 4.

# 4. Extensions

## 4.1. Declining costs and evolving demand

In this section, we extend the analysis to two periods: costs remain discontinuous but decline proportionally at a rate $\alpha$ over time. Customers are short-lived in each period with identical and independent type distributions. In this section, the seller is facing the trade-off to save the capital expenditure by delaying the installation of demand capacity while sacrificing the revenue in the first period. We show that when the seller faces a finite horizon of product viability, pricing may not be affected by anticipated cost declines up to a point; this result highlights one implication of the "capacity" nature of the cost function of the digital products we model. However, if the anticipated drop of costs is significant enough, it results in a decline in the demand capacity installed in the first period with an increase in the demand capacity installed in the second period, which leads to decreasing average price and increasing consumption of each participating buyer over time.

We have solved the single period pricing problem in section 3. Define the associated single period revenue function as $R(K(n))$. We can express the single period optimization problem 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 of this two-period problem is

$$\max_{n_1,n_2} R(K(n_1)) - C(K(n_1)) + \delta[R(K(n_1+n_2)) - \alpha C(K(n_1+n_2)) + \alpha C(K(n_1))]. \qquad (4.1)$$

The first two terms are the profits from the first period and $\delta$ is the discount rate. Terms in

the bracket are the profits earned in the second period. It should be noted that $n_2$ is defined as the incremental rather than cumulative number of block purchased in the second period. The total number of block installed in the second period is thus $n_1 + n_2$, which leads to the capital expenditure in the second period being $\alpha[C(K(n_1 + n_2)) - C(K(n_1))]$. For ease of exposition, we define the single-period optimal number of blocks in terms of the market size, $x$, as follows:

$$n(x) = \arg\max_n x \cdot R(K(n)) - C(K(n)) = \arg\max_n R(K(n)) - \frac{1}{x}C(K(n)). \tag{4.2}$$

There are two additional interpretations of $n(x)$: (1) The optimal number of blocks when the infrastructure can be reused $x$ times. (2) The optimal number of blocks when the cost is proportionally scaled down by $1/x$. By this definition, $n(1)$ is our single period solution in Section 3. The following theorem describes how this result changes when costs decline from period 1 to period 2.

**Theorem 3.** *The monopolist will install demand capacity only in the first period when the costs decline slightly. Formally, when* $\alpha \in [1/(1+\delta), 1]$, *we have*

$$n_1^* = n(1+\delta) \text{ and } n_2^* = 0.$$

*Otherwise when* $\alpha \in [0, 1/(1+\delta)]$, *the seller will defer the installation of demand capacity.*

$$n_1^* = n(\frac{1}{1+\alpha\delta}) \text{ and } n_2^* = n(\frac{1}{\alpha}) - n_1^* \ .$$

The condition that separates our two cases depends only on the discount rate, and a more patient monopolist tends not to defer installation. In contrast, the monopolist always defers

the installation of some infrastructure when the discount rate is lower. A contrast between two extreme cases is instructive. When $\delta \to 1$, costs needs to decline by more than 50% before the installation decision is influenced. In contrast, a monopolist who is very impatient ($\delta \to 0$) will almost always defer some demand capacity installation, but to a lesser degree. The reason is that when the monopolist cares less about the future, both the revenue and cost in the future are less important than the current profit. As a result, the monopolist will not procure all of the demand capacity in the first period because of the high present value of the capital cost, and always finds it profitable for to defer some infrastructure procurement to the second period.

Given the optimal number of blocks in each scenario, we can compare the market coverage, optimal quantity and pricing across cases. It can be verified that when cost declines significantly, the optimal number of blocks in the first (second) period is smaller (higher) than that of the benchmark case. More precisely, this is the case when $n(1/(1+\alpha\delta)) \leq n(1+\delta) \leq n(1/\alpha)$. Using Corollary 1, we can summarize the results in Table 4.1. Observe that declining costs always lead to an increase in market coverage, an increase in demand fulfilled, and a lower unit price. The direction of change in total price offered to each type of customer is generally ambiguous since consumers purchase more at lower unit prices. However, since the market coverage is smaller and the lowest type always gets zero quantity, we know that at least an interval of lower type customers will pay a lower total price as costs decline. The greater the decline in costs, the more significant is this trend.

### 4.2. Multimarket pricing with shared infrastructures

Our next extension considers multimarket pricing with shared IT infrastructure. In a modern business environment, most firms provide multiple products and/or services in several markets

| | Cost Down Period 1 | Cost Down Period 2 | Benchmark Case |
|---|---|---|---|
| Total number of blocks | Small | Large | Medium |
| Market coverage | Small | Large | Medium |
| Quantity to each type | Small | Large | Medium |
| Total price to low types | Low | High | Medium |
| Total price to high types | Ambiguous | Ambiguous | Ambiguous |
| Unit price to each type | Large | Low | Medium |

Table 4.1: Summary of Results

based on a shared IT infrastructure. For example, Google and Yahoo! both provide a number of services (mail, chat, spreadsheets, calendaring, social networking), and the provision of these services is based on a powerful shared IT infrastructure.

Suppose the seller offers $m$ products rather than one. For simplicity, we assume that the demand for each of these products is independent. We continue to maintain each of the assumptions made in Section 3. In the absence of an infrastructure-related constraint on fulfilling demand, denote the revenue function for product $i$ as $R_i(Q)$. This revenue function is constructed by constructing the optimal schedule with total demand $Q$ as in section 3, and can vary arbitrarily across the $m$ products.

We also allow the provision of each of the services to entail the use a varying "amounts" of the infrastructure, while assuming that the services are ranked in decreasing order of their infrastructure. For example, Google Search, Google Maps, and Google Mail may utilize the underlying shared IT infrastructure at differing intensities. Specifically, let service 1 be the most infrastructure intensive. If the shared infrastructure was *dedicated* to the provision of this product, then the $j^{th}$ block of infrastructure facilitates the fulfillment of $k(j)$ additional units of demand (this is following the development of the model in section 3.1). For any product $i > 1$, if the shared infrastructure was dedicated to the provision of this product, then the $j^{th}$ block of infrastructure facilitates the fulfillment of $k(j)/u_i$ additional units of demand, where

$u_i < 1$ captures the infrastructure needs per unit demand for product $i$ relative to product 1.

It is now straightforward to show that the profit maximization problem of the firm can be written as

$$\max_{n,Q_i} \sum_{i=1}^{m} R_i(Q_i) - C(K(n)), \tag{4.3}$$

$$\text{subject to: } \sum_{i=1}^{m} u_i \times Q_i = K(n), \tag{4.4}$$

where $Q_i$ is the total demand of product $i$, and $n$ is the number of blocks of infrastructure.

Since there is no interaction among the revenue functions, all of our results pertaining to the design of pricing schedules still apply. Therefore, the optimal pricing schedules can be derived by solving unconstrained nonlinear pricing problems with shadow variable costs $u_i \times \lambda$, associated with the demand in market $i$. In other words, the more a product uses the IT infrastructure, the higher the shadow variable cost is accrued to mitigate its demand so that the total usage of the infrastructure equals the installed demand capacity ($K(n)$).

## 5. 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 policies for information goods differ significantly from those which are optimal for goods with positive variable costs. However, we posit that information technology-based products and services – internet service, 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 provided the first general solution for the optimal nonlinear pricing of such digital goods and services. We show that the optimal nonlinear pricing schedule with discontinuous supply functions coincides with the solution to one specific instance of a constrained pricing problem (which we characterize the solution to), thus reducing a complex constrained optimization problem with a discontinuous objective function to simply identifying an optimal number of "blocks" based on a simple and intuitive rule analogous to "balancing" the marginal revenue with average "marginal cost". We show how our results differ from those based on the information goods assumption, and provide extensions that allow declining IT costs, and shared IT infrastructures.

There are many interesting directions for future research that our model suggests. First, our extension to multimarket pricing with shared infrastructures does not permit demand in these markets to be interdependent. An extension that admits this would improve our understanding of how the use of shared infrastructures affects pricing policy. Second, we do not consider the effects competition may have. 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 which is far lower than its competitors. An extension of our model to one of infrastructure-based competition would add to our understanding of how pricing power is influenced by technological capability these increasingly common shared infrastructure environments.

## 6. References


1. Afeche, P., and Mendelson, H., 2004. Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Management Science* 50 (7), 869-882.

2. Bakos, Y. and Brynjolfsson, E., 1999. Bundling information goods: pricing, profits and efficiency. *Management Science* 45 (12), 1613-1630.

3. Bhargava, H., and Choudhary, V., 2001. Information goods and vertical differentiation. *Journal of Management Information Systems* 18 (2), 89-106.

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

5. Economides, N., 2005. The economics of the Internet backbone. In Majumdar, S. et al., *Handbook of Telecommunication Economics, Volume 2.* Elsevier Publishers.

6. Geng, X., Stinchcombe, M. and Whinston, A., 2005. Bundling information goods of decreasing value. *Management Science* 51 (4), 662-667.

7. Gupta, A., Stahl, D. and Whinston, A., 1997. A stochastic equilibrium model of Internet pricing. *Journal of Economic Dynamics and Control* 21 (4-5), 697-722.

8. Ha, A., 1998. Incentive-compatible pricing for a service facility with joint production and congestion externalities. *Management Science* 44 (12), 1623-1636.

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

10. Maskin, E. and Riley, J., 1984. Monopoly with incomplete information. *Rand Journal of Economics* 15 (2), 171-196.

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

12. Mendelson, H., and Whang, S., 1990. Optimal incentive-compatible priority pricing for the $M/M/1$ queue. *Operations Research* 38 (5), 870-883.

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

14. Oren, S., Smith, S., Wilson, R., 1985. Capacity Pricing, *Econometrica*, 53 (3), 545-567.

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

16. So, K., and Song, J., 1998. Price delivery time guarantees and capacity selection. *European Journal of Operational Research* 111 (1), 28-49.

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

18. Sundararajan, A., 2004. Nonlinear pricing of information goods. *Management Science* 50 (12), 1660-1673.

19. Westland, J., 1992. Congestion and network externalities in the short run pricing of information-system services. *Management Science* 38 (7), 992-1009.

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

## 7. Appendix: Proofs

### 7.1. Proof of Lemma 1

The transformation of the objective function into the virtual profit function is standard in the literature (see, for example, Armstrong, 1996 or Sundararajan, 2004). Our next step is to show that the optimization problem (3.1) - (3.3), denoted as P1, is equivalent to (3.4) - (3.7), denoted as P2. We first show that P1 and P2 have the same feasible set and next show that their optimal solutions are the same.

Firstly, the feasible set of P2 is a subset of the feasible set of P1 when we set $q(\theta) = 0, \forall\theta < \underline{\theta}$. Secondly, any feasible solution of P1 is also a feasible solution of P2 if we define $\underline{\theta}$ by $\theta = \{\theta : q(\underline{\theta}) = 0 \text{ and } q(\theta) > 0, \forall\theta > \underline{\theta}\}$, which is well-defined and it implies $q(\underline{\theta}) = 0\ \forall\theta \leq \underline{\theta}$ because $q_1(\theta) > 0$. hence, these two problems have the same feasible set.

To show the maximum is the same, note that the only difference between these two objective functions is the profit generated from $\theta \in [0, \underline{\theta}]$, which is zero because $q(\theta) = 0, \forall\theta < \underline{\theta}$ (by assumption, $u(0, \theta) = 0$ and $u_2(0, \theta) = 0, \forall\theta$.) Consequently, the total profit from $\theta \in [0, \underline{\theta}]$ is zero and these two problems have the same optimum.

### 7.2. Proof of Lemma 2

(1) Part (a) is standard in the literature (please see Sundararajan 2004).

(2) Part (b) is an Isoperimetric problem in the dynamic programming literature. It is a well-known result that the objective function of this problem is

$$\begin{aligned} \max_{\underline{\theta},q(.)} L &= \int_{\theta=\underline{\theta}}^{1} \left[U(q(\theta),\theta) - U_2(q(\theta),\theta)H(\theta)\right] f(\theta)d\theta + \lambda[K - \textstyle\int_{\underline{\theta}}^{1} q(\theta)f(\theta)d\theta], \qquad (7.1) \\ &= \int_{\theta=\underline{\theta}}^{1} \left[U(q(\theta),\theta) - U_2(q(\theta),\theta)H(\theta) - \lambda q(\theta)\right] f(\theta)d\theta + \lambda K. \qquad (7.2) \end{aligned}$$

subject to (3.5) and (3.6). Maximizing the integrand pointwise with respect to $q(\theta)$, we have the necessary condition (3.8). Also, this is a free initial point problem of calculus of variations. By Seierstad and Sydsæster (p39, equation (41b)), the transversality condition (boundary condition) at the initial point $\underline{\theta}$ is exactly (3.13). (3.14) results from the total price equals to the consumer surplus minus the information rent offered to each type.

(3) Uniqueness. We prove this by three steps: (1) We show that (3.13) uniquely determines $\underline{\theta}$. (2) We show that the LHS of (3.7) is continuous in $\lambda$ and (3.8) uniquely determines $q$ as a function of $\lambda$. (3) LHS of (3.7) is decreasing in $\lambda$ and (3.7) uniquely determines the value of $\lambda(K)$.

**Step 1:** To show the uniqueness of $\underline{\theta}$, it can be verified that $q(\underline{\theta}) = 0$ is always a solution of (3.13). Fully differentiating the LHS of (3.13) with respect to $q$, we have

$$\begin{aligned} \frac{d(LHS)}{dq} &= [U_1 - U_{12}H(\theta)] + [U_2 - U_{22}H(\theta)]\frac{d\theta}{dq} - U_2\frac{dH}{dq}, \qquad (7.3) \\ &= \lambda + [U_2 - U_{22}H(\theta)]\frac{d\theta}{dq} - U_2\frac{dH}{dq}. \qquad (7.4) \end{aligned}$$

Note that $d(LHS)/dq$ equals to $\lambda$ at $\underline{\theta}$ because $q(\underline{\theta}) = 0$, $U_2 = 0$ and $U_{22} = 0$. Hence, $d(LHS)/dq$ is strictly greater than $\lambda$ for all $\theta > \underline{\theta}$ because the second term is positive from assumptions: $U_2 > 0$, $U_{22} < 0$, $H(\theta) > 0$, and $d\theta/dq > 0$. The last term is also positive given the fact that $H(\theta)$ is nonincreasing in $\theta$ and $q$ is nondecreasing $\theta$. Since $d(LHS)/dq > \lambda$ and $LHS > \lambda q$, the only solution of (3.13) is $\underline{\theta}$.

**Step 2:** To show that (3.7) uniquely determines $\lambda(K)$, we first define the total quantity in terms of $\lambda$ as

$$Q(\lambda) \equiv \textstyle\int_{\theta=\underline{\theta}}^{1} q(\theta,\lambda)f(\theta)d\theta. \qquad (7.5)$$

Without loss of generality, we assume that for large enough $\overline{\lambda}$, $Q(\lambda) = 0$ $\forall\lambda > \overline{\lambda}$. The reason

is that when $\lambda$ (the variable cost) is very high, the monopolist will stop selling. In (3.7) and the uniqueness of $q$ is guaranteed by the concavity assumptions. As a result, the LHS of (3.7) is a one-to-one, continuous (all functions are continuous of RHS) mapping from the compact metric space of $q$ to the compact metric space of $\lambda$. By the result from real analysis, the inverse function is also continuous. In other words, $q$ is continuous in $\lambda$.

Following standard procedures in the analysis, we can show that $Q(\lambda)$ is also continuous in $\lambda$. For any $\varepsilon > 0$,we can find $\lambda$ and $\lambda'$ close enough such that

$$|Q(\lambda) - Q(\lambda')| \leq \int_{\Theta} |q(\theta, \lambda) - q(\theta, \lambda')| f(\theta) d\theta,$$

in which the RHS equals to

$$\int_{\{\theta: |q(\lambda) - q(\lambda')| < \varepsilon\}} |q(\theta, \lambda) - q(\theta, \lambda')| f(\theta) d\theta + \int_{\{\theta: |q(\lambda) - q(\lambda')| \geq \varepsilon\}} |q(\theta, \lambda) - q(\theta, \lambda')| f(\theta) d\theta.$$

This term can be shown to be smaller than

$$\varepsilon + (\max_{\theta} |q(\theta, \lambda) - q(\theta, \lambda')|) \cdot P(\theta : |q(\theta, \lambda) - q(\theta, \lambda')| \geq \varepsilon). \tag{7.6}$$

By the demand capacity constraint, $(\max_{\theta} |q(\theta, \lambda) - q(\theta, \lambda')|)$ is bounded above by some constant. Also, $|q(\theta, \lambda) - q(\theta, \lambda')|$ can be arbitrarily small by the continuity of $q(\theta, \lambda)$. As a result, the second term disappears and we have $|Q(\lambda) - Q(\lambda')| < \varepsilon$,which completes the proof.

**Step 3:** To prove the strictly decreasing part, we apply the implicit function theorem on (3.8) and it follows that

$$\frac{dq(\theta, \lambda)}{d\lambda} = \frac{-1}{-[U_{11} - U_{112} H(\theta)]} < 0, \forall \theta. \tag{7.7}$$

The last inequality results from the global concavity assumptions on $U_{11}$ and $U_{112}$.

### 7.3. Proof of Corollary 1

**Part (a-1):** $q_2^C(\theta, K) > 0$ comes from the proof of Lemma 2, in which we show $dq_1^C(\theta, K)/d\lambda < 0$ and $dK/d\lambda < 0$.

**Part (a-2): (Total Price)** We first derive the results of the total price and next the results of the average price. We can show that the sign of the total price is ambiguous. After

differentiating (3.14), we have

$$\frac{dP^C}{dK} = \left[U_1(q^C(\theta,K),\theta)\frac{dq^C(\theta,K)}{dK} - \int_{t=\underline{\theta}(K)}^{\theta} U_{12}(q^C(t,K),t)\frac{dq^C(t,K)}{dK}dt\right] + U_2(q^C(\underline{\theta},K),\underline{\theta})\frac{d\underline{\theta}}{dK}. \tag{7.8}$$

The last term is zero by $q^C(\underline{\theta},K)=0$. Hence, the sign of LHS is the same as that of 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 the lower types are higher. We can show that the sign in general is ambiguous by showing this term is decreasing in $\theta$ and $\frac{dP^C}{dK}|_{\theta=1}$ may be negative or not. Differentiating again by $\theta$, we have

$$\frac{d^2P^C}{dKd\theta} = U_{11}(q^C(\theta,K),\theta)\frac{dq^C(\theta,K)}{dK}\frac{dq^C(\theta,K)}{d\theta}. \tag{7.9}$$

The RHS is negative and therefore the sign of $\frac{dP^C}{dK}$ depends on the value at $\theta=1$. It follows that

$$\frac{dP^C}{dK}|_{\theta=1} = \lambda \cdot \frac{dq^C(1,K)}{dK} - \int_{t=\underline{\theta}(K)}^{1} U_{12}(q^C(t,K),t)\frac{dq^C(t,K)}{dK}dt. \tag{7.10}$$

As a result, when $\lambda = 0$, it is negative.

**(Average Price)** Lastly, we show that the unit price is decreasing in $K$. It follows that

$$\frac{d(P^C(\theta)/q^C(\theta))}{dK} = \frac{q^C \cdot \frac{dP^C}{dK} - P^C \cdot \frac{dq^C}{dK}}{q^C(\theta)^2}. \tag{7.11}$$

Thus, the sign of the LHS is the same as that of the numerator. It can be verified that this term goes to 0 when $\theta$ goes to $\underline{\theta}(K)$ since $q$ and $P^C$ are both zero at $\underline{\theta}(K)$. If the numerator is decreasing in $\theta$, then it is negative for all $\theta$. Equivalently, the unit price is decreasing in $K$. The last part of the proof is to prove this claim.

$$\frac{d}{d\theta}\left[q^C \cdot \frac{dP^C}{dK} - P^C \cdot \frac{dq^C}{dK}\right] = \frac{dq^C}{d\theta}\cdot\frac{dP^C}{dK} + q^C \cdot \frac{d^2P^C}{dKd\theta} - \frac{dP^C}{d\theta}\frac{dq^C}{dK} - P^C \cdot \frac{d^2q^C}{dKd\theta}.$$

Substituting $\frac{dP^C}{dK}$, $\frac{d^2P^C}{dKd\theta}$, and $\frac{dP^C}{d\theta}$ into this equation, we have

$$\frac{dq^C}{d\theta}\cdot\left[-\int_{t=\underline{\theta}(K)}^{\theta} U_{12}(q^C(t,K),t)\frac{dq^C(t,K)}{dK}dt + q^C \cdot U_{11} \cdot \frac{dq^C}{dK} - P^C \cdot \frac{dq^C}{dK}\right] < 0.$$

This inequality results from the fact that $q>0, U_{11}<0, \frac{dq^C}{dK}>0$, and $U_{12}>0$.

**Part (b)** From the proof of Lemma 2, By $dq/d\lambda < 0$ and $q(\underline{\theta})=0$, we can conclude that $d\underline{\theta}/d\lambda > 0$,which is equivalent to $d\underline{\theta}/dK < 0$.

### 7.4. Proof of Theorem 3

Solving backward, we solve $n_2^*$ in (4.1) by maximizing the following two terms.

$$\max_{n_2} \delta[R(K(n_1+n_2)) - \alpha C(K(n_1+n_2))]. \tag{7.12}$$

In fact, it is exactly the single-period problem with cost scaled by $\alpha$. The solution is

$$n_2^* = 0, \text{ when } n_1^* > n(1/\alpha). \tag{7.13}$$

$n_2^* = 0$ because the marginal revenue from adding additional block is already smaller than the marginal cost. Otherwise,

$$n_2^* = n(1/\alpha) - n_1^*, \text{ when } n_1^* \leq n(1/\alpha), \tag{7.14}$$

which means in the second period, the seller will add blocks ($n_2^*$) until $n_1^*$, where marginal revenue "equals" average cost.

Back to the first period's decision of $n_1$, we solve the problem by considering two cases.

**Case (1)** If $n_1^* > n(1/\alpha)$, we already know that $n_2^* = 0$, which means the monopolist does not defer installation decisions. After simplifying (4.1), we have

$$\max_{n_1} (1+\delta)R(K(n_1)) - C(K(n_1)). \tag{7.15}$$

The solution is thus $n_1^* = n(1+\delta)$. To satisfy $n_1^* > n(1/\alpha)$, we need $(1+\delta) > (1/\alpha) \Leftrightarrow 1/(1+\delta) < \alpha$, which is the condition in the theorem.

**Case (2)** If $n_1^* \leq n(1/\alpha)$, $n_1$ does not affect the two terms involving $n_2$ in (4.1) since the second period's optimal decision will fill up the gap to second period's optimal level. As a result, we only need to consider the other three terms in (4.1), which can be simplified to another single-period problem.

$$\max_{n_1} R(K(n_1)) - C(K(n_1)) + \alpha\delta C(K(n_1)) = R(K(n_1)) - (1-\alpha\delta)C(K(n_1)). \tag{7.16}$$

The solution of $n_1^*$ is thus $n(1/(1-\alpha\delta))$. In order to satisfy the constraint, we need $n(1/(1-\alpha\delta)) \leq n(1/\alpha)$, which is equivalent to $1/(1-\alpha\delta) \leq 1/\alpha \Leftrightarrow 1/(1+\delta) \geq \alpha$. This completes our proof.