"Pay as you Go" or "All You can Eat"? Pricing Methods for Computing and Information Services*

Hemant K. Bhargava
hemantb@ucdavis.edu
Graduate School of Management
University of California Davis
Davis CA 95616

Manish Gangwar
manish_gangwar@isb.edu
Indian School of Business
ISB Hyderabad (TG)
India 500032

Abstract

This paper examines the ubiquitous "Pay as you Go" (per unit pricing) and "All you can Eat" (buffet pricing) schemes widely adopted for SAAS and other computing services, and compares these against three-part tariff (3PT), widely used in cell phone industry. We establish the conditions under which one plans works substantially better than other. The 3PT dominates when usage heterogeneity is low and value heterogeneity is high. Under normal conditions, a single 3PT produces higher profit than even a menu of per-unit and buffet plans. The per-unit plan is best on market coverage, but buffet pricing generally creates the highest consumer surplus.

1 Introduction

Advances in technology have helped firms to adopt a new model of selling. For software and other digital goods, especially, the traditional model of selling for ownership has transformed into one where firms charge for access to a service (Rappa 2004; Ma 2007). The method of *pricing* this software-as-a-service (SAAS) system has become an important research question (Jain and Kannan 2002; Lehmann and Buxmann 2009), and is the focus of this paper.[1] On the pricing front, "Pay as You Go" has become a popular mantra, representing the idea that buyers of a product should pay only for use, rather than pay to "own" the product regardless of whether, what, or how much they use (Armbrust et al. 2010). It has become an essential participant in the servicification of products, especially for information and computing goods. Consistent with its philosophy, "Pay as You Go" is implemented in practice with a linear pricing plan, in which total payment is a multiple of consumption quantity and a constant per-unit rate. Numerous firms have adopted this pricing model, including vendors of software products and web-based computing services (e.g., Amazon Web Services, and Kareo EMR).

At the other pole of SAAS pricing is unlimited-use or flat-fee pricing (also known as subscription or membership pricing). Here, a member pays a periodic flat fee for unlimited access during the period (often a month). Such pricing is commonly observed in both consumer oriented services (e.g., Netflix) and enterprise variations of SAAS. AutoDesk and Adobe are prominent examples of software vendors that have adopted this approach. Going further, a few firms combine these two methods into a two-part tariff (a fixed membership or access fee, plus a per-unit fee for actual usage). Two-part tariffs were common in telephony, with a fixed monthly access (or rental) fee, plus a per-minute rate for calls. Finally, some firms employ three-part tariffs (a two-part tariff, but with an allowance or "bucket" in return for the access fee). Three-part tariffs have become popular for wireless calling and data services which charge a monthly fee, provide a base allowance, and charge a per-unit rate on exceeding the allowance.

The pricing schemes and examples mentioned above represent an effort by SAAS firms to offer "simple" pricing, i.e., a single price plan presented to all customers (Essegaier et al. 2002). The case for simple pricing is amplified for firms that separately monetize multiple product "features" (e.g., EMR service providers charge separately for login, access to patient files, creation of claims etc.; similarly, Amazon Web Services has a separate per-unit price for over a hundred services, covering computing, analytics, networking, storage, applications, deployment and management, databases, and mobile services). Indeed, for per-unit pricing (and similarly for buffet pric-

*Authors gratefully acknowledge financial support from SRITNE (Sreeni Raju Center for IT and the Networked Economy) at the Indian School of Business.

[1] We use the term in a general sense, covering variations such as platform as a service, infrastructure as a service etc.

1530-1605/16 $31.00 © 2016 IEEE
DOI 10.1109/HICSS.2016.648

ing), a firm can necessarily offer only one plan, because with two per-unit prices, no buyer would pick the higher price. Even with two-part tariffs and three-part tariffs, while firms can potentially offer multiple plans, several choose to offer just a single plan. This practice finds support in the theoretical literature on non-linear pricing, where researchers have observed that a well-designed single plan can already capture a majority of the potential revenue available from more fine-grained price discrimination (Wilson 1993; Miravete 2007; Schlereth et al. 2010). Several firms find this small sacrifice in revenue worthwhile, in exchange for the incredible simplicity and sense of fairness offered by a single plan.

With this multiplicity of possible pricing schemes available to implement simple pricing, the critical question then is which pricing plan to pursue? Researchers and practitioners have examined the relative performance and trade-offs between these schemes. A recent handbook on pricing research provides useful summaries of non-linear pricing (Iyengar and Gupta 2009) and pricing of services (Shoemaker and Mattila 2009). Much of the extant work has compared two-part tariffs and the extreme cases of per-unit pricing and buffet pricing (e.g., Fishburn et al. 2000; Essegaier et al. 2002; Sundararajan 2004). One bedrock of this comparative research is to evaluate how the nature of consumer heterogeneity affects the relative performance of alternative schemes. We extend this research in three ways. First, we develop an analytical framework that allows us simultaneously to vary two forms of consumer heterogeneity: usage heterogeneity and value heterogeneity. We execute this framework to generate a series of insights regarding how heterogeneity affects relative performance of different pricing schemes. Second, we extend the analysis to cover three-part tariffs. Despite the trivial foreknowledge that a 3PT would dominate other schemes, it is nevertheless useful to examine "how much" of an improvement can be made, and under what conditions a 3PT would make substantial improvement to justify the additional complexity in pricing. Third, we extend the comparative evaluation to multiple performance metrics, specifically, firm's profit, market coverage, and consumer surplus. This multi-metric evaluation is crucial because the alternative schemes operate in fundamentally different ways with respect to the seller's conflicting desires for profitability, market coverage (which may help create economies of scale and network benefits), and sufficient consumer surplus (e.g., to manage customer retention).

We adopt the following notation. A 3PT is denoted by a triple $\tau{=}(F,Q,s)$ (with F the access fee, Q the allowance or bucket size, and s the rate for over-allowance consumption). Then, a per-unit plan corresponds to $F{=}Q{=}0$, a subscription or unlimited use plan to $s=0$, and a two-part tariff to $Q=0$. We designate the 3PT with τ, the per-unit plan with its rate r, and a buffet plan with its fee U for unlimited consumption.

2 Literature Review

The striking advantage of both per-unit pricing and buffet pricing is their utter simplicity: the firm can communicate its pricing plan with just a single parameter and can effectively differentiate themselves by offering different pricing schemes (Choudhary 2010). The two pricing techniques differ in an important way. With buffet pricing, all users pay the same regardless of consumption, hence light users face a substantial entry barrier while heavy users enjoy substantial quantity discount, at a cost to light users (Schlereth et al. 2010). In contrast, heavy users pay proportionately higher prices under per-unit pricing and light users face no entry barrier, but per-unit pricing is unable to implement quantity discounts, which are fundamental to improving profit when the firm faces heterogeneous consumers. Which pricing scheme works better depends on the nature of consumer heterogeneity, and whether consumers are more heterogeneous in their value per unit or in quantity demanded.

The choice of pricing scheme depends on the levels of marginal costs and transaction or monitoring costs. Intuitively, buffet pricing dominates when monitoring costs are high because it can avoid these costs where usage-based (or metered) pricing cannot (Sundararajan 2004; Levinson and Odlyzko 2008). Conversely, buffet pricing is less efficient as marginal costs rise because it promotes more consumption and leads to higher costs. However, modern information technology has made both these arguments less relevant by enabling an end-to-end digital infrastructure which makes marginal costs negligible, lowers monitoring and transaction costs, and eliminates resale. Moreover, the *difference* in monitoring costs between per-use and buffet plans has nearly been eliminated due to other reasons. Today, firms monitor and collect data not for determining the customer's payment level, but to enable data analytics and customer relationship management, as well as to satisfy legal and security requirements. Hence, they would incur these costs even under buffet pricing. In line with these arguments, our analytical framework is set up to exclude a decisive role for both marginal costs and transaction costs.

Capacity and congestion are economic cousins of marginal costs. Cachon and Feldman (2011) examine how congestion affects the firm's choice of buffet pricing vs. per-unit pricing. They find, as expected, that per-unit pricing works better when customers are vastly heterogeneous in consumption quantity. Surprisingly, though, buffet pricing (which increases consumption levels) becomes more useful when customers dislike congestion. In a related direction, Essegaier et al. (2002) examine the role of capacity constraints in a firms' choice among a buffet plan, per-use pricing and two part tariff. They found that when light users are more valuable, the firm's optimal strategy is to use two part tariff or flat fee plan depending on capacity constraint (tight constraints, or high marginal costs, tilt the balance in favor of usage pricing). We extend the analysis of various nonlinear pricing schemes in two ways: first, to include the more general 3PT plan; and secondly, we use a more general framework to capture nuances of consumer heterogeneity across two dimensions namely willingness to pay (valuation heterogeneity) and the rate of satiation (usage heterogeneity). Our framework does not include capacity constraints (i.e., marginal costs) but the impact of capacity constraints remains the same, to tilt the balance in favor of use-based fees.

3 Modeling Framework

Following the discussion in §2, this paper focuses on how the choice of pricing scheme depends on customer heterogeneity, rather than factors such as marginal costs, capacity constraints and transaction costs. Iyengar and Gupta (2009) also identify consumer heterogeneity as the most important factor influencing nonlinear pricing design. Accordingly, we seek a model structure that enables easy analysis of the two forms of heterogeneity, in usage and value. We start with the classical demand structure for a product that has multi-unit demand from each consumer. Let $v(x,q) \geq 0$ be type x consumer's *marginal* valuation for the q^{th} unit. Let G represent the distribution of the type variable x over set X (ordered such that $v(x,q)$ is increasing in x). Marginal valuations and the distribution function are subject to the following assumption that guarantee the Spence-Mirrlees single-crossing property and non-decreasing demand elasticity (Lariviere 2006).

Assumption 1 (Demand). *(i)* $v(x,q) \geq 0$, *(ii)* $\frac{\partial v}{\partial q} < 0$, *(iii)* $\frac{\partial^2 v}{\partial x \partial q} \geq 0$, *(iv)* G *is log concave.*

Many researchers have adopted specific forms for the v and G functions to facilitate analysis. For 3PT plans, which are hardest to analyze among the space of pricing schemes considered here, the commonly adopted formulation of marginal value is $v(x,q) = x - \beta q$, where β can be set to 1 without loss of generality by appropriately adjusting the measurement units for q. This leads to a quadratic form of the total value function (Lambrecht et al. 2007; Iyengar and Gupta 2009; Schlereth et al. 2010). This formulation accounts for marginal diminishing utility from increasing consumption, and it captures consumer heterogeneity. However, it does so in a restricted form because marginal valuation (or demand) curves of different consumers are parallel (i.e., marginal value diminishes at the same rate for all consumers), and hence satiation levels (which measure usage heterogeneity) vary the same way as valuation for the first unit.

3.1 Exploring the role of Heterogeneity

Generalization requires considering the cases where marginal valuations decline at different rates (either increasing with x or decreasing with x), so that satiation levels can either vary in a proportion greater than or less than x. Why is this important? Consider an example of a restaurant, different consumers may value the same meal quite differently, but will not vary hugely in how much they can eat in a single setting. Here, the satiation levels of high value consumers are not very different from low value consumers (leftmost panel in Fig. 1). In contrast, for consumption of Internet data, consumption patterns can demonstrate immense variation (rightmost panel in Fig. 1). Indeed, Altmann and Chu (2001) and other studies have found that high-usage consumers can consume 100 times the number of units consumed by typi-

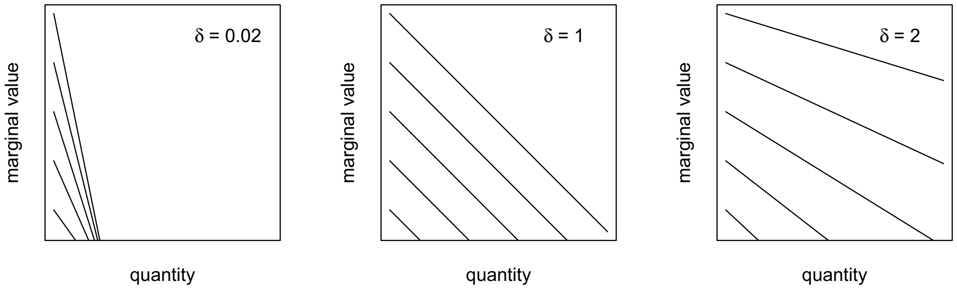

Figure 1: Level of usage heterogeneity increases with δ, moving from the left panel to the right.

cal low-usage customers. Hui et al. (2012) termed this "appetite" heterogeneity. As the two examples demonstrate, valuation heterogeneity and usage heterogeneity are separate effects, and cannot be captured on just one dimension. Our pricing analysis therefore considers both, the first via the distribution G, and the second by adding a usage heterogeneity parameter δ which represents variation in satiation levels across consumers. To explicitly incorporate δ into the analysis, we generalize the linear marginal value function ($v(x,q) = x - \beta q$) employed in past literature, to:

$$v(x,q) = x{-}\beta(1{-}x\ln(\delta))q \quad ... \ (\delta \in (0, \mathrm{e})). \qquad (1)$$

We assume excess units can be disposed freely, so that marginal valuation is zero, $v(x,q){=}0 \ \forall \ q{>}q_0(x)$, for quantities above the satiation level. Since β is just a scaling factor, it is normalized to 1. For G, we pick three distributions that have varying locations of the mode: (i) the widely used uniform distribution, where every type is equally likely, so that there is no clustering of users and heavy users have the same mass as light users, (ii) a triangle distribution where the density of users is inversely proportional to their "heaviness", so that customers are clustered towards the lower end of valuations, and (iii) a distribution where there is a greater density or clustering of users in the middle. These distributions can be generated as special cases of a $\mathcal{B}(\nu,\omega)$ distribution with parameters (1,1), (1,2), (2,2) respectively (see Fig. 2).

This setup enables us to account for both forms of heterogeneity (consumer usage heterogeneity and traditional market valuation heterogeneity), covering all scenarios of interest to our research questions. The parameter δ measures the degree of usage heterogeneity. The parallel demand curves formulation ($v(x,q) = x{-}q$) corresponds to $\delta = 1$, while marginal valuations converge when $\delta < 1$ and diverge when $\delta > 1$ (see Fig. 1). Similarly, varying G affects the proportion of consumer types, nature of value heterogeneity. Together, the marginal value function and the variations on G fully account for the various scenarios of relevance to our research question.

3.2 Pricing Schemes

In the sequel we examine design and properties of i) per-unit pricing (identified by rate r), ii) buffet pricing (fee U), and iii) three part tariff (3PT) $\tau{=}(F,Q,s)$. To start the analysis we compute each customer's participation and consumption decision under each plan. For a customer x, conditional on purchasing the plan, the optimal consumption quantity $q^*(x)$ is computed as follows. For per-unit rate r, it is the level at which $v(x,q^*) = r$, hence $q^*(x,r) = \mathrm{v}^{\text{-}1}(x,r)$, the inverse of the marginal value function for x. For the specific function $\tilde{v}{=}x{-}q(1{-}x\ln(\delta))$, $q^*(x,r) = \frac{x-r}{1-x\ln(\delta)}$ (for $x \geq r$). Under buffet pricing, every customer will simply consume at their satiation level $q^*(x,0)$. For $\tilde{v}$, this is $q^*(x,0) = \frac{x}{1-x\ln(\delta)}$. Under a 3PT plan (F,Q,s), the optimal level is the higher of Q and $q^*(x,s)$, but no more than the satiation level for x. Formally, $q^* = \max\{\mathrm{v}^{\text{-}1}(x,s), \min\{Q, q^*(x,0)\}\}$. For $\tilde{v}$, $q^* = \max\{\frac{x-s}{1-x\ln(\delta)}, \min\{Q, \frac{x-s}{1-x\ln(\delta)}\}\}$. The optimal consumption quantity can be plugged into the total valuation function $V(x,q) = \int_0^q v(x,q)\mathrm{d}q$ to get the maximum surplus customer x would obtain conditional on purchasing the plan. These surplus terms are summarized in Table 1.

The participation or purchase decisions are as follows. A per-unit plan with rate r is purchased by all $x \geq r$. Hence the indifferent customer is $\hat{x}(r) = r$. A buffet plan with price U is bought by x such that $V(x, q^*(x)) \geq U$. Finally, a 3PT plan is purchased by all consumers whose gross valuation at the free allowance level is at least as high as the fixed fee. Solving these participation constraint yields the marginal buyer under each plan. Then the firm's profit functions under the three pricing schemes are:

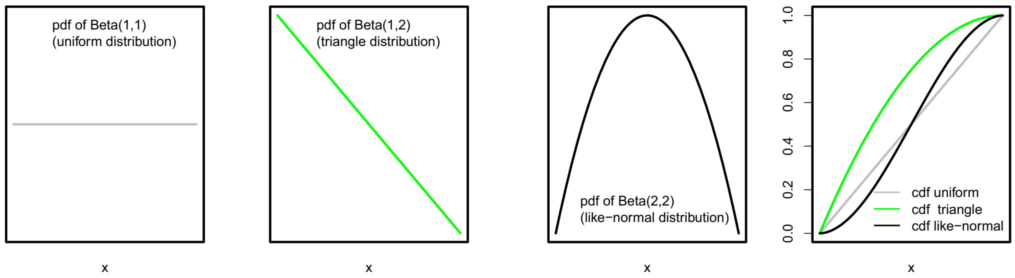

Figure 2: Three distributions that explore different extent and location of clustering of consumers. The final right panel compares the cumulative density functions for the three distributions.

Table 1: Customer Behavior and Valuations

	$Plan$		
	r	U	$\tau{=}(F,Q,s)$
$q^*(x)$	$\frac{x-r}{1-x\ln(\delta)}$	$\frac{x}{1-x\ln(\delta)}$	$\begin{cases} \frac{x}{1-x\ln(\delta)} & \text{if } \frac{x}{1-x\ln(\delta)} < Q \\ \frac{x-s}{1-x\ln(\delta)} & \text{if } Q < \frac{x-s}{1-x\ln(\delta)} \\ Q & \text{otherwise} \end{cases}$
$V(x,q^*)$	$\frac{x^2-r^2}{2(1-x\ln(\delta))}$	$\frac{x^2}{2(1-x\ln(\delta))}$	$\begin{cases} \frac{x^2}{2(1-x\ln(\delta))} & \text{if } \frac{x}{1-x\ln(\delta)} < Q \\ \frac{x^2-s^2}{2(1-x\ln(\delta))} & \text{if } Q < \frac{x-s}{1-x\ln(\delta)} \\ xQ-\frac{Q^2}{2}-\frac{bxQ^2}{2} & \text{otherwise} \end{cases}$

- Per-unit Pricing:

$$\begin{aligned} \Pi &= r\int_r^{\bar{X}} q^*(x,r)g(x)\mathtt{d}x \qquad (2) \\ &= r\int_r^{\bar{X}} \left(\frac{x-r}{1-x\ln(\delta)}\right) g(x)\mathtt{d}x. \end{aligned}$$

- Buffet Pricing:

$$\Pi = U(1-G(\hat{x})) = \frac{\hat{x}^2\,(1-G(\hat{x}))}{2(1-\hat{x}\ln(\delta))} \qquad (3)$$

  where $\hat{x}$ is the value at which the participation constraint binds ($V(\hat{x},q^*(\hat{x},0)) = U$), i.e., solves the quadratic equation $\hat{x}^2 = 2(1-\hat{x}\ln(\delta))U$.

- 3PT Plan:

$$\Pi = F(1{-}G(\hat{x})){+}s\int_y^1 (q^*(x){-}Q)g(x)\mathtt{d}x \qquad (4)$$

  where $\hat{x}$ is the marginal consumer for the 3PT (the participation constraint binds, $V(\hat{x},Q) = F$, i.e., $xQ{-}\frac{Q^2}{2}{-}\frac{bxQ^2}{2} = F$) and y is the first consumer who is willing to consume Q units at rate s (i.e., $v(y,Q) = s$; $y = \frac{s+Q}{1+Q\ln(\delta)}$).

3.3 Optimal Plan Design

We computed the profit-maximizing plan design for each pricing plan using the following procedure. First, we solve separately for each distribution. Second, even within each distribution, the first-order conditions involve transcendental terms and have order higher than two. However, we are able to uniquely identify the optimal solution for each level of $\delta \in (0,\mathrm{e})$ by eliminating certain candidate solutions. Since intermediate terms and final expressions are very messy and non-intuitive, we suppress them due to space limitations, and instead focus on outlining the properties of these solutions in the next section.

4 Impact on Heterogeneity

We examine how the plan design and properties vary across a spectrum of values or δ and for all three distributions. We present the results visually, to facilitate comparison along the two dimensions of heterogeneity. Note that the total value under the demand curve changes with variation in value and usage heterogeneity. Hence the plan metrics must be normalized to ensure a meaningful comparison. Specifically, profit and consumer surplus are divided by the maximum trade surplus available in the market; fixed fee and allowance (in the 3PT plan) are divided by the average of maximum consumption across all consumers; while the usage fee component remains the same (between 0 and 1 by construction).

4.1 Impact on Plan Design

Fig. 3 presents how δ (usage heterogeneity) and G (value heterogeneity) shape the optimal design of each pricing scheme. As δ increases, there is an increase in both r (in per-unit plan) and U (buffet pricing). Both are explained by the fact that an increase in δ has greater impact on V of higher x customers. Hence the margin volume trade-off is tilted towards higher margin, hence higher fees. Intuitively, the 3PT, should combine the two effects. However, while s does increase, F falls in the 3PT, indicating that as usage heterogeneity increases, the 3PT design places more prominence on the usage or overage fee than on the fixed fee. For the impact of value heterogeneity, compare prices across the three distributions. The per-unit price (r and s) is highest in the uniform distribution. It drops for the normal and triangle distributions, because the shift in cumulative density makes it attractive to target customers at the lower end.

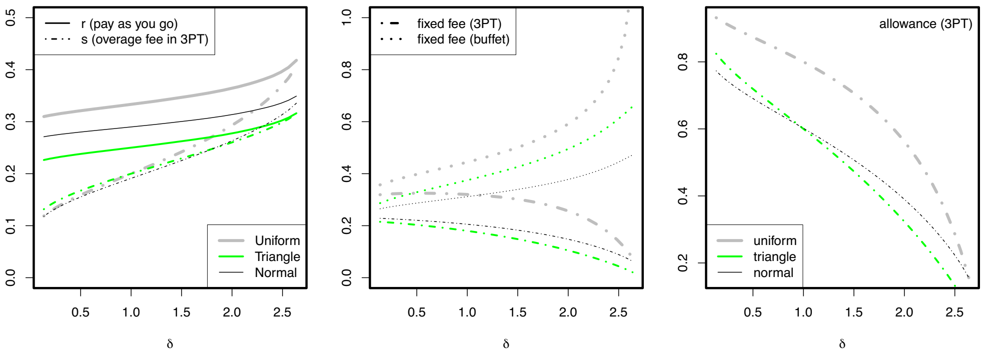

Figure 3: Optimal plan design: Impact of usage and value heterogeneity. The three distributions are differentiated by line thickness (uniform is thickest, like-normal is thinnest), and the three pricing techniques by line type (dotted for buffet, dashed for 3PT, and regular for pay-as-you-go).

Fixed fees are also highest in the uniform distribution, even after normalizing the fees in the other two distributions, accounting for lower total value.

Now consider the effect of heterogeneity on *relative* design of these plans. The first panel in Fig. 3 reveals that the 3PT plan has lower usage fee than the per-unit plan, because the former can also monetize through the fixed access fee (buffet plan is absent because it has zero marginal price). The first and second panels also explain the effect of value heterogeneity. Consider any value of δ (i.e., take a vertical slice), and compare the price levels when consumer types are uniformly distributed vs. bunched in the mid-value range (normal distribution) vs. bunched in the low-value range (triangle distribution). Not surprisingly, both usage fees and fixed fees follow the shift in valuations and decrease as the mass of consumers shifts towards lower value. The left panel also demonstrates that the difference between r and s is highest for low values of δ (low usage heterogeneity), and for the uniform distribution (high value heterogeneity); this difference shrinks as usage heterogeneity increases. Moreover, F and Q also fall as usage heterogeneity increases. These observations are summarized as follows.

Result 1. *The optimal 3PT sharply differs from the optimal per-unit plan under high value heterogeneity and low usage heterogeneity, but resembles it when usage heterogeneity is very high. A 3PT is most consequential over buffet pricing when usage heterogeneity is high.*

To examine how these results relate to pricing techniques in practice, consider pricing of Amazon Web Services (AWS). Because computing cycles are a commodity product, clients of AWS have similar reservation prices, hence value heterogeneity is quite low. However, usage heterogeneity is high because different users can have vastly different needs for computing capacity. Hence, Results 1 and 3 suggest that a per unit plan will be as efficient as 3PT in capturing most of the profits, consistent with the pricing technique employed by Amazon. In contrast, 3PT is employed for more specialized computing services such as EMR (electronic medical record) service.

The second panel of Fig. 3 compares fixed fees under the 3PT plan and buffet pricing. The latter obviously has higher fees because it has zero marginal price. The figure reveals that the difference in fixed fees increases with usage heterogeneity. With buffet pricing, the firm has only one lever, the flat fee, with which to address the changed demand environment. But with 3PT, it can choose to either employ the fixed fee lever or, conversely, place less weight on fixed fees (and, correspondingly, the allowance, see panel 3) and more on usage fees. As users become more diverse in demand quantity, the firm deploys the usage fee lever to monetize this demand. The optimal 3PT plan has a lower normalized fixed fee and allowance as δ increases.

Result 2. *Compared with the optimal per-unit plan, the optimal 3PT features a higher per-unit rate F/Q for the included Q units, but a lower rate s for overage consumption.*

As noted earlier, the 3PT plan has lower usage fees than the per-unit plan, because the former also monetizes with fixed fees. However, interestingly, the implied per-unit rate for the fixed fee (i.e., $\frac{F}{Q}$) is higher than the per-unit rate r. This property is true regardless of the nature of value heterogeneity. In essence a 3PT plan provides a low usage (over-allowance) fee in exchange for a guaranteed payment level by the user, and the commitment is secured by offering an allowance. This feature makes the 3PT plan more attractive to heavy users, and

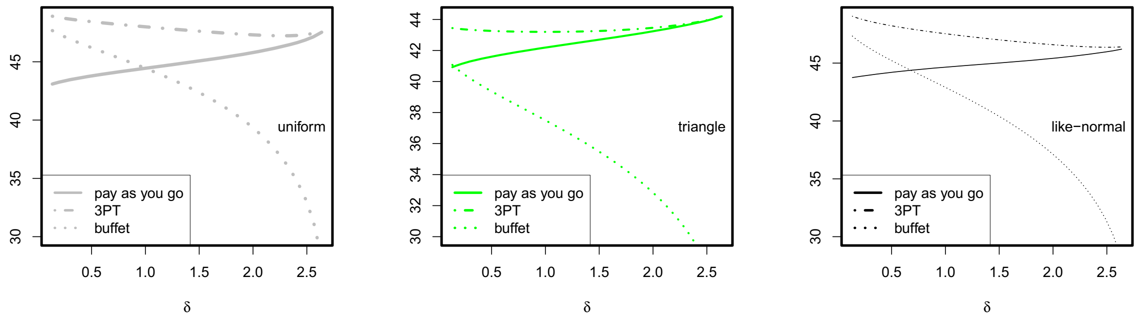

Figure 4: Impact of value and usage heterogeneity on profitability (% of total profit captured).

less desirable to light users. This is a crucial property of the 3PT, and enables it to leverage diminishing marginal valuations. The higher F/Q value reflects that the first Q units are worth more to each buyer. The lower rate s is able to encourage additional consumption, responding to the lower worth consumers place on these units. This illuminates the essence of how 3PT plan works. Due to diminishing marginal value, per-unit pricing necessarily leaves some surplus to the consumer. The 3PT plan can collect the entire net consumer surplus of the marginal consumer as the fixed fee, yet enable more consumption with s. The per-unit rate alone lacks such versatility.

4.2 Impact on Plan Performance

Next we examine the relative performance of each plan, and how it is impacted by value and usage heterogeneity. We focus on two aspects of performance: profit and market coverage. Fig. 4 displays profits for all pricing schemes under different scenarios. The per-unit plan works best as usage heterogeneity increases, because the unit fee can address the high variation in quantity demanded and generate revenue proportional to consumption. Buffet pricing displays the opposite behavior, and works best when usage levels are more homogeneous, because the fixed fee is tied to the consumer surplus of marginal buyers. The three-part tariff employs a mechanism that combines the effect of per-unit and buffet pricing. It works best when usage heterogeneity is low, because the fixed fee works more efficiently when usage levels are more homogeneous. This efficiency reduces as usage heterogeneity grows, because the firm now has to rely more on the usage fee to address the greater variation in quantity demanded. Finally, when usage heterogeneity becomes very high, the profit performance improves because the pricing scheme basically mimics per-unit pricing for high value consumers.

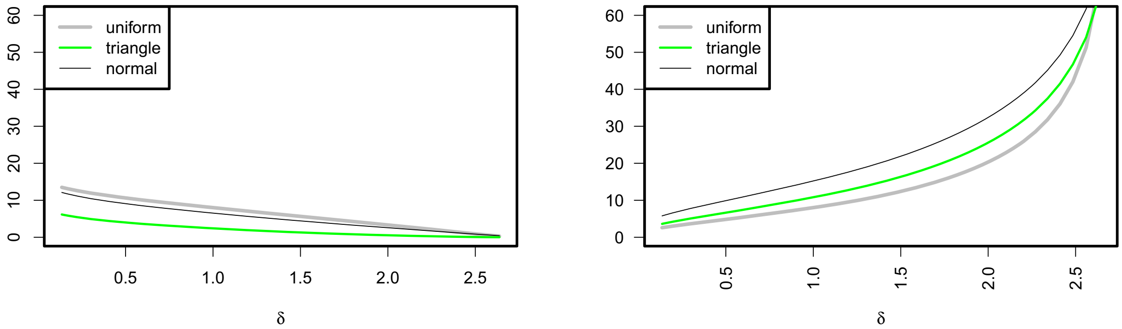

Figure 5: 3PT plan: Profit increase relative to per-unit (left panel) and buffet pricing (right panel).

Result 3. *Per-unit pricing is better at profit extraction when consumers are very heterogeneous in usage lev-*

els, while buffet pricing works better when usage is more homogeneous. The 3PT is best under low usage heterogeneity and maintains performance by mimicking the per-unit pricing when usage heterogeneity is high.

Being a more general pricing scheme than both per-unit and buffet pricing, a 3PT naturally produces higher profit. However, a 3PT is marginally more complex than the other two plans. It is therefore useful to examine how the market demand environment impacts the *magnitude* of profit increase obtained by using a 3PT plan. Fig. 5 presents the result with respect to the nature of value and usage heterogeneity. Because per-unit and buffet pricing work in diametrically opposite ways, the 3PT plan's profit advantage over these plans also follows this pattern. The profit advantage of a 3PT plan over per-unit pricing is highest when usage heterogeneity is low. In this case, the 3PT plan is very efficient at re-collecting the excess consumer surplus that would be lost under per-unit pricing. As usage heterogeneity grows, the potential revenue from usage pricing exerts a greater weight. Then, the 3PT plan begins to resemble per-unit pricing with proportionally lower fixed fees and allowance, hence produces only a modest increase in profit.

The comparison flips in the case of buffet pricing, except that there is a much larger profit difference between the two plans. The 3PT plan handily beats buffet pricing when usage heterogeneity is high. This result might be surprising because intuition suggests that buffet pricing should be most useful in this case. Instead, when usage heterogeneity is high, buffet pricing leaves firms with a very stark contrast between volume and margin. Higher market coverage (i.e., lower $\hat{x}$) requires a lower fixed fee, thereby sacrificing substantial potential revenue from higher-x users. In contrast, the 3PT plan is able to add revenue through the overage fee. Conversely, when there is very little usage heterogeneity, then the simpler buffet pricing is able to achieve most of the profit obtained from the 3PT plan.

Result 4. *The 3PT plan's dominance over the per-unit plan is highest when value heterogeneity is high and usage heterogeneity is low. Conversely, its dominance over buffet pricing is highest when usage heterogeneity is high and value heterogeneity is low.*

Next, consider market share. Because use fees impose greater total costs on high usage customers while fixed fees cause light users to subsidize heavy users, a buffet plan and 3PT plan favor high-usage customers, working in a strikingly different way than a per-unit plan, which favors low-usage customers. Hence, market coverage is highest under a per-unit plan. In our framework it approaches $\frac{2}{3}$ when there is no usage heterogeneity, far exceeding the $\frac{2}{5}^{th}$ under the other two plans (see Fig. 6). However, as usage heterogeneity increases, the per-unit plan faces a starker tension between volume and margin, and market coverage drops. Market coverage falls even more rapidly for the buffet plan, because besides this tension, the buffet plan is already ill-suited to high levels of usage heterogeneity. The 3PT, however, does not face as stark a tension because it can leverage both the fixed and usage fee components. The gap in market coverage between the 3PT and per-unit plan shrinks as δ increases.

However, besides market share, another metric for evaluating impact on consumers is consumer surplus. This metric paints a different picture, because now the buffet plan (which has lowest market coverage) encourages all buyers to consume up to their satiation level, creating additional surplus for consumers. Naturally, this

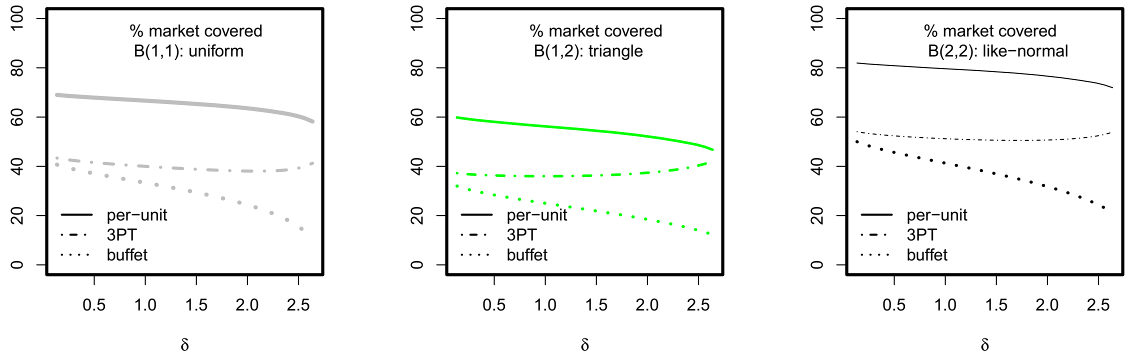

Figure 6: Market share across pricing schemes, and impact of heterogeneity.

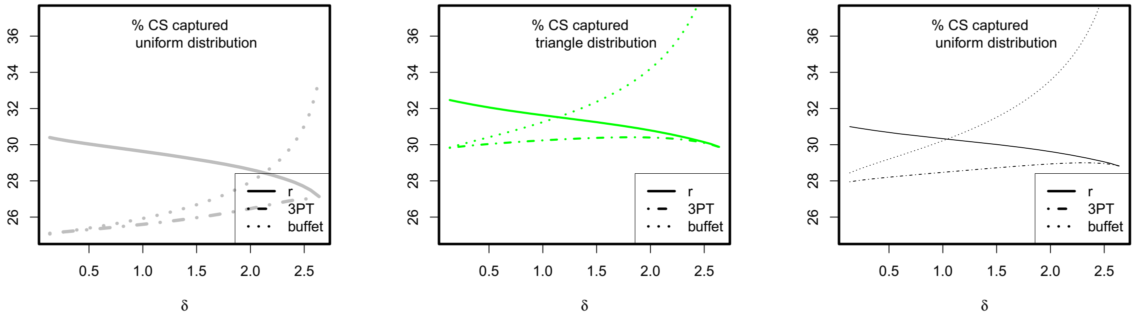

Figure 7: Percentage of Consumer surplus captured under different prices schemes and heterogeneity.

effect is more pronounced for heavy users. Consequently, the buffet plan performs quite well when usage heterogeneity is high. Conversely, the per-unit plan performance degrades as δ increases, because the use price places a heavy tax on consumption and surplus creation. The 3PT, being able to use both F and s improves modestly with usage heterogeneity. Fig. 7 visualizes these observations, and the results on market coverage and surplus are summarized below.

Result 5. *Per-unit pricing results in maximum market coverage, and buffet pricing the least. However, buffet pricing leads to greater consumption, and creates more consumer surplus when usage heterogeneity is high.*

Our results demonstrate, while the profit from the simple pricing schemes, per-unit pricing and buffet pricing can be close to the 3PT profit under some market scenarios, it can be substantially lower in others. One implication of this is that if a marketer decides to use a simple plan—per-unit or buffet—then, picking the wrong one, given the market conditions, could cause a big sacrifice in profit. The results also demonstrated that the two simpler pricing schemes work in starkly contrasting ways. One captures most of the potential profit (relative to 3PT) when value heterogeneity is high, the other when usage heterogeneity is high. This suggests that a marketer could get the best of both by simply combining the two schemes, offering a menu with one per-unit price and one unlimited-use price. The per-unit plan can cater to the low-value customers, while the buffet price can be used to lure high-usage customers. Intuitively, then, this menu should produce both higher market coverage and higher profit than a three-part tariff.

We examine this conjecture, pitting a single 3PT against a menu of per-unit and buffet prices, and varying both usage heterogeneity and value heterogeneity. The results are presented in Fig. 8, which displays the percentage increase (or decrease) in profit by using a 3PT vs. the menu. Surprisingly, the 3PT beats the menu under most conditions. Specifically, the best relative performance of 3PT occurs when value heterogeneity involves clustering towards mid-value consumers. It also performs well even when clustering is towards low-value customers. When there is no clustering (uniform distribution), then the 3PT performs well under low usage heterogeneity but not when usage heterogeneity is high.

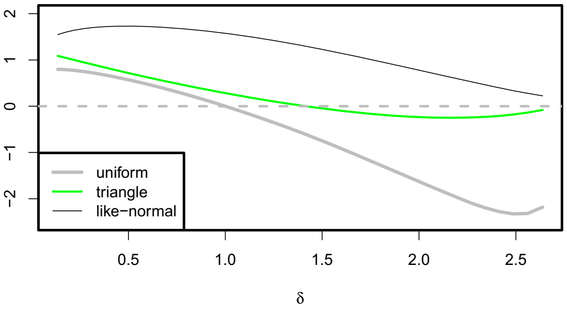

Figure 8: 3PT vs. Menu of Per-Unit and Buffet Prices.

Result 6. *The single 3PT produces higher profit than the optimal menu of per-unit and buffet prices, except when consumers are very heterogeneous on appetite, and highly scattered on per-unit valuation.*

5 Conclusion

In this paper we analyzed three most widely used pricing plans for computing services, and the impact of value and appetite heterogeneity on their relative performance. Generally, both forms of heterogeneity favor a "Pay as you Go" plan relative to an "All you can Eat" buffet plan. A three-part tariff (being more general) beats both on profit, although per-unit pricing yields higher market coverage and buffet pricing yields maximum consumer surplus. We found that usage heterogeneity has larger role to play in deciding optimal plan rather than traditional value heterogeneity. The 3PT profit beats the buffet plan handily in all conditions, while its advantage over a per-unit plan is highest under high value heterogeneity and low usage heterogeneity. This explains why AWS, facing low value heterogeneity and high usage heterogeneity market, may choose to stay with simpler per unit pricing. Nonetheless, the 3PT is more versatile: the overage fee, serves as a device to meter high type consumers, but when market conditions favor per-unit pricing, it essentially acts as one.

We also examine a menu of "Pay as you Go" and "All you can Eat" pricing plan's performance against a 3PT plan. The menu intuitively should exhibit the better performance because it combines the positive, but contrasting, qualities of both pricing schemes. In the menu, per unit plan serves low valuation consumers and the buffet plan extracts additional consumer surplus from high valuation consumers. Yet, surprisingly, the 3PT exceeds the profit from this menu under most conditions. This further extends (Bagh and Bhargava 2013)'s finding on the efficiency of price discrimination via three-part tariffs. One limitation of the current framework is that it does not formally includes the capacity constraints (i.e. marginal cost); intuitively the impact of capacity constraints will remain the same, to tilt the balance in favor of per unit pricing.

References

Altmann, Jörn and Karyen Chu (2001). "How to charge for network services–flat-rate or usage-based?" *Computer Networks* 36.5, pp. 519–531.

Armbrust, Michael et al. (2010). "A View of Cloud Computing". *Commun. ACM* 53.4, pp. 50–58.

Bagh, Adib and Hemant K Bhargava (2013). "How to Price Discriminate When Tariff Size Matters". *Marketing Science* 32.1, pp. 111–126.

Cachon, Gérard P and Pnina Feldman (2011). "Pricing Services Subject to Congestion: Charge Per-Use Fees or Sell Subscriptions?" *Manufacturing & Service Operations Management* 13.2, pp. 244–260.

Choudhary, Vidyanand (2010). "Use of Pricing Schemes for Differentiating Information Goods". *Info. Sys. Res.* 21.1, pp. 78–92.

Essegaier, Skander, Sunil Gupta, and Z John Zhang (2002). "Pricing Access Services". *Marketing Science* 21.2, pp. 139–159.

Fishburn, Peter C, Andrew M Odlyzko, and Ryan C Siders (2000). "Fixed fee versus unit pricing for information goods: competition, equilibria, and price wars". *Internet publishing and beyond: The economics of digital information and intellectual property*, pp. 167–189.

Hui, W, B Yoo, V Choudhary, and K Y Tam (2012). "Sell by bundle or unit?: Pure bundling versus mixed bundling of information goods". *Decis. Support Syst.*

Iyengar, Raghuram and Sunil Gupta (2009). "Nonlinear Pricing". In: *Handbook of Pricing Research in Marketing*. Ed. by Vithala Rao. Northampton, MA: Edward Elgar Publishing, Inc.1, pp. 355–383.

Jain, Sanjay and P K Kannan (2002). "Pricing of Information Products on Online Servers: Issues, Models, and Analysis". *Management Science* 48.9, pp. 1123–1142.

Lambrecht, Anja, Katja Seim, and Bernd Skiera (2007). "Does Uncertainty Matter? Consumer Behavior under Three-Part Tariffs". *Marketing Science* 43 (5), pp. 698–710.

Lariviere, Martin (2006). "A Note on Probability Distributions with Increasing Generalized Failure Rates". *Operations Research* 54.3, pp. 602–604.

Lehmann, Dipl-Wirtsch-Ing Sonja and Peter Buxmann (2009). "Pricing Strategies of Software Vendors". *Bus. Inf. Syst. Eng.* 1.6, pp. 452–462.

Levinson, David and Andrew Odlyzko (2008). "Too expensive to meter: the influence of transaction costs in transportation and communication". *Philos. Trans. A Math. Phys. Eng. Sci.* 366.1872, pp. 2033–2046.

Ma, Dan (2007). "The business model of" software-as-a-service"". In: *Services Computing, 2007. SCC 2007. IEEE International Conference on*. ieeexplore.ieee.org, pp. 701–702.

Miravete, Eugenio J. (2007). *The Limited Gains From Complex Tariffs*. Tech. rep. Working paper, University of Texas at Austin, Department of Economics.

Rappa, Michael A (2004). "The utility business model and the future of computing services". *IBM Syst. J.* 43.1, pp. 32–42.

Schlereth, Christian, Tanja Stepanchuk, and Bernd Skiera (2010). "Optimization and analysis of the profitability of tariff structures with two-part tariffs". *Eur. J. Oper. Res.* 206.3, pp. 691–701.

Shoemaker, Stowe and Anna Mattila (2009). "Pricing in Services". In: *Handbook of Pricing Research in Marketing*. Ed. by Vithala Rao. Northampton, MA: Edward Elgar Publishing, pp. 355–383.

Sundararajan, A (2004). "Nonlinear pricing of information goods". *Management Science*.

Wilson, Robert B. (1993). *Nonlinear pricing*. New York: Oxford University Press.