

Usage-based versus flat pricing for e-business services with differentiated QoS

Z. Liu L. Wynter* C. Xia*

Abstract—Design of e-commerce services that are competitive, in such a quickly responding market, requires the analyses of prices and price structures. We present a general model of an e-commerce market that allows us to analyze optimal price structures, both flat and usage-based. Based on the price structure of a major web hosting provider, we consider both single-tier and two-tier (burst-rate) pricing, and our result suggests that the more complex two-tier structure may not be worth the marketing effort, as the firm’s equilibrium profits will not increase through the use of this structure. An essential feature of our approach is that we model explicitly the *spread of price-QoS tradeoffs* across the end-user population.

Keywords: E-commerce, web hosting, content distribution, quality of service (QoS), Nash equilibrium, value of time

I. INTRODUCTION

The pricing of electronic goods, and, in particular, e-commerce services, such as web hosting, has received considerable attention in the literature. See, for example, [5], [7], [10], [11], [9], [13], [8], [15]. Some studies into optimal pricing have been viewed from the perspective of a single firm, or in the context of a monopoly. (See, for example, [6], [14]). This approach is appropriate if one is optimizing pricing choices for a particular provider in the very short term; in that setting, one could take other firms price choices as fixed, and given those prices, along with a model of consumer behavior, compute optimal prices for the firm in question.

It is however of interest to develop models of pricing behavior of more than one provider in the electronic marketplace. Indeed, in the market for e-commerce services, other firms can adjust their price schedules rapidly in response to that of a competitor. Then, the question for any one provider is no longer how to set prices when other firms’ price choices are given, but rather whether the joint setting of prices by all providers will tend

towards an equilibrium, and, in the affirmative, what are the properties of the equilibrium.

Fishburn and Odlyzko [3] explored the Nash equilibrium that would result across two firms competing in an e-service market, one charging a fixed, per-period, fee, and the other charging on a per-transaction basis, where the per-transaction fee is linear. The authors concluded that, with the exception of a few special forms of the clients’ demand distribution, competitive equilibria of this type resulted in the trivial solution of each firm’s price tending towards zero. This result may be seen, however, as natural, given that the two firms in the model of [3] were competing solely on the basis of price, the capacity of each firm was unlimited, and no product differentiation was introduced. In that setting, it can be seen as an instance of a classic Bertrand duopoly, which is known to result in similar economics to that of perfect competition, the latter leading clearly to zero profits for all firms in the market.

In the commerce of electronic goods, there is generally some product differentiation that is naturally present or can easily be introduced. While spatial factors do not play a role with respect to the Internet, other variations in the quality of service do exist, such as host server and network speeds or response times, availability, reliability, In short, these variations lead to product differentiation that can justify the presence of non-zero revenues across firms.

In this paper, we concentrate on a particular characterization of quality of service (QoS) that is of importance in e-services, namely, response time, or delay. We formulate a model of competitive Nash equilibrium across two firms providing e-services, such as web hosting, where each firm is characterized by the price it charges, and by the quality of service it offers.

Gibbens, Mason, and Steinberg [4], studying product differentiation in the context of the Internet also consider the two parameters of price and delay, focusing on a market of two service providers. They consider in particular the choice of each provider to offer one or two QoS classes. We compare and contrast their results with ours when applicable in this work.

Other work on the pricing of information goods

IBM Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532, USA

zhenl@us.ibm.com
wynterl@us.ibm.com
cathyx@us.ibm.com

and services has advocated differentiated services and bundling [14], or flat pricing, due to its simplicity [1], [11] but have not made use of an equilibrium framework explicitly in their arguments.

While we model here customer behavior dispersion due to price and QoS differences (and their perception thereof), it is possible through our framework to isolate one or the other effect by considering, in the first case, that both providers offer the same price structure (then only QoS-dispersion effects are present). Similarly, one can set the QoS of both providers equal, thereby focusing on the price structure difference. Note, however, that in the latter case, our model reduces to that of [3].

In the next section, we present a case study into the pricing structure of a major international web hosting service and optimize for the user what strategy to follow given that only that service provider is available. Based upon prices derived from the user-optimal strategy for that first provider, we consider in Section 3 a new entrant into the market and develop and solve an equilibrium model across these two providers. A number of different variants are considered. In the first part of Section 3, we introduce the notion of a value of QoS parameter that varies continuously throughout the population of users, allowing us to model the universe of choices made by users with respect to the cost-QoS tradeoff. Depending upon whether this random variable is uniformly or exponentially distributed, we obtain qualitatively different results. We consider the case where the new entrant to the market offers a lower QoS (along with, presumably, a lower price) as well as the converse. Section 4 presents the case of two-tier pricing, composed of a base rate and a higher burst-rate price, where the cut-off between the two prices is defined by the user. Section 5 concludes with a few recommendations for further study on this theme.

II. PROVIDER 1: THE MARKET LEADER

The pricing structures of the major web hosting sites render the user's choices quite complex. Users typically must decide on several factors simultaneously, namely which provider to subscribe to, what quantity of transactions to allot to each web hosting site. In some cases, users must provide a prediction of the quantity that will be needed in the next time period, with the prices they pay a function of that prediction.

Since the way in which these parameters are used in calculating final user prices varies across web hosting sites, the choice faced by the user becomes very difficult, and may result in sub-optimal decisions, for both the users and the web hosting sites.

The purpose of this section is to present an analysis of the currently used pricing structures of a major web hosting site, and to evaluate the optimal decisions of users under that pricing structure. We then follow up with an analysis of how a competitor could enter the market, what its pricing structure should be, and what, if any, equilibrium would result.

A. Pricing policy: Flat or two-tier pricing

A standard approach to pricing the use of a web hosting site is to charge a flat unit fee per megabyte per second, where the per unit fee decreases with increasing quantities. This marginal fee structure, p' , then looks like a step function, where the staircase decreases to the right. The total price function, p , is then a discretization of a concave increasing curve. However, due to the granularity of the typical discretizations, the total cost curves are generally *not* concave.

A two-tier structure exists when a primary fee is charged for a pre-determined usage level, and a second, higher, charge is imposed for flow levels above that pre-set quantity. The higher charge is typically some constant $1 < q \leq 2$ times the primary, per unit fee, and is also known as the *burst-rate* price. This pricing structure protects the web hosting site, by ensuring a minimum revenue, regardless of the true usage level, and also by charging for bursts, or use above that level, at a higher rate.

In order to give the users some incentive to pay this pre-determined rate, and also pay for bursts, the user is given the choice to provide to the hosting site an estimate of its expected usage level for the period, μ . The primary fee is a function of this expected level, and is computed using the staircase structure described above. This parameter also offers the provider an a priori estimate of the likely usage level of the client, which is of use in capacity planning.[12]

The secondary fee is typically not charged based upon the total usage level but rather on some percentile of it, e.g., B_α , where $\alpha = 95\%$, of the total usage over the period. This also gives some leeway to the user, who is allowed to have some burstiness in his traffic without incurring a charge for it.

Then, the marginal price charged to a user over a given period $p'(x)$, using this pricing structure, would be

$$c(x, \mu) = p'(\mu) + qp'(B_\alpha(x) - \mu). \quad (1)$$

B. User-optimal strategies for provider 1

The question each user must answer is then how to optimally set his committed rate μ , and how to allocate x if there is more than one web hosting site available.

To illustrate the first choice, that of determining a continuous rate schedule obtained by taking a concave envelope of the step function in Figure 1.

Then, we have the following property:

Proposition 1: Suppose that $p' : \mathbb{R}_+ \mapsto \mathbb{R}_+$ satisfies the following properties: $p'(x) \geq 0$ for all $x \geq 0$, $p''(x) \geq 0$ for all $x \geq 0$, and $p'''(x) \leq 0$ for all $x \geq 0$ and $p' \in C^2$, that is, is twice continuously differentiable. Then, the user's price $c(x)$ achieves its minimum at $\mu = B_\alpha(x)$. That is, the user pays only the base rate and the premium.

Proof: The marginal cost $c(x, \mu) = p'(\mu)$ if $\mu > x$ and $c(x, \mu) = p'(\mu) + qp'(B_\alpha(x) - \mu)$ otherwise. We have that $c''(x, \mu) \leq 0$ for $x \in [0, B_\alpha(x)]$, and therefore it attains a minimum at an extreme point. Then, $c(x, 0) = qp'(B_\alpha(x))$, and $c(x, B_\alpha(x)) = p'(B_\alpha(x))$. Along with $q > 1$, it follows that c' is minimal when the second term is null, that is $\mu = B_\alpha(x)$.

Figure 3; in the minimal price is obtained when the user declares an expected flow *higher* than his true expected flow.

This illustrates that, theoretically, the optimal user strategy is to declare a committed usage level that avoids paying the burstable rate, in this case equal to B_α . However, the discretization of the rate schedule can lead to anomalies, in which that level is no longer optimal for the user, but that the user should declare an even higher level, thereby paying less.

In both cases, whether the committed usage level is chosen optimally or not, since the price is a unit quantity times the committed level, users are in effect paying (for a large range of quantities) a flat rate, similar to subscription-based services. We shall therefore focus on comparing the equilibrium that would occur in a market where one provider used a subscription-based rate, as the above reduces to, in practice, and another provider wished to enter the market.

Price schedule when unit price is a discrete

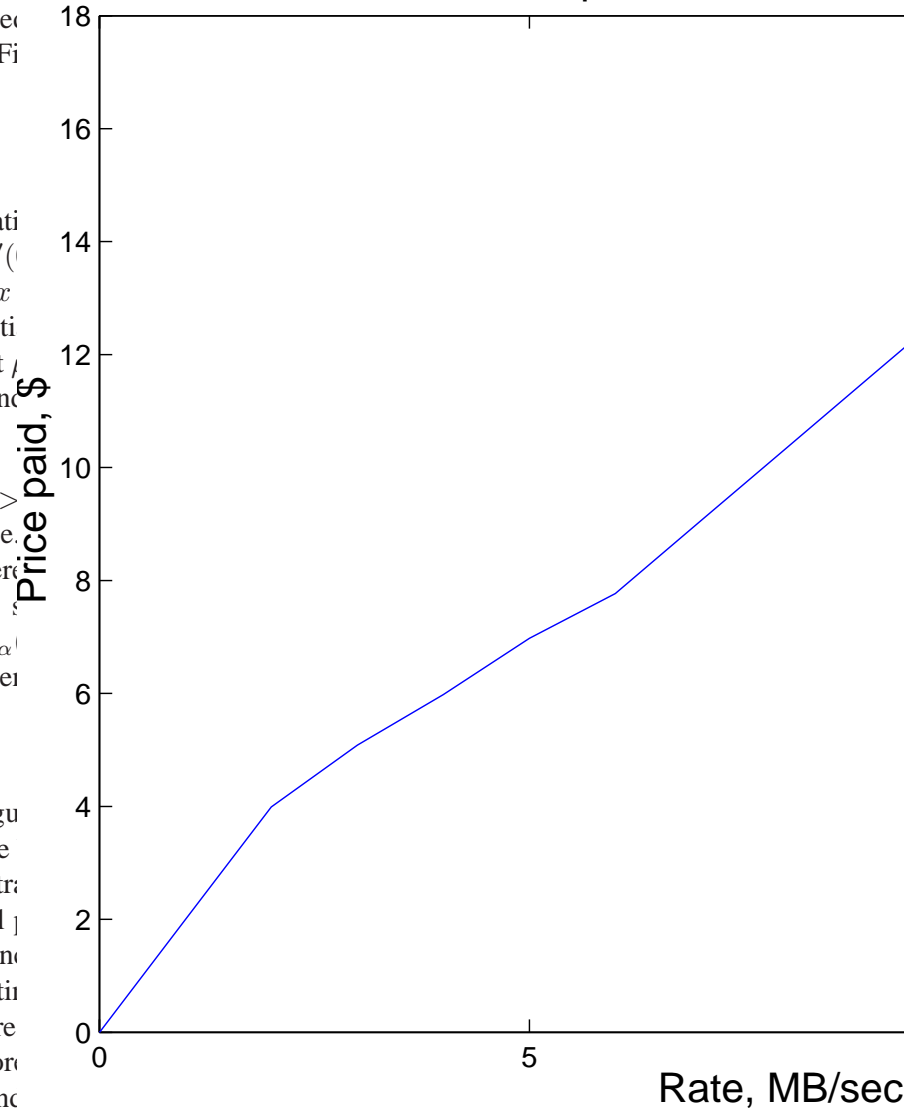


Fig. 1. Discrete rate schedule, unit price and total price, respectively, as a function of rate (MB-sec)

III. NEW ENTRANT INTO THE MARKET: USER DIFFERENTIATION AND MULTIPLE SERVICE CHARACTERISTICS

We now turn towards a new entrant into the market. We consider therefore a setting similar to that of [3], in which potential customers' usage rates, or requested capacities, are defined by a probability density function μ , that is, $\int_0^\infty \mu(x)dx = 1$, where the argument x is the desired rate of a potential user.

Contrary to the model of [3], we suppose that the e-service offered by firm $i = 1, \dots, n$ is characterized by a 2-tuple, $(p_i(x), d_i)$, where $p_i : \mathbb{R}_+^n \mapsto \mathbb{R}_+$ is the price function charged for use of the service, which depends upon the usage level, x , and d_i the quality of service. Indeed, we know from that reference [3] as well as from

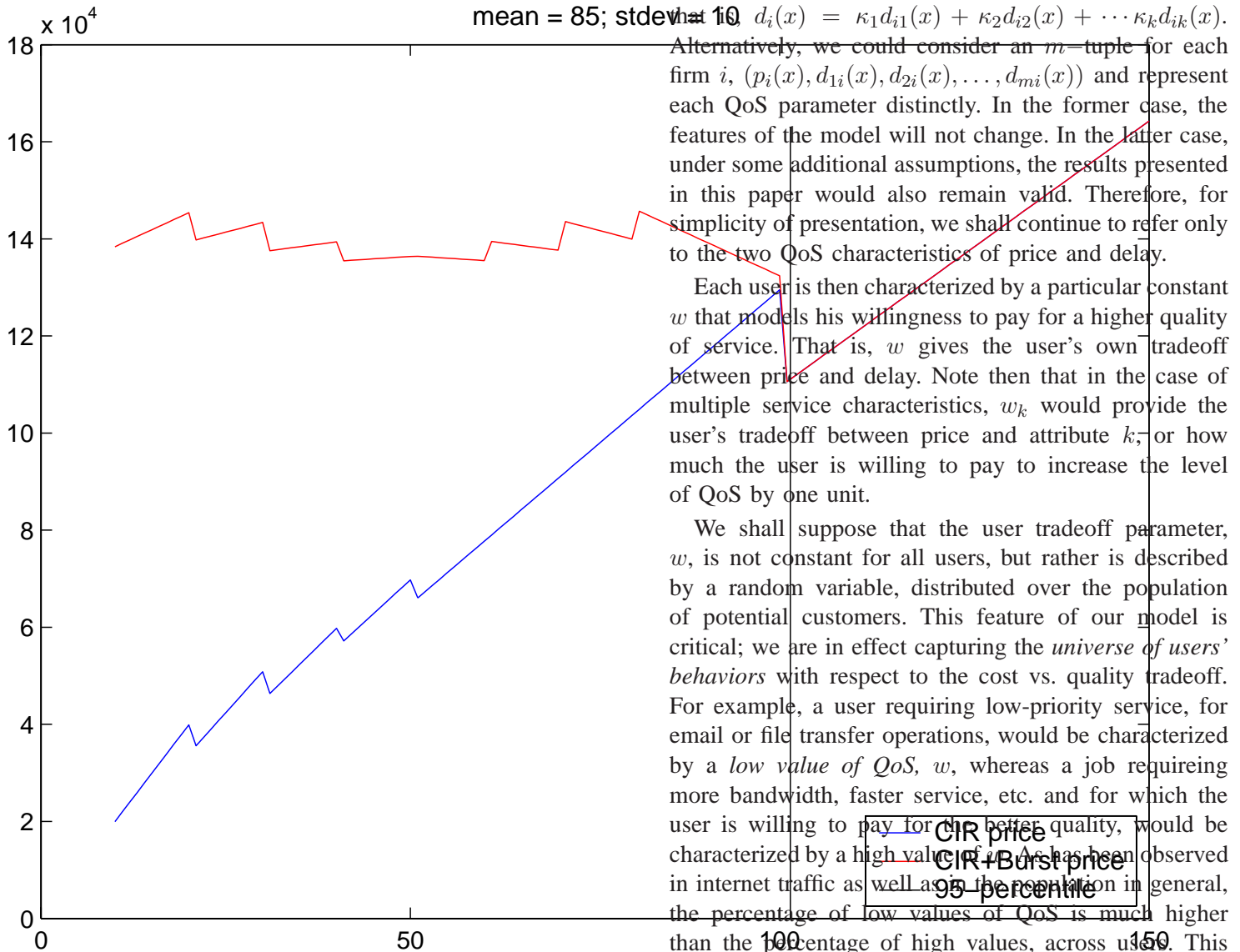


Fig. 2. Optimal committed usage level: $\mu^* = B_\alpha$

classical Bertrand competition, that competition based only upon price leads to zero profits for all firms. Service differentiation, through a QoS parameter, can remedy that ruinous result.

The quality of service will be taken in the remainder of this paper to be some measure of service performance, namely, the delay incurred on a typical request. Note that we are *not* considering the situation in which $d_i = d_i(x)$; in this work, we shall restrict ourselves to the simpler setting of usage-independent delays.

It is possible to extend this framework to more than two (possibly usage-dependent) service characteristics in two ways: (i) we could consider that the quality of service includes more than one physical characteristic (delay, reliability, etc.) and model the choice as a function of price and a composite characteristic,

$d_i(x) = \kappa_1 d_{i1}(x) + \kappa_2 d_{i2}(x) + \dots + \kappa_k d_{ik}(x)$. Alternatively, we could consider an m -tuple for each firm i , $(p_i(x), d_{1i}(x), d_{2i}(x), \dots, d_{mi}(x))$ and represent each QoS parameter distinctly. In the former case, the features of the model will not change. In the latter case, under some additional assumptions, the results presented in this paper would also remain valid. Therefore, for simplicity of presentation, we shall continue to refer only to the two QoS characteristics of price and delay.

Each user is then characterized by a particular constant w that models his willingness to pay for a higher quality of service. That is, w gives the user's own tradeoff between price and delay. Note then that in the case of multiple service characteristics, w_k would provide the user's tradeoff between price and attribute k , or how much the user is willing to pay to increase the level of QoS by one unit.

We shall suppose that the user tradeoff parameter, w , is not constant for all users, but rather is described by a random variable, distributed over the population of potential customers. This feature of our model is critical; we are in effect capturing the *universe of users' behaviors* with respect to the cost vs. quality tradeoff. For example, a user requiring low-priority service, for email or file transfer operations, would be characterized by a *low value of QoS*, w , whereas a job requiring more bandwidth, faster service, etc. and for which the user is willing to pay for the better quality, would be characterized by a high value of w . As has been observed in internet traffic as well as the population in general, the percentage of low values of QoS is much higher than the percentage of high values, across users. This observation has an impact on the *form* of the distribution of the tradeoff parameters, w , as we shall discuss in this paper.

Then, the probability distribution on user's rate levels becomes a joint distribution of rate levels and cost-QoS tradeoff parameters:

$$\int_{x=0}^{\infty} \int_{w=0}^{\infty} \mu(x, w) dw dx = 1.$$

Consider one potential user or, equally, one potential usage decision. Note that a user may make several decisions, as as we consider the mass of all such decisions, one may interpret each atomic decision as a single user, or a single choice, where a user may make several choices. Then, given each value of the tradeoff parameter, w (for each usage choice), and the desired usage level, x , he will optimize, for each choice, his choice of provider, among the n firms, by choosing the

one that minimizes his combined cost:

$$i^* \in \arg \min_i \{p_i(x) + wd_i\}. \quad (2)$$

If prices are exactly equal across providers, then one may assume that the market is split equally across those providers. Given that we work in real numbers, such an outcome is highly unlikely. Note that not only does w have the behavioral representation of the user's cost-QoS tradeoff, but, as it is expressed in units of dollars per time, it permits summing the two criteria, $p_i(x)$ and d_i .

Up to now, we have not specified the forms of the prices offered by each provider, $p_i(\cdot)$. In order to specify fully the model, we must make some assumptions about the rate structures, p_i .

Suppose, as in [3], that $n = 2$ providers, and that μ is continuously differentiable in its arguments. Let further $p_1(x) = p_1$ and $p_2(x) = p_2x$. That is, provider 1 charges a flat (subscription-based) fee while provider 2 charges a simple (linear) usage-based fee.

In this case, a user characterized by two-tuple (x, w) chooses provider 1 if

$$p_1 + wd_1 \leq p_2x + wd_2, \quad (3)$$

and chooses provider 2 otherwise.

Let us suppose initially that $p_1 > p_2$ and $d_1 < d_2$; that is, the supplier offering the flat rate offers a better quality of service (lower delay) as well.

It is clear then that there are thresholds in x and w for which one or the other supplier is cost-effective for a user. Specifically, for $x(w) \geq p_1/p_2$, supplier 1 is cheaper. Since in this example, supplier 1 also has a better QoS in that the delay it offers is lower, users with $x(w) \geq p_1/p_2$ will choose supplier 1 for all w . Similarly, when $x \leq p_1/p_2$ and $w \geq \hat{w} = (p_1 - p_2x)/(d_2 - d_1)$, supplier 1 is chosen. Supplier 2 is chosen for all other values of w, x .

For succinctness, let us refer to the vector (p_1, p_2) as p , henceforth. Then, the revenues of providers 1 and 2 can be expressed by:

$$R_1(p) = p_1 \left[\int_{\frac{p_1}{p_2}}^{\infty} \int_0^{\infty} \mu(x, w) dw dx + \int_0^{\frac{p_1}{p_2}} \int_{\hat{w}(x)}^{\infty} \mu(x, w) dw dx \right], \quad (4)$$

$$R_2(p) = p_2 \int_0^{\frac{p_1}{p_2}} \int_0^{\hat{w}(x)} x \mu(x, w) dw dx, \quad (5)$$

where, as before,

$$\hat{w}(x) = \frac{p_1 - p_2x}{d_2 - d_1}. \quad (6)$$

In that case, we can write out the first order conditions for Nash equilibrium, that is, $\partial R_1(p)/\partial p_1 = 0$ and $\partial R_2(p)/\partial p_2 = 0$.

While we do not include a fixed portion of the usage-based cost for provider 2, it is clearly the case that one could add such a cost, in which case, the revenue for provider 2 would have two integrals, where the limits would be the same, but only the second one would include the variable x . The constants would be the fixed fee and p_2 , respectively. The interpretation would be then that, in order to have service from provider 2, one needs to pay some upfront fee, and thereafter a usage-dependent price.

Then, the question of interest is whether this system has a nontrivial solution, that is, one in which $p_i \neq 0$, $i = 1, 2$ for different assumptions on the forms of the distribution $\mu(x, w)$, and, if so, what are the properties of that equilibrium.

A. Simplified model: homogeneous usage levels and uniformly distributed values of QoS

We may first consider a simplified model in which we do not maintain a distribution of usage levels x , but rather examine the equilibrium in which all users are defined by a unique, constant, usage level, x . Then, $\mu(x, w) = \mu(w)$.

To further simplify, let the distribution of delay-cost tradeoff constants, w , be uniform on the interval $[0, 1]$. Then, assuming still that $d_2 - d_1 \geq 0$, provider 1 will obtain $1 - \hat{w}(x)$ of the market when $p_1 \geq p_2x$. Note that if $p_1 \leq p_2x$, provider 1 obtains the entire market, since both price and delay are less than that offered by provider 2. Therefore, consider the former setting; we have that

$$R_1(p) = p_1(1 - \hat{w}(x)) = p_1 \left[1 - \frac{p_1 - p_2x}{d} \right], \quad (7)$$

$$R_2(p) = p_2x\hat{w}(x) = p_2x \left[\frac{p_1 - p_2x}{d} \right]. \quad (8)$$

Solving, we obtain that the equilibrium prices are

$$p_1^* = 2d/3, \quad (9)$$

$$p_2^*x = d/3, \quad (10)$$

where $d = d_2 - d_1$. The equilibrium threshold, $\hat{w}^*(x)$, for choosing provider 2 is

$$\hat{w}^*(x) = 1/3.$$

The two prices are equal only in the case where the two providers offer the same QoS, that is, $d_1 = d_2$, and, indeed, both would be zero. Provider 1, offering the flat, subscription-based price structure always has the larger market share, with 2/3 over provider 2's 1/3.

However, we observe here that the use of a uniformly-distributed value of QoS parameter, w , can lead to misleading conclusions. By taking the distribution, $\mu(w)$ to be uniform, we assume that there are as many economical-minded users as there are 'big-spending' users. High values of w signify a high willingness to pay for an improved QoS. Common knowledge, and empirical data, tell us otherwise, however – the proportion of users who opt for cheaper, lower QoS, service configurations is generally much larger than the proportion who pay highly for the best QoS. Typically, an exponential, or log-normal distribution should be used to model such tradeoff parameters over the population. Through the example, we see that the simplification of uniformly distributing that parameter may lead to the misleading conclusion that the market share for provider 2, appealing to users wishing for cheaper, lower QoS, service, would always be fixed at 1/3; were the distribution more realistic, as we shall see, this conclusion will no longer be valid.

Remark 1: Both providers offer flat, subscription-based services Note further that this result is unchanged if both providers charge flat, subscription-based fees. In that case, the equilibrium prices are simply $p_1 = 2d/3$, $p_2 = d/3$, where provider 2's price is now flat rather than multiplied by the usage level, x , and the threshold $\hat{w}^*(x) = \hat{w}^* = 1/3$.

This type of result contrasts with that of Gibbens, Mason, and Steinberg [4][Prop.1], who consider two service providers, each offering one type of service, defined by price and delay, and, as in this example, a uniformly-distributed value of QoS parameter, w . In their model, delay is linear in usage for both providers, and prices for both providers are flat (subscription-rather than usage-based). They conclude that the unique equilibrium in this case occurs when both providers' prices are equal. However, the assumption that both providers offer services with the same delay level is clearly a strong one; we saw from the equilibrium we computed in our above simple example, that, if delays of both providers are equal, ($d_1 = d_2$ so that $d = 0$), then prices are indeed equivalent for the two providers, but they are also zero. Furthermore, the use of a uniformly-distributed QoS parameter, as we shall confirm, leads to biased results.

1) *New Entrant with Better QoS*: Let us return again to a market situation in which provider 1 charges a flat fee and provider 2 a usage-based fee, p_2x . Assume this time that provider 2's QoS can be better than that of provider 1, i.e., $d_2 < d_1$. Then a user (x, w) will choose provider 2 if $p_1 \geq p_2x$, or if $p_1 < p_2x$ and $w \geq \hat{w}(x)$, where $\hat{w}(x)$ is given by (6). In the latter setting, provider

1 will then get $\hat{w}(x)$ of the market when $p_1 < p_2x$.

Therefore, when assuming all users have a single usage level x , the revenue of the two providers are respectively,

$$R_1(p) = p_1 \hat{w}(x) = p_1 \frac{p_2x - p_1}{d_1 - d_2}, \quad (11)$$

$$R_2(p) = p_2x(1 - \hat{w}(x)) = p_2x \left[1 - \frac{p_2x - p_1}{d_1 - d_2} \right]. \quad (12)$$

Solving, we obtain that the equilibrium prices are

$$p_1^* = -d/3 = \frac{d_1 - d_2}{3} \geq 0, \quad (13)$$

$$p_2^*x = -2d/3 = \frac{2(d_1 - d_2)}{3} \geq 0, \quad (14)$$

since, as before, $d = d_2 - d_1$, which is now negative, and

$$\hat{w}^*(x) = 1/3.$$

That is, provider 2, now having a better QoS, will get 2/3 of the market in equilibrium.

B. Exponentially-distributed usage levels and value of QoS parameters

Suppose now that both the usage levels, x , and the values of QoS, w , are distributed according to exponential distributions, each with its own mean, $1/a$ and $1/b$, respectively, where $a, b > 0$. Then, $\mu(x, w) = g(x)h(w)$, with $g(x) = ae^{-ax}$ and $h(w) = be^{-bw}$.

For usage levels, this hypothesis is a well-motivated one, since it represents the usual Poisson arrivals into the system. For describing the dispersion of the value of QoS parameter over the population, the exponential distribution seems also well-justified since it possesses a shape close to the commonly used log-normal distribution.

The system (4)–(5) then simplifies due to the separability of the distributions on x and w . Evaluating the expression for $R_1(p)$, we obtain

$$R_1(p) = Qp_1e^{-\frac{bp_1}{a}} + (1 - Q)p_1e^{-\frac{ap_1}{p_2}},$$

where

$$Q = \frac{a}{a - \frac{bp_2}{d}}.$$

The expression for provider 2's revenue is less compact. We obtain

$$R_2(p) = ap_2 \int_0^{p_1/p_2} x \left[e^{-ax} - e^{-\frac{p_1 b}{d}} e^{-\left[a - \frac{bp_2}{d}\right]x} \right] dx, \quad (15)$$

$$= \frac{p_2}{a} \int_0^{ap_1/p_2} x e^{-x} dx - \frac{ap_2 e^{-\frac{p_1 b}{d}}}{\left[a - \frac{bp_2}{d}\right]^2} \int_0^{\left[a - \frac{bp_2}{d}\right] \frac{p_1}{p_2}} x e^{-x} dx \quad (16)$$

$$= \frac{p_2}{a} \left[1 - \left(1 + \frac{ap_1}{p_2} \right) e^{-\frac{ap_1}{p_2}} \right] - \frac{ap_2 e^{-\frac{p_1 b}{d}}}{\left[a - \frac{bp_2}{d}\right]^2} \left[1 - \left(1 + \frac{p_1}{p_2} \left[a - \frac{bp_2}{d} \right] \right) e^{-\frac{p_1}{p_2} \left[a - \frac{bp_2}{d} \right]} \right]. \quad (17)$$

Now, defining dummy variables u and z and letting $u = ap_1/p_2$ and $v = [a - bp_2/d](p_1/p_2) = u - bp_1/d$, and then noting that $\partial/\partial p_2 = (\partial u/\partial p_2)(\partial/\partial u) = -(ap_1/p_2^2)(\partial/\partial u)$ and that $\partial z/\partial u = 1$, we obtain in terms of u and w :

$$\frac{\partial R_1}{\partial p_1} = \frac{u}{z} (1 + z - u) e^{z-u} + \left(1 - \frac{u}{z} \right) (1 - u) e^{-u} \quad (18)$$

$$\frac{\partial R_2}{\partial p_2} = \frac{1}{a} \left[1 + \frac{u^2}{z^2} \left(1 - \frac{2u}{z} \right) e^{z-u} - \left(1 + u + u^2 - \frac{2u^3}{z^3} \right) e^{-u} \right] \quad (19)$$

It is not possible to solve the system (18)–(19) analytically; however, we may examine the system numerically for simultaneous solutions (i.e. interior solutions). The Figure 4 provides a graphical illustration of the zeros of each function, $\frac{\partial R_1(p)}{\partial p_1}$, which we shall denote $dR1$, and $\frac{\partial R_2(p)}{\partial p_2}$, denoted $dR2$. In the figure, when $dR1$ or $dR2$ is negative, at a point (u, z) , the corresponding point in the space (u, z) is black. When $dR1(u, z) > 0$, then the point (u, z) is colored red, and when $dR2(u, z) > 0$, the point (u, z) is colored green. The Figure 5 then illustrates those two graphs super-imposed, providing the simultaneous zero of the two equations. Note however, that the roles of the colors are reversed in the Figure 5; black represents positive and colors the negative zones.

Two equilibria were identified; that is, two points in the (u, z) plane were found at which black, green, and red meet (i.e., simultaneous zeros of both $dR1$ and $dR2$). The first point is $(u, z) = (2.5701, 1.96)$,

The other solution was found at $(u, z) = (2.133, -2.64)$.

Figures 6 and 7 provides a graphical study of the uniqueness of the (non-trivial) equilibrium solution. Figure 6 provides the zeros of each derivative function, $dR1$

and $dR2$, while Figure 7 superimposes them to illustrate possible equilibrium points. The left image of Figure 7 set examines larger values of u , where it can be seen that the curve becomes asymptotic to the singularity; no other solution exists in that direction. The second of the plots in Figure 7 illustrates negative values of z , namely z does from $[-4, 4]$.

In terms of extremal, solutions, we would not expect interesting solutions in which either p_1 or p_2 were zero.

Solving then for the equilibrium prices, p_1^* and p_2^* , when $(u, z) = (2.5701, 1.96)$, we obtain:

$$p_1^* = 0.61d \frac{1}{b}, \quad (20)$$

$$p_2^* = 0.24d \frac{a}{b}. \quad (21)$$

When $(u, z) = (2.133, -2.64)$, we obtain:

$$p_1^* = 4.77d \frac{1}{b}, \quad (22)$$

$$p_2^* = 2.23d \frac{a}{b}. \quad (23)$$

Analyzing the revenue of the two providers at this solution, we consider a continuum of values for the two means, a , the mean on-client usage levels, and b the mean value of users values of QoS, and a number of possible QoS differences, $d = d_2 - d_1$. The revenues depend upon the QoS differences, and increase as d increases, as can be seen from Figure 8. However, the ratio of R_2 to R_1 stays constant at around 1.1; that is, provider 2 always has a higher revenue than that of provider 1. This contrasts with the results obtained when the distribution governing the value of QoS parameter was uniform.

1) *Both providers charge flat fees:* Suppose now that both providers 1 and 2 choose to charge flat, subscription-based fees. Then, assuming still that $d_2 > d_1$ and $p_2 \leq p_1$, the revenues of the two providers are given by

$$R_1(p) = p_1 \int_{\hat{w}}^{\infty} \mu(w) dw, \quad (24)$$

$$R_2(p) = p_2 \int_0^{\hat{w}} \mu(w) dw. \quad (25)$$

With exponentially-distributed value-of-QoS parameters, we have that

$$R_1(p) p_1 \left(e^{-b\hat{w}} \right) = p_1 e^{-b \left(\frac{p_1 - p_2}{d} \right)}. \quad (26)$$

$$R_2(p) = p_2 \left(1 - e^{-b\hat{w}} \right) = p_2 \left(1 - e^{-b \left(\frac{p_1 - p_2}{d} \right)} \right). \quad (27)$$

Solving, we have that

$$\frac{\partial R_1}{\partial p_1} = \left(\frac{1 - bp_1}{d}\right)e^{-b\hat{w}} = 0 \quad (28)$$

$$\frac{\partial R_2}{\partial p_2} = 1 - e^{-b\hat{w}}\left(1 + \frac{bp_2}{d}\right) \quad (29)$$

Then, as opposed to the case in which provider 2 chooses usage-based pricing, when both providers choose flat pricing, there is no positive solution to the system. Indeed, we have that $p_1 = d/b$ and, setting $Z = bp_2/d$, p_2 solves $e^{Z-1}(1 - Z) = 1$. However, as we see from Figure 9 there is no solution to that system, as the function does not equal 1 for any set of positive parameter values. This parallels the finding of Fishburn and Odlyzko [3] that price wars would result from price competition in the sector.

2) *New Entrant with Better QoS*: Assume now that $d_2 < d_1$, that is, provider 2, who charges a usage-based fee, offers a better QoS. Then a user (x, w) will choose provider 2 if $p_1 \geq p_2x$, or if $p_1 < p_2x$ and $w \geq \hat{w}(x)$, where $\hat{w}(x)$ is given by (6). The corresponding revenues of providers 1 and 2 can be expressed by:

$$R_1(p) = p_1 \int_{\frac{p_1}{p_2}}^{\infty} \int_0^{\hat{w}(x)} \mu(x, w) dw dx, \quad (30)$$

$$R_2(p) = p_2 \left[\int_0^{\frac{p_1}{p_2}} \int_0^{\infty} x\mu(x, w) dw dx + \int_{\frac{p_1}{p_2}}^{\infty} \int_{\hat{w}(x)}^{\infty} x\mu(x, w) dw dx \right], \quad (31)$$

where $\hat{w}(x)$ is given by (6).

Again, suppose that both the usage levels and the values of time are distributed according to exponential distributions, each with its own mean, $1/a$ and $1/b$, respectively, where $a, b > 0$. Then, $\mu(x, w) = g(x)h(w)$, with $g(x) = ae^{-ax}$ and $h(w) = be^{-bw}$. Evaluating the system (30)–(31) at the given exponential distributions, we obtain

$$R_1(p) = (1 - Q)p_1 e^{-\frac{ap_1}{p_2}}, \quad (32)$$

$$R_2(p) = \frac{p_2}{a} \left[\int_0^{\frac{ap_1}{p_2}} xe^{-x} dx + Q^2 e^{-\frac{p_1 b}{d}} \int_{\frac{ap_1}{p_2}}^{\infty} xe^{-x} dx \right] \quad (33)$$

, where, as before, $d = d_1 - d_2$ (but now negative) and $Q = \frac{a}{a - bp_2}$.

Define, as before, dummy variables u and w and letting $u = ap_1/p_2$ and $z = [a - bp_2/d](p_1/p_2) = u - bp_1/d$, we obtain in terms of u and z :

$$\frac{\partial R_1}{\partial p_1} = \left(1 - \frac{u}{z}\right)(1 - u)e^{-u}, \quad (34)$$

$$\frac{\partial R_2}{\partial p_2} = \frac{1}{a} \left[1 - \left(1 + u + u^2 - \frac{u^3}{z} + \frac{u^2}{z^2}(1 + z)\left(1 - \frac{2u}{z}\right)\right) e^{-u} \right] \quad (35)$$

Solving the above system then gives the equilibrium: $u^* = 1$, and $z^* = 2.5227$, which is the only real solution of $(3 - e)z^3 - z - 2 = 0$.

The equilibrium prices of the two providers are thus

$$p_1^* = -1.5227d\frac{1}{b} \geq 0, \quad (36)$$

$$p_2^* = ap_1^*. \quad (37)$$

Note that both equilibrium prices are therefore non-negative, since here, $d = d_2 - d_1 \leq 0$.

IV. TWO TIER PRICING

Let us now consider the question of two-tier pricing, as is practiced by the provider described in Section 2. Suppose that provider 2 now charges using a two-tier, or burst-rate, pricing structure, where

$$p_2(x) = \begin{cases} p_2x, & \text{if } x \leq T, \\ q_2x, & \text{otherwise,} \end{cases}$$

and $q_2 > p_2$. That is, provider 2 still charges a (linear) usage-based fee but the burst rate will be higher when the usage level is above a given threshold T .

With respect to Section 2, T here represents the 95 percentile of total period usage, referred to as α for a particular user. Since we consider here the entire distribution of such users, x , this model represents at an aggregate level the price structure model of Section 2.

Assume that provider 1 still charges a flat fee, p_1 , and that the delays are the same as in the previous section, i.e. $d_2 - d_1 \geq 0$.

For a user characterized by two-tuple (x, w) , his strategies can be summarized by the following three cases:

- 1) If $T \geq \frac{p_1}{p_2}$, a user would choose provider 1 if either $x \geq \frac{p_1}{p_2}$ or $x < \frac{p_1}{p_2}$ and $w \geq \hat{w}(x)$, where $\hat{w}(x)$ is given by (6) as before. In this case, the revenues of the two providers R_1 and R_2 are the same as in (4) and (5), and are independent of q_2 .
- 2) If $\frac{p_1}{p_2} > T \geq \frac{p_1}{q_2}$, users would choose provider 1 if either $x \geq T$ or $x < T$ and $w \geq \hat{w}(x)$. Then, the corresponding revenues are

$$R_1(p) = p_1 \left[\int_T^{\infty} \int_0^{\infty} \mu(x, w) dw dx + \int_0^T \int_{\hat{w}(x)}^{\infty} \mu(x, w) dw dx \right] \quad (38)$$

$$R_2(p) = p_2 \int_0^T \int_0^{\hat{w}(x)} x\mu(x, w) dw dx, \quad (39)$$

if $\frac{p_1}{q_2} > T$, users would choose provider 1 in one of the following three sub-cases: i) $x \geq \frac{p_1}{q_2}$; ii) $T \geq x$

and $w \geq \hat{w}(x)$; and iii) $T \leq x < \frac{p_1}{q_2}$ and $w \geq \hat{w}_q(x)$, where

$$\hat{w}_q(x) = \frac{p_1 - q_2x}{d}. \quad (40)$$

The corresponding revenues are given by

$$R_1(p) = p_1 \left[\int_{\frac{p_1}{q_2}}^{\infty} \int_0^{\infty} \mu(x, w) dw dx + \int_T^{\frac{p_1}{q_2}} \int_{\hat{w}_q(x)}^{\infty} \mu(x, w) dw dx + \int_0^T \int_{\hat{w}(x)}^{\infty} \mu(x, w) dw dx \right], \quad (41)$$

$$R_2(p) = p_2 \int_0^T \int_0^{\hat{w}(x)} x \mu(x, w) dw dx + q_2 \int_T^{\frac{p_1}{q_2}} \int_0^{\hat{w}_q(x)} x \mu(x, w) dw dx. \quad (42)$$

A. Uniform distributed QoS

As before, we begin with the simplest model based on a single usage level, x , and assume that the distribution of delay-cost tradeoff constants, w , is uniform on the interval $[0, 1]$. Assume still that $d_2 - d_1 \geq 0$.

Note that when $x \leq T$, the user strategy is independent of q_2 . By checking through the three cases introduced above, one can easily identify that provider 2 will obtain the portion $\hat{w}(x)$ of the market when $p_1 \geq p_2x$, and 0% of market otherwise. In this case, the revenues of the two providers are the same as in Section III-A, hence the same equilibrium exists as before, namely,

$$p_1^* = 2d/3, \quad \text{and} \quad p_2^* = d/3x$$

and $w^* = \frac{1}{3}$.

When $x > T$, however, the unit price using Provider 2 will be q_2 . So provider 2 will obtain the portion $\hat{w}_q(x)$ of the market when $p_1 \geq q_2x$, and 0% of the market otherwise. Hence

$$R_1 = p_1(1 - \hat{w}_q(x)) = p_1 \left[1 - \frac{p_1 - q_2x}{d} \right], \quad (43)$$

$$R_2 = q_2x\hat{w}_q(x) = q_2x \left[\frac{p_1 - q_2x}{d} \right], \quad (44)$$

which can be easily solved to obtain

$$p_1^* = 2d/3, \quad \text{and} \quad q_2^* = d/3x$$

and $w^* = \frac{1}{3}$.

That is, assuming the delay-cost tradeoff w is uniformly distributed, the multi-tier pricing structure does not change the Nash equilibrium.

This implies the following, perhaps counter-intuitive result: it may not be worth the effort for a firm to engage in convincing users to subscribe to two-tiered prices, if customer willingness is already low, since equilibrium

profits will not be higher with that more complex price structure. Naturally, if there are other reasons for using the burst-rate structure, such as obtaining a priori estimates of customer usage levels, e.g. for capacity planning, these would have to be weighed with the simplicity gained from eliminating the two tiers. Other customer preference-revealing methods may also exist without resorting to a two-tiered structure, as well.

V. PERSPECTIVES

We have proposed a model for analyzing markets for electronic goods, that takes into account the stochastic nature of user demand, as well as the spread of tradeoffs between cost and quality of service across the population of end users. As a by product of our general model of a firm selling electronic services, we demonstrated that the pricing of e-services need not result in a ruinous game, as suggested by [3]. We also demonstrated that the nature of the market equilibrium, in terms both of prices and of market share, depends heavily on the assumptions made on user behavior. The simplified model of uniformly-distributed value of QoS parameters, for example, leads to a fundamentally different conclusion than the more well-motivated exponentially distributed value.

As is practiced by a major web hosting firm, we considered single-tier versus two-tier (burst-rate) pricing strategies. We showed that the equilibrium that results, under the simplified assumption of uniformly-distributed values of QoS, is identical to the single-tier equilibrium, which leads one to the preliminary conclusion that the more complex burst-rate pricing, if it is not embraced by users, may not be worth the effort, in terms of profits. This conclusion is preliminary, however, and a further numerical study of this model in the presence of exponentially- or log-normally-distributed values and/or other measures of QoS is a worthwhile subject for future research, in particular, for testing that initial result. Furthermore, one advantage of the burst-rate structure is that it provides the supplier with the firms' apriori estimates of their usage levels, which has value in capacity planning.

Another interesting topic of future study would be incorporating usage-dependent values of QoS parameters, as the resulting model is significantly more complex, both theoretically and numerically, but would allow sensitivity analysis in terms of demand (and profitability) increases as a function of improved (or diminished) QoS.

Finally, it is of substantial interest to develop more complex definitions of QoS in this framework. One such effort in this direction is the recent work [2], which considers delay as a function of provider capacity, through

explicit queueing relationships. However, the complexity introduced by the capacity-delay dependencies prevents us from modeling the price structure complexities found in this work. That is, prices in the latter references are flat, or subscription-based, rather than usage-dependent. On the other hand, we consider additional complexity in [2] through a bilevel, or Stackelberg, framework that optimizes capacity decisions for a particular supplier, when prices are determined by Nash equilibria. Other more sophisticated definitions of QoS are envisageable as well: loss probability, reliability, delay variance (rather than expected value, etc.)

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Parijat Dube, of IBM Watson Research Center, for noticing a typo in a formula, and to an anonymous referee for numerous helpful comments which improved the presentation of this paper.

REFERENCES

- [1] L. Anania and R.J. Solomon. "Flat- The Minimalist Price", in Lee W. McKnight and Joseph P. Bailey, editors, *Internet Economics*, MIT Press, (1997) pp. 91–118.
- [2] P. Dube, Z. Liu, L. Wynter, and C. Xia, "Optimal capacity planning, QoS, and pricing in a competitive market", submitted.
- [3] P. C. Fishburn and A. M. Odlyzko, "Competitive pricing of information goods: Subscription pricing versus pay-per-use", *Economic Theory* 13 (1999), pp. 447–470.
- [4] R. Gibbens, R. Mason, and R. Steinberg, "Internet service classes under competition", *IEEE Journal on Selected Areas in Communications* 18, 12 (2000), 2490–2498.
- [5] R.J. Gibbens and F.P. Kelly, "Resource Pricing and the Evolution of Congestion Control", *Automatica* 35, (1999), pp. 1969–1985. Available from URL <http://www.statslab.cam.ac.uk/~frank/evol.html>
- [6] D. Hurlley, B. Kahin, and H. Varian, Eds. *Internet publishing and beyond: The economics of digital information and intellectual property*, MIT Press, Cambridge, MA, (1997).
- [7] F.P. Kelly, A.K. Maulloo, and D.K.H. Tan, "Rate Control for Communication Networks: Shadow Prices, Proportional Fairness, and Stability", *Journal of the Operational Research Society* 49 (1998) pp.237–252. Available from URL <http://www.statslab.cam.ac.uk/~frank/rate.html>
- [8] J. Mackie-Mason and H. Varian, "Pricing congested network resources," *IEEE Journal on Selected Areas of Communications* 13:7, 1141–1149, 1995.
- [9] J. A. van Mieghem, "Price and Service Discrimination in Queuing Systems – Incentive Compatibility of Gcu Scheduling", *Management Science* 46 (9) (2000) pp. 1249–1267.
- [10] A.M. Odlyzko, "Paris Metro Pricing for the Internet", *Proc. ACM Conference on Electronic Commerce (EC'99)*, ACM, 1999, pp. 140–147.
- [11] A.M. Odlyzko, "Internet pricing and the history of communications", *Computer Networks*, **36** (5–6) (2001) pp. 493–517.
- [12] G. Paleologo, IBM Watson Research Center, private communication.
- [13] S. Shenker, D. D. Clark, D. Estrin, and S. Herzog, "Pricing in Computer Networks: Reshaping the Research Agenda", *ACM Computer Communication Review* 26 (1996) pp. 19–43.
- [14] H. Varian, "Pricing Information Goods" , presented at the Research Libraries Group Symposium on "Scholarship in the New Information Environment", Harvard Law School, May 2–3, (1995). available at <http://www.sims.berkeley.edu/hal/people/hal/papers.html>
- [15] Q. Wand and J.M. Peha, "State-Dependent Pricing and its Economic Implications," *Telecommunication Systems* 18:4, 2001, 315–329.

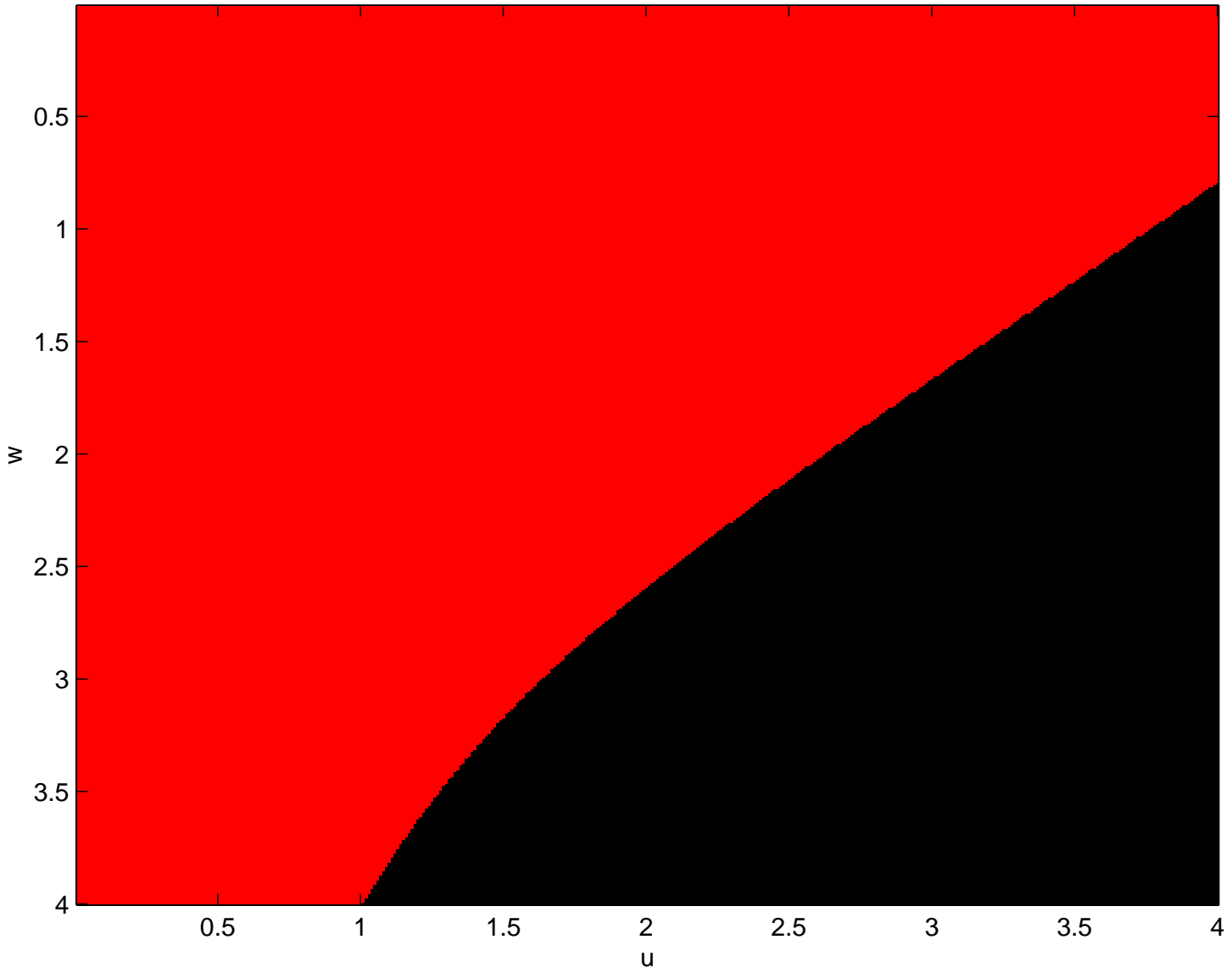


Fig. 3. The user should not declare a committed usage level of $\mu = B_\alpha$

Fig. 4. Zeros of each derivative function, $dR1$ and $dR2$ in the space of dummy variables (u, z)

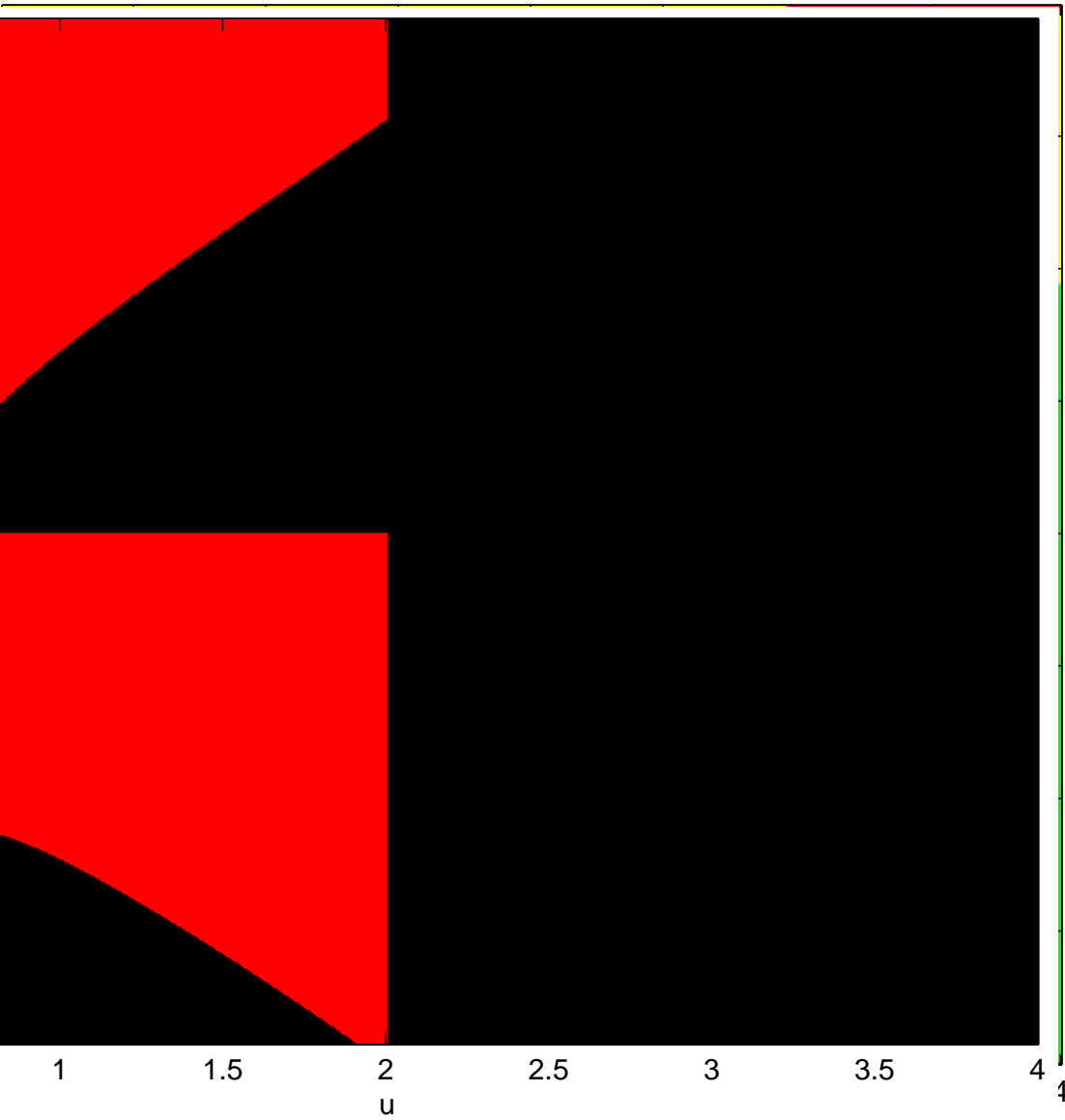


Fig. 5. The two plots of Figure 4 super-imposed to illustrate the interior Nash equilibrium solutions in the space of dummy variables (u, z)

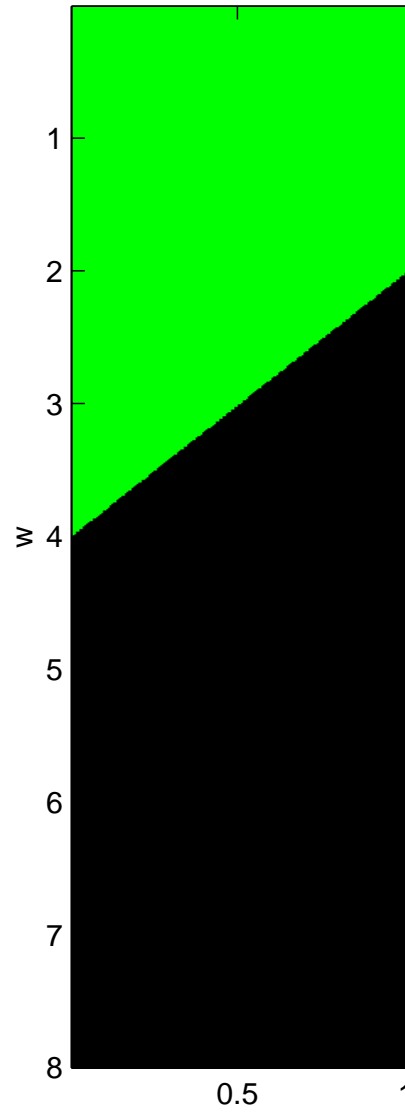


Fig. 6. Zeros of each derivative function, $dR1$ and $dR2$ on a larger area of the space (u, z)

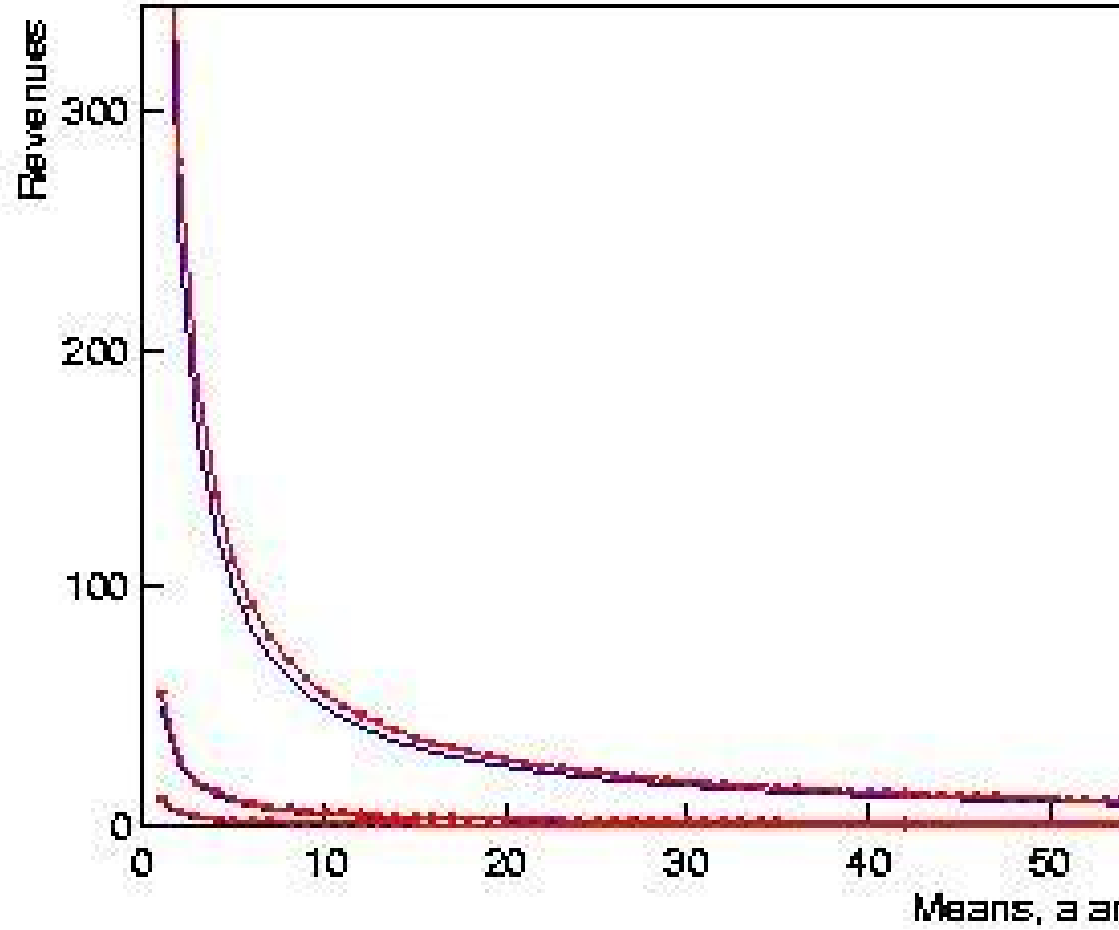
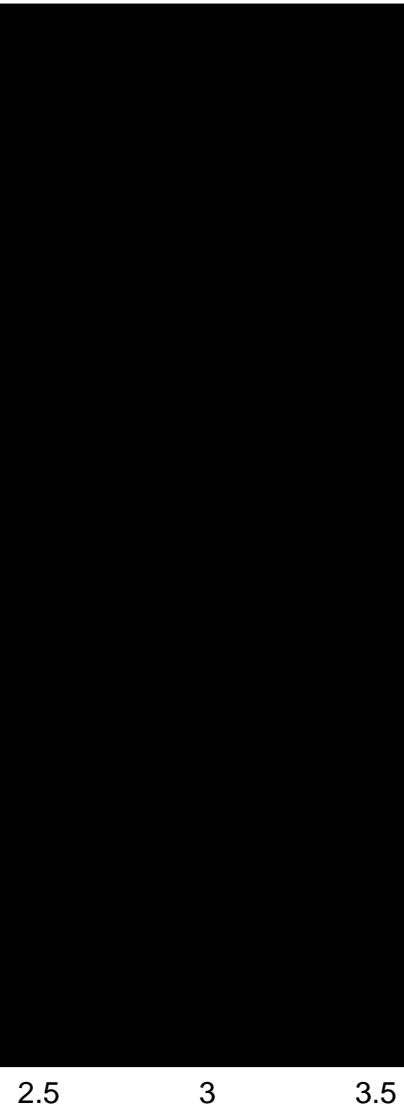


Fig. 7. Graphical examination of the solution space on a larger domain; no other equilibria appear

Fig. 8. Revenues of provider 1 (blue) and provider 2 (green) as d increases on the x-axis

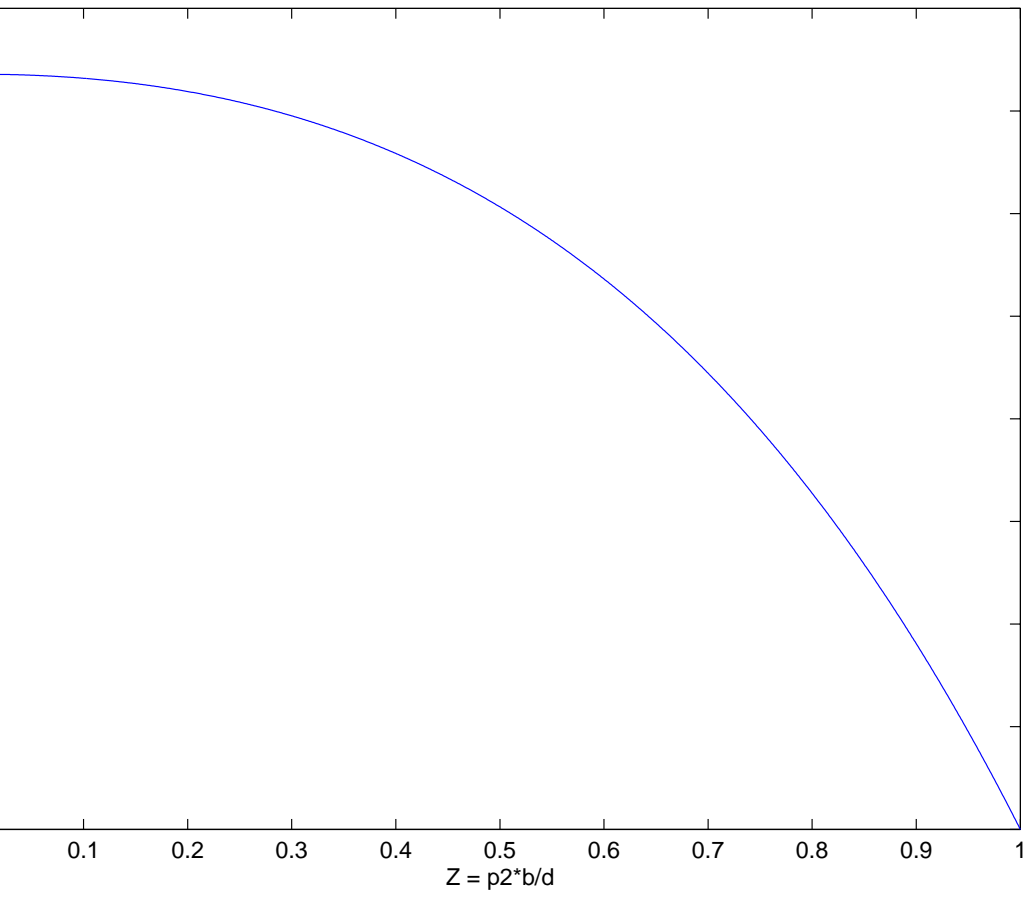


Fig. 9. There is no equilibrium solution in the case of both providers offering flat prices