ECONOMICS BULLETIN

Nonlinear pricing of a congestible network good

Khaïreddine Jebsi
LEGI, Ecole Polytechnique, La Marsa and Faculté de Droit et des Sciences Economiques et Politiques de Sousse, Tunisia

Lionel Thomas
CRESE, Université de Franche Comté, IUT Besançon Vesoul, Besançon, France

Abstract

This paper analyzes optimal nonlinear pricing of a congestible network good. In contrast to the traditional efficiency−on−the−top result in the standard screening model, we show that the presence of a delay cost borne by users leads to upward distortions in consumption for the high−demand users.

Citation: Jebsi, Khaïreddine and Lionel Thomas, (2005) "Nonlinear pricing of a congestible network good." *Economics Bulletin*, Vol. 4, No. 2 pp. 1−7
Submitted: November 30, 2004. **Accepted:** January 15, 2005.
URL: http://www.economicsbulletin.com/2005/volume4/EB−04D80013A.pdf

1. Introduction

In many networks, consumers' utility decreases with the number of users who are linked to the same network. Indeed, when these consumers share a fixed amount of capacity, the increase in the aggregate consumption decreases the utility of users (i.e. there is a negative externality).[1] This congestion externality characterizes, among others, the delay systems in which all users receive a service at a quality that degrades with aggregate consumption.[2]

For this type of systems, the delay can arise from the development of queues, flow congestion or crowding. We are interested in this paper in the delay that takes the form of a waiting time to satisfy the need of the user. This kind of delay concerns the industries in which the aggregate demand is served at the same time (all customers have the same service order) such as computer loads, electronic mail systems... The time that an E-mail arrives at one's destination and the duration for loading a file provide appropriate examples of this delay. In other terms, the delay is considered as the timing of delivery.[3] In this systems, each user bears a delay cost, interpreted as a congestion cost, that depends on the aggregate consumption and on the amount of the shared capacity.

The textbook congestion pricing which allows an efficient use of the resource shows that each user should use the system until the marginal benefit from his usage equals the marginal cost of production plus the marginal cost of congestion that he imposes on other users. Scotchmer (1985) examines the implications of this rule to determine the optimal number of firms that supply shared facilities having a natural fixed size, such as golf courses, ski scopes, etc... As for MacKie-Mason and Varian (1995), they consider resources such as an ftp server, a router, a Web site, etc... and explore the implications of this textbook pricing for capacity expansion in centrally planned, competitive, and monopolistic environments. These analyses are considered in a context of complete information.

Our scope in this paper is to analyze the implications of the textbook congestion pricing when a monopoly offering a delay system with a single capacity,[4] shared by some users who bear a delay cost, faces asymmetric information on consumers' preferences.[5]

Comparing the optimal consumption in complete information with the optimal consumption in incomplete information, we show that the standard under-consumption in incomplete information does not hold anymore.

Generally, in nonlinear pricing, leaving a surplus to consumers is costly to the monopoly. The latter reduces this rent by inducing users to consume much less than their first-best optimal quantities (see Maskin and Riley,1984). Thus the aggregate consumption is lower in asymmetric information. This has important implications in the context of congestible

[1] We use the term aggregate consumption to designate the total use of the network resource.

[2] We are not interested in this paper in the loss systems which provide an all-or-nothing service at a constant quality such as electricity and natural gas distribution. For this type of systems see Wilson (1989).

[3] We are not interested in the delay that consists in delivery delays due to queues. This concerns the industries in which demand is backlogged in queues for service such as a service industry in which each worker serves one customer at a time or the industries of the service sector that rent capacity or servers to customers... . For this kind of delay see also Wilson (1989).

[4] We consider a vertically integrated monopoly, that is, we omit the interdependence among congestible elements in a network.

[5] Oren et al. (1985) propose a model of nonlinear pricing for both usage and capacity where the customers' demand is represented by the load duration curve. They show that the capacity charges induce the buyers to truncate their purchase sets in order to reduce their charge for capacity.

network resources. Indeed, when the amount of the capacity is fixed, it implies that the level of congestion and its marginal cost are greater in complete information. So, a new phenomenon appears: a feedback-in-demand; the decrease of consumption of the low-demand consumers leads to a reduction of congestion inciting the high-demand users to increase their consumptions above their first-best quantities. So, the optimal nonlinear tariff leads to upward distortions in the consumption of the high-demand users. The well-known efficiency-on-the-top result in the standard screening model no more holds when we consider a congestible network good.[6]

In section 2, we describe the model. In section 3, we study the properties of the optimal pricing in symmetric and in asymmetric information in order to analyze the refereeing between rent and efficiency.

2. The model

A monopolistic firm produces a congestible network good with a fixed amount of capacity $K > 0$. To sell it, it designs a nonlinear pricing for a continuum of better-informed consumers. Consumers have taste (i.e.: the type) for the good characterized by $t, t \in \left[t_L, t^H\right]$. The firm is not able to observe a given consumer's type, but has prior beliefs over the distribution summarized by a density function $g(.) > 0$. The cumulative distribution function is noted $G(.)$. Let $\Gamma(t)$ be the inverse of the hazard rate.

Consumer $t \in \left[t_L, t^H\right]$ who consumes $x > 0$ units of the resource gets a gross surplus $S(t, x) > 0$. Let subscript j denote partial derivative with respect to the j^{th} argument. We make the following standard assumptions. Gross customer's surplus is an increasing concave function; $S_2(.) > 0; S_{22}(.) < 0$. Higher taste results in higher surplus for a given quantity; $S_1(.) > 0$. The gross surplus function is also characterized by the single crossing property; $S_{12}(.) > 0$, $\forall t, \forall x$.

Moreover, if user t consumes $x(t)$ units of the resource, the aggregate consumption is:

$$X = \int_{t_L}^{t^H} x(t)g(t)dt \tag{1}$$

Since the capacity is shared by the consumers, the latter suffer from congestion and its level is given by the ratio $\frac{X}{K}$.[7] This ratio measures the intensity of use of the network resource. To simplify the analysis, we assume that the users have the same unit cost of waiting (or the same willingness to pay in order to avoid congestion), so that, they bear the same delay cost due to congestion. This delay cost is given by the function $D(\frac{X}{K})$ and it is supposed to be common information, increasing and convex.

Like Scotchmer (1985) and MacKie-Mason and Varian (1995) we assume that the consumers have separable preferences. So, the utility function is:

$$V(t) = S(t, x(t)) - P(x(t)) - D\left(\frac{X}{K}\right)$$

where $P(x(t))$ is the payment associated to the consumption of units. We assume that each user does not take into account his own contribution to congestion when deciding

[6] The efficiency-on-the-top is also violated in other microeconomic contexts: when we take into account the call externality in the nonlinear tariff of a two-way telecommunication service market (Hahn 2003) or when a positive externality depends on the type of the user (Sundararajan 2004).

[7] See Reitman (1991) for the formulation of congestion functions of different forms of congestion in particular for the processor sharing.

his optimal quantity to consume. So, he consumes a quantity such as:

$$S_2(t, x(t)) = P'(x(t)) = p(x) \tag{2}$$

where $p(x)$ is the marginal price which the user must dispense to get an extra unit of the good.

The monopoly constructs a menu of optional tariffs $\left\{x(\widetilde{t}), P(x(\widetilde{t}))\right\}$ which offers consumer $\widetilde{t}$ the quantity $x(\widetilde{t})$ at price $P(x(\widetilde{t}))$.

Without loss of generality, we assume that the service supply cost is zero. So, the problem of the monopoly is to:

$$\max_{x(.),P(.)} \int_{t_L}^{t^H} P(x(t))g(t)dt$$

subject to: (1) and $\forall t, \widetilde{t} \in t \in \left[t_L, t^H\right]$

PC: $V(t) \geq 0$

IC: $V(t) \equiv V(t,t) \geq V(\widetilde{t}, t) = S(t, x(\widetilde{t})) - P(x(\widetilde{t})) - D\left(\frac{X}{K}\right)$

PC guarantees the participation of consumer . IC is for truth-telling by the agent. From the well-known standard arguments (see Guesnerie and Laffont 1984), the problem can be reformulated as follows:

PC and IC are satisfied if $V'(t) = S_1(t, x(t))$ and $V(t_L) = 0$, so the problem of the monopoly leads to:

$$\max_{x(.)} \int_{t_L}^{t^H} \left[S(t, x(t)) - \Gamma(t) S_1(t, x(t))\right] g(t)dt - D\left(\frac{X}{K}\right)$$

subject to $x'(t)$ and (1), where $\Gamma(t)S_1(t, x(t))$ is the expected informational rent of user t.

In order to ensure that an optimal fully separating pricing exists (i.e.: different types are served at different quantities), we make the following assumptions. The firm's revenue, $S(t, x(t)) - \Gamma(t)S_1(t, x(t))$, is strictly concave in $x, \forall t$. The expected marginal rent $\Gamma(t)S_{12}(t, x(t))$ is non increasing in t, and $S_{211}(t, x) \leq 0, \forall x, \forall t.$[8]For the sake of simplicity, we assume that K is sufficiently high so that all types, even t_L, are willing to purchase.

3. The properties of the optimal pricing

Before presenting the optimal selling strategy of the monopoly, let us briefly remind the benchmark result of Maskin and Riley (1984) in the case of a non-congestible good. The optimal selling strategy is such that every type (except for those with the strongest preference) faces a higher marginal price in a context of incomplete information. So users consume less than as in the full information setting. Indeed, in the second-best allocation, the monopoly seeks to reduce informational rent for consumers with strong willingness to pay.

Now, we present the optimal pricing for a good that presents congestion.

3.1 Optimal pricing

As we will see, the presence of asymmetric information don't modify the principle of the textbook congestion pricing, but it has a direct effect on the surplus and an indirect one on the congestion.

[8] Standard specifications of the function S(.) and g(.) lead to check these assumptions.

Proposition 1 *Optimal pricing for a congestible good is such as:*

$$p(x) = D'\left(\frac{X}{K}\right)\frac{1}{K} \text{ \textit{in complete information}}$$

$$p(x) = \Gamma(t)S_{12}(t,x) + D'\left(\frac{X}{K}\right)\frac{1}{K} \text{ \textit{in incomplete information}}$$

Proof. See appendix A. ■

As it might be expected, in complete information, optimal pricing implies that the marginal surplus of each consumer equals the marginal cost of congestion. We obtain then the textbook congestion pricing, which gives the first-best optimal quantity. It tells us that the consumer uses the resource until the marginal utility from his usage equals the marginal cost of congestion that he imposes on other users.

In a context of incomplete information, the marginal price is distorted in such a way that virtual marginal surplus, that is marginal surplus minus expected marginal informational rent of user, becomes equal to the marginal cost of congestion. The firm modifies its marginal pricing so as to take into account the informational rent that it must leave to users.

With the presence of the delay cost, the asymmetric information (hereafter AI) has then a direct effect on the surplus and an indirect one on the congestion. First, the price depends on the virtual marginal surplus which influences the quantity consumed by the users (direct effect). The individual quantity is thus adjusted according to the first-best quantity, and so is the aggregate consumption. Second, this change of the aggregate consumption modifies the level of the congestion and its marginal cost (indirect effect).

Having observed the two effects, it seems interesting to study the properties of the pricing under symmetric and asymmetric information in order to analyze the distortions between first and second-best allocations. For the rest of the paper, we adopt the following notations: subscripts i and c denote respectively the optimal values in incomplete and in complete information.

3.2 Comparison of the equilibrium allocations

Before comparing the equilibrium allocations, let us compare the congestion levels.

Proposition 2 *The congestion level is reduced in the incomplete information setting.*

Proof. See appendix B. ■

This proposition shows that the aggregate consumption decreases in the incomplete information setting. The well-known interpretation of this result is that the seller tends to reduce consumption in order to limit informational rent. His aggregate supply is then reduced in the incomplete information setting. But, since the capacity is fixed, the level of congestion is greater in full information. It implies that the delay cost borne by the users is higher under symmetric information, and so is the marginal cost of congestion.

Let us note the difference between the marginal costs of congestion by:

$$\Delta MC = \frac{1}{K}\left[D'\left(\frac{X_c}{K}\right) - D'\left(\frac{X_i}{K}\right)\right] > 0 \tag{3}$$

We can present the following proposition.

Proposition 3 *The distortion between the first and the second best allocations is expressed as:*

$$p(x_c(t)) \gtreqless p(x_i(t)) \text{ if } t \gtreqless t^*$$

$$\text{with } t^* \in \left]t_L, t^H\right[\text{ such that } \Gamma(t^*)S_{12}(t^*, x(t^*)) = \Delta MC$$

Proof. For a given t, the use of proposition 1 gives:

$$p(x_c(t)) \gtreqless p(x_i(t)) \Leftrightarrow \Gamma(t)S_{12}(t, x_i(t)) \lesseqgtr \Delta MC > 0$$

where the last inequality follows from (3). Since $\Gamma(t)S_{12}(t, x_i(t))$ is monotonic, t^* is unique and must exist since $X_c > X_i$. ■

The presence of congestion implies over-consumption for customers with strong preferences with regard to their consumptions under complete information. It is explained by the occurrence of a feedback-in-demand. Because the informational rent is costly to the monopoly, this one tends to reduce the quantity more for low-t consumers (direct effect of AI). This decrease of the consumption, involving a reduction of the congestion level, incites the high-t consumers to increase their demands above their first-best levels (indirect effect of AI). Hence, if the expected marginal informational rent, i.e.: $\Gamma(t)S_{12}(t, x_i(t))$, is lower than the difference of the marginal costs of congestion ΔMC, the user gets a greater marginal gross surplus in AI. Such a user must consequently consume a higher amount of units and the marginal price must be lower in the asymmetric information setting.

As the expected marginal rent decreases with the willingness to pay, the highest types are served at a lower marginal price. On the other hand, the lowest types receive a higher amount of units in complete information. The type served at the first-best level is thus strictly interior when expected marginal rent exactly offsets the difference of the marginal costs of congestion.

4. Conclusion

We have shown in this paper that the well-known efficiency-on-the-top in the standard nonlinear pricing models no more holds when we consider a congestible network good. The presence of a delay cost borne by users leads to a feedback in demand. The high demand users consume above their first-best quantities and the type served at his first best level is strictly interior.

Nevertheless, in our model, it has been assumed that all users have the same unit cost of waiting. The analysis could be extended till taking into account the differences among users concerning their unit costs of waiting. A multidimensional asymmetry of information clearly appears.[9] Moreover, the model developed here assumes a given time during which all users simultaneously consume. More generally delay may depend on peak utilization or the variance of utilization. There is then another possible extension of the model.

[9] Jebsi and Thomas (2004) have determined the optimal pricing for a congestible good when the user has unit-demand and his unit waiting cost is perfectly correlated to his valuation.

Appendix A

In order to use standard control theory optimization, let us introduce the following variable $z(t) = \int_{t_L}^{t} x(s)g(s)ds$, so that

$$z(t_L) = 0 \; ; \; z(t^H) = X \tag{A.1}$$

and by differentiation:

$$z'(t) = x(t)g(t) \tag{A.2}$$

Then, (1) is replaced by (A.1) and (A.2).

Ignoring for instance the constraint $x'(t) \geq 0$, we have an optimal control problem with a scrap value where $z(t)$ is a state variable and $x(t)$ is a control. The Hamiltonian is:

$$H(t) = [S(t, x(t)) - \Gamma(t)S_1(t, x(t))] \, g(t) + \lambda(t)x(t)g(t)$$

where $\lambda(t)$ is the costate variable associated to the state.

Necessary conditions are (see Seierstadt and Sydsaeter, theorem 3, page 182):

$$[S_2(t, x(t)) - \Gamma(t)S_{12}(t, x(t))] \, g(t) + \lambda(t)g(t) = 0 \tag{A.3}$$

$$\lambda'(t) = 0 \tag{A.4}$$

$$\lambda(t^H) = -D'\left(\frac{X}{K}\right)\frac{1}{K} \tag{A.5}$$

From (A.4) and (A.5), $\lambda(t) = -D'\left(\frac{X}{K}\right)\frac{1}{K}, \forall t$. So (A.3) gives:

$$S_2(t, x(t)) = \Gamma(t)S_{12}(t, x(t)) + D'\left(\frac{X}{K}\right)\frac{1}{K} \tag{A.6}$$

From Seierstadt and Sydsaeter, theorem 4, page 182, we check that necessary conditions are sufficient because $D(.)$ is convex and $H(t)$ is independent in $z(t)$ and strictly concave in $x(t)$. Last, by totally differentiating (A.6), we can make sure that the ignored constraint is verified under assumption of the model.

Solving the complete information problem follows the same arguments except that $V(t) = 0, \forall t \in \left[t_L, t^H\right]$. So the expected informational rent is null and (A.6) becomes:

$$S_2(t, x(t)) = D'\left(\frac{X}{K}\right)\frac{1}{K}$$

Using (2) completes the proof.

Appendix B

Adopting our convention on subscripts i and c, we check that $X_i \geq X_c$ is not feasible. $X_i \geq X_c$ implies, from the convexity of the disutility, that:

$$D'\left(\frac{X_i}{K}\right)\frac{1}{K} \geq D'\left(\frac{X_c}{K}\right)\frac{1}{K}$$

Proposition 1 implies that:

$$p(x_i(t)) - p(x_c(t)) \geq 0 \tag{B.1}$$

since $\Gamma(t)S_{12}(t,x(t)) \geq 0$.

From (2), $p'(x) = P''(x) = S_{22}(t,x) < 0$, so (B.1) implies that:

$$x_c(t) > x_i(t) \forall t \text{ if } X_i > X_c$$

or

$$x_c(t) \geq x_i(t) \forall t \text{ (with equality holding at } t^H\text{) if } X_i = X_c$$

These conditions lead to a contradiction with the initial assumption $X_i \geq X_c$. Hence, we must have $X_c > X_i$ and so, $\frac{X_c}{K} > \frac{X_i}{K}$.

References

Guesnerie, R. and J.J. Laffont (1984), "A Complete solution to a class of principal-agent problems with an application to the control of a self-managed firm" *Journal of Public Economics* **25**, 329-369.

Hahn, J-H. (2003) "Nonlinear pricing of telecommunications with call and network externalities" International Journal of Industrial Organization 21, 949-967.

Jebsi, K. and L. Thomas (2004) "Optimal pricing for selling a congestible good with countervailing incentives" *Economics Letters* **83**, 251-256.

MacKie-Mason, J. and H. Varian (1995) "Pricing congestible network resources" *IEEE Journal of Selected Areas in Communications* **13**, 1141-1149.

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

Oren, S., Smith, S. and R. Wilson (1985) "Capacity pricing" *Econometrica* **53**, 545-566.

Reitman, D. (1991) "Endogenous quality differentiation in congested markets" *The Journal of Industrial Economics* **39**, 621-647

Sundararajan, A. (2004) "Nonlinear pricing and type-dependent network effects" *Economics Letters* **83**, 107-113.

Scotchmer, S.(1985) "Two-tier pricing of shared facilities in a free-entry equilibrium" *Rand Journal of Economics* **16**, 456-472.

Seierstad, A. and K. Sydsaeter (1987) *Optimal control theory with economic applications,* North-Holland: Amsterdam.

Wilson, R., 1989, Efficient and competitive rationing, Econometrica **57**, 1-40.