THREE ESSAYS ON ECONOMIC ISSUES IN TELECOMMUNICATIONS

by

Douglas Robert Bergman

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF ECONOMICS

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2001

UMI Number: 3031370

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

THE UNIVERSITY OF ARIZONA ®
GRADUATE COLLEGE

As members of the Final Examination Committee, we certify that we have read the dissertation prepared by DOUGLAS ROBERT BERGMAN

entitled THREE ESSAYS ON ECONOMIC ISSUES IN TELECOMMUNICATIONS

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of DOCTOR OF PHILOSOPHY

John Wooders	7/13/01 Date
William Horrace	7/13/01 Date
Stanley Reynolds	7/13/01 Date
	Date
	Date

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copy of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Dissertation Director
John Wooders

7/13/01.
Date

STATEMENT BY AUTHOR

SIGNED: ______________________

Acknowledgement

Special thanks to Bill Horrace, Stan Reynolds, and Ed Zajac, my esteemed professor, and to John Wooders, my dissertation director and a real trooper, without whose inspiration, stewardship, and support, this would be even more of a mess than it is.

DEDICATION

This is dedicated to my parents, whose constant support, in many forms, made this possible.

Table of Contents

Table of Contents—*Continued*

List of Figures

List of Tables

ABSTRACT

In Chapter 1, we construct a model to illustrate conditions under which a government-controlled Post, Telephone, and Telegraph ministry (PTT), which is a monopolist in multiple demand-complementary markets, can increase its profit by exiting some of its markets. The PTT may increase profit by exiting one market, provided that a foreign firm can supply the market at lower cost and the PTT retains market power in a complementary market, where it recovers the difference. The PTT will generally earn greater income by regulating and taxing the abandoned market than by allowing it to become competitive. However, consumer welfare would be greater if the PTT were to permit competition in the abandoned market.

In Chapter 2, we use Heckman's model for consistent estimation on selected data, modified to allow for group dummies in the second-stage regression, to estimate supply of wireless telephone subscription, on a panel of data. The modification enables us to control for country-specific effects, and to adjust for the penetration that would otherwise exist in years and countries where wireless is unavailable. We find substantial bias in the estimate of the supply function. That is, equivalent economic conditions in countries where wireless is not yet available will likely result in lower levels of supply than those where wireless is available. The quantity of wireless telephones supplied is explained by time, the number of existing fixed telephone lines, and telephone company revenues, but not by prices.

In Chapter 3, we construct a means of resource allocation on a data network when bandwidth becomes scarce. Our approach extends Elwalid and Mitra's (1992) model so that two users may send streams of information to a router, which employs an

auction mechanism to award priority to one stream when the router is congested. The high bidder in this auction enjoys the right to transmit data without risk of loss, whereas the low bidder loses data during congested periods. The second-price auction offers its property of incentive compatibility in this real-time framework. Allocations arising from this mechanism are more economically efficient than those in which information is discarded without regard to economic value.

Chapter 1

The Benefit of a Competitive Equipment Market to a Government-Controlled Telecommunications Monopoly

1.1 Introduction

In many countries, all communications services are owned by a self-regulating government agency, frequently referred to as a Post, Telephone, and Telegraph Ministry (PTT). Until recently, the typical PTT operated an antiquated, high-cost, and low-penetration network as a monopoly. Moreover, as a regulator, the PTT can enforce barriers to entry. The purpose of this paper is to provide an explanation for the surprising fact that PTTs around the world have been welcoming competition, particularly in the form of cellular communications services.

We use a model to show that the PTT actually benefits by exiting one of the markets it had been serving as a monopoly. This is the case whether the PTT is replaced in the exited market by a firm in which the PTT regulates price or by a competitive market. In the first alternative, the PTT exploits the regulated firm's lower costs and recovers its profit with a tax. In the second, the competition results in a lower price for the product, which complementarily enhances demand for the PTT's other products. We find that the PTT will always prefer regulating the exited market to allowing competition, but that consumers will prefer the competitive market to regulation. We also find that if the PTT is replaced by another unregulated monopolist in the exited market, it will necessarily be worse off than if it had allowed competition.

1.1.1 Theories of Deregulation

Peltzman [1976] and Becker [1976] propose that agencies such as PTTs have deregulated themselves chiefly as the result of political pressure for better service. An alternative explanation is cited by Conner [1992], Griffin [1982], Rohlfs [1978], and Taylor [1994], who note that an increased network externality provides the PTT with an incentive to keep the prices of some services low. Our results are independent of either of these arguments. Rather, we consider the possibility that PTTs have begun to exploit the low cost and ease of distribution of new wireless technology simply to increase their own profits. As the cost of wireless communications has fallen to accessible levels, PTTs have been permitting well-capitalized and technically competent foreign firms to create and distribute the new technology, in order to enhance demand for the services over which they retain monopoly control. Because the financial and organizational costs of handing a customer a pocket-sized telephone over a retail counter are likely lower than those of installing wiring between his house and a central switching office, waiting lists for telephones that had been years long in many countries have been reduced to days or even minutes. This larger customer base results in higher demand for more easily scaleable services, such as international calls, for which PTTs tend to retain monopolies.

1.1.2 Stylized Facts

In the last several years, the erosion of trade barriers around the world and the growth of market economies have caused the demand for communications services to increase dramatically. Political actions to introduce competition into communications services markets, which would assuredly benefit consumers and the numerous industries dependent upon communications services, have increased in strength. Examples include the passage of the Telecommunications Act of 1996, the World Trade Organization's Agreement on Basic Telecommunications Services [1997], and the introduction of nearly complete deregulation in Australia, Britain, Chile, Guatemala, Hong Kong, and New Zealand, as observed by Spiller and Cardilli [1990].

This situation seems to be the case in many other countries as well. Sunarno [1997] describes the case of Telekomunikasi Indonesia (TELKOM), the Indonesian PTT, which had a telephone line penetration rate in 1996 of 1.7 per 100 citizens. In order to expedite its growth, TELKOM has chosen to form partnerships with foreign investors experienced in the provision of wireless and satellite services, including France Telecom, US West, Telstra (Australia), Nippon Telephone & Telegraph, Telekom Malaysia Berhad, and Singapore Telecom. While TELKOM's plans for growth in many service sectors are extensive, it retains regulatory control and substantial stock ownership in each enterprise. By retaining regulatory power and partial ownership in monopolies across several markets, TELKOM will be able to control prices. Indeed, TELKOM has announced that it will permit competition in some markets for the very purpose of enhancing demand in its monopolized markets. The ministry cites increased and improved services, as well as profit enhancement, as its objectives.

Kubasik [1997] notes that the emerging countries of Eastern Europe are also expanding their networks by taking similar, albeit more decentralized, courses of action. Since 1990, Poland has been selling its entire telecommunications infrastructure to private investors and is welcoming competition in many sectors, including local wireline, long distance, and mobile services. The Czech Republic's PTT has encouraged competition in areas of low service density – presumably the same areas in which provision of additional wiring and equipment is most costly. In each of these countries, the PTTs retain the lucrative privilege of providing international service, which they will continue to serve as monopolies.

We model this phenomenon by first considering the scenario that the PTT acts as a monopolist in markets for two complementary products, which one might think of as telephone wireline equipment and installation, and minutes of international calls. We then compare this to three alternative scenarios in which the PTT relinquishes its presence in the equipment market, where its costs are high.

In one such alternative, which reflects the case of Indonesia, the PTT regulates a modern foreign supplier in the market. In this case, the PTT enjoys the greatest possible surplus, essentially by recovering the profit in both markets, while exploiting the foreign supplier's low costs. In the next alternative, modeled after Poland and the Czech Republic, the PTT allows foreign suppliers to compete in the equipment market. The PTT benefits in this case because the resulting low price in the equipment market enhances demand in the market for international calls. Finally, we consider the case in which the PTT is replaced in the equipment market by an unregulated monopolist with low costs. We find that the PTT's surplus is highest in the Indonesian case, while consumers are best off in the Eastern European case.

1.2 Model

Our model considers two goods, "Equipment" and "Service." A unit of equipment is one connection to a customer plus his or her customer premises equipment, and could be a wireline, wireless, or other type of connection and telephone. We assume that any of these is a perfect substitute for any other. We can think of a unit of service as a telephone call billed for a unit of time.

Consider the following symmetric system of demand functions that relate the prices p_e and p_s, of equipment and service, respectively, to the quantities q_e and q_s demanded of each:

$$
\begin{aligned}
q_e(p_e, p_s) &= A - \alpha p_e - \beta p_s \qquad (1.1)\\
q_s(p_e, p_s) &= A - \alpha p_s - \beta p_e
\end{aligned}
$$

where α and β are positive. The demand system is symmetric for the sake of tractability. This system defines relationships between the quantities and the prices in the service (minutes of telephone calls billed) and the equipment (wiring and CPE) markets, respectively. Alternatively, we can imagine the equipment price p_e as a monthly line charge.

Then (1) defines a system in which service and equipment are complements; that is, a higher price for either good reduces the quantity demanded for the other.

The PTT can produce units of equipment at a constant marginal cost c_e^P. One or more foreign firms can produce units of equipment at a constant marginal cost c_e^F. To focus the analysis on the issue at hand, the marginal cost of service is zero.

We make the following assumptions:

ASSUMPTION 1.1 Each commodity's own-price effect exceeds its cross-price effect; i.e.

$$\alpha > \beta.$$

As we shall see, Assumption 1.1 ensures that a solution to the PTT's profit maximization problem exists and has a maximum.

ASSUMPTION 1.2 Marginal costs are not too high; i.e.

$$\begin{aligned} 0 &< c_e^P < \frac{A}{\alpha} \\ 0 &< c_e^F < \frac{A}{\alpha}. \end{aligned}$$

Assumption 1.2 ensures that the relevant profit-maximizing prices yield quantities greater than zero.

In the interest of brevity, we shall neglect any network externality component; that is, a demand parameter for the number of telephones connected to the network. By including it our results would likely be strengthened, as suggested by work on the network externality by, e.g., Griffin [1982] and Rohlfs [1978].

We now consider four separate economic regimes and compare the PTT's welfare in each. In Case (i), the PTT retains monopolies in both the equipment and service markets. In Case (ii), the PTT retains its monopoly in the service market, and exits the equipment market but continues to regulate a foreign equipment supplier. In Case (iii), the PTT retains its monopoly in the service market, and exits the equipment market but allows it to become competitive. In Case (iv), the PTT retains the service market, and exits the equipment market, which is then served by another monopolist.

Case (i). PTT Holds Monopolies in Both Markets

Let us assume that the PTT initially is a single, integrated firm that holds monopolies in both the service and the equipment markets. The PTT's problem, then, is to choose prices (p_e, p_s) for equipment and service, respectively, that maximize its profit

$$\begin{aligned}\Pi_i(p_e, p_s) &= (p_e - c_e^P) q_e(p_e, p_s) + p_s q_s(p_e, p_s) \\ &= (p_e - c_e^P)(A - \alpha p_e - \beta p_s) + p_s(A - \alpha p_s - \beta p_e), \end{aligned} \tag{1.2}$$

where service is provided at zero marginal cost. The first term in the sum is the profit from the equipment market; the second term is the profit from the service market.

PROPOSITION 1.1. The PTT's profit-maximizing prices (p_e^i, p_s^i) when it serves both the equipment and the service markets as a monopoly are

$$(p_e^i, p_s^i) = \left(\frac{1}{2}\left(\frac{A}{(\alpha+\beta)} + c_e^P\right), \frac{A}{2(\alpha+\beta)}\right). \tag{1.3}$$

The PTT's corresponding levels of output are

$$\begin{aligned} q_e^i &= q_e(p_e^i, p_s^i) = \frac{1}{2}(A - \alpha c_e^P) \\ q_s^i &= q_s(p_e^i, p_s^i) = \frac{1}{2}(A - \beta c_e^P) \end{aligned}$$

and profit is

$$\Pi_i(p_e^i, p_s^i) = \frac{1}{4\alpha}(A - \alpha c_e^P)^2 + \frac{A^2}{4\alpha}\left(\frac{\alpha-\beta}{\alpha+\beta}\right).$$

Case (ii). PTT Regulates Equipment Market

Now suppose the PTT learns that another firm - say, a foreign cellular telephone operator - is able to provide equipment. Since the PTT is a government agency, we assume it has the power to regulate the equipment price, even when equipment is provided by another firm. Thus, the PTT chooses the price p_e of equipment sold by the foreign supplier.

Furthermore, suppose the firm has a simple cost function: zero fixed cost and the constant marginal cost c_e^F per unit of equipment. If the PTT decides to withdraw from the equipment market entirely, it now wants to maximize its own profit function for the service market. (Let us abstract away from the possibility that the PTT could remain in competition with other firms, yet retain the market power necessary to earn an economic profit in the equipment market. This again would cannibalize the service market, and complicate our argument.)

If the PTT abandons the equipment business altogether, its problem is to choose p_e, p_s, and a (possibly negative) equipment tax τ that maximizes the profit function

$$\Pi_{ii}\left(p_e, p_s, \tau\right) = \tau q_e(p_e, p_s) + p_s q_s\left(p_e, p_s\right). \tag{1.4}$$

Here, the PTT will set the equipment price p_e equal to the foreign firm's cost c_e^F plus the tax τ, so that the foreign supplier's participation constraint

$$(p_e - c_e^F - \tau)q_e(p_e, p_s) \geq 0 \tag{1.5}$$

is just satisfied whenever $q_e > 0$.

PROPOSITION 1.2. If $(p_e^i, p_s^i, \bar{\tau})$ maximizes the PTT's profit subject to the foreign supplier's equipment constraint, then

$$\bar{\tau} = p_e^i - c_e^F.$$

That is, the foreign supplier's participation constraint is binding, and the tax is exactly the foreign supplier's markup. In this case, the PTT's objective function (1.4) simplifies to the unconstrained function

$$\Pi_{ii}\left(p_e, p_s, \bar{\tau}\right) = (p_e - c_e^F)q_e\left(p_e, p_s\right) + p_s q_s\left(p_e, p_s\right). \tag{1.6}$$

This is identical, except in the cost of equipment, to the profit function (1.2) for which the PTT does not relinquish the equipment market. Consequently, the PTT will leave the equipment market to be served by competitive foreign suppliers if and only if it faces a higher cost structure than the foreign suppliers.

The tax is a cash transfer from the foreign supplier to the PTT. It effectively recovers for the PTT any economic profit a foreign supplier may earn in the equipment market. However, if the PTT benefits by subsidizing equipment in order to enhance the demand for service, $\bar{\tau}$ will actually be negative. That is, the foreign supplier would be subsidized, rather than taxed, for each unit of equipment it sells.

PROPOSITION 1.3. The PTT chooses the equipment and service prices p_e^{ii} and p_s^{ii}

$$\left(p_e^{ii}, p_s^{ii}\right) = \left(\frac{1}{2}\left(\frac{A}{(\alpha+\beta)} + c_e^F\right), \frac{A}{2(\alpha+\beta)}\right). \tag{1.7}$$

The foreign supplier provides the quantity of equipment

$$q_e^{ii} = q_e\left(p_e^{ii}, p_s^{ii}\right) = \frac{1}{2}\left(A - \alpha c_e^F\right)$$

and the PTT provides the quantity of service

$$q_s^{ii} = q_s\left(p_e^{ii}, p_s^{ii}\right) = \frac{1}{2}\left(A - \beta c_e^F\right).$$

The foreign equipment supplier breaks even, and the PTT enjoys profit

$$\Pi_{ii}\left(p_e^{ii}, p_s^{ii}\right) = \frac{1}{4\alpha}\left(A - \alpha c_e^F\right)^2 + \frac{A^2}{4\alpha}\left(\frac{\alpha-\beta}{\alpha+\beta}\right).$$

Thus, the PTT will choose to allow a foreign supplier whenever $\Pi_{ii}\left(p_e^{ii}, p_s^{ii}\right) > \Pi_i\left(p_e^i, p_s^i\right)$. This is the case if and only if $c_e^F < c_e^P$; that is, the foreign supplier enjoys a cost advantage over the PTT in producing equipment.

COROLLARY 1.3.1. The tax $\bar{\tau}$ is positive, if and only if

$$c_e^F < \frac{A}{\alpha+\beta}.$$

Finally, we note that the PTT's choices of service price p_s^i and p_s^{ii}, from expressions (1.3) and (1.7), are equal. The intuition behind this is as follows. Proposition 1.2 shows that the profit functions (1.2) and (1.6), in the exit and non-exit scenarios, respectively, are identical, except for the fact that the PTT's equipment cost c_e^P appears in (1.2), and the foreign firm's equipment cost c_e^F appears in (1.6). In either case, the price of service is independent of the cost of equipment, so one price of service must be equal to the other.

Case (iii). Competitive Equipment Market

In several political and regulatory regimes, the PTT has yielded control of the equipment market to foreign suppliers that are free to compete with one another. However, the PTT retains a monopoly in a service or infrastructure bottleneck. For example, in Poland, the cellular equipment sector exhibits the usual properties of innovation and falling prices typical of a competitive market, but the PTT retains control of the lucrative market for international calls, and does not subsidize or tax the firms in the equipment market. In order to predict the effects on price and sales, we can apply the demand model (1.1) to this situation.

In this case of our model, the PTT exits the equipment market and allows it to be supplied by competing foreign equipment suppliers. We assume the foreign suppliers all bear the same marginal cost c_e^F per unit of equipment. The PTT's problem is now to choose the service price p_s that maximizes the profit function

$$\Pi_{iii}(p_e, p_s) = p_s q_s(p_e, p_s) \tag{1.8}$$

subject to the foreign suppliers' participation constraint

$$(p_e - c_e^F) q_e(p_e, p_s) \geq 0.$$

Our assumption of competition implies that the market equipment price p_e^{iii} is equal to c_e^F. Since the PTT's profit is decreasing in p_e, the PTT benefits from this being as low as possible. This simplifies the PTT's problem to the unconstrained maximization problem

$$\max_{p_s} \Pi_{iii}(c_e^F, p_s). \tag{1.9}$$

ASSUMPTION 1.3. The foreign suppliers' equipment cost is not too large; i.e.

$$c_e^F \leq A\left(\frac{2\alpha - \beta}{2\alpha^2 - \beta^2}\right).$$

This strengthens the restriction on c_e^F made in Assumption 1.2, to ensure that the quantity of equipment q_e is nonnegative in case (iii).

PROPOSITION 1.4. The PTT chooses the service price p_s^{iii}

$$p_s^{iii} = \frac{1}{2\alpha}\left(A - \beta c_e^F\right). \tag{1.10}$$

The foreign suppliers provide the quantity of equipment

$$q_e^{iii} = q_e\left(p_e^{iii}, p_s^{iii}\right) = \frac{2\alpha - \beta}{2\alpha}\left(A - \left(\frac{2\alpha^2 - \beta^2}{2\alpha - \beta}\right) c_e^F\right)$$

and the PTT provides the quantity of service

$$q_s^{iii} = q_s\left(p_e^{iii}, p_s^{iii}\right) = \frac{1}{2}\left(A - \beta c_e^F\right).$$

The foreign equipment supplier breaks even, and the PTT enjoys profit

$$\Pi_{iii}\left(p_e^{iii}, p_s^{iii}\right) = \frac{1}{4\alpha}\left(A - \beta c_e^F\right)^2. \tag{1.11}$$

COROLLARY 1.4.1. The PTT earns more profit in Case (iii) than in Case (i) whenever

$$\left(A - \beta c_e^F\right)^2 - \left(A - \alpha c_e^P\right)^2 > A^2\left(\frac{\alpha - \beta}{\alpha + \beta}\right). \tag{1.12}$$

The pairs $\left(c_e^F, c_e^P\right)$ that satisfy (1.12) are bounded by an hyperbola, shown as the region above the dotted curve in Figure 1.1. This is the set of cost pairs for which the PTT is better off by allowing competition in the equipment market than it is by serving the equipment market as a monopoly.

Corollary 1.4.1 suggests that, in some situations, the PTT will resist competition in the service market even when the foreign suppliers face significantly lower costs. This is in fact the case when the PTT's equipment cost is either very low or very high.

When the PTT has a low equipment cost and monopolizes both markets (Case (i)), it can extract nearly as much profit in the equipment market as in the service market (where its cost is zero). Thus, the cost advantage for a foreign firm must be strong enough to offset this loss in profit to make the switch to a competitive equipment market with foreign suppliers (Case (iii)) worthwhile.

Meanwhile, if the PTT has a high equipment cost and monopolizes both markets (Case (i)), it subsidizes equipment in order to stimulate service demand. If the PTT switches to Case (iii), it is no longer subsidizing equipment. Thus, the equipment price paid by the consumer, which is equal to the foreign supplier's cost, may or may not be lower than the subsidized price the consumer paid when the PTT sold equipment. If the equipment price rises significantly as a result of the switch, demand for service will be adversely affected. This is the reason that the the foreign firm must also have a strong cost advantage in this case for the PTT to view the switch as beneficial.

COROLLARY 1.4.2. *Ceteris Paribus,* for $c_e^F \neq \frac{A}{\alpha+\beta}$, the PTT always earns more profit in Case (ii) than in Case (iii); i.e.

$$\Pi_{ii}\left(p_e^{ii}, p_s^{ii}\right) > \Pi_{iii}\left(p_e^{iii}, p_s^{iii}\right).$$

For $c_e^F = \frac{A}{\alpha+\beta}$, the PTT earns the same profit in Cases (ii) and (iii).

Corollary 1.4.2 implies that, the marginal cost c_e^F held constant, a PTT earns at least as much profit in Case (ii), by regulating the equipment market, than it does in Case (iii), with a competitive equipment market. Indeed, as long as the marginal cost c_e^F is not exactly $\frac{A}{\alpha+\beta}$, the PTT will earn more in Case (ii) than in Case (iii). We can infer from this that a PTT will tend to use its political leverage to retain the ability to regulate a market from which it divests.

COROLLARY 1.4.3. It is necessary, but not sufficient, that the foreign suppliers' cost c_e^F be lower than the PTT's cost c_e^P for the PTT to prefer Case (iii) to Case (i). That is, if $\Pi_{iii}\left(p_e^{iii}, p_s^{iii}\right) > \Pi_i\left(p_e^i, p_s^i\right)$, then $c_e^F < c_e^P$.

This corollary states that a necessary condition for the PTT to earn more profit in Case (iii) than it would in Case (i) is that the foreign supplier has a lower cost than the PTT. However, this is not a sufficient condition, since equipment sales generates a significant amount of profit for the PTT when it faces a low equipment cost. For example, suppose that the PTT's cost of equipment is zero; that is, $c_e^P = 0$. In Case (i), the PTT's profit results from sales of both products equally, as we might expect for the symmetric demand construction. In Case (iii), by allowing a foreign firm with the same cost $c_e^F = 0$ to produce equipment, the PTT raises its service price from $\frac{A}{2(\alpha+\beta)}$ to $\frac{A}{2\alpha}$, as the quantity of service sold remains constant at $\frac{A}{2}$, but this does not offset the loss in profit from the equipment market.

Case (iv). Independent Monopolist Serves Equipment Market

We now consider the situation in which the PTT abandons the equipment market in favor of another monopolist. The analysis of this case is similar to Cournot's duopoly analysis. The PTT sets the service price to maximize profit in the service industry, while the independent monopoly supplier of equipment sets the price of equipment to maximize profit in that market. Both products are complementary in demand, as in the original model.

The PTT's problem is to maximize profit in the service market

$$\Pi_{iv}(p_e, p_s) = p_s(A - \alpha p_s - \beta p_e). \tag{1.13}$$

The equipment supplier simultaneously faces the problem of maximizing profit in the equipment market

$$\Pi_F(p_e, p_s) = p_e(A - \alpha p_e - \beta p_s). \tag{1.14}$$

Assumption 1.3 ensures that quantities in both markets are nonnegative.

PROPOSITION 1.5. The equipment monopolist and PTT respectively choose the equipment and service prices

$$\left(p_e^{iv}, p_s^{iv}\right) = \left(\frac{1}{2\alpha+\beta}\left(A + \frac{2\alpha^2}{2\alpha-\beta}c_e^F\right), \frac{1}{2\alpha+\beta}\left(A - \frac{\alpha\beta}{2\alpha-\beta}c_e^F\right)\right).$$

The equipment monopolist provides the quantity of equipment

$$q_e\left(p_e^{iv}, p_s^{iv}\right) = \frac{\alpha}{2\alpha+\beta}\left(A - \frac{2\alpha^2-\beta^2}{2\alpha-\beta}c_e^F\right)$$

and the PTT provides the quantity of service

$$q_s\left(p_e^{iv}, p_s^{iv}\right) = \frac{\alpha}{(2\alpha+\beta)}\left(A - \frac{\alpha\beta}{2\alpha-\beta}c_e^F\right).$$

The equipment monopolist earns profit from the equipment market

$$\Pi_F\left(p_e^{iv}, p_s^{iv}\right) = \frac{\alpha}{(2\alpha+\beta)^2}\left(A - \frac{2\alpha^2-\beta^2}{2\alpha-\beta}c_e^F\right)^2$$

and the PTT earns profit from the service market

$$\Pi_{iv}\left(p_e^{iv}, p_s^{iv}\right) = \frac{\alpha}{(2\alpha+\beta)^2}\left(A - \frac{\alpha\beta}{2\alpha-\beta}c_e^F\right)^2.$$

COROLLARY 1.5.1. Given identical marginal costs for foreign equipment suppliers, The PTT will always prefer competition to another monopoly in the equipment market. That is, for c_e^F satisfying Assumption 1.3,

$$\Pi_{iii}\left(p_e^{iii}, p_s^{iii}\right) > \Pi_{iv}\left(p_e^{iv}, p_s^{iv}\right). \tag{1.15}$$

In other words, whenever the PTT finds it rational to abandon the equipment market, the PTT will find a competitive equipment market more hospitable than an equipment market controlled by another monopolist firm.

1.3 Comparison of Consumer Welfare Effects

When the PTT exits the equipment market and regulates it, as in Case (ii), the price of equipment falls and the price of service does not change. Thus, this has a positive effect on consumer welfare. When the PTT exits the equipment market and allows it to become competitive, as in Case (iii), the prices of each good may rise or fall, depending on the firms' costs and the demand parameters, so claims on consumer welfare effects require stronger specifications on cost and parameters. As we noted in Corollary 1.4.2, if the PTT benefits by exiting the equipment market and can decide between regulating price in the equipment market and permitting competition in the equipment market, it will always prefer to regulate. This raises the policy question of whether consumers' interests are aligned with those of the PTT; that is, if consumers prefer the regulated equipment market to a competitive equipment market.

Examining prices does not suffice to answer this question, since the prices necessarily move in opposite directions if the regime were to change from Case (ii) to Case (iii). Suppose the foreign supplier's equipment cost c_e^F is less than $\frac{A}{\alpha+\beta}$. Then the regulated equipment price p_e^{ii} will be higher than the competitive equipment price p_e^{iii}, while the regulated service price p_s^{ii} will be lower than the competitive service price p_s^{iii}. On the other hand, if the equipment cost c_e^F is greater than $\frac{A}{\alpha+\beta}$, Case (ii) has the lower equipment price and higher service price.

We use Hicks' equivalent variation (EV) and compensating variation (CV) as proxies for consumer utility. One can show that EV and CV are both equal to the change in net consumer surplus ΔCS, defined as the sum of the changes in areas above prices and to the left of the demand curves in each market as prices change from (p_e^0, p_s^0) to (p_e^1, p_s^1):

$$\Delta CS\left(\left(p_e^0,p_s^0\right),\left(p_e^1,p_s^1\right)\right)=\int_{p_e^1}^{p_e^0} q_e\left(p_e,p_s^0\right)dp_e+\int_{p_s^1}^{p_s^0} q_s\left(p_e^1,p_s\right)dp_s \tag{1.16}$$

Because this equality property is treated at length in Varian [V 1996], for example, and is considered standard economic theory, we shall not prove the result here.[1] When only the equipment (service) price changes, the integral over the quantity of service (equipment) is equal to zero.

PROPOSITION 1.6. The PTT's preference of regulation of the equipment market (with equipment and service prices (p_e^{ii},p_s^{ii})) to a competitive equipment market (at prices (p_e^{iii},p_s^{iii})) is necessarily contrary to consumers' interests. That is, if there were a change from Case (ii) to Case (iii), consumers would necessarily be better off:

$$\begin{aligned}\Delta CS\left(\left(p_e^{ii},p_s^{ii}\right),\left(p_e^{iii},p_s^{iii}\right)\right) &= \frac{3(\alpha-\beta)(\alpha+\beta)}{8\alpha}\left(\frac{A}{\alpha+\beta}-c_e^F\right)^2\\ &> 0.\end{aligned}$$

Proposition 1.6 implies that, with the marginal cost c_e^F held constant, consumers gain a higher surplus in Case (iii), when the equipment market is competitive, than

[1]The equality of CV, EV, and ΔCS assumes that a representative consumer has a quasilinear utility function, such as

$$U\left(q_e,q_s,q_0\right)=q_0+\frac{A}{\alpha+\beta}\left(q_e+q_s\right)+\frac{\alpha}{\alpha^2-\beta^2}\left(q_e^2+q_s^2\right)-\frac{\beta}{\alpha^2-\beta^2}q_eq_s,$$

where q_0 is money spent on other goods and services. In this case, increases in consumer wealth beyond a basic level of consumption result exclusively in higher q_0, and the consumption of equipment and service have no wealth effects.

in Case (ii), when the equipment market is regulated. This is due to the following: Suppose there is a change from Case (ii) to Case (iii). If the foreign suppliers' cost c_e^F is small (that is, less than $\frac{A}{2(\alpha+\beta)}$), the increase in the equipment price is large, in absolute terms, relative to the decrease in service price, so the net change in surplus is positive. Alternatively, if c_e^F is large, the equipment price *decreases* by a small amount, relative to the *increase* of the service price, so the net change in surplus is again positive.

1.4 Conclusion

We have provided a simple mathematical explanation for the otherwise counterintuitive phenomenon that many monopolist telephone operators have welcomed competitive equipment providers into their markets. In short, as long as a PTT retains monopoly control of some high-margin bottleneck, such as circuits to other countries, it can actually benefit by eliminating its high-cost lines of business, such as providing the equipment necessary to connect new customers. Rather, the PTT permits other low-cost suppliers to serve those markets instead.

An alternate supplier's lower cost structure is a necessary, but not sufficient, requirement for the PTT to consider jettisoning the equipment market if it does not have the authority to tax the supplier's profit. However, if the PTT is able to tax away the low-cost suppliers' profit, or subsidize a low-cost supplier to sell below cost, it could be as profitable as if the low-cost supplier were integrated with the PTT into a single firm, and will always choose to do so. The low-cost suppliers connect many more customers to the network, thereby increasing demand for the service over which the PTT continues to serve as a monopoly.

Consumers benefit from the lower prices resulting from the more efficient provider of equipment. However, consumers will necessarily benefit more if the PTT chooses not to tax or subsidize the equipment market. That is, in many cases, consumers will be better off by paying a bit more for (unsubsidized) telephones, as this will cause the PTT to lower its prices for service. Thus, the interests of the PTT are aligned with those of the consumers in regard to the exit decision, but differ with respect to the rules of the new regime. This suggests that governments and citizens should encourage privatization and divestiture of equipment businesses, but should also be wary of advertising campaigns for free or discounted telephones, since they are likely to come with high tariffs for service.

Empirical evidence suggests that PTTs do indeed continue to charge prices above cost in markets for monopolized services, such as international calls. Tariffs on competitive international routes, such as between the United States and Canada or the United Kingdom, approach marginal cost and are low – often around 10 cents per minute. On other routes, PTTs set monopoly prices. For example, the price of a call from the United States – a competitive market – to China, Germany, or Italy – countries with monopolized telephone services and excess capacity – is over 50 cents per minute. To be sure, growth of the Internet or other competing means of communication may eventually erode profit margins in these markets. For the time being, however, margins in bottleneck services remain strong, and may well be increasing as PTTs enhance demand for those services.

Moreover, this effect comes into play in many other industries where goods are priced to exploit complementarities and manufacturers enjoy some market power on complementary items. Hewlett Packard tends to price its DeskJet color printers at cost, and makes its profit on its expensive ink cartridges. Any automobile owner seeking to make repairs has faced high prices on dealer parts. Newspapers, television, and Internet service providers offer their products to end users at low competitive prices, but all of these entities generate income by also selling advertising in their pages, commercial spots, or portal sites, respectively. Hotels may price rooms based upon competitive market forces, but enjoy near monopolies on ancillary products such as room service items.

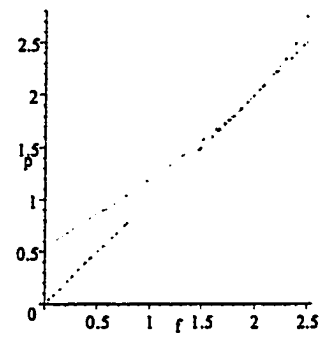

FIGURE 1.1: Boundary for which PTT prefers Case (iii) to Case (i)

The PTT prefers a competitive equipment market to retaining its monopoly in equipment for all points above the curved line. In this picture, p denotes the equipment cost c_e^p faced by the PTT; f denotes the equipment cost c_e^f faced by competitive foreign suppliers.

Chapter 2

Estimating Wireless Telephone Supply, while Accounting for Selection and Country-Specific Effects

2.1 Introduction

From 1985 to 1998, the International Telecommunication Union [1999] (ITU) recorded that the number of wireless telephone users in the world increased from fewer than 100,000 to more than 300 million, and is widely expected to exceed 1 billion by 2004. In light of the fact that there have been surprisingly few empirical studies of this explosive industry, this paper is an attempt to estimate a supply function for the wireless telephone market from a panel of data across 124 countries from 1989 to 1998. However, estimating the function using prices and other industry-related variables by usual regression methods may result in two types of misspecification. The first is due to sample selection bias, since our wireless price and quantity data exist only for countries where the product is actually available, and cannot account for the potential penetration rates that might otherwise exist in countries where wireless is not in fact available. The second is due to country-specific effects that our data do not explain. We address these issues jointly by developing a two-stage procedure, based upon Heckman's [1979] model, for consistent estimation on selected data, that also controls for fixed effects in the second-stage regression model. We retain information from this probability model for use in the second-stage supply regression model.

We specifically seek estimates of per-capita wireless telephone supply, as a linear function of time; the number of existing fixed-line telephones per capita; the prices of connection and a 3-minute local call; and other variables.

Because the government of any country may or may not choose to allow wireless communications within its borders, we are concerned that estimates on our data from 72 countries that offer wireless suffer from sample selectivity bias. That is, they measure supply only in countries where the service is already offered, which may differ from the supply that might otherwise exist in countries where it is not offered. Moreover, our explanatory data are not sufficient to explain fully the conditions in each country, so we would like to control for individual country effects.

Our modification of Heckman's two-stage "Heckit" procedure provides a method for the consistent estimation of production in light of these problems. The first stage of our model estimates the probability that wireless is available in a country in a particular year, based upon explanatory variables from a larger set of 124 countries, some of which have deployed wireless from their earliest presence of the sample, some of which adopt wireless during the course of the sample, and some of which remain without wireless through the entire sample. We then estimate a supply regression that controls for fixed (country-specific) effects, on the "selected subsample" of data; that is, on observations for countries in years during which wireless is available. By inserting the probability estimates from the first estimation into the second estimation, we control for the bias resulting from the sample selectivity issue.

We first introduce the data, including the explanatory variables and their relevance to supply. The next section of the paper proposes the theory for for the modified "Heckit" regression model and corresponding two-stage least squares model that control for both fixed effects and sample selectivity. We follow the theory with the actual estimation procedures. Finally, we present our statistical results with some discussion of their economic implications, and offer some ideas for alternate methods.

2.2 Observable Determinants of Wireless Supply

We consider a variety of relevant available explanatory variables in our model. We begin with variables for our selection model, which may affect whether or not wireless is provided in a country. For countries in which wireless service is available, we initially suspect some relationship between prices and quantity supplied. We can supplement this with more subtle and indirect variables, such as the price of a substitute and other market conditions.

Since wireless penetration in a country will only exist provided that wireless service is available, we would like to account for the bias that may arise from selection in our subsample. In our first-stage selection model, we estimate the probability of the existence of wireless in a country as a function of other variables that exist whether or not a country offers wireless service. Our binary indicator variable ($CELLYN$) takes the value of 1 or 0 when wireless is or is not offered in a country i during the observation year t, respectively. As explanatory criteria for the selection model (2.1), we use variables that we feel are unequivocally exogenous, and do not require the existence of wireless. For each country-year in the unselected data set, we include: a constant term, the year ($YEAR$), total population (POP), and income per capita (GDP), both of which are likely to influence consumption. We include the number of main, or fixed, telephone lines per capita (ML), in the selection decision because existing infrastructure may lower the incremental cost of deploying a new wireless network, or may provide some network externality to new wireless subscribers. The charges for fixed service connection ($FIXCONN$), monthly service

($FIXMONTH$), and a three-minute call ($FIX3$), as weighted averages of residential and business users, give an idea for prices of a substitute product that exists in all observed countries. Finally, we include the size of a country's telecommunications staff per capita ($STAFF$), given the notions that telecommunications workers may be needed to install wireless infrastructure, or employees as an interest group may actively resist competition to the existing fixed-line network.

Our dependent variable in the supply model $CELSUBS$ is the number of wireless subscribers per capita for a given country i in a given year t.

For country-years in which wireless service is available, we presume the aforementioned variables $YEAR$, ML, and $STAFF$, enter the supply equation, whereas the others are demand components. In addition, the supply curve may have other components specific to wireless-enabled countries. These include the charges for analog wireless conection ($ANCONN_{it}$) and a three-minute call ($AN3_{it}$), telephone revenue ($REVENUE_{it}$), and investment ($INVEST_{it}$). Finally, the volume of minutes of incoming and outgoing international calls in a country is a function of the total number of telephones, and is a high value-adding service provided by one or more telephone authorities. Thus, we include it in the supply equation.

Another wireless-specific variable that may affect the demand side, and thus may be useful as an instrument for endogenous variables, is $WAIT_{it}$, the length of the waiting list per capita.

All countrywide quantities in our data set ($CELSUBS, ML, WAIT, STAFF$) are converted to numbers *per capita* by dividing by population. All prices are deflated to 1995 local currency and then converted to 1995 U.S. dollars according to exchange rates from the International Monetary Fund [1998]. The countrywide economic variables GDP, $REVENUE$, and $INVEST$, are deflated to 1995 U.S. dollars *per capita.* Our data are from the International Telecommunication Union's *World Telecommunication Indicators* Database [1999]. The Database contains responses from annual surveys of PTT regulators on many different quantifiable telecommunications indicators for 209 countries since 1975, and also for the years 1960, 1965, and 1970. We initially extract a panel of observations of all 209 countries between the years 1981 and 1998, the first and last years for which reliable data are available. After deleting incomplete observations, we are left with an unbalanced panel of 585 observations covering 124 countries. The set of observations range from those of the wealthiest and earliest of wireless adopters (Belgium and Denmark, 1989), to poor countries in which the technology has never become available (Syria and Sierra Leone, 1998). Of these, 376 observations are of country-years with wireless technology. We discard eight observations because they are singletons, and we are thus unable to take fixed effects for those groups. We are left with the selected subsample of 368 observations covering 55 countries. The remaining 217 observations are of countries in years without wireless.

2.3 Econometric Theory

2.3.1 Two-Stage Model for Selected Data with Fixed Effects

The ITU database provides information on countries that do and do not offer wireless telephony to their citizens. This enables us to control for selection in our model. Specifically, we wish to model wireless supply in countries that have adopted the technology, and then to extrapolate this information to countries where service was not available during the period of time covered by the ITU survey.

Heckman [1979] constructs a model for consistent, albeit inefficient, estimation of selected data. The "Heckit" estimator first estimates a model of selection, and then uses the hazard rates from the selection model in the estimable regression model. However, the Heckit is not defined for models with fixed effects, or within-group dummy variables. Since our set of explanatory variables cannot hope to capture all of the relevant social, political, and economic issues in the countries included in our data set, we would like to use country-specific dummies in the regression model. Thus, we propose the following modification to the Heckit model for consistent estimation of selected data that accounts for fixed effects in the second stage.

Failure-time models are frequently used to model the adoption of a technology. Hannan and McDowell [1984] and Levin, Levin, and Meisel [1987] respectively model the adoptions of automatic teller machines by banks, and optical scanners by supermarkets, using exponentially-distributed adoption processes. Unfortunately, it would be difficult to incorporate such a model into a subsequent regression model as in a Heckman-style framework. Because the process is exponentially, rather than normally, distributed, we cannot use Heckman's approach of applying the theorem of Johnson and Kotz [1974] (noted below as Lemma 1) to account for the adoption process in the supply regression. Second, they generally apply to cross-sectional models of pooled data, making it difficult to account for both adoption in a certain year within a country, and the evolution of supply in that country over time.

More progress has been made in the literature of sample selectivity in panel data, often referred to as "Attrition," due to the fact that the methods largely were developed for the purpose of controlling for early exit of subjects from surveys of workers over periods of several years. The seminal paper is that of Hausman and Wise [1979], who construct a random-effects maximum-likelihood (ML) model that accounts for the bias resulting from systematic subject attrition in Period 2 of a stylized two-period panel. Their model is generalized to allow for arbitrary entry and exit over a longer panel period by Ridder [1990]. Keane, Moffitt, and Runkle [1988] apply the ML approach of Hausman and Wise to show that the tendency of workers to become unemployed during recessions is greater among high-wage workers than among low-wage workers in a random-effects model, but find no such evidence in a fixed-effects model. Chamberlain notes that in the fixed effects model, if the number of time periods over which an individual is observed is small, the fixed effect parameters will be inconsistent; however, Heckman [1981] provides experimental data to suggest that this bias is not significant for panels with at least eight observations per individual. Verbeek [1990] finesses the issue by providing a likelihood function for a transformation of the fixed-effects model as a least-squares deviation form model.

We address these issues by considering the adoption problem as a sample selection issue, estimating the bias of the selection, and then including it in a supply regression on the panel. We apply the methodology of Heckman to the problem proposed in passing by Hausman and Wise, as suggested in a footnote by Keane, Moffitt, and Runkle, by employing a two-stage procedure consistently to estimate fixed effects in the presence of sample selection. The first stage is a probit selection model on a panel of data, and the second stage incorporates a selection criterion from the first stage in a regression model on the selected subset of data. While this necessitates that we presume some information about all observations in our panel, our goal is to consistently estimate parameters and test hypotheses only on the selected subset of observations. This procedure may be used as a substitute for Verbeek's likelihood method when an unstable data set or other issues create problems in maximum likelihood estimation. Alternatively, as Heckman suggested for his model, a two-stage method provides starting values that are relatively close to the likelihood function's maximum.

We begin by specifying the selection criteria for panel data

$$z_{it}^{*} = W_{it}\gamma + u_{it}, \tag{2.1}$$

where

$$\begin{aligned} E\left(u_{it}\right) &= 0 \\ E\left(u_{it}^{2}\right) &= 1 \\ E\left(u_{it}u_{js}\right) &= 0, \text{ if } i \neq j \text{ or } s \neq t. \end{aligned}$$

We specify the selection criteria for groups $i = 1, ..., N$, and, within each group, time periods $t \in \Re_i$. The overall panel may be unbalanced; however, we assume that the observations omitted from the unbalanced panel (due to incomplete or non-reporting) are *non-systematically* omitted.

The (systematic) selectivity issue comes into play as follows. We cannot observe the dependent variable z_{it}^{*}, but we do observe across both selected and omitted observations the explanatory variables W_{it} and a binary variable

$$z_{it} = \begin{cases} 1, \text{ if } z_{it}^{*} > 0 \\ 0, \text{ if } z_{it}^{*} < 0. \end{cases}$$

Then for all $z_{it} = 1$, we observe

$$y_{it} = \alpha_i + x_{it}\beta + \varepsilon_{it}, \tag{2.2}$$

where

$$\begin{aligned}
E(\varepsilon_{it}) &= 0 \\
E(\varepsilon_{it}^2) &= \sigma_\varepsilon^2 \\
E(\varepsilon_{it}\varepsilon_{js}) &= 0, \text{ if } i \neq j \text{ or } s \neq t \\
E(\varepsilon_{it}u_{it}) &= \sigma_{\varepsilon u} \;\; \text{(covariance)} \\
\rho &= \frac{\sigma_{\varepsilon u}}{\sigma_\varepsilon} \;\; \text{(correlation)}
\end{aligned}$$

The selected subsample of data in general forms an unbalanced panel, over groups $i = 1, ..., n$, where $n \leq N$, and, for each group i of the selected subsample, time periods $t \in \Im_i$, where $\Im_i \subseteq \Re_i$. That is, the set of selected observations is a subset of all observations subject to the selection criteria. We define T_i as the number of periods in the panel for group i; that is, T_i is the count of elements in the set $\Im_i$.

ASSUMPTION 2.1. For all groups i in the selected subsample of the panel, we need at least two time periods in the unbalanced panel in each group. That is,

$$T_i \geq 2, \text{ for all } i.$$

The assumption is necessary for estimation of fixed effects, which are not defined for single observations in groups.

We now evaluate the moments of the regression error in group deviation form ε_{it}^D, where

$$\varepsilon_{it}^D = \varepsilon_{it} - \frac{1}{T_i}\sum_{s \in \Im_i} \varepsilon_{is},$$

and the covariance and correlation of ε_{it}^D with the selection error u_{it}.

For purposes of statistical inference, we are interested in the distribution of the dependent variable y_{it}, *conditional* upon observation *it* being in the selected subsample. This necessitates understanding the distribution of the error term ε_{it}:

LEMMA 2.1.[1] (Johnson and Kotz [1974]) The moments of the distribution of ε_{it}, conditional upon observation *it* being included in the selected subsample, are

$$\begin{aligned} E\left(\varepsilon_{it}|\text{selection}\right) &= \rho\sigma_{\varepsilon}\lambda_{it} \\ Cov\left(\varepsilon_{is},\varepsilon_{jt}|\text{selection}\right) &= \begin{cases} \sigma_{\varepsilon}^{2}\left(1-\rho^{2}\delta_{it}\right), \text{ if } i=j \text{ and } s=t \\ 0, \text{ if } i\neq j \text{ or s} \neq t \end{cases} . \end{aligned}$$

where, for standard normal probability density and cumulative distribution functions $\phi(\cdot)$ and $\Phi(\cdot)$, respectively,

$$\begin{aligned} \lambda_{it} &= \frac{\phi\left(W_{it}\gamma\right)}{\Phi\left(W_{it}\gamma\right)} \\ \delta_{it} &= \frac{\partial\lambda_{it}}{\partial\gamma} = \lambda_{it}\left(W_{it}\gamma+\lambda_{it}\right). \end{aligned}$$

The value λ_{it}, a hazard rate commonly known as the Inverse of Mill's Ratio (IMR), is the change in the probability of selection as explained by the selection variables W_{it}. The notable insight, attributed to Johnson and Kotz, is that the coefficient c of the variable λ_{it} is exactly the correlation ρ between the selection and regression models, multiplied by the standard deviation σ_{ε}, or square root of the variance, of the regression model.

The lemma implies directly that

$$E\left(y_{it}|\text{selection}\right) = \alpha_i + x_{it}\beta + c\lambda_{it},$$

[1]This Lemma, which is not proved here, is also written as Theorem 22.7 in Greene [1995].

where

$$c = \rho\sigma_\varepsilon. \tag{2.3}$$

The specification is completed by adding an error term v_{it}:

$$(y_{it}|\text{selection}) = \alpha_i + x_{it}\beta + c\lambda_{it} + v_{it}. \tag{2.4}$$

That is, $(y_{it}|\text{selection})$ is a linear combination of the explanatory variables x_{it} and the IMR selection term λ_{it}, plus an error component v_{it}. In addition to the error component v_{it}, the model we estimate contains another source of error, due to the fact that it contains an *estimated* IMR, rather than the true IMR. We must account for this estimated IMR in the covariances that we use for inference.

Heckman [1979] and Greene [1995] note the moments of the error v_{it}, which are summarized in the following corollary and are not proved in this paper.

COROLLARY 2.0.1. (Heckman [1979], Greene [1995]) Assume the model (2.4) is derived from (2.2) with independent errors ε_{it} conditionally distributed as in Lemma 2.1. Then the moments of the added error term v_{it}, conditional upon observation *it* being included in the selected subsample, are

$$\begin{aligned} E(v_{it}) &= 0 \\ Cov(v_{is}, v_{jt}) &= \begin{cases} \sigma_\varepsilon^2 (1 - \rho^2\delta_{it}), \text{ if } i = j \text{ and } s = t \\ 0, \text{ if } i \neq j \text{ or } s \neq t \end{cases}. \end{aligned} \tag{2.5}$$

As in Verbeek [1990], we are interested in testing hypotheses only on β and c. We can sidestep estimation of the α_i group dummy variables by taking equation (2.4) as a least-squares group deviation form model:

$$\begin{aligned}
&\left(y_{it} - \frac{1}{T_i}\sum_{s\in\Im_i} y_{is}|\text{selection}\right) \\
&= \left(x_{it} - \frac{1}{T_i}\sum_{s\in\Im_i} x_{is}\right)\beta + c\left(\lambda_{it} - \frac{1}{T_i}\sum_{s\in\Im_i}\lambda_{is}\right) \\
&+ \left(v_{it} - \frac{1}{T_i}\sum_{s\in\Im_i} v_{is}\right).
\end{aligned}$$

Using the notation defined above, we would like the distribution of $\left(y_{it}^D|\text{selection}\right)$, where the superscript D indicates a variable in groupwise deviation form:

$$\left(y_{it}^D|\text{selection}\right) = x_{it}^D\beta + c\lambda_{it}^D + v_{it}^D. \tag{2.6}$$

Note that the means about which the deviation-form variables vary are within-country averages of the observations in the selected subsample.

Since ordinary least-squares (OLS) parameter estimates of (2.6) are consistent (unlike their respective standard errors), we can recover consistent estimates of ρ and σ_ε by first obtaining an estimate of c by running an OLS regression on the model (2.6), and then substituting our estimate of c in the variance identity (2.3). This procedure is discussed at length in the next section.

PROPOSITION 2.1. The moments of the error in deviation form v_{it}^D, for observations in the selected subsample, are

$$E\left(v_{it}^D|\text{selection}\right) = 0$$

$$\begin{aligned}
&Cov\left(v_{is}^D, v_{jt}^D|\text{selection}\right) \\
&= \begin{cases} \frac{\sigma_\varepsilon^2}{T_i}\left[(T_i - 1) - \rho^2\left((T_i - 2)\,\delta_{it} + \bar{\delta}_i\right)\right], \text{ for } i = j \text{ and } s = t \\ -\frac{\sigma_\varepsilon^2}{T_i}\left[1 - \rho^2\left(\delta_{is} + \delta_{it} - \bar{\delta}_i\right)\right], \text{ for } i = j \text{ and } s \neq t \\ 0, \text{ for } i \neq j \end{cases}
\end{aligned} \tag{2.7}$$

where $\bar{\delta}_i$ is the mean value of δ_{is} across all observations s in group i that are in the selected subsample; that is,

$$\bar{\delta}_i = \frac{1}{T_i} \sum_{s \in \Im_i} \delta_{is}.$$

The covariances (7) form a block-diagonal covariance matrix of dimension$(\sum_i T_i) \times (\sum_i T_i)$, with n block submatrices along the main diagonal, of size $T_i \times T_i$ for the ith block-diagonal.

Practically speaking, we do not know the true IMRs λ_{it}. Rather, we use estimates $\hat{\lambda}_{it}$ provided by the Probit regression, produced in most packages by the Probit estimation procedure. That is, we actually estimate

$$\left(y_{it}^D|\text{selection}\right) = x_{it}^D \beta + c\hat{\lambda}_{it}^D + v_{it}^D. \tag{2.8}$$

Thus, inference requires that we account for the errors $e_{\lambda it}$ resulting in the estimation of the IMRs. We add to the model (2.4) the term

$$e_{\lambda it} = c\left(\hat{\lambda}_{it} - \lambda_{it}\right),$$

so that our estimable model in deviation form (2.6) becomes

$$\left(y_{it}^D|\text{selection}\right) = x_{it}^D \beta + c\lambda_{it}^D + v_{it}^D + e_{\lambda it}^D.$$

PROPOSITION 2.2. The moments of the second error term $e_{\lambda it}^D$ are

$$E\left(e_{\lambda it}^D\right) = 0$$

$$
\begin{aligned}
& Cov\left(e^D_{\lambda is}, e^D_{\lambda jt}\right) \\
= \;& c^2\left[\omega_{is,jt} - \frac{1}{T_i}\sum_{q\in\Im_i}\omega_{iq,jt} - \frac{1}{T_j}\sum_{r\in\Im_j}\omega_{is,jr} + \bar{\omega}_{ij}\right].
\end{aligned}
$$

where

$$
\begin{aligned}
\omega_{is,jt} &= \delta_{is}\delta_{jt}W_{is}\Sigma_\gamma W'_{jt} \\
\bar{\omega}_{ij} &= \frac{1}{T_iT_j}\sum_{q\in\Im_i}\sum_{r\in\Im_j}\omega_{iq,jr},
\end{aligned}
$$

and Σ_γ is the covariance matrix of γ from the Probit selection model.

Let us combine the error terms as the variable e_{it}, defined

$$
e_{it} = v_{it} + e_{\lambda it}.
$$

COROLLARY 2.2.1. The total expectation of the composite error in deviation form e^D_{it}, given the selection, is zero. That is,

$$
E\left(e^D_{it}|\text{selection}\right) = 0.
$$

The total variance is

$$
\begin{aligned}
& Cov\left(e^D_{is}, e^D_{jt}|\text{selection}\right) \\
= \;& \rho^2\sigma^2_\varepsilon\left[\omega_{is,jt} - \frac{1}{T_i}\sum_{q\in\Im_i}\omega_{iq,jt} - \frac{1}{T_j}\sum_{r\in\Im_j}\omega_{is,jr} + \bar{\omega}_{ij}\right] \qquad (2.9) \\
+ \;& \begin{cases} \frac{\sigma^2_\varepsilon}{T_i}\left[(T_i-1) - \rho^2\left((T_i-2)\,\delta_{it} + \bar{\delta}_i\right)\right], & \text{for } i=j \text{ and } s=t \\ -\frac{\sigma^2_\varepsilon}{T_i}\left[1 - \rho^2\left(\delta_{is}+\delta_{it}\right) - \bar{\delta}_i\right], & \text{for } i=j \text{ but } s\neq t \\ 0, & \text{for } i\neq j \end{cases}.
\end{aligned}
$$

For exposition, we abbreviate the covariance matrix with elements as in (2.9) as Ω. We write y^* as the vector of deviation-form dependent variables y_{it}^D, X^* as the horizontally concatenated matrix of deviation-form explanatory variables x_{it}^D and IMRs λ_{it}^D , β^* as the vertically concatenated matrix of β and c, and e as the vector of deviation-form composite errors e_{it}^D. That is,

$$y^* = X^* \beta^* + e^*,$$

where

$$\begin{aligned} y^* &= \left[y_{it}^D\right]_{t \in \Im_i; i \in N} \\ X^* &= \begin{bmatrix} x_{it}^D & \lambda_{it}^D \end{bmatrix}_{t \in \Im_i; i \in N} \\ \beta^* &= \begin{bmatrix} \beta \\ c \end{bmatrix} \\ e^* &= \left[e_{it}^D\right]_{t \in \Im_i; i \in N} \end{aligned}$$

Let $\breve{\beta}^*$ be an estimator for β^*, defined

$$\breve{\beta}^* = \left(X^{*\prime} X^*\right)^{-1} X^{*\prime} y^*.$$

COROLLARY 2.2.2. The moment matrices of the parameter estimator $\hat{\beta}^*$ are

$$\begin{aligned} E\left(\breve{\beta}^*\right) &= \beta^* \\ Cov\left(\breve{\beta}^*\right) &= \left(X^{*\prime} X^*\right)^{-1} X^{*\prime} \Omega X^* \left(X^{*\prime} X^*\right)^{-1}, \end{aligned}$$

where Ω is the block diagonal matrix whose elements are defined by (2.9).

2.3.2 Two-Stage Least Squares Model

Our set of explanatory variables may be subject to endogeneities. For example, as a country's waiting list for fixed-line telephones lengthens, potential customers may opt for a wireless substitute which may be available immediately. Meanwhile, as customers purchase wireless telephones, they may remove themselves from the queue for wireline telephones, resulting in a shorter waiting list. We shall address endogeneity effects by means of a two-stage least squares (2SLS) estimator, which we define as follows.

We partition matrix X^* into submatrices Y and X of the endogenous and exogenous explanatory variables, respectively.[2] We also partition vector β^* into respective components ψ and θ to correspond with Y and X. Since we assume that the IMRs λ_{it} are exogenous, the λ^D vector is included in the X submatrix. Let $\tilde{\beta}^*$ be the 2SLS estimator for β^* in our model, written in matrix form

$$\begin{aligned} y^* &= X^*\beta^* + e^* \\ &= Y\psi + X\theta + e^* \end{aligned}$$

In accordance with the usual 2SLS procedure, we first obtain estimates of the endogenous variables in Y by fitting them to an OLS regression on the exogenous variables X. That is, for variable y_k, we estimate

$$y_k = X\eta_k + \nu_k,$$

[2]Note that the matrices Y and X do not correspond to the matrices of vectors y_{it}^D and x_{it}^D. The matrix Y is the submatrix of all endogenous variables in X^*, and the matrix X is the submatrix of endogenous variables in X^*. The vector y_{it}^D of dependent variables is included in the 2SLS model as y^*. The IMRs λ_{it}^D are necessarily included in X, because they are assumed to be exogenous. The parameter vectors $\psi$and θ correspond to matrices Y and X, respectively.

where

$$Y = \begin{bmatrix} y_1 & \cdots & y_K \end{bmatrix}$$

and ν_k is a vector of errors. The fitted value for y_{jk} is

$$\hat{y}_k = X\hat{\eta}_k,$$

where $\hat{\eta}_k$ is the OLS estimator for η_k; that is,

$$\hat{\eta}_k = (X'X)^{-1} X'y_k.$$

Finally, we estimate β with the IV estimator $\tilde{\beta}^*$, defined as

$$\tilde{\beta}^* = \left(\hat{X}^{*\prime}X^*\right)^{-1} \hat{X}^{*\prime}y,$$

where

$$\begin{aligned} \hat{X}^* &= \begin{bmatrix} \hat{Y} & X \end{bmatrix} \\ \hat{Y} &= \begin{bmatrix} \hat{y}_1 & \cdots & \hat{y}_K \end{bmatrix}. \end{aligned}$$

The following proposition states the moments of the 2SLS estimator.

PROPOSITION 2.3. The 2SLS estimator $\tilde{\beta}^*$ for the parameters in the instrumented model has the expectation and covariance matrix

$$\begin{aligned} E\left(\tilde{\beta}^*\right) &= \beta^* \\ Cov\left(\tilde{\beta}^*\right) &= \left(\hat{X}^{*\prime}X^*\right)^{-1} \hat{X}^{*\prime}\Omega\hat{X}^* \left(\hat{X}^{*\prime}X^*\right)^{-1}, \end{aligned}$$

where Ω is defined as in (2.9).

2.3.3 Estimation Procedure

We begin by estimating the selection model (2.1) as a Probit, from which we retain the parameter estimates $\hat{\gamma}$, estimated parameter covariance matrix $\hat{\Sigma}_{\gamma}$, and estimated IMRs $\hat{\lambda}_{it}$.

For observations it in the selected subsample of data, we estimate the derivative of the IMRs $\hat{\delta}_{it}$ from the IMRs, selection criteria, and parameter estimates:

$$\hat{\delta}_{it} = \hat{\lambda}_{it}\left[\hat{\lambda}_{it} + W_{it}\hat{\gamma}\right].$$

Using only the selected subsample of data, we then run the OLS regression

$$y_{it}^{D} = x_{it}^{D}\beta + \hat{\lambda}_{it}^{D}c + e_{it}^{D}. \tag{2.10}$$

As above, the variables y_{it}^{D}, x_{it}^{D}, and λ_{it}^{D} are in deviation form. We retain parameter estimates $\hat{\beta}$ and $\hat{c}$, as well as OLS residuals e_{it}^{D}. The estimates of parameter variances from OLS are not valid for hypothesis testing; rather, we use the covariance matrix Ω that incorporates the variances of v_{it} and $c\left(\hat{\lambda}_{it}^{*} - \lambda_{it}^{*}\right)$ from Proposition 2.2.

To estimate our covariance matrix, we require estimates of the variance σ_{ε}^{2} and correlation ρ. They are consistently estimated by a weighted average of errors

$$\begin{aligned}
\hat{\sigma}_{\varepsilon}^{2} &= \frac{\sum_{i\in n}\left[\sum_{t\in\Im_i}\left(\frac{T_i}{T_i-1}\right)\left(e_{it}^{2} + \hat{c}^{2}\left(1-\frac{1}{T_i}\right)^{2}\hat{\delta}_{it}\right)\right]}{\sum_{i\in n}T_i} \\
\hat{\rho} &= \frac{\hat{c}}{\hat{\sigma}_{\varepsilon}},
\end{aligned}$$

where, as in regression equation (2.10) on the selected subsample of data, e_{it}^{D} is the residual for observation it, and $\hat{c}$ is the estimate of the coefficient on the $\boldsymbol{\lambda}$ vector. Here, we have derived $\hat{\sigma}_{\varepsilon}^{2}$ by substituting c^{2} for $\rho^{2}\sigma_{\varepsilon}^{2}$, and then solving (2.5) for σ_{ε}^{2}, in the spirit of Greene's [1995] discussion of estimation of the Heckit model.

From the Probit model, we use parameter estimates $\hat{\gamma}$ and IMRs $\hat{\lambda}_{it}$. We also use $\hat{\delta}_{it} = \hat{\lambda}_{it}\left(\hat{\lambda}_{it} + W_{it}\hat{\gamma}\right)$ as an estimator for δ_{it}, and take group means $\bar{\hat{\delta}}_i = \frac{1}{T_i}\sum_{t\in\Im_i}\hat{\delta}_{it}$ to estimate $\bar{\delta}_i$.

We then construct the block-diagonal covariance matrix estimator $\hat{\Omega}$, in which the submatrix of estimated covariances corresponding to variables in group i is of size $T_i \times T_i$:

$$\hat{Cov}\left(error_{is}, error_{jt} | is,\ jt \text{ are both in the selected subsample}\right) =$$
$$\begin{cases} \hat{\sigma}_\varepsilon^2\left(1-\frac{1}{T_i}\right)\left[1-\hat{\rho}^2\left(\left(1-\frac{2}{T_i}\right)\hat{\delta}_{it}+\frac{\bar{\hat{\delta}}_i}{T_i}-\left(1-\frac{1}{T_i}\right)\hat{\delta}_{it}^2 W_{it}\hat{\Sigma}_\gamma W_{it}'\right)\right], \\ \text{for } i=j \text{ and } s=t \\ -\frac{\hat{\sigma}_\varepsilon^2}{T_i}\left[1-\hat{\rho}^2\left(\hat{\delta}_{is}+\hat{\delta}_{it}-\bar{\hat{\delta}}_i\right)\right], \text{ for } i=j \text{ but } s\neq t \\ 0, \text{ for } i\neq j. \end{cases} \tag{2.11}$$

For a single-equation model without endogenous variables, the asymptotically consistent estimated covariance matrix, which we may use for inference, is

$$\hat{Cov}\left(\hat{\beta}^*\right) = \left(X^{*\prime}X^*\right)^{-1} X^{*\prime}\hat{\Omega}X^*\left(X^{*\prime}X^*\right)^{-1}.$$

The corresponding 2SLS estimated covariance matrix for a model with estimated endogenous variables is

$$\hat{Cov}\left(\hat{\beta}^*\right) = \left(\hat{X}^{*\prime}X^*\right)^{-1} \hat{X}^{*\prime}\hat{\Omega}\hat{X}^*\left(\hat{X}^{*\prime}X^*\right)^{-1}.$$

2.4 Specifications and Results

2.4.1 First-Stage Probit Model

The first stage is a Probit, with variables as explained in the preceding section. This results in the model

$$\Pr\left(CELLYN_{it}=1\right)=\Phi\left(\begin{array}{c}\gamma_0+\gamma_1 TIME_{it}+\gamma_2 FIXCONN_{it}+\gamma_3 FIXMONTH_{it}\\ +\gamma_4 FIX3_{it}+\gamma_5 ML_{it}+\gamma_6 GDP_{it}+\gamma_7 STAFF_{it}\end{array}\right).$$

Our supply specification depends on whether the selection bias is indeed significant. However, we can only determine this by first specifying and running the two-stage regression discussed previously and testing for significance the coefficient on the deviation form IMRs (henceforth noted in the discussion of the application as $LAMBDA^D$). If we reject the hypothesis of a zero coefficient on $LAMBDA^D$, we conclude that our data suffers from selection bias, and the proper model is the two-stage regression. If we cannot reject the hypothesis and thus assume that the $LAMBDA^D$ selection term is insignificant, we can conclude that our data, as modeled, does not exhibit selection bias, so a straightforward least-squares model on the selected data in deviation form consistently estimates parameters on the explanatory variables. (see, e.g., Greene [1995])

Consequently, we must run the two stage regression in any case, and then run the least-squares model in the case that the $LAMBDA^D$ coefficient is insignificant.

We report results of the Probit selection model in Table 2.1a. On average, the model predicts 85 percent of observations accurately. It predicts 329 of 368, or 89 percent, of observations with wireless, and 173 of 217, or 79 percent, of observations without wireless. We find that the existence of wireless is very well explained by the time, income, and fixed-line telephone variables $YEAR$, GDP, and ML. These variables are all positive and statistically significant. That is, as one would expect, countries with high incomes and well-developed fixed-line telephone infrastructure are likely to offer wireless service, particularly in the later years of the panel. None of the coefficients on prices for fixed-line service are significant. The negative and significant constant variable corrects for the high values resulting from the significant and large time variables, which are narrowly clustered within the years from 1981 to 1998.

A second Probit regression, reported in Table 2.1b, also includes the length of the waiting list for a fixed-line phone $WAIT$ as an explanatory variable. The results of this regression are nearly identical, with the exception that $WAIT$ is negative and significant. That is, a long waiting list suggests the unavailability of wireless.

For the remainder of the results, we use the IMRs in groupwise deviation form $LAMBDA^D$ from the Probit model reported in Table 2.1a, in which the waiting list variable is excluded. The models are not materially different from one another, and the first model's predictions are similar to those of the second.[3] Moreover, our procedure requires exogenous explanatory variables in the selection model. Because there is a risk of undetected endogeneity in the waiting list variable, since the availability of wireless would likely affect the number of people in a country who would wait to obtain fixed-line telephones, we use the IMRs from the Probit that excludes it.

2.4.2 Endogeneity Tests for Second-Stage Model

We surmise that some of the explanatory variables in our model may be endogenous.

[3]The Probit reported in Table 1a has two fewer incorrectly predicted wireless observations (type I errors), but one more correctly predicted non-wireless observations (type II errors).

To address this issue, we would like to test variables in each equation for endogeneity, using the Wald test proposed by Hausman and procedurally simplified by Spencer and Berk [1981]. Since the test statistic, distributed as chi-squared, is a sum of squared normals, it naturally assumes that variances are positive-valued, as is to be expected. However, the subtracted term in the diagonal elements of the regression covariance matrix Ω raises the possibility that some of these elements may be negative. Indeed, Greene [1981], in a comment on Heckman's [1979] paper, notes that the possibility of negative elements in the correcting covariance matrix may be a necessary cost of the selection bias correction. This results in a statistic whose distribution is unknown, since some numbers are subtracted when the statistic should be a sum of (positive) squared numbers, and thus is not useful for hypothesis testing. In attempting to apply Hausman tests to our model with the correcting covariance matrix, we repeatedly obtained negative statistics.

As a substitute, we test our supply model for endogeneity using OLS regressions, rather than the generalized least-squares methodology described above, which produces meaningful statistics. We take the IMRs $LAMBDA_{it}^{D}$ from the first stage as a coefficient in an OLS regression with other second-stage variables in regression form, and test the null hypotheses that each parameter suspected of feedback effects is exogenous. We then retest those variables for which we have rejected the hypotheses of exogeneity, to see if they retain endogeneity given that other variables are also endogenous. Thus the procedure to test for the null hypothesis that an individual dependent variable is exogenous is as follows:

1. Choose a test variable and a list of variables to be used as instruments. Ideally, certain variables used as instruments should be tested for exogeneity themselves. If we unequivocally cannot reject the hypothesis of exogeneity of a variable, we

may choose to use it as an instrument.

2. Regress the test variable on a set of variables assumed to be exogenous, and keep the instrumented version of the test variable.

3. Execute two regression models. The first is an OLS or 2SLS regression of $CELSUBS^D$ on $LAMBDA^D$ and the other variables, using the uninstrumented test variable and any instrumented variables that are assumed to be endogenous. The second is a 2SLS regression of $CELSUBS^D$ on $LAMBDA^D$ and the other variables, using instrumented versions of the test variable(s) and the same assumed-endogenous instrumented variables.

4. The sum of the squared differences of the estimated coefficients and their covariances is distributed as a Chi-square with one degree of freedom. If this sum is greater than the critical value, we reject the hypothesis of exogeneity for the variable.

5. Repeat the procedure for all other variables suspected of endogeneity, assuming even that variables whose tests we have rejected are exogenous.

6. Once we have identified variables that individually appear to be endogenous, we must retest them bilaterally, trilaterally, and so forth, given that other variable(s) for which we also rejected the hypothesis of exogeneity are assumed to be endogenous. This will enable us to identify exactly which variables suffer from endogeneity. Otherwise, we may over-instrument in the system, which could result in collinearity problems.

The instruments that we unequivocally assume to be exogenous are $LAMBDA^D$, $TIME^D$, POP^D, ML^D, and $STAFF^D$. The list of supply variables alternately to be used as instruments and test variables includes $REVENUE^D$, $INVEST^D$, $INTL^D$, $ANCONN^D$, and $AN3^D$. We also include in the list of instruments demand variables GDP^D and $WAIT^D$.

The tests and results are presented in Table 2.2. In general, we test the hypothesis that an explanatory variable is endogenous in the equation of dependent variable $CELSUBS^D$, given that the other explanatory variables are exogenous. The test statistic is shown below the variable subject to testing. We explain the tests as follows:

In the first round, we test the null hypotheses that variables are individually exogenous. We reject the null hypotheses for variables $REVENUE^D$ and $INTL^D$.

In the second round, we test hypotheses that variables are exogenous given that only $REVENUE^D$ is endogenous. In this round of tests, $INTL^D$ again is the only variable for which we reject the hypothesis of exogeneity, given that $REVENUE^D$ is endogenous.

We then test variables, including $REVENUE^D$, given that only $INTL^D$ is endogenous. With this assumption, we cannot reject the null for any variable that we test.

Hence, we conclude that only the volume of international calls $INTL^D$ is endogenous. This conforms to the expectation that international calling volume rises with the number of wireless telephones available from which people can make them.

We show the instrumenting regression for $INTL^D$ in Table 2.3. The significant coefficients include $LAMBDA^D$, ML^D, $STAFF^D$, $REVENUE^D$, and GDP^D, all of which are positive.

2.4.3 Second-Stage Supply Model

Incorporating into the matrix of supply variables an instrumentation of the international calling volume $INTL^D$, we are left with the estimable second-stage supply specification

$$\begin{aligned} CELSUBS_{it}^D &= \beta_0^* LAMBDA_{it}^D + \beta_1^* YEAR_{it}^D + \beta_2^* POP_{it}^D + \beta_3^* ML_{it}^D \\ &+\beta_4^* STAFF_{it}^D + \beta_5^* ANCONN_{it}^D + \beta_6^* AN3_{it}^D + \beta_7^* REVENUE_{it}^D \\ &+\beta_8^* INVEST_{it}^D + \beta_9^* IN\hat{T}L_{it}^D + e_{it}. \end{aligned}$$

Based upon our original model (2.8), we have the identities $\beta_0^* = c$, and $\beta_k^* = \beta_k$ for $k = 0, ..., 9$.

Results of the supply regression model are presented in Table 2.4. We show results with coefficients and t-statistics using ordinary least-squares (OLS), generalized least-squares (GLS) using the Ω covariance matrix described in (9), two-stage least squares (2SLS) using the IV estimator for international calling volume $INTL^D$, and two-stage generalized least squares (2SGLS) using the IV estimator for $INTL^D$ and corresponding covariance matrix Ω.

The Heckit model, as well as this adaptation of it, does not guarantee that its estimates of the variance and correlation parameters σ_ε^2 and ρ will be positive and on the unit disc $[-1, 1]$, respectively. This could result in erroneous estimators and test statistics. In fact, the procedure adapted from Greene [1995] to estimate ρ resulted in correlation coefficients above 1. We suspect this may be due to the small sample size. Thus, we report the estimates with these calculated correlation coefficients (with "Unrestricted Rho," denoted URGLS and UR2SGLS in tables), in addition to estimates with in which the values for ρ are bound to 1 (with "Restricted Rho," denoted RRGLS and RR2SGLS in tables). In the UR2SGLS regression, two parameters actually resulted in negative variance parameters, so t-statistics could not be computed. This particular regression's results diverge significantly from the others', which are largely similar to one another.

Results are as follows:

- In all but the UR2SGLS regression, the coefficient on $LAMBDA^D$ is positive and significant. That is, *ceteris paribus,* countries in which wireless is not yet available are not likely to see the same levels of supply as those in which it is available, should it become so in the future. In the UR2SGLS regression, the t-statistic could not be computed, because the variance for the coefficient on $LAMBDA^D$ was negative, as noted above.

- In all but the UR2SGLS regression, the coefficient on $YEAR^D$ is positive and significant. That is, supply reliably increases with time.

- In all but the UR2SGLS regression, the coefficient on POP^D is negative, and is significant in both the GLS and 2SGLS regressions where rho is restricted to 1. That is, a large population is likely to result in lower per-capita supply, presumably because a small population is easier to serve than a large population.

- In all regressions, the coefficient on staff of the incumbent telephone provider per capita $STAFF^D$ is positive and significant. This supports the notion that employees of incumbent telephone companies add value in providing connectivity to other networks, and refutes the claim that employees use their power as an interest group to undermine deployment of wireless, which would serve as competition to incumbent fixed-line service.

- The prices $ANCONN^D$ and $AN3^D$, for connection to the wireless network and a three-minute call, respectively, are unequivocally insignificant. That is, supply does not appear to be significantly sensitive to price in any regression.

- In all regressions, the coefficient on telephone company revenue per capita $REVENUE^D$ is positive and statistically significant, perhaps because this variable is correlated with national income. However, $REVENUE^D$ may also be correlated with other indicators not observed by our data set, such as privatization and management skills, which are likely to encourage wireless.

- Per-capita volume of minutes of international calls $INTL^D$ has positive and significant coefficients in all regressions in which it is not instrumented, but has negative coefficients in all 2SLS regressions. In the straight 2SLS regression, its coefficient is significant, but this is not the case in either of the 2SGLS regressions. An increase in the number of wireless telephones enables more citizens to place and receive international calls. Thus, the coefficient on the instrumented variable suggests that the volume of international calls has minimal effect on suppliers' decisions. This may be due to the fact that, in most countries, the incumbent fixed-line telephone supplier retains a monopoly on connection to foreign providers, and thus is able to extract the entire economic benefit resulting from the provision of the service.

- Per-capita telephone investment $INVEST^D$ is significant in all of the non-2SLS regressions, but in none of the 2SLS regressions. That is, when we erroneously take $INTL^D$ as exogenous, investment results in increased supply. However, when we more appropriately take $INTL^D$ as endogenous, investment is no longer a significant factor in wireless supply. The instrumenting regression shows that an increase in $INVEST^D$ has an (insignificantly) negative effect on the instrumented $INTL^D$, and thus may dampen the effect of $INVEST^D$ on wireless supply.

2.5 Discussion

We have provided a procedure to estimate parameters on selected data, based upon Heckman's two-stage model, that controls for fixed effects in the second-stage regression. We then employed this procedure to estimate a supply function for wireless telephone subscribership in 124 countries, in deviation form with respect to within-country means.

Some results appear to be robust to the variety of specifications presented here, with the exception in some cases of the UR2SGLS model, which results in negative variances. The supply specification, if applied only to a data set of country-years in which wireless is available, would suffer from sample selectivity bias; such a model applied only to wireless-enabled countries apparently would exaggerate the quantity of wireless supplied as a function of other explanatory variables. Time enables wireless deployment, while a large population hinders it. A large employment base in the telecommunications industry generally facilitates deployment, and does not stifle it, as some economists have suggested. Price has no appreciable effect on the quantity supplied. However, telephone company revenue, which does not itself appear to be a function of wireless deployment, has a significantly positive effect on quantity supplied.

A potential theoretical issue that arises in our specification concerns the first stage in the model. In any country, barring extraordinary circumstances, once a wireless provider builds an infrastructure and a subscriber base, the probability that the country will revert to one in which no wireless service is offered is essentially nil. A proper selection model should thus assume that the probability that a country will have wireless, given that wireless has existed in the past, is one. However, in order to use Heckman's framework specifically, the selection model must necessarily have normally distributed error terms that are correlated with the regression model's error terms, as in Section 1. An exponential duration model, for example, does not have this characteristic.

We employ the Probit model as a practical, albeit imperfect, selection model with normally-distributed error terms. Another approach to this problem may be to compare parameter estimates from the Probit regression to those of a model of data grouped by country. In such a model, we might estimate the probability of adoption in a particular country as the proportion of all observations for that country for which wireless is available, as a function of the means of explanatory variables across time. We note that relationships between this duration model and the probit model remain to be explored. Moreover, this alternative will only provide us with a single observation for each country, so we would not be able to retain estimated IMRs from the model, necessary for the second-stage regression in a Heckman-style procedure. An alternative is the random effects model, described below.

Another technical tradeoff in our model is the necessity to revert to OLS-based tests for endogeneity. Since the theoretical distribution of quadratics of errors using this modified Heckit model are not known, we cannot conduct specification tests on them. Instead, we use Hausman's test, which uses quadratics of OLS residuals, which are known to have a chi-squared distribution, and thus are meaningful in specification tests. We view this as a favorable alternative to either a haphazard process of guessing which variables are endogenous, or ignoring endogeneity altogether.

We also encountered the practical issue of estimating a correlation coefficient ρ at a value outside the interval $[-1, 1]$, which resulted in negative varances for certain parameters, and thus made statistical inference impossible. However, substituting the nearest value within the interval (namely 1) for our excessively large ρ seemed to solve this problem without materially changing t-statistics for other parameters. We conclude that this appears to be a reasonable solution to the problem; however, we note that the statistical properties of this substitution are not known.

2.5.1 Future Topics

One other model that could be explored by a researcher with access to ample computing power is a dummy variable model, in which the group coefficients α_i are estimated. This would simplify the process of estimating the covariance matrix Ω significantly, because it does not require computation of deviation-form variables.

There are a few alternative approaches to the issue of selection, in light of the "No-exit" nature of technology adoption. Hausman and Wise [1979] develop theory for a maximum-likelihood method to measure attrition bias, or the estimation bias that results from survey subjects entering a panel after the first year the survey is taken, and/or exiting before the last year. Since their survey subjects are sampled from a large pool of individuals, they estimate a random effects model, in which the individual-specific term is measured as an error term constant across observations for the individual drawn from a pool, rather than a fixed effect, in which the individual-specific constant does not have the properties of an error (i.e., a normal distribution centered at zero). Rather than using a panel for sample selectivity, Hausman and Wise assume all individuals are included in the panel in Year 1 of a stylized two-year panel, and then use grouped data to estimate the probability of appearing in Year

2. They extract the IMRs from the probability model for individuals who appear in Year 2, and include them in a regression on the Year 2 data, to give estimates conditional upon survival to Year 2. They then generalize the theory for multiple years by specifying individual attrition equations for each year.

Ridder [1990] generalizes the random effects model of Hausman and Wise by allowing subjects to enter and exit a panel repeatedly. An individual's inclusion in the panel is specified as a Markov chain, in which he enters, exits, or remains in or out of the survey between any two survey periods. Much of Ridder's discussion is devoted to the intuition behind attrition. He notes that individuals with the largest negative random effect, i.e. those who are most "reluctant to participate," exit the survey sample first, leaving the sample with a greater average random effect. He proves this in a proposition, with the following results: First, if an individual is more likely to be included in the sample in a given year, his or her individual effect decreases. Second, by easing restrictions on inclusion in year t, the as a pool of individuals is more likely to be included in a given year t based upon the selection criteria, the error in selection in year t becomes smaller, and the error in prediction in another year s becomes larger.

A random-effects model to estimate supply would employ a reverse-attrition phenomenon. One could execute a random-effects probit regression on the pool of selected and non-selected data using the methodology of Ridder. In such a specification, we include a random effect in the first-stage probit. The model for an ideal data set is

$$z_{it}^{*} = W_{it}\gamma + \alpha_i + \varepsilon_{it}$$

(where α_i is an error term), and the set of explanatory selection variables W_{it} includes:

- A complete lagged history of exogenous variables $w_{it}, w_{i(t-1)}, \ldots, w_{i1}$;
- A variable d_{it} that denotes the number of years before adoption whenever $z_{it}^{*} < 0$, or is equal to zero whenever $z_{it}^{*} > 0$;[4] and
- A variable $\left(1 - z_{i(t-1)}\right)$ which is equal to one in the year of adoption.

This produces IMRs dependent upon the time remaining until adoption, which can then be inserted into a subsequent random-effects regression model. This model may be better-specified than a fixed-effects (or groupwise deviation form) model; however, we note the following caveats: The large number of coefficients on the lagged historical variables $w_{i(t-s)}$ require considerably more data than our original specification (10) for well-conditioned estimation, and the issue of the groupwise error term discussed above may result in biased estimates in the selection model.

[4] The years-to-adoption variable d_{it} implies that a country necessarily does adopt by the end of the panel. However, since most countries in the world have some form of wireless by the time of this writing, one could assume an upper bound of 2001 less the first year of the survey for those countries which never adopt by the end of the survey.

Table 1a: Selection equation

VARIABLE NAME	ESTIMATED COEFFICIENT	ASYMPTOTIC STANDARD ERROR	T-RATIO	ELASTICITY AT MEANS	WEIGHTED AGGREGATE ELASTICITY
YEAR	0.41304	4.11E-02	10.053	124.11	234.23
FIXCONN	0.4187	0.27283	1.5347	1.23E-02	1.96E-02
FIXMONTH	-14.77	11.882	-1.2431	-1.69E-02	-2.20E-02
FIX3	-169.66	836.96	-0.20271	-2.32E-03	-3.99E-03
ML	6126.2	1831.6	3.3447	0.17055	0.12256
GDP	0.16016	4.61E-02	3.4729	0.18856	9.35E-02
STAFF	-1.63E+05	1.26E+05	-1.2934	-4.00E-02	-5.21E-02
CONSTANT	-823.93	81.951	-10.054	-124.21	-234.37

LOG-LIKELIHOOD FUNCTION -183.62
LOG-LIKELIHOOD (0) = -3 85.78
LIKELIHOOD RATIO TEST = 404.323 WITH 7 D.F.

		actual 0	actual 1
predicted	0	173	39
	1	44	329
		217	368
		80%	89%

Table 1b: with Wait variable for comparison

VARIABLE NAME	ESTIMATED COEFFICIENT	ASYMPTOTIC STANDARD ERROR	T-RATIO	ELASTICITY AT MEANS	WEIGHTED AGGREGATE ELASTICITY
YEAR	0.42405	4.21E-02	10.068	108.31	234.02
FIXCONN	0.47206	0.28338	1.6658	1.18E-02	2.13E-02
FIXMONTH	-18.432	11.699	-1.5755	-1.80E-02	-2.68E-02
FIX3	-483.26	851.29	-0.56768	-5.61E-03	-1.10E-02
ML	7085.7	1932.3	3.667	0.16767	0.13608
GDP	0.14855	4.73E-02	3.1409	0.14865	8.31E-02
STAFF	-1.04E+05	1.30E+05	-0.79849	-2.18E-02	-3.22E-02
WAIT	-8394.8	3211.4	-2.614	-1.61E-02	-4.39E-02
CONSTANT	-845.8	84.004	-10.069	-108.38	-234.14

LOG-LIKELIHOOD FUNCTION -179.62
LOG-LIKELIHOOD (0) =-3 5.78
LIKELIHOOD RATIO TEST = 412.32 WITH 8 D.F.

		actual 0	actual 1
predicted	0	174	41
	1	43	327
		217	368
		80%	89%

Table 2: Hausman Tests

Round		Test variable (supply variables) Revenue	Invest	Intl	AnConn	An3
1		10.8253	1.424	14.74	1.8242	0.044
	Reject?	x		x		
2r			0.942	5.414	0.8505	0.39
	Reject?			x		
2i		0.4948	0.412		0.1956	0.58
	Reject?					

Round 1: All variables except test variable are assumed exogenous.
Reject Ho:Exogeneity for Revenue, Intl.

Round 2r: Revenue is assumed endogenous;
all other variables except test variable are assumed exogenous.
Reject Ho:Exogeneity for Intl.

Round 2i: Intl is assumed endogenous;
all other variables except test variable are assumed exogenous.
Reject Ho:Exogeneity for Revenue. Conclude that only Intl is endogenous.

Table 3
Instrumenting regression for Intl

Instrument	Coefficient	T-stat
lambda	0.016152	2.05
year	0.000794	0.44747
pop	-0.00285	-1.64
ml	693.7042	6.336
fixconn	0.002842	0.4306
fixmonth	0.070224	0.06319
fix3	-56.0127	-0.8562
staff	33572.41	3.263
telrev	0.105713	2.296
invest	-0.01876	-1.003
gdp	0.007592	4.8
wait	440.8565	1.59
anconn	-0.00032	-0.1553
an3	-0.03753	-0.1673

Other variables have t-stats in [-1,1] (to be filled in later)

R-squared 0.625248

Table 4: Supply model

Variable	OLS coeff	OLS t-stat	GLS t-stat	Rho=1 GLS t-stat	2SLS coeff	2SLS t-stat	2SGLS t-stat	Rho=1 2SGLS t-stat
lambda	5.79	6.48999 *		11.144	8.1	6.174 *		8.21
year	1.944	9.8287	6.005	6.72	2.207	8.46	0.347	5.73
pop	-0.22	-1.19 *		-2.54	-0.406	-1.649 *		-2.17
main lines	-30644	-2.5996	-1.833	-1.89	22504	1.046	0.509	0.63
staff	3906356	3.449	2.37	2.48	6507818	4.0232	2.33	2.63
analog conn chg	-0.07546	-0.309	-0.209	-0.22	-0.1595	-0.516	-0.33	-0.35
analog 3-min call ch	12.1795	0.46	0.309	0.32	11.38	0.343	0.2154	0.23
telco revenue	24.24	4.8558	3.259	3.46	42.65	5.149	3.273	3.49
int'l calls (IV)	19.6466	2.84	2.14	2.119	-66.33	-2.4898	-1.23	-1.514
telco investment	5.299	2.407	2.89	2.03	3.73	1.326	1.57	1.05
Sigma2			12.367	12.367		21.768	21.768	21.768
Rho2			2.712	1		3.0144	3.0144	1
Rho			1.646815	1		1.73620275	1.736202753	1
sqrt(sigma2*rho2)			5.791313	3.51667		8.10046043	8.100460431	4.66561893
C			5.791	5.791		8.1	8.1	8.1
R-squared		0.674				0.7779		

CHAPTER 3

MANAGING BANDWIDTH WITH A SECOND-PRICE AUCTION

3.1 Introduction

As they are designed today, most telecommunications networks suffer from the *tragedy of the commons,* in which one consumer's use of a shared resource degrades the quality of the resource for others. High-value users – and in some cases *all* users – may at times find the network unusably slow or unreliable, due to congestion and the resulting loss of information. This paper is a preliminary attempt at a design for a data router that discriminates by means of price in the presence of congestion.

Elwalid and Mitra (EM) [1992] propose an algorithm for multiplexing information from independent sources of data connected to a network, each of which sends information of multiple priority classes to a router located at a common node. The router either dispatches arriving information immediately toward its destination, or stores it in a buffer until it can be routed downstream in the network. Their methodology has the restriction that each source be stochastically homogeneous, and, in particular, each send information of the same set of priority classes to the buffer. Thus, the router treats each source's information identically, so it cannot discriminate by means of a price mechanism.

We extend a case of the model of EM [1992] to allow different sources to send information of different priority classes. This allows the incremental throughput of the high-priority source to be rationed independently of the remaining default throughput rate that is common to all users. We propose the auctioning of the resource using a second-price auction of Vickrey [1961]. The auction mechanism gives priority to users based upon their respective willingness to pay, and has the desirable feature that users have an incentive to bid the amount that they actually value their use of the resource.

3.1.1 Background

A source of digital information in a network, such as a wireless telephone or a web server, passes digitized information to a recipient of that information, such as another telephone or a user's computer, elsewhere in the network. The data are encapsulated into standardized packets with destination addresses contained in the packet headers. Thus, when the source sends a packet to a nearby node, a router reads the packet's address and dispatches it to another node, and so on, en route to the packet's ultimate destination. Networks eliminate the need for every information sender to be connected with every information recipient by means of a "hub-and-spoke" system, and can be optimized to handle a statistically predicted subset of all potential users at any given time. Digital packet-switched networks are more efficient than traditional circuit-switched networks because they can *multiplex* data: when one source is not sending information to a nearby node, such as when a person speaking on a telephone pauses, the available resource can be used to route someone else's information. A circuit-switched network, meanwhile, provides a complete end-to-end connection between sender and recipient at all times. For this reason, a circuit-switched telephone network can accomodate only a small fraction of the traffic volume that would be possible on a packet-switched network of equivalent size.

The inherent tradeoff, of course, is that too many users independently will make the decision to send (or to receive) data at the same time. A router is designed with this possibility in mind, and contains a buffer – memory – to store information when it arrives at a node faster than the router can dispatch it. During periods of heavy congestion, such as a peak hour of telephone use, the router will need to buffer much of the data it receives. In extreme situations, the buffer will fill to capacity. In this case, the buffer can only accept additional packets at the rate the router can dispatch packets already stored in the buffer. Any information that arrives in excess of this rate cannot be stored in the overflowing buffer, and thus is lost forever. The result is that the destination does not receive all of the packets sent by the source. This results, for example, in choppy, intermittent silences – and disconnections – during a conversation over a congested digital cellular network.

Frequently, during the peak hour, the quality of service on the typical digital wireless network erodes to the point that a conversation is not possible. Meanwhile, users who are billed by the minute continue to be charged for their calls. This results in the curious situation that the provider charges more money, while adding less economic value for its customers, than at any other time during the day.

As the telecommunications service industry becomes more competitive, a provider may need to find a solution to this problem in order to survive. The issue will exacerbated by the migration of services away from fixed communications media, such as optical fiber, and toward the capacity-constrained airwaves.

In this paper, we provide the theory that enables a simple second-price auction mechanism to be used to manage data traffic into a buffer on a per-call basis. We begin by discussing some relevant work in the fields of network engineering and auctions. We discuss the model of EM [1992], upon which our work is based, beginning with a description of the individual sources and continuing with the development of the buffer architecture, which treats sources as stochastically homogeneous with respect to priority. We then explain our innovation, which extends EM [1992] by allowing sources to send information that is heterogeneous with respect to priority. This enables us to apply a single-unit second-price auction to the property right to send data during periods of congestion. The low bidder in the auction, meanwhile, faces a specified rate of packet loss during those same periods. We note that the second-price auction's familiar properties, such as demand revelation, apply to the auction for bandwidth. We conclude with welfare implications, implementation issues, and opportunities for future research.

3.1.2 Seminal Papers in Real-Time Data Traffic

Data flowing into a router are often modeled as *fluid processes*, or streams of intermittent bursts of information, much like faucets that are repeatedly being turned ON and OFF. The router's buffer, in turn, is analogous to a sink with a drain, so that the buffer fills whenever the total rate of inflowing information exceeds the constant rate at which the router is able to dispatch information to its destination.

Anick, Mitra, and Sondhi (AMS) [1982] construct a model in which data arrive as fluid processes and fill a buffer of infinite length. In this seminal work, the streams of data intermittently turn ON and OFF according to stochastic processes. The lengths of ON and OFF periods are exponentially distributed and the buffer drains at a constant rate. AMS deduce and solve differential equations that define probabilities for any possible quantity of data in the buffer.

Elwalid and Mitra [1992] (EM) recognize that different types of traffic have different sensitivities to loss and delay. In their second-generation model, each packet of data is associated with a priority class, and each source sends packets of one or more classes. When the buffer fills beyond a specified threshold, data of the lowest admitted class are no longer accepted, and thus are lost. We discuss their model in greater detail in the next section.

EM's contribution has taken root and has been incorporated into the specification of the Asynchronous Transfer Mode (ATM) protocol for traffic management with quality-of-service guarantees. However, EM do not address the problem of determining priority classes, but do assume traffic will be partitioned into classes appropriately. Rather than examine the problem of determining economically efficient routing, they consider the need to prioritize traffic at a single node in a network when loss of certain data is inexpensive, and can thus be categorized as "low-priority." They cite the example of real-time digitized video, in which a small amount of stochastic data loss will have no noticeable effect on picture quality, and note that a few classes of priority is acceptable for such traffic.

The risk, of course, is that an end-user will encroach on the public nature of the network by using software that tags all of her information as high-priority. The ATM Forum addresses these issues by establishing "Policing" specifications, but sidesteps the issue of contract negotiation, and particularly the process by which a user negotiates the contract for her source's data on a per-call basis.

3.1.3 Seminal Papers in Auctions and Auction-Based Management

While the advent of packet-switching technology has increased the efficiency of data networking, it also adds a great deal of complexity to the process of pricing services. It necessitates that we associate prices for transportation with individual packets, or rates of packet flow. The supply of equipment to route packets will generally be constant, but the demand for this equipment – the flow of packets themselves – will fluctuate dramatically over the course of a day or even a minute at any given point in a network. This in large part is the reason data services have traditionally been provided at fixed prices without regard to congestion. Setting individual prices requires that users be able to express their willingness to pay for rates of packet flow in real time. During most periods, this may not be problematic; however, it becomes very difficult during peak periods, when much information is delayed and even lost. Demand in some market segments has been sufficiently met by technological change. In other markets, such as that of wireless telephone traffic during rush hours, supply remains overwhelmed.

Vickrey [1961] proposes the "Second-Price" competitive simultaneous sealed-bid auction for a single indivisible good, in which the highest bidder purchases the good, but pays the second-highest price. The second-price auction is said to be demand-revealing because a bidder will find it in her own best interest to submit a bid of her actual value of the good. This is due to the fact that increasing a bid raises the probability of winning, since the bidder is more likely to be the high bidder, without resulting in the usual auction tradeoff of increasing the price she will pay, since her price is actually determined by the second-highest bid. Vickrey compared the second-price auction with other auctions, such as decreasing-clock Dutch auctions and the first-price simultaneous sealed-bid auction. His theorem of revenue equivalence states that the predicted prices in second-price and (non-demand-revealing) first-price auctions are equal.

Naor [1969] is the first to use a congestion price to alliviate congestion in an M/M/1 queue. In his model, identical customers arrive to a queue as a Poisson process, and are served by a single server according to a first-in, first out (FIFO) discipline. Each customer gains a fixed economic value by being served, but incurs a waiting cost. A customer joins the queue if the value from service exceeds her expected waiting cost, given the length of the line she encounters upon arrival. Naor specifies the equilibrium queue length below which an arriving customer should join the queue, and above which she should not bother waiting for service. He calculates the waiting tolls that maximize the joint benefits to waiting customers and the toll collector, and to a monopolist toll collector acting in self interest.

Mendelson and Whang (MW) [1990] enrich Naor's work by considering an arriving customer to be a member of one of several priority classes. Their goal is to design a server to serve customers of high priority before customers of low priority. Since a customer's inherent priority class is known only to the customer herself, MW elicit self-selection in a Pareto-optimal separating equilbrium by charging a higher price for immediate service, which is paid only by high-priority customers.

Hassin [1995] takes a similar approach, but allows customers to bid their own prices. He shows that this enables customers to be served in the socially optimal sequence.

Mackie-Mason and Varian (MMV) [1994] propose the second-price auction for Internet access. They provide a conceptual framework for bidding and routing, but do not detail the engineering architecture necessary to build such a network. They were the first to raise the issue of economic value to the engineering community.

Gupta, Stahl, and Whinston (GSW) [1994] construct a model of an economically efficient computer network. Assuming that end-users act as consumers, processors act as producers, and servers act as market-clearing agents, they use prices to allocate costly and time-consuming processing resources efficiently. In their model, each processor services tasks according to priority class, in which high-priority tasks preempt low-priority tasks, and charges a competitive market-clearing price according to processing time and priority class. A typical end-user is presumed to have a shopping agent that offers her a menu of priority classes at different prices, from which she selects a portfolio of services to maximize a utility function. While GSW contribute to the development of a theoretical topology of a computer network, and provide a framework for allocative efficiency using market-clearing prices on the network, they treat computing tasks as individual jobs with specfic service times that can be organized into queues, rather than (infinitely divisible) packets, as information is modeled by EM [1992].

Later, GSW [1997] construct a theoretical model of the Internet that specifically addresses the issue of aggregating costs across network resources. They conduct a simulation of the Internet and project that a lack of prices would result in deadweight losses of \$10 to \$20 billion in 1999.

Kelly [1997] uses the concept of *effective bandwidth* – a transmission threshold rate, below which a network is ensured a low probability of data loss – to develop contract specifications for users on a network. Kelly notes that statistical multiplexing demands specific knowledge of the parameters of traffic that each data source sends to the network, and responds to this need by developing a mechanism by which users have an incentive to anticipate and reveal their expectations of bandwidth consumption on the shared resource. Specifically, the network requires the user to submit a limit price, from which the network deduces the mean rate and effective bandwidth for the user's transmission of information. This paper contributes to the literature by establishing a method by which a network congestion management system can estimate a ratio of buffer capacity to the total number of connections of specified throughput rates, and may assist the operator in deciding how many users to permit on a network at a given time. However, while their mechanism encourages users to report their average send rates accurately and enables the network operator to estimate the probability of network overflow, it does not address the issue of allocating priority when the overflow actually occurs. In this case, the network operator may wish to supplement this price with a congestion price that allocates the right to this priority. For example, a voice telephony user of a congested integrated-services network that is also used for large file transfers may have relatively low bandwidth requirements. Her real-time conversation would be more demanding with regard to minimizing data loss, since her end-user equipment's software could not offer the luxury of multiple redundancy checks when her information is lost.

Wang, Peha, and Sirbu (WPS) [1997] develop a monopoly pricing model for an integrated-services network with performance guarantees. Their model assumes the network will admit or reject calls based upon some criterion. A source whose call is accepted sends individual packets to a buffer as a poisson process. WPS calculate an optimal pricing policy by solving a bandwidth investment optimization problem, determining the optimal price for guaranteed service on the optimized bandwidth, and then provide spot-pricing for low-priority service. This model provides the significant contribution of a network resource allocation mechanism that offers service guarantees, with prices established at the beginning of a call and fixed for the duration of the call. It also assumes arrivals of cells of fixed size. This differs from our model in that we price congestion in real time, and assume arrivals of information as fluid bursts of random duration.

Our explanation of the theory of the buffer begins with the statistical properties of the sources that pass information to the router. We continue with the theoretical development of the buffer, following AMS [1982] and EM [1992] in spirit, but with modifications that enable the buffer to isolate priority classes by auction. This analysis culminates with the quantified specification of the item to be auctioned.

The development of the buffer is followed by the application of the Vickrey auction. We conclude with discussion of practical implications and ideas for future research.

3.2 Model

We model two potential users of a network uploading (or downloading) data through a single particular node within the network. Each user's data must pass through a router, which dispatches packets toward their destinations. When the router receives information faster than it is able to dispatch it, it must store the information in the buffer until it is free to dispatch the information.

Consider a single user, equipped with a data-sending device, such as a wireless telephone or a computer, which we shall refer to as a *source*. At any point t in time, the source owned by user i ($i = 1, 2$) will alternately be in an ON-condition or an OFF-condition. The state σ_{it} of source i at time t is denoted

$$\sigma_{it} = \begin{cases} 0, & \text{if Source } i \text{ is OFF at time } t \\ 1, & \text{if Source } i \text{ is ON at time } t. \end{cases}$$

When ON, each of these sources passes information to the router. A contiguous ON-period is often referred to in the literature as a *burst* of information.

Let us associate a *bid* b_i ($i = 1, 2$) with each source. Because the nature of bidding depends directly on the behavior of sources and the buffer, we refrain from explaining the bidding process until we have completed the introduction of the engineering, and for now take the bids as given.

We define the *ordered source* $< i >$ as the source associated with the ith highest bid. We analogously define the *ordered state* $\sigma_{\langle i \rangle t}$ as the state of the source associated with the ith highest bid at the moment t in time.

We write the vector Σ_t of states of both sources at a point t in time as

$$\Sigma_t = \left(\sigma_{<1>t}, \sigma_{<2>t}\right).$$

For example, the vector of states $\Sigma_t = (1, 0)$ indicates that source $< 1 >$ associated with the highest bid is ON and the source $< 2 >$ associated with the second-highest bid is OFF at instant t in time.

3.2.1 Assumptions and Rules

Without any way to predict exactly which sources will be ON at which times and for what durations, we merely assume that a source alternates between ON- and OFF-states, and the duration of each ON-period and OFF-period are random. These random lengths are typically modeled as exponentially distributed, as we shall do here. Thus, we make the following assumptions about OFF- and ON-periods of sources:

Assumption 3.1. The length τ of an OFF-period is a random variable, independently and identically distributed (i.i.d.) exponentially with parameter λ; i.e., the cumulative distribution function (c.d.f) of the length of the OFF-period is

$$\begin{aligned} \Pr(\tau \leq t) &= \int_0^t \lambda e^{-\lambda\tau} d\tau \\ &= 1 - e^{-\lambda t}. \end{aligned}$$

The length τ of an ON-period is also a random variable, i.i.d. exponentially distributed with parameter 1; i.e., the c.d.f. of the length of the ON-period is

$$\begin{aligned} \Pr(\tau \leq t) &= \int_0^t e^{-\tau} d\tau \\ &= 1 - e^{-t}. \end{aligned}$$

ASSUMPTION 3.2. During an ON-state, a source passes information to the router at the rate of 1 unit of information per unit time (ui/ut). During an OFF-state, it passes no information to the buffer at all; or, equivalently, it passes information at the rate of 0 ui/ut.

The buffer has a maximum capacity of B units of information; $B \in (0, \infty)$. At any moment t in time, the buffer contains a quantity $X_t \in [0, B]$ of information, or *content* (we use these terms interchangeably).

The buffer passes information to the router to be dispatched at its *dispatch rate* of c ui/ut, whenever $X_t > 0$; that is, the buffer drains at constant rate c whenever it contains information.

ASSUMPTION 3.3. The dispatch rate c is not too large. Specifically, $0 < c < 2$.

As long as $c < 2$, the buffer can fill and congestion is a possibility. For $c > 2$, the buffer always passes information to the router faster than it can receive information; in this case the buffer never fills at all and consequently is of no use.

When neither source is on and the buffer has a positive quantity of content, the buffer content decreases at a rate of c ui/ut, since it is receiving no information but dispatches at the constant rate c ui/ut. When exactly one source is ON, the buffer content X_t decreases at the rate of $1 - c$ ui/ut, because it is receiving at the rate of 1 ui/ut and dispatching at the rate of c ui/ut. Similarly, when both sources are ON, it increases at the rate of $2 - c$ units of information per unit time.

During an interval $[t, t + \Delta t]$ of time, the buffer content's *drift* Δx is the change in the buffer content X_t.

In order to allocate priority during periods of congestion, we set rules by which the buffer accepts data to be dispatched, as follows:

RULE 3.1. The buffer content X_t has a threshold quantity B_1 of information, above which it does not accept additional information from the source associated with the low bid. That is, there exists a quantity $B_1 \in (0, B)$ such that whenever $X_t > B_1$, the buffer accepts only information from the source $< 1 >$ associated with the high bid. In this case, any information from the source $< 2 >$ associated with the low bid is discarded, or *lost*, and will not be dispatched by the router. If $X_t = B_1$, the buffer accepts only a portion of the information from the source associated with the low bid, so that the buffer does not fill beyond B_1.

RULE 3.2. Whenever the buffer is full; that is, when $X_t = B$, it will continue to accept information from sources at the dispatch rate c, according to the convention that the source $< 1 >$ associated with the high bid is served first. That is, the buffer will accept information from the source associated with the high bid when it is full up to the dispatch rate c. The buffer will then accommodate information from the source associated with the low bid such that the total inflow rate from the two sources does not exceed the dispatch rate c. The remainder of the information sent when the buffer is full will be lost.

We refer to the condition in which the buffer is full at least to the threshold quantity B_1 as "congested."

Finally, we add an assumption on the threshold level B_1:

ASSUMPTION 3.4. For $1 < c < 2$, the threshold B_1 is equal to the total buffer capacity B.

We include Assumption 3.4 because the threshold level B_1 of content in the buffer is irrelevant for $1 < c < 2$. If the buffer were to refuse information from one source above a threshold B_1, it would fill at a rate of at most $1-c < 0$; that is, it necessarily would be decreasing in content. Thus, it could never fill beyond B_1, rendering a portion of the buffer of magnitude $B - B_1$ useless.

To summarize, the router receives information passed to it by the data sources. The router dispatches information at a constant rate toward its destination. If the the router can route the information as quickly as it receives the information, it will do so. Otherwise, it must store the information in its buffer until it is free to dispatch the information. We model the buffer as a "Leaky bucket" in the spirit of AMS [1982] and EM [1992]. Thus, when sources pass data to the router faster than it can dispatch information, the buffer fills. When the sources pass data to the router at a slower rate than it is able to dispatch information, the buffer will empty until it is completely empty. When the buffer is empty and information arrives at a slower rate than the router can dispatch information, the newly arriving information can be dispatched immediately. However, if the buffer is full and new information arrives more quickly than the router can dispatch information, the excess information beyond the incoming throughput will be lost. As in EM [1992], information stored in the buffer is routed in accordance with a FIFO service discipline.

Rule 3.1 defines the first priority rule for the buffer. The buffer accepts and stores information from the source associated with the high bid without regard to its content X_t. It will accept and store information from the source associated with the low bid provided the buffer content is less than the threshold quantity B_1. In the case that the dispatch rate $c < 1$, if X_t is exactly equal to B_1, information from either source causes the buffer to fill. Hence, if $c < 1$ and $X_t = B_1$, the buffer accepts information from source $< 2 >$ if and only if source $< 1 >$ is OFF, and then only a portion c ui/ut of the information it generates.

In the case that $c > 1$, if the buffer content $X_t = B_1 = B$ (by Assumption 3.4) and source $< 1 >$ is ON, the buffer can accommodate a portion c ui/ut of the information from source $< 2 >$ while remaining full exactly to the threshold level B_1. In the case that $c > 1$, $X_t = B_1 = B$, and source $< 1 >$ is OFF, the buffer is able to accept any and all information from source $< 2 >$. Even when the buffer accepts information from source $< 2 >$ at the rate of 1 ui/ut, it dispatches more quickly than information arrives, and the content X_t decreases from the capacity level B toward zero.

Rule 3.2 is a policy to be used when the buffer is in a state of extreme congestion; that is, when it is full to the point of capacity, i.e. $X_t = B$, *and* sources continue to send information, resulting in overflow (i.e. loss of valuable information). In the case that $c < 1$, when $X_t = B$ and source $< 1 >$ is ON, the buffer will store a portion c units of its information per unit time, and the remaining $1-c$ units of information per unit time that it sends will be lost. There is no way to avoid this possibility of loss of information from the high bidder; however, we can lower its probability by either increasing the buffer capacity B or decreasing the threshold level B_1. The former option requires the purchase of more memory for the router. The latter option, which is tantamount to a more conservative concept of congestion, results in the tradeoff of a greater probability of loss of information from source $< 2 >$.

Note that Rule 1 precludes information from source $< 2 >$ ever from being accepted when the buffer is full. We impute Rule 3.2 to the case that $1 < c < 2$. In this case, then, information from the source $< 2 >$ associated with the low bid will be lost only when the buffer is full and both sources are on, and in this case only the fraction $c - 1 < 1$ units of information per unit time will be lost. When only the source

$< 2 >$ is on, it sends information to the buffer at a rate slower than the dispatch rate c, so the buffer content X_t actually decreases, and the source's information can be dispatched in its entirety at all times.

3.2.2 Behavior of Sources

We now derive some basic results about the sources' off- and on-periods.

PROPOSITION 3.1. The expected lengths of OFF- and ON-periods are $\frac{1}{\lambda}$ and 1, respectively. A source will be in an OFF-state approximately a fraction $\frac{1}{1+\lambda}$ of the time and in an ON-state approximately a fraction $\frac{\lambda}{1+\lambda}$ of the time. The source-state vector at any moment will be in one of four states, each with probability in accordance with the following table:

$$\Sigma_t = \begin{cases} (0,0) & \text{with probability } \frac{1}{(1+\lambda)^2} \\ (0,1) & \text{with probability } \frac{\lambda}{(1+\lambda)^2} \\ (1,0) & \text{with probability } \frac{\lambda}{(1+\lambda)^2} \\ (1,1) & \text{with probability } \frac{\lambda^2}{(1+\lambda)^2} \end{cases}$$

The following lemma articulates the exponential distribution's well-known property of *memorylessness*, or independence of history. It states that the length of time that an ON- or OFF-period will continue to last beyond a point in time, is independent of the length of time that the period has existed up until that point in time.

LEMMA 3.1. (The "Memoryless" Property) Let the length of time that a source will be OFF be described by the random variable τ, distributed exponentially with parameter λ. The probability that the source will remain OFF until the conclusion of a small interval $(t, t + \Delta t)$ of time, given that it is OFF at the beginning of the interval t, is equal to the probability that it will be OFF for the duration of an interval Δt. That is,

$$\Pr(\tau > t + \Delta t | \tau > t) = \Pr(\tau > \Delta t)$$

for the OFF-period τ. The result for ON- periods is similar: if θ is a random variable that describes the length of time that a source will be ON, and is exponentially distributed with parameter 1, the probability that the source will remain on through $(t, t + \Delta t)$ is equal to the probability that it will be ON for the duration of an interval Δt; i.e.

$$\Pr(\theta > t + \Delta t | \theta > t) = \Pr(\theta > \Delta t).$$

The memoryless nature of the exponential distribution implies that the probability that a source that will be ON (OFF) at some moment in the future depends only on its present state, and is independent of the length of the current ON-spell (OFF-spell).

We now evaluate a source's OFF- and ON-hazard rates – the probability that a source that is OFF at the "present" moment t will be ON, and the probability that a source that is ON will be OFF – at the end of a short interval Δt of time. We shall see that the hazard rates are proportional to the length of the interval. Specifically, the probability that an OFF-source will be ON is approximately $\lambda \Delta t$, and the probability that an ON-source will be OFF is Δt. The probabilities that these events do not occur are $1 - \lambda \Delta t$ and $1 - \Delta t$, respectively.

PROPOSITION 3.2. The hazard rate of the random variable τ that describes the length of an OFF-period, or probability that the OFF-period will end before a moment $t + \Delta t$, given that it has not yet ended by a moment t, is approximately $\lambda \Delta t$. The hazard rate of the random variable θ that describes the length of an ON-period is approximately Δt. Specifically

$$\begin{aligned} \Pr(\tau < t + \Delta t | \tau > t) &= \lambda \Delta t + o(\Delta t) \\ \Pr(\theta < t + \Delta t | \theta > t) &= \Delta t + o(\Delta t), \end{aligned}$$

where

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0.$$

That is, the $o(\cdot)$ terms become negligible for sufficiently small Δt.

Proposition 3.2 implies that the terms of order of Δt higher than 1 rapidly decrease to zero as the interval of time Δt becomes small. Thus, for brief intervals of time, the hazard rate is a function only of the term of order 1.

COROLLARY 3.2.1. The probability that the length of an OFF-period exceeds an interval of length Δt is approximately $1 - \lambda \Delta t$, and the probability that the length of an ON-period exceeds an interval of length Δt is approximately $1 - \Delta t$; i.e.,

$$\begin{aligned} \Pr(\tau > t + \Delta t | \tau > t) &\approx 1 - \lambda \Delta t \\ \Pr(\theta > t + \Delta t | \theta > t) &\approx 1 - \Delta t. \end{aligned}$$

Next, we show that, for Δt small, compound events have approximately zero probability. That is, the probability that more than one source will change states, or a single source will change states more than once, within a small interval Δt, is $o(\Delta t)$.

COROLLARY 3.2.2. Let τ_1 and τ_2 be random variables that describe the lengths of the current states (ON or OFF) of sources 1 and 2.

For Δt small, compound events have negligible probability. That is, for any $\Delta t > 0$, compound events have positive probability, but only of order $o(\Delta t)$; i.e.

$$\Pr\left(\tau_1 < t + \Delta t, \tau_2 < t + \Delta t | \tau_1 > t, \tau_2 > t\right) = 0 + o(\Delta t),$$

where

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0.$$

To recap, we have modeled sources that intermittently turn ON and OFF. Regardless of whether how long a source has been OFF (ON), its probability of turning ON (OFF) is constant over time. When ON, sources generate information and pass it to the router at a rate of 1 ui/ut.

3.2.3 Behavior of Buffer Content

The state of the buffer at a particular moment is defined by the quantity of content in the buffer, the number of sources from which the buffer accepts information, and the resulting rate of change of the quantity of content. Given the state of the buffer, it could change to any one of other states with a specific probability, all of which form a distribution for the momentary state of the buffer. Here, we shall focus on a distribution that converges to a time-stationary distribution.

We now derive a set of cumulative probability distributions for the buffer content X_t, for each possible condition of the vector of sources. Specifically, these are differential equations that govern the rates the probabilities of source states and content drift within the buffer, and do not change with respect to time. The method follows procedures of AMS [1982] and EM [1992] closely. A time-stationary probability distribution gives the cumulative distribution of buffer content X_t for each source-vector condition Σ_t, and does not change with t. For the case $c < 1$, we use the notation

$$\pi^{(\ell)}_{\sigma_{<1>t},\sigma_{<2>t}}(x) = \Pr\left(\Sigma_t = \left(\sigma_{<1>t}, \sigma_{<2>t}\right), X_t \leq x\right)$$

to denote the cumulative probability that at any moment t in time,

- The source associated with the high bid is in state $\sigma_{<1>t}$ (ON or OFF);
- The source associated with the low bid is in state $\sigma_{<2>t}$ (ON or OFF);
- The buffer content is at most x;
- The buffer is in a condition of congestion ℓ, where

$$\ell = \begin{cases} 0, \text{ if } X_t < B_1 \\ 1, \text{ if } X_t \geq B_1 \end{cases}$$

  and

- The probabilities are stationary with respect to time; i.e.

$$\frac{\partial}{\partial t}\Pr\left(\Sigma_t = \left(\sigma_{<1>t}, \sigma_{<2>t}\right), X_t \leq x\right) = 0.$$

In the case $1 < c < 2$, the buffer has no threshold level B_1, so we use the notation

$$\pi_{\sigma_{<1>t},\sigma_{<2>t}}(x)$$

to denote the same probability, but without reference to the condition of congestion ℓ.

For example, in the case that $c < 1$, the notation $\pi^{(0)}_{0,1}(2)$ is the probability that the high bidder's source is OFF, the low bidder's source is ON, and the buffer content is less than or equal to 2 units of information (ui), where $B_1 > 2$ ui. Similarly, for the case that $1 < c < 2$, $\pi_{0,1}(2)$ is the probability that the high bidder's source is OFF, the low bidder's source is ON, and the buffer content is less than or equal to 2 ui, where $B > 2$ ui.

The buffer behavior is fully explained by cumulative probability distributions of buffer content X_t and source states Σ_t. To understand this, we note the *drift*, or rate of change of quantity X_t of content in the buffer, due to the constant drain rate out of the buffer, inflows of information from the two sources, and the diminishing effects of inflows due to Rules 1 and 2.

Because, as in EM [1992], the probability distributions that we are considering do not change with time, the time variable becomes superfluous. Hence, we drop the time variable from our cumulative probability distributions.

PROPOSITION 3.3. For the case $c < 1$, when the buffer is not congested (i.e. $X_t \leq B_1$), the steady-state probabilities for the buffer are defined by the system of differential equations

$$\begin{bmatrix} \frac{d}{dx}\pi_{0,0}^{(0)}(x) \\ \frac{d}{dx}\pi_{1,0}^{(0)}(x) \\ \frac{d}{dx}\pi_{0,1}^{(0)}(x) \\ \frac{d}{dx}\pi_{1,1}^{(0)}(x) \end{bmatrix} = \begin{bmatrix} \frac{2\lambda}{c} & -\frac{1}{c} & -\frac{1}{c} & \\ \frac{\lambda}{1-c} & -\frac{1+\lambda}{1-c} & & \frac{1}{1-c} \\ \frac{\lambda}{1-c} & & -\frac{1+\lambda}{1-c} & \frac{1}{1-c} \\ & \frac{\lambda}{2-c} & \frac{\lambda}{2-c} & -\frac{2}{2-c} \end{bmatrix} \begin{bmatrix} \pi_{0,0}^{(0)}(x) \\ \pi_{1,0}^{(0)}(x) \\ \pi_{0,1}^{(0)}(x) \\ \pi_{1,1}^{(0)}(x) \end{bmatrix}. \tag{3.1a}$$

When $c < 1$ and the buffer is congested (i.e. $X_t > B_1$), the steady-state probabilities for the buffer are defined by the system of differential equations

$$\begin{bmatrix} \frac{d}{dx}\pi_{0,0}^{(1)}(x) \\ \frac{d}{dx}\pi_{1,0}^{(1)}(x) \\ \frac{d}{dx}\pi_{0,1}^{(1)}(x) \\ \frac{d}{dx}\pi_{1,1}^{(1)}(x) \end{bmatrix} = \begin{bmatrix} \frac{2\lambda}{c} & -\frac{1}{c} & -\frac{1}{c} & \\ -\frac{\lambda}{c} & \frac{1+\lambda}{c} & & -\frac{1}{c} \\ \frac{\lambda}{1-c} & & -\frac{1+\lambda}{1-c} & \frac{1}{1-c} \\ & \frac{\lambda}{1-c} & \frac{\lambda}{1-c} & -\frac{2}{1-c} \end{bmatrix} \begin{bmatrix} \pi_{0,0}^{(1)}(x) \\ \pi_{1,0}^{(1)}(x) \\ \pi_{0,1}^{(1)}(x) \\ \pi_{1,1}^{(1)}(x) \end{bmatrix}. \tag{3.1b}$$

For the case $1 < c < 2$, the steady-state probabilities for the buffer are defined by the system of differential equations

$$\begin{bmatrix} \frac{d}{dx}\pi_{0,0}(x) \\ \frac{d}{dx}\pi_{1,0}(x) \\ \frac{d}{dx}\pi_{0,1}(x) \\ \frac{d}{dx}\pi_{1,1}(x) \end{bmatrix} = \begin{bmatrix} \frac{2\lambda}{c} & -\frac{1}{c} & -\frac{1}{c} & \\ \frac{\lambda}{1-c} & -\frac{1+\lambda}{1-c} & & \frac{1}{1-c} \\ \frac{\lambda}{1-c} & & -\frac{1+\lambda}{1-c} & \frac{1}{1-c} \\ & \frac{\lambda}{2-c} & \frac{\lambda}{2-c} & -\frac{2}{2-c} \end{bmatrix} \begin{bmatrix} \pi_{0,0}(x) \\ \pi_{1,0}(x) \\ \pi_{0,1}(x) \\ \pi_{1,1}(x) \end{bmatrix}. \tag{3.2}$$

AMS [1982] show the procedure for the solutions to systems (3.1a-b) and (3.2) of eight and four differential equations, provided that we also have eight and four boundary conditions, respectively. Some of these equations are dependent on whether the drain rate c is less or greater than 1.

We note boundary conditions as follows, using the methodology of EM [1992]; i.e., by observing the behavior of drift when the buffer is nearly full ($X_t = 0$), nearly empty ($X_t = B$), or near its threshold capacity ($X_t = B_1$).

In the case that $c < 1$, we note the buffer's drift behavior at its boundaries as follows:

- When both sources are OFF, the buffer content changes at the rate $-c$ ui/ut (where $-c < 0$). Thus, the buffer cannot ever be full for more than an instant when both sources are OFF. This implies that the probability that both sources are OFF and the buffer content is below B is merely the probability that both sources are OFF; i.e.,

$$\begin{aligned} \pi_{0,0}^{(0)}(B) &= \Pr(\Sigma_t = (0,0))\,|_{\frac{\partial \Pr(\cdot)}{\partial t}=0} \\ &= \frac{1}{(1+\lambda)^2}, \end{aligned}$$

where the last equality is due to the independence of the distributions of sources, the memoryless property of the exponential distribution, and Proposition 3.1. Since this is the case regardless of whether or not the buffer is discriminatorily congested, we also have

$$\pi_{0,0}^{(1)}(B) = \frac{1}{(1+\lambda)^2}.$$

- When either source is ON, the the buffer fills at the rate $1-c$ ui/ut (where $0 < 1-c < 1$). When both are ON, the buffer fills at the rate $2-c$ ui/ut (where $1 < 2-c < 2$). Thus, the buffer cannot ever be empty for more than an instant whenever one or more sources are ON. We have

$$\begin{aligned} \pi_{0,1}^{(0)}(0) &= 0 \\ \pi_{1,0}^{(0)}(0) &= 0 \\ \pi_{1,1}^{(0)}(0) &= 0 \end{aligned}$$

  again due to independence, memorylessness, and Proposition 3.1.

- Priority Rule 3.1 has no effect as long as source $<2>$ is OFF. Whenever both sources are OFF or only source $<2>$ is ON, the rate of change of the drift is constant as the level of content passes through the threshold level B_1. Thus

$$\begin{aligned} \pi_{0,0}^{(0)}(B_1) &= \pi_{0,0}^{(1)}(B_1) \\ \pi_{1,0}^{(0)}(B_1) &= \pi_{1,0}^{(1)}(B_1). \end{aligned}$$

The other atoms of probability for $c < 1$ are due to the buffer being empty or full, as noted. In the case that $1 < c < 2$, we include four boundary conditions that are derived from the fact that the buffer is always draining (and thus is necessarily less than full) when fewer than two sources are ON and always filling (and thus is necessarily not empty) when exactly two sources are ON. In addition, we include a fifth boundary condition, which is a degenerate form of a fifth differential equation

$$0\frac{\partial}{\partial x}\pi_{1,1}(x) = \lambda\left(\pi_{0,1}(x) + \pi_{1,0}(x)\right) - 2\pi_{1,1}(x) \tag{3.3}$$

at the single value $x = B$. This equation arises from the fact that whenever both sources are ON and the buffer is full (i.e. $x = B$), it will remain full as long as both sources remain ON.

We list all of the boundary conditions in the following proposition:

PROPOSITION 3.4. For $c < 1$, a set of boundary conditions for the differential equations is

$$\begin{aligned}
\pi_{0,0}^{(0)}(B) &= \frac{1}{(1+\lambda)^2} \\
\pi_{0,1}^{(0)}(B) &= \frac{\lambda}{(1+\lambda)^2} \\
\pi_{0,1}^{(0)}(0) &= 0 \\
\pi_{1,0}^{(0)}(0) &= 0 \\
\pi_{1,1}^{(0)}(0) &= 0 \\
\pi_{0,0}^{(0)}(B_1) &= \pi_{0,0}^{(1)}(B_1) \\
\pi_{1,0}^{(0)}(B_1) &= \pi_{1,0}^{(1)}(B_1) \\
\pi_{1,1}^{(0)}(B_1) &= \pi_{1,1}^{(1)}(B_1)
\end{aligned}$$

For $1 < c < 2$, a set of four boundary conditions is

$$\begin{aligned}
\pi_{0,0}(B) &= \frac{1}{(1+\lambda)^2} \\
\pi_{0,1}(B) &= \frac{\lambda}{(1+\lambda)^2} \\
\pi_{1,0}(B) &= \frac{\lambda}{(1+\lambda)^2} \\
\pi_{1,1}(0) &= 0.
\end{aligned}$$

Also for $1 < c < 2$, we add the single atom of probability

$$\pi_{1,1}(B) = \frac{\lambda^2}{(1+\lambda)^2},$$

which follows from (3.3).

Based upon the system of differential equations and set of boundary conditions, we are able find the steady-state cumulative distributions in closed form as functions of the buffer capacity B, the threshold B_1, average ON and OFF periods 1 and $\frac{1}{\lambda}$, respectively, and the drain rate c. The procedure for deriving these distributions is explained in detail in AMS [1982].[1] As in EM [1992], we shall represent them using

[1] The solution to a matrix system of differential equations

$$\frac{d}{dx}\pi^{(\ell)}(x) = M\pi^{(\ell)}(x)$$

is

$$\sum_{\sigma_{<1>}=0}^{1} \sum_{\sigma_{<2>}=0}^{1} a^{(\ell)}_{\sigma_{<1>},\sigma_{<2>}} \phi^{(\ell)}_{\sigma_{<1>},\sigma_{<2>}} \exp\left(z_{\sigma_{<1>},\sigma_{<2>}} x\right)$$

where $a^{(\ell)}_{\sigma_{<1>},\sigma_{<2>}}$ and $\phi^{(\ell)}_{\sigma_{<1>},\sigma_{<2>}}$ each are respectively one of the four eigenvalues and eigenvectors for the matrix equation

$$z^{(\ell)}_{\sigma_{<1>},\sigma_{<2>}} \phi^{(\ell)}_{\sigma_{<1>},\sigma_{<2>}} = M\phi^{(\ell)}_{\sigma_{<1>},\sigma_{<2>}}$$

the π-notation we introduced above. We collapse the cumulative distributions into four expressions (for each case) by considering the condition of congestion ℓ in the following corollary:

COROLLARY 3.4.1. For $c < 1$, the cumulative probability distributions $\pi_{\sigma_{<1>},\sigma_{<2>}}(x)$ are

$$\begin{aligned}
\pi_{0,0}(x) &= \pi_{0,0}^{(0)}(x)\,,\ 0 \le x \le B \\
\pi_{0,1}(x) &= \begin{cases} \pi_{0,1}^{(0)}(x)\,,\ 0 \le x < B_1 \\ \pi_{0,1}^{(1)}(x)\,,\ B_1 < x \le B \end{cases} \\
\pi_{1,0}(x) &= \begin{cases} \pi_{1,0}^{(1)}(x)\,,\ 0 \le x < B \\ \frac{\lambda}{(1+\lambda)^2}\,,\ x = B \end{cases} \\
\pi_{1,1}(x) &= \begin{cases} \pi_{1,1}^{(0)}(x)\,,\ 0 \le x \le B_1 \\ \pi_{1,1}^{(1)}(x)\,,\ B_1 < x < B \\ \frac{\lambda^2}{(1+\lambda)^2}\,,\ x = B \end{cases}
\end{aligned}$$

For $1 < c < 2$, the cumulative probability distributions $\pi_{\sigma_{<1>},\sigma_{<2>}}(x)$ are

$$\begin{aligned}
\pi_{0,0}(x)\,,\ 0 &\le x \le B \\
\pi_{0,1}(x)\,,\ 0 &\le x \le B \\
\pi_{1,0}(x)\,,\ 0 &\le x \le B \\
\pi_{1,1}(x) &= \begin{cases} \pi_{1,1}(x)\,,\ 0 \le x < B \\ \frac{\lambda^2}{(1+\lambda)^2}\,,\ x = B \end{cases}
\end{aligned}$$

corresponding to the state of buffer congestion $\ell \in \{0, 1\}$ and source-state vector $(\sigma_{<1>}, \sigma_{<2>})$.

As in EM [1992], we note that in certain circumstances the cumulative probability $\pi^{(\ell)}_{\sigma_{<1>},\sigma_{<2>}}(x)$ has "jumps," or atoms of probability. As the buffer content increases or decreases through a particular quantity, it momentarily pauses and remains at that quantity for a positive amount of time, until there is a change in source-states. For the case $c < 1$, atoms exist at $\pi_{0,0}(0)$, $\pi_{0,1}(B_1)$, $\pi_{1,0}(B)$, and $\pi_{1,1}(B)$. For the case $1 < c < 2$, atoms exist at $\pi_{0,0}(0)$, $\pi_{0,1}(0)$, $\pi_{1,0}(0)$, and $\pi_{1,1}(B)$. Most of these are trivial and occur when the buffer drains (fills) until it is empty (full), at which point it remains that way until a sufficient number of sources turn ON (OFF). However, this issue merits further discussion for $\pi_{0,1}(B_1)$ in the case $c < 1$.

To determine the cumulative distribution of buffer content in source-state vector $(0,1)$, in which only source $< 2 >$ sends information to the buffer, we note the following. If $X_t < B_1$, the discriminatory rules do not apply, so the buffer content drifts at the rate $1 - c$ ui/uc, which is a net fill. However, if $X_t > B_1$, Rule 3.1 causes the buffer to refuse information from source $< 2 >$. Hence, the buffer content in this region drifts at the rate $-c$ ui/uc, which is a net drain. As in EM [1992], when the buffer content increases (decreases) toward B_1, it must remain stationary upon reaching B_1, since it will begin to decrease (increase) when it becomes infinitesimally larger (smaller) than B_1. This *confluence of drifts* persists until the state of sources Σ_t changes.

3.2.4 Throughput Rates

To impute economic value to a series of bursty data, some of which may be lost, we need to know the *throughput* for each source.

We formally define the **throughput rate** $T^{<i>}$ for the source associated with the ith highest bid as the rate at which data from source $< i >$ is dispatched by the router successfully.

The throughput rate necessarily excludes information that is lost due to buffer overflow. We assume all information that is dispatched successfully provides equal value to the owner of the source from which it came, and information that is lost provides no value. Thus, we can determine the throughput rate for a source by adding disjoint cumulative probabilities for periods during which the buffer dispatches information from that source, weighted by the rates at which the buffer dispatches information from the source. Similarly, we determine the loss rates by adding disjoint weighted cumulative probabilities for periods during which the buffer is full.

PROPOSITION 3.5. For the case $c < 1$, the throughput rates for sources $< 1 >$ and $< 2 >$, respectively associated with high and low bids, are

$$\begin{aligned} T^{<1>} &= \pi_{1,0}^{(1)}(B) + \pi_{1,1}^{(1)}(B) + c\left[\frac{\lambda}{(1+\lambda)} - \pi_{1,0}^{(1)}(B) - \pi_{1,1}^{(1)}(B)\right] \\ T^{<2>} &= \pi_{1,0}^{(0)}(B_1) + \pi_{1,1}^{(0)}(B_1) + c\left[\pi_{1,0}^{(1)}(B_1) - \pi_{1,0}^{(0)}(B_1)\right]. \end{aligned}$$

The quantities of information lost per unit time from sources $< 1 >$ and $< 2 >$ are

$$\begin{aligned} L^{<1>} &= [1-c]\left[\frac{\lambda}{(1+\lambda)} - \pi_{1,0}^{(1)}(B) - \pi_{1,1}^{(1)}(B)\right] \\ L^{<2>} &= \frac{\lambda}{(1+\lambda)} - \pi_{1,0}^{(1)}(B_1) - \pi_{1,1}^{(1)}(B_1) + [1-c]\left[\pi_{1,0}^{(1)}(B_1) - \pi_{1,0}^{(0)}(B_1)\right] \end{aligned}$$

For the case $1 < c < 2$, the throughput rates are

$$\begin{aligned} T^{<1>} &= \frac{\lambda}{(1+\lambda)} \\ T^{<2>} &= \frac{\lambda}{(1+\lambda)^2} + \pi_{1,1}(B) + (c-1)\left[\frac{\lambda^2}{(1+\lambda)^2} - \pi_{1,1}(B)\right] \end{aligned}$$

and the loss rates are

$$\begin{aligned} L^{<1>} &= 0 \\ L^{<2>} &= [2-c]\left[\frac{\lambda^2}{(1+\lambda)^2} - \pi_{<1,1>}(B)\right]. \end{aligned}$$

Note that for both cases, the sums $T^{<i>} + L^{<i>}$, for $i \in \{1,2\}$, always total $\frac{\lambda}{(1+\lambda)}$, which is the probability that a source is ON, as noted in Proposition 3.1. In other words, any unit of information sent by a source is either routed or lost.

These are complete closed-form expressions for throughput rates as functions of the drain rate c, the source parameter λ, the buffer capacity B, and, in the case $c < 1$, the threshold B_1.

In either case, the resource to be auctioned is quantified explicitly as the difference between the throughput rates for the high and low bidders; i.e.

$$T^A = T^{<1>} - T^{<2>}.$$

3.3 Auction

In order to allocate the throughput T^A by means of a second-price auction, we make some assumptions in regard to the way users value the resource.

First, we assume that users have constant valuations of throughput. That is, a user values units of information per unit time that are successfully dispatched by the router, without regard to their time of arrival to the buffer, time of dispatch, or delay due to buffering. Information that is lost as a result of overflow is assumed to have zero value to the user.

Second, we assume these valuations are private (i.e. a user's valuation is known only to that user) independent random variables.

These assumptions imply that users' valuations each reduce to a single parameter that can be multiplied by the auctioned throughput T^A.

We employ the **second-price** auction of Vickrey [1961]. In this auction, both players simultaneously submit sealed bids as amounts of money they will pay per unit of information per unit time. The highest bidder wins the auction and pays the second-highest bid per unit of information per unit time, sent by her source during periods of congestion.

The high bidder pays the money to a *Recipient*, the only specification for which is that neither player has an interest in it. That is, the change in the high bidder's welfare arising from payment to the recipient should be a decrease exactly equal to the payment.

This allows us to itemize all of the elements necessary for the formal definition of a Bayesian game:

- Two PLAYERS, owners of sources, denoted Player 1 and Player 2.

- ACTIONS, or decisons, to be selected by each player, as bids b_1 and b_2 for players 1 and 2, respectively.

- INFORMATION that each player knows about himself or herself, in the form of a valuation v_i that player i has for each unit of the auctioned resource, and some notion or distribution of her rival's valuations v_{-i} of the resource.[2] We assume

[2]Because the value T^A represents a rate of throughput, or units of information dispatched per unit time, and we assume that the winner values each unit of throughput equally, we conclude that the total value of the resource is the average money value v_i per unit of information per unit time, multiplied by the rate T^A of units of information per unit time. Similarly, the total cost of the resource is $b_{<2>}T^A$, where $b_{<2>}$ is the low bid.

that v_i and v_{-i} are i.i.d. random variables from some continuous distribution over the positive real line.

- PAYOFFS for each of the players, defined by the function

$$u_i(b_i, b_{-i}) = \begin{cases} (v_i - b_{-i})\, T^{<A>}, \text{ if } b_i > b_{-i}; \\ 0, \text{ if } b_i < b_{-i}, \end{cases}$$

  where b_{-i} denotes the bid of player i's rival.[3] (We assume $b_i = b_{-i}$ with negligible probability.) If player i's bid b_i exceeds her rival's bid b_{-i} ($i, -i \in \{1, 2\}$), she wins the auctioned resource but pays her rival's bid, so her payoff is $(v_i - b_{-i})$ per unit of throughput, and her rival's payoff is zero. If player i's bid is less than that of her rival, her payoff is zero and her rival's payoff is $(v_{-i} - b_i)$ per unit of throughput.

If bid b_i is higher than bid b_{-i}, source i becomes ordered as the high-bid source $< 1 >$, and source $-i$ becomes ordered as the low bid source $< 2 >$. We assume $b_i = b_{-i}$ with zero probability.

We define a player's **strategy** as a bidding rule that maps her valuation v_i into a bid b_i.

We often write a bidding strategy as a function of a player's own valuation. Some examples of strategies include:

[3] In addition to the marginal benefit awarded only to the winner of the auction, *each* player i receives her private benefit $v_i T^{<2>}$ from the information that can be routed whenever the buffer is not congested. We could have used an equivalent auction that includes this additional payoff. In this case, the winning bidder i would gain her full payoff $v_i T^{<1>} - b_{-i} T^{<A>}$, which includes her private benefit of both auctioned and non-auctioned buffer resources, and the losing bidder $-i$ would gain payoff $v_{-i} T^{<2>}$, which is his private benefit of non-auctioned resources.

- Bid\ding half of one's own valuation; i.e. $b_i(v_i) = \frac{v_i}{2}$.
- Bidding a number, regardless of one's own valuation; i.e. $b_i(v_i) = 2$.

In particular, a player's strategy *cannot* be a function of her rival's private information.

The BEST RESPONSE for player i to her rival's strategy $b_{-i}(v_i)$ is her strategy $b_i^*(v_i)$ that provides her with the greatest payoff $u_i(b_i^*(v_i), b_{-i}(v_{-i}))$; that is,

$$u_i(b_i^*(v_i), b_{-i}(v_{-i})) \geq u_i\left(b_i'(v_i), b_{-i}(v_{-i})\right)$$

for all $b_i'(v_i) \neq b_i^*(v_i)$.

We define a WEAKLY DOMINANT STRATEGY as a bid strategy that results in payoffs to a player that in all cases are at least equal to payoffs resulting from any other bid, and in at least one case is strictly greater than a payoff resulting from any other bid.

The popularity of the Vickrey auction arises from its property of *revelation*. That is, bidding one's true valuation is a weakly dominant strategy for each player.

PROPOSITION 3.6. (Vickrey [1961]) The strategy for player i of bidding her own value is a weakly dominant strategy, i.e., for any bid $b_i \neq v_i$ we have

$$u_i(v_i, b_{-i}) \geq u_i(b_i, b_{-i})$$

for all of her rival's potential bids b_{-i}, and

$$u_i(v_i, b_{-i}) > u_i(b_i, b_{-i})$$

for at least one of her rival's bids b_{-i}.

Because each player has a weakly dominant strategy to bid her true valuation, including if her rival also bids his true valuation, we call the combination of both players bidding their true valuations an **equilibrium in weak dominant strategies.** That is, in this game as defined, neither player has any incentive, strategic or otherwise, to deviate from her strategy of bidding her valuation, so we assume that neither player will change such a strategy once she has selected it.

3.4 Welfare Analysis

Proposition 3.6 has important implications for economic welfare. Because no user can do better than to reveal her valuation truthfully, we can reasonably expect that all users will truthfully reveal their valuations for the resource. This would mean that the high-valuing user will win the auction and be awarded use of the resource, which is economically efficient. In other words, a player is individually incentivized to act in a way that is efficient for the entire group.

We summarize the economic benefit to each party with the following corollary.

COROLLARY 3.6.1. The equilibrium arising from both players bidding their true valuations results in the total economic benefit, inclusive of both auctioned and non-auctioned throughput, of

$$v_{<1>}T^{<2>} + (v_{<1>} - v_{<2>})\,T^{<A>},\ \text{for the high bidder}$$

$$v_{<2>}T^{<2>},\ \text{for the low bidder}$$

$$v_{<2>}T^{<A>},\ \text{for the recipient.}$$

The total surplus is

$$v_{<1>}\left(T^{<2>} + T^{<A>}\right) + v_{<2>}T^{<2>}.$$

An alternative scheme assumes that the buffer discriminates by priority, but priorities are determined by some methodology other than the real-time auction. For example, priority among employees in a business may be determined by seniority within the firm. We use the term "Inherited priority" to refer to the general situation that priorities are determined by some means other than the real-time second price auction. In either the auction or the inherited-priority case, the buffer recognizes information from one of the two sources as high-priority, and provides it with the additional throughput $T^{<A>}$ above the default level $T^{<2>}$, to which the low-priority user's source is entitled.

Let us compare the economic value of the system when priorities are assigned by this alternative. We would like to quantify the improvement in economic efficiency that results by changing from an institution of inherited priority to the use of the auction to allocate bandwidth when the buffer becomes congested. The following proposition shows that the auction allocation mechanism is equal or superior to an inherited-priority scheme, when the metric is *total* economic welfare.

PROPOSITION 3.7. The auction allocation mechanism produces total economic welfare that equals or exceeds the welfare that results from an inherited priority scheme.

We remark that the auction mechanism (excluding the possibility of non-auction bargaining transactions) is *not* necessarily Pareto-superior to a inherited-priority regime. In particular, a user that has inherited the property right to high-priority treatment can only become worse off if she is suddenly compelled to pay for that right.

Rawls [1971] uses an oft-cited definition of economic justice, in which an institution is presumed to be just if a participant would choose it before knowing her own type or preferences. His thought experiment of an "original position behind a veil of ignorance" is a natural way to determine whether a user would voluntarily choose an auction over an allocation of property rights by some non-economic means.

Suppose a user *a priori* knows neither her nor her rival's valuation, and is also ignorant of the priority class to which she would be entitled if the scarce resource were to be allocated by inherited priority. *Ceteris paribus* and without this information, which mechanism would the user prefer? One could think of a situation in which a user is choosing among wireless calling plans, in which one of the options uses the auction allocation mechanism described above during periods of congestion.

The answer to this question depends on the distributions of the valuations and the corresponding probability of entitlement to high-priority status. Consider the stylized example that each user's valuation is drawn independently from a uniform distribution over the interval $[0, 1]$. For this example, let us assume that priority classes are determined *independently* of valuation: User 1 is assigned high-priority status (and, equivalently, User 2 is assigned low-priority status) with probability φ. User 1 (User 2) is assigned low-priority (high-priority) status with the converse probability $1 - \varphi$. Before learning the results of these random draws, the user weighs the expected benefits of the auction and the entitled priority standing. Figure 3.1 depicts a decision tree for this scenario.

Let us first consider User 1's expected benefit when the resource is allocated by auction and she assumes that both she and her rival will use the Nash-equilibrium strategy of bidding one's value. In this case, her priority status is irrelevant; her benefit is determined by her valuation and the outcome of the auction. Her expected benefit of participating in the auction is her incremental benefit from winning the auction, less her payment, weighted by the probability that she has the high valuation.

Alternatively, the user may choose to have the resource allocated by inherited priority. Since she *a priori* does not know whether the buffer will treat the information her source sends as low- or high-priority, her decision depends on expected values. Her expected benefit of selecting inherited priority is equal to her incremental benefit of inheriting the high priority class, weighted by the probability of that inheritance.

Proposition 3.8 states the expected benefits of the buffer with allocation by auction and inherited priority, respectively. The decision rule is summarized in the corollary that follows.

PROPOSITION 3.8. Before learning her valuation, a user's expected benefit of allocation by auction is

$$\frac{1}{2}\left(T^{<2>} + \frac{1}{3}T^{<A>}\right).$$

Before learning either her valuation or her priority class, her expected benefit of allocation by inherited priority is

$$\frac{1}{2}\left(T^{<2>} + \varphi T^{<A>}\right).$$

COROLLARY 3.8.1. Before learning either her valuation or her priority class, a user would prefer allocation by auction provided the probability that she will inherit high-priority status is small. Specifically, the auction's expected benefit exceeds that of inherited priority if and only if

$$\varphi < \frac{1}{3}.$$

In other words, a user that expects to be given high-priority status only occasionally would benefit by choosing an auction-based provider over the alternative of one that allocates scarce resources by another means of priority.

3.5 Discussion

EM [1992] provide a means of allocating resources by priority class, which may be efficient for some uses, but they sidestep the issue of assigning priority class. We have refined their model by providing a means of assigning priority class on a connection-by-connection basis using a second-price auction. Such an auction could be scaled for more than two users, and then practicably implemented for a single-node network by requiring users to bid for priority at the beginning of each connection establishment. Because the proof of Proposition 3.6 does not require users to have any knowledge of the distributions of their rival users' valuations, they need only have some understanding of the parameters of throughput and overflow, which could be simplified into, say, willingness to pay for a specified level of quality per minute of service of telephone use.

For example, a wireless telephone user could could set her telephone to default to a bid of zero for most calls, and then enter a bid of $1.00 per minute for an urgent call during a peak hour. While she will not be permitted to know how much she was billed until the call ends, she knows that she will definitely pay less than $1.00 per minute.

The limit-price nature of this auction is similar to that of an Ebay auction, in which a user enters a maximum willingness to pay for an item, and the system incrementally "Proxy Bids" up to the limit price as others bid on the item.

This methodology may also be applied to auction scarce bandwidth for prices at multiple nodes in a network. A downstream node could transmit its current market price (the highest rejected bid) to an upstream node, which must then add the downstream node to its current rejected price. In fact, a "Smart" router could choose from several alternative downstream paths by selecting the downstream node with the lowest price, and then pass this price upstream to users in real time. This would provide the lowest-cost path in a network and implicitly guarantee a high quality of service for a specified (fluctuating) price.

We noted in the welfare analysis that if overflow is a rare event, and the distribution of valuations is such that their difference is relatively small, the economic benefit of the auction may go entirely to the recipient, and implementation of the auction may indeed result in a loss to one or both of the users. We emphasize that this conclusion is based upon the assumption that *all other economic costs are equal.* In practice, we expect long-run benefits to accrue to users. If the recipient of the auction is a telecommunications provider (rather than some third party), its profits from the use of the auction mechanism will serve as a signal to potential competitors. Since barriers to entry in data communications are much lower than they had been in the past, competitors are likely to enter the market and increase the supply of bandwidth. This will ultimately reduce congestion and lower prices.

There are some potential problems with this method of auctioning bandwidth. First, the auction assumes that a user bids to send data from a single source with no knowledge of other sources' valuations or impact on other users' bids. If a user bids on behalf of multiple sources, she may be able to disrupt the competitive nature of the auction by colluding among bids for sources to lower the competitive price for bandwidth. While this may not have any efficiency implications for our design immediately, it will have an effect on "market prices," which, as noted above, serve as signals to service providers to enter the market.

Second, the auction allocates priority by price, rather than sensitivity to delay. Some digital media, such as email or file downloads, may be sufficiently tolerant to delay not to require priority. However, we can simply address this by programming enduser software to assign a valuation of zero to any such document. A user could then prioritize a truly time-sensitive file, such as an urgent voice conversation, as needed by bidding a price above zero.

Our decision rule for the conditions in which a user would prefer an auction to inherited priority refers to a specific case, in which valuations are independent and identically uniformly distributed; and the probability of inheriting high priority is discrete and independent of valuation. In reality, valuations are almost certainly not uniform. Recall that our concept of valuation represents a *willingness to pay for priority, should there be congestion*, and not the absolute inherent value of throughput to a user. A typical user presumably would value the default level of throughput $T^{<2>}$ such that she normally would be willing to pay zero for the incremental level of throughput, but may find that her valuation increases significantly at certain times. In this case, her willingness to pay is usually zero, but is occasionally positive. Moreover, valuation may be correlated with other demographic variables, such as income.

Furthermore, we have assumed for the sake of tractability that inherited priority is not correlated with valuation. This assumption would not hold in many cases. If a user has an option to purchase the right to the high-priority throughput, her demand for this product may indeed increase with income, as would her real-time demand.

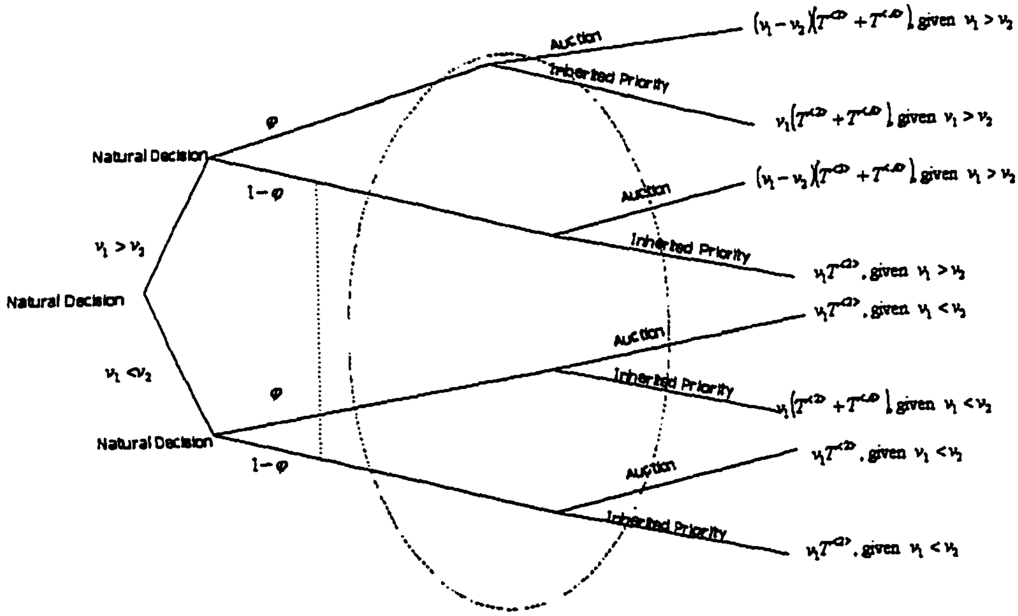

Figure 3.1: Rawlsian Decision Tree

Chapter 4

Proofs of Results

Proof of Proposition 1.1. The PTT searches for a maximum according to the first-order conditions

$$\begin{aligned}\frac{\partial \Pi_i}{\partial p_e} &= A - 2\alpha p_e - 2\beta p_s + \alpha c_e^P = 0 \\ \frac{\partial \Pi_i}{\partial p_s} &= -2\beta p_e + \beta c_e^P + A - 2\alpha p_s = 0.\end{aligned}$$

The prices (p_e^i, p_s^i) are the solution to this system. The quantities (q_e^i, q_s^i) and profit $\Pi_i\left(p_e^i, p_s^i\right)$ follow by substitution.[1]

The second-order condition sufficient for concavity of the profit function (1.2) is that its Hessian matrix is negative definite, which is true provided that

$$4\left(\alpha+\beta\right)\left(\alpha-\beta\right) > 0.$$

In other words, a profit maximum exists provided that each commodity's own price effect exceed its cross-price effect, as in Assumption 1.1. □

[1] The quantities $q_e\left(p_e, p_s\right)$ and $q_s\left(p_e, p_s\right)$ must be nonnegative. We evaluate at $\left(\hat{p}_e, \hat{p}_s\right)$:

$$\begin{aligned}q_e\left(\hat{p}_e, \hat{p}_s\right) &= \frac{1}{2}(A - \alpha c_e^P) \geq 0 \\ q_s\left(p_e, p_s\right) &= \frac{1}{2}\left(A - \beta c_e^P\right) \geq 0.\end{aligned}$$

This necessitates the restriction in Assumption 1.2,

$$c_e^P \leq \frac{A}{\alpha}. \tag{4.1}$$

The restriction on q_s will also be true provided Assumptions 1.1 and 1.2 are satisfied.

Proof of Proposition 1.2. To show that the participation constraint (1.5) is binding in the profit maximation problem, we first show that the unconstrained profit function is not bounded from above. The unconstrained profit function (1.7) has the Hessian matrix of second partial derivatives

$$\begin{bmatrix} -2\alpha & -\beta & -\beta \\ -\beta & 0 & -\alpha \\ -\beta & -\alpha & 0 \end{bmatrix}.$$

The determinants of its submatrices are

$$\begin{aligned} -2\alpha &< 0 \\ \begin{vmatrix} -2\alpha & -\beta \\ -\beta & 0 \end{vmatrix} &= -\beta^2 < 0 \\ \begin{vmatrix} -2\alpha & -\beta & -\beta \\ -\beta & 0 & -\alpha \\ -\beta & -\alpha & 0 \end{vmatrix} &= 2\alpha(\alpha+\beta)(\alpha-\beta) > 0. \end{aligned}$$

(The latter expression is positive in accordance with Assumption 1.) Hence, (4) is an indefinite quadratic and consequently is not bounded from above without an upper bound on τ. This implies that the foreign supplier's participation constraint is binding (5). As long as the foreign supplier provides a positive quantity of equipment, we must have that

$$\bar{\tau} = p_e - c_e^F.$$

□

Proof of Proposition 1.3. The result follows from Proposition 1.2 exactly as in the proof of Proposition 1.1. □

Proof of Corollary 1.3.1. We determine whether or not the tax is positive by substituting p_e^{ii} from expression (1.7) into the expression for $\bar{\tau}$ to get

$$\bar{\tau} = \frac{1}{2}\left(\frac{A}{\alpha+\beta} - c_e^F\right).$$

The expression is positive whenever

$$c_e^F < \frac{A}{\alpha+\beta}.$$

□

Proof of Proposition 1.4. Substituting $p_e = c_e^F$ in (1.8) and differentiating with respect to p_s, we find that (1.10) optimizes (1.8), with the corresponding quantities of service and equipment, and profit to the PTT. □

Proof of Corollary 1.4.1. The PTT will choose to exit the equipment market whenever $\Pi_{iii}\left(p_e^{iii}, p_s^{iii}\right) > \Pi_i\left(p_e^i, p_s^i\right)$, which directly implies (1.12). □

Proof of Corollary 1.4.2. The PTT earns greater in Case (ii) than in Case (iii) provided that $\Pi_{ii}\left(p_e^{ii}, p_s^{ii}\right) > \Pi_{iii}\left(p_e^{iii}, p_s^{iii}\right)$. To establish this, we note that

$$\begin{aligned} \Pi_{ii}\left(p_e^{ii}, p_s^{ii}\right) - \Pi_{iii}\left(p_e^{iii}, p_s^{iii}\right) &= \frac{1}{4\alpha}\left(\left(A - \alpha c_e^F\right)^2 - \left(A - \beta c_e^F\right)^2 - A^2\left(\frac{\alpha-\beta}{\alpha+\beta}\right)\right) \\ &= \frac{(\alpha-\beta)(\alpha+\beta)}{4\alpha}\left(c_e^F - \frac{A}{\alpha+\beta}\right)^2 \end{aligned}$$

which is at least zero for any α and β satisfying Assumption 1.1, and any c_e^F. The expression is exactly zero when $c_e^F = \frac{A}{\alpha+\beta}$. □

Proof of Corollary 1.4.3. The hyperbola that bounds the set of prices satisfying (1.12) is defined by the equation

$$\left(\beta c_e^F + A\right)^2 - \left(\alpha c_e^P + A\right)^2 = A^2\left(\frac{\alpha-\beta}{\alpha+\beta}\right). \tag{4.2}$$

Assuming that c_e^P is a function of c_e^F (which is true under Assumption 1.2), we differentiate (4.2) implicitly to get the equation

$$-2\beta\left(A-\beta c_e^F\right)+2\alpha\left(A-\alpha c_e^P\right)\frac{\partial c_e^P}{\partial c_e^F}=0. \tag{4.3}$$

We solve for $\frac{\partial c_e^P}{\partial c_e^F}$ and equate the identity

$$\frac{\partial c_e^P}{\partial c_e^F}=\frac{\beta\left(A-\beta c_e^F\right)}{\alpha\left(A-\alpha c_e^P\right)}$$

to 1. The solution to this resulting equation gives us the points on the hyperbola that are parallel to the $c_e^P = c_e^F$ line, and simplifies to

$$c_e^P=\frac{\beta^2}{\alpha^2}c_e^F+\frac{\alpha-\beta}{\alpha^2}A. \tag{4.4}$$

Substituting (4.4) for c_e^P in (4.3), we find that the only point on the hyperbola and parallel to the $c_e^P = c_e^F$ line that lies in the feasible region is the point $\left(c_e^F, c_e^P\right) = \left(\frac{A}{\alpha+\beta}, \frac{A}{\alpha+\beta}\right)$, which is exactly on the $c_e^P = c_e^F$ line. Differentiating (4.3) again and rearranging, we find that the second derivative of (4.3) is

$$\frac{\partial^2 c_e^P}{\partial\left(c_e^F\right)^2}=\frac{2\left(\alpha\left(\frac{\partial c_e^P}{\partial c_e^F}\right)-\beta\right)\left(\alpha\left(\frac{\partial c_e^P}{\partial c_e^F}\right)+\beta\right)}{A\left(2\alpha+\frac{1}{\alpha+\beta}\right)}. \tag{4.5}$$

We have already established that $\frac{\partial c_e^P}{\partial c_e^F}$ is equal to 1 at the point $\left(c_e^F, c_e^P\right) = \left(\frac{A}{\alpha+\beta}, \frac{A}{\alpha+\beta}\right)$. Substituting into (4.5), we find that $\frac{\partial^2 c_e^P}{\partial(c_e^F)^2}$ is positive about the point $\left(c_e^F, c_e^P\right) = \left(\frac{A}{\alpha+\beta}, \frac{A}{\alpha+\beta}\right)$. Hence, the hyperbola is concave upward in the neighborhood of this point.

In the interest of rigor, we check for points of inflection by setting $\frac{\partial^2 c_e^P}{\partial(c_e^F)^2}$ equal to zero in (4.5). This implies that $\frac{\partial c_e^P}{\partial c_e^F}$ is equal to either $-\frac{\beta}{\alpha}$ or $\frac{\beta}{\alpha}$. These two possibilities respectively result in the identities

$$\begin{aligned} c_e^P &= -\frac{1}{\alpha}\left(2A - \beta c_e^F\right) \\ c_e^P &= \frac{\beta}{\alpha} c_e^F. \end{aligned}$$

Neither of these identities describes any points on the hyperbola (4.2). Thus, the hyperbola has no points of inflection.

This confirms that the hyperbola is tangent to the $c_e^P = c_e^F$ line at the point $c_e^P = c_e^F = \frac{A}{\alpha+\beta}$, and $c_e^P > c_e^F$ on either side of $c_e^F = \frac{A}{\alpha+\beta}$.We conclude that the set that satisfies Assumptions 1.2 and 1.3 and Inequality (12) is such that $c_e^P \geq c_e^F$. □

Proof of Proposition 1.5. The PTT and equipment monopolist simultaneously choose p_s and p_e to maximize (1.13) and (1.14), respectively. The first-order conditions are respectively

$$\begin{aligned} \frac{\partial \Pi_{iv}}{\partial p_s} &= A - 2\alpha p_s - \beta p_e = 0 \\ \frac{\partial \Pi_F}{\partial p_e} &= A + \alpha c_e^F - 2\alpha p_e - \beta p_e = 0. \end{aligned}$$

The Bertrand-Nash equilibrium solution to the system is the set of prices (p_e^{iv}, p_s^{iv}), and we come to the quantities $q_e\left(p_e^{iv}, p_s^{iv}\right)$ and $q_s\left(p_e^{iv}, p_s^{iv}\right)$ and levels of profit $\Pi_{iv}\left(p_e^{iv}, p_s^{iv}\right)$ and $\Pi_F\left(p_e^{iv}, p_s^{iv}\right)$ by substitution. □

Proof of Corollary 1.5.1. Assumption 1.3 states that

$$c_e^F \leq A\left(\frac{2\alpha - \beta}{2\alpha^2 - \beta^2}\right).$$

Since the expression $\beta\left(\frac{4\alpha^2-\beta^2-2\alpha^2}{2\alpha-\beta}\right)$ is positive, the inequality still holds if we multiply both sides by it to get

$$\beta\left(\frac{4\alpha^2 - \beta^2}{2\alpha - \beta}\right) c_e^F - \beta\frac{2\alpha^2}{2\alpha - \beta} c_e^F < A\beta.$$

The expression in parentheses simplifies to $(2\alpha+\beta)$. Adding the quantity $2A\alpha - \beta(2\alpha+\beta)c_e^F$ to each side gives us

$$2\alpha\left(A - \frac{\alpha\beta}{2\alpha-\beta}c_e^F\right) < (2\alpha+\beta)\left(A - \beta c_e^F\right). \tag{4.6}$$

Meanwhile we can show that

$$\frac{\alpha\beta}{2\alpha-\beta} < \frac{2\alpha^2-\beta^2}{2\alpha-\beta}.$$

Using Assumption 1.3 and transitivity, this implies that

$$\frac{\alpha\beta}{2\alpha-\beta}c_e^F \leq A,$$

which implies that the left-hand side of (4.6) is nonnegative. By squaring and multiplying each side of (4.6) by the positive constant $\frac{1}{4\alpha(2\alpha+\beta)^2}$, we arrive at the result. □

Proof of Proposition 1.6. We show the result by evaluation:

$$\begin{aligned}\Delta CS\left(\left(p_e^{ii},p_s^{ii}\right),\left(p_e^{iii},p_s^{iii}\right)\right) &= \int_{p_e^{iii}}^{p_e^{ii}} q_e\left(p_e,p_s^{ii}\right)dp_e + \int_{p_s^{iii}}^{p_s^{ii}} q_s\left(p_e^{iii},p_s\right)dp_s \\ &= \left(A - \beta p_s^{ii} - \frac{\alpha}{2}\left(p_e^{ii}+p_e^{iii}\right)\right)\left(p_e^{ii}-p_e^{iii}\right) \\ &\quad + \left(A - \beta p_e^{iii} - \frac{\alpha}{2}\left(p_s^{ii}+p_s^{iii}\right)\right)\left(p_s^{ii}-p_s^{iii}\right)\end{aligned}$$

We substitute from (1.7) and (1.10) for p_e^{ii}, p_s^{ii}, p_e^{iii}, and p_s^{iii}. This gives

$$-\frac{\beta}{\alpha}\left(p_e^{ii}-p_e^{iii}\right) = \left(p_s^{ii}-p_s^{iii}\right) = \frac{-\beta}{2\alpha}\left(\frac{A}{\alpha+\beta}-c_e^F\right)$$

and thus from (4.5)

$$
\begin{aligned}
& \Delta CS\left(\left(p_e^{iii}, p_s^{iii}\right), \left(p_e^{ii}, p_s^{ii}\right)\right) \\
= & \frac{1}{2}\left[\frac{A}{\alpha+\beta} - c_e^F\right] \\
& \cdot \left[\begin{array}{c} A\left(1-\frac{\beta}{\alpha}\right) - \beta\left(p_s^{ii} - \frac{\beta}{\alpha} p_e^{iii}\right) \\ -\frac{\alpha}{2}\left(p_e^{ii} + p_e^{iii}\right) + \frac{\beta}{2}\left(p_s^{ii} + p_s^{iii}\right) \end{array}\right] \\
= & \frac{1}{2}\left[\frac{A}{\alpha+\beta} - c_e^F\right] \\
& \cdot \left[\begin{array}{c} A\left(\frac{\alpha-\beta}{\alpha}\right) + \left(\frac{-\beta}{2}\frac{A}{\alpha+\beta} + \frac{\beta^2}{\alpha} c_e^F\right) \\ -\frac{\alpha}{4}\left(\frac{A}{\alpha+\beta} + 3c_e^F\right) + \frac{\beta}{2}\left(\frac{2\alpha+\beta}{2\alpha}\left(\frac{A}{\alpha+\beta}\right) - \frac{\beta}{2\alpha} c_e^F\right) \end{array}\right] \\
= & \frac{1}{2}\left[\frac{A}{\alpha+\beta} - c_e^F\right] \\
& \cdot \left[\begin{array}{c} \frac{A}{\alpha+\beta}\left(\frac{4\alpha^2 - 4\beta^2 - 2\alpha\beta - \alpha^2 + 2\alpha\beta + \beta^2}{4\alpha}\right) \\ + c_e^F\left(\frac{4\beta^2 - 3\alpha^2 - \beta^2}{4\alpha}\right) \end{array}\right] \\
= & \frac{3(\alpha-\beta)(\alpha+\beta)}{8\alpha}\left(\frac{A}{\alpha+\beta} - c_e^F\right)^2 .
\end{aligned}
$$

□

Proof of Proposition 2.1. We evaluate the moments of v_{it}^D from definitions:

$$
\begin{aligned}
E\left(v_{it}^D\right) &= E\left[v_{it} - \frac{1}{T_i}\sum_{s\in\Im_i} v_{is}\right] \\
&= E\left(v_{it}\right) - \frac{1}{T_i}\sum_{s\in\Im_i} E\left(v_{is}\right) \\
&= 0.
\end{aligned}
$$

$$
\begin{aligned}
Cov\left(v_{is}^{D}, v_{jt}^{D} | \text{selection}\right) &= E\left[\left(v_{is}^{D} - E\left(v_{is}^{D}\right)\right)\left(v_{jt}^{D} - E\left(v_{jt}^{D}\right)\right)\right] \\
&= E\left[v_{is}^{D} v_{jt}^{D}\right] \\
&= E\left[\left(v_{is} - \frac{1}{T_i}\sum_{q\in\Im_i} v_{iq}\right)\left(v_{jt} - \frac{1}{T_j}\sum_{r\in\Im_j} v_{jr}\right)\right] \\
&= E\left(v_{is}v_{jt}\right) - \frac{1}{T_i}\sum_{q\in\Im_i} E\left(v_{iq}v_{jt}\right) \qquad (4.7) \\
&\quad -\frac{1}{T_j}\sum_{r\in\Im_j} E\left(v_{is}v_{rt}\right) + \frac{1}{T_iT_j}\sum_{q\in\Im_i}\sum_{r\in\Im_j} E\left(v_{iq}v_{jr}\right)
\end{aligned}
$$

Because of the result in Corollary 2.0.1 that $E\left(v_{is}v_{jt}\right) = 0$ for $i \neq j$ or $s \neq t$, we need only count the covariance terms in expression (4.7) that are of the form v_{it}^2, whose expectation is also shown in Corollary 2.0.1.

CASE 1: ($i = j$ and $s = t$) The first term in (4.7) is equal to $E\left(v_{it}^2\right)$ and the single-summation components each contain a single term of the form $E\left(v_{it}^2\right)$ (or equivalently $E\left(v_{is}^2\right)$). The double-summation component altogether contains T_i terms of values $E\left(v_{i1}^2\right), E\left(v_{i2}^2\right), \ldots, E\left(v_{iT_i}^2\right)$. Hence

$$
\begin{aligned}
& Cov\left(v_{is}^{D}, v_{jt}^{D}\right) \\
&= E\left(v_{it}^2\right) - \frac{2}{T_i}E\left(v_{it}^2\right) + \frac{1}{T_i^2}\sum_{r\in\Im_i} E\left(v_{ir}^2\right) \\
&= \sigma_\varepsilon^2\left[\left(1 - \frac{2}{T_i}\right)\left(1 - \rho^2\delta_{it}\right) + \frac{1}{T_i^2}\sum_{r\in\Im_i}\left(1 - \rho^2\delta_{ir}\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{T_i}\left[\left(T_i - 1\right) - \rho^2\left(\left(T_i - 2\right)\delta_{it} + \bar{\delta}_i\right)\right],
\end{aligned}
$$

with

$$
\bar{\delta}_i = \frac{1}{T_i}\sum_{r\in\Im_i}\delta_{ir}.
$$

CASE 2: ($i = j$ but $s \neq t$) The first term in (4.7) has an expectation of 0. The single-summation components each contain a single term of the form v_{is}^2 and v_{it}^2, respectively. The double-summation component is as in Case 1. Hence

$$\begin{aligned}
& Cov\left(v_{is}^D, v_{jt}^D\right) \\
&= 0 - \frac{1}{T_i}\left[E\left(v_{is}^2\right) + E\left(v_{it}^2\right)\right] + \frac{1}{T_i^2}\sum_{r\in\Im_i} E\left(v_{ir}^2\right) \\
&= \sigma_\varepsilon^2\left[-\frac{1}{T_i}\left(1-\rho^2\left(\delta_{is}+\delta_{it}\right)\right) + \frac{1}{T_i^2}\sum_{r\in\Im_i}\left(1-\rho^2\delta_{ir}\right)\right] \\
&= -\frac{\sigma_\varepsilon^2}{T_i}\left[1-\rho^2\left(\delta_{is}+\delta_{it}-\bar{\delta}_i\right)\right].
\end{aligned}$$

CASE 3: ($i \neq j$) In each component of (4.7), the two errors are independent of each other. Thus the covariance is zero.

The resulting covariances are

$$\begin{aligned}
& Cov\left(v_{is}, v_{jt}\right) \\
&= \begin{cases}
\frac{\sigma_\varepsilon^2}{T_i}\left[\left(T_i-1\right)-\rho^2\left(\left(T_i-2\right)\delta_{it}+\bar{\delta}_i\right)\right], & \text{for } i=j \text{ and } s=t \\
-\frac{\sigma_\varepsilon^2}{T_i}\left[1-\rho^2\left(\delta_{is}+\delta_{it}-\bar{\delta}_i\right)\right], & \text{for } i=j \text{ and } s\neq t \\
0, & \text{for } i\neq j
\end{cases}.
\end{aligned}$$

□

Proof of Proposition 2.2. We evaluate the expectation of the error resulting from the estimation of the IMRs

$$\begin{aligned}
E\left(e_{\lambda it}^D\right) &= E\left(c\left(\hat{\lambda}_{it}-\lambda_{it}\right) - \frac{1}{T_i}\sum_{s\in\Im_i}\left[c\left(\hat{\lambda}_{is}-\lambda_{is}\right)\right]\right) \\
&= c\left[\left(E\left(\hat{\lambda}_{it}\right)-\lambda_{it}\right) - \frac{1}{T_i}\sum_{s\in\Im_i}\left(E\left(\hat{\lambda}_{is}\right)-\lambda_{is}\right)\right] \\
&= 0.
\end{aligned}$$

To determine the covariance, we refer to Heckman [1979], who notes that the variance of the error $e_{\lambda is}e_{\lambda it}$ is

$$\begin{aligned} Cov\left(e_{\lambda is}, e_{\lambda jt}\right) &= E\left[\left(\hat{\lambda}_{is} - \lambda_{is}\right)\left(\hat{\lambda}_{jt} - \lambda_{jt}\right)\right] \\ &= \delta_{is}\delta_{jt}W_{is}\Sigma_{\gamma}W_{jt}'. \end{aligned}$$

Denote $\delta_{is}\delta_{jt}W_{is}\Sigma_{\gamma}W_{jt}'$ by $\omega_{is,jt}$. Then

$$\begin{aligned} & Cov\left(e^{D}_{\lambda is}, e^{D}_{\lambda jt}\right) \\ =\; & E\left[\begin{array}{c} \left(c\left(\hat{\lambda}_{is} - \lambda_{is}\right) - \frac{1}{T_i}\sum_{q\in\Im_i}\left[c\left(\hat{\lambda}_{iq} - \lambda_{iq}\right)\right]\right) \\ \cdot\left(c\left(\hat{\lambda}_{jt} - \lambda_{jt}\right) - \frac{1}{T_j}\sum_{r\in\Im_j}\left[c\left(\hat{\lambda}_{jr} - \lambda_{jr}\right)\right]\right) \end{array}\right] \\ =\; & c^2 E\left[\begin{array}{c} \left(\hat{\lambda}_{is} - \lambda_{is}\right)\left(\hat{\lambda}_{jt} - \lambda_{jt}\right) - \frac{1}{T_i}\sum_{q\in\Im_i}\left(\hat{\lambda}_{iq} - \lambda_{iq}\right)\left(\hat{\lambda}_{jt} - \lambda_{jt}\right) \\ -\frac{1}{T_j}\sum_{r\in\Im_j}\left(\hat{\lambda}_{is} - \lambda_{is}\right)\left(\hat{\lambda}_{jr} - \lambda_{jr}\right) \\ +\frac{1}{T_iT_j}\sum_{q\in\Im_i}\sum_{r\in\Im_j}\left(\hat{\lambda}_{iq} - \lambda_{iq}\right)\left(\hat{\lambda}_{jr} - \lambda_{jr}\right) \end{array}\right] \\ =\; & c^2\left[\omega_{is,jt} - \frac{1}{T_i}\sum_{q\in\Im_i}\omega_{iq,jt} - \frac{1}{T_j}\sum_{r\in\Im_j}\omega_{is,jr} + \frac{1}{T_iT_j}\sum_{q\in\Im_i}\sum_{r\in\Im_j}\omega_{iq,jr}\right] \\ =\; & c^2\left[\omega_{is,jt} - \frac{1}{T_i}\sum_{q\in\Im_i}\omega_{iq,jt} - \frac{1}{T_j}\sum_{r\in\Im_j}\omega_{is,jr} + \bar{\omega}_{ij}\right] \end{aligned}$$

Proof of Corollary 2.2.1

$$\begin{aligned} E\left(e^{D}_{it}|\text{selection}\right) &= E\left(v_{it} + e_{\lambda it}|\text{selection}\right) \\ &= E\left(v_{it}|\text{selection}\right) + E\left(e_{\lambda it}\right) \\ &= 0, \end{aligned}$$

where the second equality is due to the fact that the IMR errors $e_{\lambda it}$ are unconditional, and the third is due to Propositions 2.2 and 2.3.

$$\begin{aligned}
Cov\left(e_{is}^{D}, e_{jt}^{D}|\text{selection}\right) &= E\left[\left(e_{is}^{D} - E\left(e_{is}^{D}|\text{selection}\right)\right)\left(e_{jt}^{D} - E\left(e_{jt}^{D}|\text{selection}\right)\right)|\text{selection}\right] \\
&= E\left[e_{is}^{D} e_{jt}^{D}|\text{selection}\right] \\
&= E\left[\left(v_{is}^{D} + e_{\lambda is}^{D}\right)\left(v_{jt}^{D} + e_{\lambda jt}^{D}\right)|\text{selection}\right] \\
&= E\left[v_{is}^{D} v_{jt}^{D}|\text{selection}\right] + E\left[v_{is}^{D} e_{\lambda jt}^{D}|\text{selection}\right] \\
&\quad + E\left[e_{\lambda is}^{D} v_{jt}^{D}|\text{selection}\right] + E\left[e_{\lambda is}^{D} e_{\lambda jt}^{D}|\text{selection}\right] \\
&= Cov\left(v_{is}^{D}, v_{jt}^{D}|\text{selection}\right) + E\left[v_{is}^{D} e_{\lambda jt}^{D}|\text{selection}\right] \\
&\quad + E\left[e_{\lambda is}^{D} v_{jt}^{D}|\text{selection}\right] + Cov\left(e_{\lambda is}^{D}, e_{\lambda jt}^{D}\right) \\
&= Cov\left(v_{is}^{D}, v_{jt}^{D}|\text{selection}\right) + Cov\left(e_{\lambda is}^{D}, e_{\lambda jt}^{D}\right)
\end{aligned}$$

The fifth equality uses the fact that the errors due to estimation of the IMRs are unconditional. The final equality follows from zero covariance between $e_{\lambda it}$ and v_{it}, as in Greene [1981].

Proof of Corollary 2.2.2 We want the expectation and covariance matrices of $\breve{\beta}^{*}$ given the selected data:

$$\begin{aligned}
E\left(\breve{\beta}^{*}|\text{selection}\right) &= E\left[\left(X^{*\prime}X^{*}\right)^{-1} X^{*\prime}\left(X^{*}\beta^{*} + e^{*}\right)|\text{selection}\right] \\
&\quad \beta^{*} + \left(X^{*\prime}X^{*}\right)^{-1} X^{*\prime} E\left(e^{*}|\text{selection}\right) \\
&= \beta^{*},
\end{aligned}$$

where the final equality is due to the expectation portion of Corollary 2.0.1. Likewise:

$$\begin{aligned}
Cov\left(\breve{\beta}^{*}|\text{selection}\right) &= E\left[\left(\breve{\beta}^{*} - \beta^{*}\right)\left(\breve{\beta}^{*} - \beta^{*}\right)'|\text{selection}\right] \\
&= \left(X^{*\prime}X^{*}\right)^{-1} X^{*\prime} E\left(e^{*}e^{*\prime}|\text{selection}\right) X^{*}\left(X^{*\prime}X^{*}\right)^{-1} \\
&= \left(X^{*\prime}X^{*}\right)^{-1} X^{*\prime}\Omega X^{*}\left(X^{*\prime}X^{*}\right)^{-1},
\end{aligned}$$

where the final equality is due to the covariance portion of Corollary 2.2.1.

Proof of Proposition 2.3 Denoting the vector $\left[\varepsilon_{it}^{D}\right]$ by ε^{D}, we evaluate the expectation of $\tilde{\beta}^{*}$

$$\begin{aligned} E\left(\tilde{\beta}^{*}\right) &= E\left[\left(\hat{X}^{*\prime}X^{*}\right)^{-1}\hat{X}^{*\prime}\left(X^{*}\beta^{*}+\varepsilon^{D}\right)\right] \\ &= \beta^{*}+\left(\hat{X}^{*\prime}X^{*}\right)^{-1}\hat{X}^{*\prime}E\left(\varepsilon^{D}\right) \\ &= \beta^{*} \end{aligned}$$

where the third equality uses the facts that the errors in $\hat{X}^{*\prime}$are orthogonal to those in ε^{D}, and that $E\left(\varepsilon^{D}\right)=0$ from Proposition 2.1.

We evaluate the covariance from the definition

$$\begin{aligned} Cov\left(\tilde{\beta}^{*}\right) &= E\left[\left(\tilde{\beta}^{*}-\beta^{*}\right)\left(\tilde{\beta}^{*}-\beta^{*}\right)'\right] \\ &= \left(\hat{X}^{*\prime}X^{*}\right)^{-1}\hat{X}^{*\prime}E\left(\varepsilon^{D}\left(\varepsilon^{D}\right)'\right)\hat{X}^{*}\left(\hat{X}^{*\prime}X^{*}\right)^{-1} \\ &= \left(\hat{X}^{*\prime}X^{*}\right)^{-1}\hat{X}^{*\prime}\Omega\hat{X}^{*}\left(\hat{X}^{*\prime}X^{*}\right)^{-1}. \end{aligned}$$

□

Proof of Proposition 3.1. The expected length of an OFF-period can be solved using integration by parts:

$$\int_{0}^{\infty}\tau\lambda e^{-\lambda\tau}d\tau=\frac{1}{\lambda}.$$

The expected length of an ON-period is found similarly, using $\lambda=1$. Since the off- and on-states are disjoint and exhaustive possibilities for a source i, the probabilities that it will be on and off are then

$$\begin{aligned} \Pr\left(\sigma_{it}=0\right) &= \frac{E\left(\text{length of off-period}\right)}{E\left(\text{length of off-period}\right)+E\left(\text{length of on-period}\right)} \\ &= \frac{\lambda}{1+\lambda} \end{aligned}$$

and

$$\Pr(\sigma_{it} = 1) = \frac{1}{1+\lambda}$$

by similar argument.

The off-periods and on-periods of both sources are independent, so

$$\begin{aligned} \Pr(\sigma_t = (\delta_{1t}, \delta_{2t})) &= \Pr(\sigma_{1t} = \delta_1, \sigma_{2t} = \delta_2) \\ &= \Pr(\sigma_{1t} = \delta_1) \Pr(\sigma_{2t} = \delta_2). \end{aligned}$$

This leads directly to the results. □

Proof of Lemma 3.1. Let τ be a random variable, exponentially distributed with parameter 1, as in Assumption 3.1. Then the probability that a source will stay OFF past $t + \Delta t$, given that it was OFF at t, is

$$\begin{aligned} \Pr(\tau > t + \Delta t | \tau > t) &= \frac{\Pr(\tau > t + \Delta t, \tau > t)}{\Pr(\tau > t)} \text{ (Bayes' Theorem)} \\ &= \frac{\Pr(\tau > t + \Delta t)}{\Pr(\tau > t)} \text{ (since } \tau > t \text{ whenever } \tau > t + \Delta t) \\ &= \frac{1 - \Pr(\tau < t + \Delta t)}{1 - \Pr(\tau < t)} \\ &= \frac{e^{-\lambda(t+\Delta t)}}{e^{-\lambda t}} \\ &= e^{-\lambda \Delta t} \\ &= \Pr(\tau > \Delta t). \end{aligned}$$

The proof for an ON-period is similar, with $\lambda = 1$. □

Proof of Proposition 3.2. To show that the hazard rate for an OFF-period is proportional to the interval, we first apply Lemma 3.1:

$$\begin{aligned} \Pr(\tau < t + \Delta t | \tau > t) &= 1 - \Pr(\tau > t + \Delta t | \tau > t) \\ &= 1 - e^{-\lambda \Delta t}. \end{aligned} \tag{4.8}$$

Note that the taylor series for $e^{-\lambda\Delta t}$ is

$$\begin{aligned} e^{-\lambda\Delta t} &= 1-\lambda\Delta t+\frac{(\lambda\Delta t)^2}{2!}-\frac{(\lambda\Delta t)^3}{3!}+\ldots \\ &= 1-\lambda\Delta t+o(\Delta t), \end{aligned} \tag{4.9}$$

since all but the first two terms contain factors of Δt of order 2 or higher. By substituting (4.9) into (4.8), we arrive at the result:

$$\begin{aligned} \Pr(\tau<t+\Delta t|\tau>t) &= 1-[1-\lambda\Delta t+o(\Delta t)] \\ &= \lambda\Delta t+o(\Delta t). \end{aligned}$$

The proof for the hazard rate of an ON-period is identical, with the substitution $\lambda=1$. □

Proof of Corollary 3.2.1. By Proposition 3.2,

$$\begin{aligned} \Pr(\tau>t+\Delta t|\tau>t) &= 1-\Pr(\tau>t+\Delta t|\tau>t) \\ &\approx 1-\lambda\Delta t, \end{aligned}$$

where the approximate equality is also due to Proposition 3.2. The proof for the ON-period variable is similar. □

Proof of Corollary 3.2.2. Suppose τ_1 and τ_2 are distributed i.i.d. exponentially with parameters λ_1 and λ_2, respectively, where $\lambda_i\in\{1,\lambda\}$, for $i=1,2$. That is, $\lambda_i=1$ if

τ_i describes an ON-period, and $\lambda_i = 2\lambda$ if τ_i describes an OFF-period. Then,

$$\begin{aligned}
& \Pr(\tau_1 < t+\Delta t, \tau_2 < t+\Delta t | \tau_1 > t, \tau_2 > t) \\
= \; & \Pr(\tau_1 < t+\Delta t | \tau_1 > t) \Pr(\tau_2 < t+\Delta t | \tau_2 > t) \\
= \; & [\lambda_1 \Delta t + o(\Delta t)] [\lambda_2 \Delta t + o(\Delta t)] \qquad (4.10) \\
= \; & \lambda_1 \lambda_2 \Delta t^2 + (\lambda_1 \Delta t) \, o(\Delta t) + (\lambda_2 \Delta t) \, o(\Delta t) + o(\Delta t) \, o(\Delta t) \\
= \; & o(\Delta t),
\end{aligned}$$

where the first equality is due to independence of the random variables, and the second equality is due to Proposition 3.2. Since all of the terms in each set of brackets in expression (4.10) contain a factor of Δt of order 1 or greater, multiplying the two bracketed factors together ensures that every term in the result contains a factor of Δt of order 2 or greater. Thus, every term will be $o(\Delta t)$, and will quickly converge to zero as Δt becomes small. □

Proof of Proposition 3.3. We prove the first equation in system (3.1a); the proofs of the other equations are similar, as explained below.

We first identify upper and lower bounds for the cumulative probability

$$\Pr(\Sigma_{t+\Delta t} = (0,0), X_{t+\Delta t} \leq x).$$

Suppose that at moment t, the content of the buffer is given by $x - \Delta x$, and the buffer's quantity changes by an amount Δx in an interval of time of length Δt. Thus, at moment $t + \Delta t$, the buffer content is equal to x. To identify the stationary distributions, we use the conditional probabilities $\Pr(\Sigma_t = (\sigma_{<1>}, \sigma_{<2>}), X_t \leq x - \Delta x)$ of each aggregate source-state Σ_t at moment t.

If, at time t, both sources are OFF and they both remain OFF through $t + \Delta t$, the change Δx in buffer content during the interval would be $-c\Delta t$, the drain rate $-c$

when both sources are OFF, multiplied by the duration Δt of the interval. In this case, we have the identity

$$x - \Delta x = x + c\Delta t.$$

Hence, we have the equality

$$\Pr\left(\Sigma_t = (0,0), X_t \leq x - \Delta x\right) = \Pr\left(\Sigma_t = (0,0), X_t \leq x + c\Delta t\right).$$

Now consider the cases that exactly one source is ON at moment t but is OFF at moment $t + \Delta t$. These are the cases $\Sigma_t = (1,0)$ and $\Sigma_t = (0,1)$. In either of these cases, the change in buffer content would be at least $-c\Delta t$, and at most $(2-c)\Delta t$.[2] In these cases, we have the inequalities

$$x - (2-c)\Delta t \leq x - \Delta x \leq x + c\Delta t.$$

Since these cumulative probabilities are increasing in x, we have the inequalities

$$\begin{aligned}
\Pr\left(\Sigma_t = (0,1), X_t \leq x - \Delta x\right) &\leq \Pr\left(\Sigma_t = (0,1), X_t \leq x + c\Delta t\right) \\
\Pr\left(\Sigma_t = (0,1), X_t \leq x - \Delta x\right) &\geq \Pr\left(\Sigma_t = (0,1), X_t \leq x - (2-c)\Delta t\right)
\end{aligned}$$

and

$$\begin{aligned}
\Pr\left(\Sigma_t = (1,0), X_t \leq x - \Delta x\right) &\leq \Pr\left(\Sigma_t = (1,0), X_t \leq x + c\Delta t\right) \\
\Pr\left(\Sigma_t = (1,0), X_t \leq x - \Delta x\right) &\geq \Pr\left(\Sigma_t = (1,0), X_t \leq x - (2-c)\Delta t\right).
\end{aligned}$$

[2]The change nears its lower bound when the switch occurs immediately following t, so that both sources are OFF for most of the interval. Conversely, the change nears its upper bound when the source that is OFF turns ON shortly after t, and both sources remain OFF until moments just prior to $t+\Delta t$, at which point both turn OFF, so that two sources are ON for most of the interval.

If we think of the interval $[t, t+\Delta t]$ as a length of time so short that only a single event can occur during the interval, the drift is constrained further, and falls within $[-c\Delta t, (1-c)\Delta t]$.

We can now identify an upper bound for the cumulative probability

$$\Pr\left(\Sigma_{t+\Delta t}=(0,0)\,,X_{t+\Delta t}\leq x\right)$$

that both sources are OFF and the content is at most x at a moment $t+\Delta t$ in time, expressed as the sum of disjoint conditional probabilities, each multiplied by the probability of its respective condition:

$$\begin{aligned}
&\Pr\left(\Sigma_{t+\Delta t}=(0,0)\,,X_{t+\Delta t}\leq x\right)\\
\leq\; &\Pr\left(\Sigma_{t}=(0,0)\,,X_{t}\leq x+c\Delta t\right)\\
&\cdot\Pr\left(\text{both sources stay off between } t \text{ and } t+\Delta t\right)\\
&+\Pr\left(\Sigma_{t}=(1,0)\,,X_{t}\leq x+c\Delta t\right)\\
&\cdot\Pr\left(\text{high bidder's source turns off, low bidder's source stays off}\right)\\
&+\Pr\left(\Sigma_{t}=(0,1)\,,X_{t}\leq x+c\Delta t\right)\\
&\cdot\Pr\left(\text{high bidder's source stays off, low bidder's source turns off }\right)\\
&+\sum_{\sigma_{<1>}=0}^{1}\sum_{\sigma_{<2>}=0}^{1}\left[\begin{array}{c}\Pr\left(\Sigma_{t}=(\sigma_{<1>},\sigma_{<2>})\,,X_{t}\leq x-\Delta x\right)\\ \cdot\Pr\left(\text{Compound events between } t \text{ and } t+\Delta t\right)\end{array}\right].
\end{aligned}$$

The compound events, such as both sources changing states, and/or either source changing states more than once, all have probability $o(\Delta t)$, as shown in Corollary 3.2.2. Substituting results from Corollary 3.2.1, we have

$$\begin{aligned}
&\Pr\left(\Sigma_{t+\Delta t}=(0,0)\,,X_{t+\Delta t}\leq x\right)\\
\leq\; &\Pr\left(\Sigma_{t}=(0,0)\,,X_{t}\leq x-\Delta x\right)(1-\lambda\Delta t)\,(1-\lambda\Delta t)\\
&+\Pr\left(\Sigma_{t}=(1,0)\,,X_{t}\leq x+c\Delta t\right)(\Delta t)\,(1-\lambda\Delta t)\\
&+\Pr\left(\Sigma_{t}=(0,1)\,,X_{t}\leq x+c\Delta t\right)(1-\lambda\Delta t)\,(\Delta t)\\
&+\sum_{\sigma_{<1>}=0}^{1}\sum_{\sigma_{<2>}=0}^{1}\Pr\left(\Sigma_{t}=(\sigma_{<1>},\sigma_{<2>})\,,X_{t}\leq x-\Delta x\right)\cdot o(\Delta t)\,.
\end{aligned}$$

By rearranging and subtracting $\Pr(\Sigma_t = (0,0), X_t \leq x - \Delta x)$, we have

$$\begin{aligned} & \Pr(\Sigma_{t+\Delta t} = (0,0), X_{t+\Delta t} \leq x) - \Pr(\Sigma_t = (0,0), X_t \leq x - \Delta x) \\ \leq\ & -2\lambda\Delta t \Pr(\Sigma_t = (0,0), X_t \leq x + c\Delta t) + \Delta t \Pr(\Sigma_t = (1,0), X_t \leq x + c\Delta t) \\ & +\Delta t \Pr(\Sigma_t = (0,1), X_t \leq x + c\Delta t) + o(\Delta t). \end{aligned}$$

Adding and subtracting $\Pr(\Sigma_t = (0,0), X_t \leq x)$ from the left-hand side and multiplying by $\frac{1}{\Delta t}$ gives

$$\begin{aligned} \Rightarrow\ & \frac{\Pr(\Sigma_{t+\Delta t} = (0,0), X_{t+\Delta t} \leq x) - \Pr(\Sigma_t = (0,0), X_t \leq x)}{\Delta t} \\ & + \frac{\Pr(\Sigma_t = (0,0), X_t \leq x) - \Pr(\Sigma_t = (0,0), X_t \leq x - \Delta x)}{\Delta t} \\ \leq\ & -2\lambda \Pr(\Sigma_t = (0,0), X_t \leq x + c\Delta t) + \Pr(\Sigma_t = (1,0), X_t \leq x + c\Delta t) \\ & + \Pr(\Sigma_t = (0,1), X_t \leq x + c\Delta t) + \frac{o(\Delta t)}{\Delta t}. \end{aligned}$$

Recall that the drift $\Delta x = -c\Delta t$ when $\Sigma_t = (0,0)$. Thus, the inequality becomes

$$\begin{aligned} & \frac{\partial}{\partial t} \Pr(\Sigma_t = (0,0), X_t \leq x) - c\frac{\partial}{\partial x} \Pr(\Sigma_t = (0,0), X_t \leq x) \qquad (4.11) \\ \leq\ & -2\lambda \Pr(\Sigma_t = (0,0), X_t \leq x) + \Pr(\Sigma_t = (1,0), X_t \leq x) \\ & + \Pr(\Sigma_t = (0,1), X_t \leq x). \end{aligned}$$

We now identify the lower bound for the cumulative probability

$$\Pr(\Sigma_{t+\Delta t} = (0,0), X_{t+\Delta t} \leq x).$$

In accordance with the methodology used in identifying the upper bound, we use the bounds for each of the disjoint conditionals:

$$\begin{aligned}
& \Pr\left(\Sigma_{t+\Delta t}=(0,0), X_{t+\Delta t}\leq x\right) \\
\geq\ & \Pr\left(\Sigma_t=(0,0), X_t\leq x-\Delta x\right) \\
& \cdot\Pr\left(\text{both sources stay off between } t \text{ and } t+\Delta t\right) \\
& +\Pr\left(\Sigma_t=(1,0), X_t\leq x-(2-c)\,\Delta t\right) \\
& \cdot\Pr\left(\text{high bidder's source turns off, low bidder's source stays off}\right) \\
& +\Pr\left(\Sigma_t=(0,1), X_t\leq x-(2-c)\,\Delta t\right) \\
& \cdot\Pr\left(\text{high bidder's source stays off, low bidder's source turns off }\right) \\
& +\sum_{\sigma_{<1>}=0}^{1}\sum_{\sigma_{<2>}=0}^{1}\left[\begin{array}{c}\Pr\left(\Sigma_t=(\sigma_{<1>},\sigma_{<2>}), X_t\leq x-\Delta x\right) \\ \cdot\Pr\left(\text{Compound events between } t \text{ and } t+\Delta t\right)\end{array}\right].
\end{aligned}$$

Continuing with the same analysis that we used to show the upper bound, we find that

$$\begin{aligned}
& \frac{\partial}{\partial t}\Pr\left(\Sigma_t=(0,0), X_t\leq x\right)-c\frac{\partial}{\partial x}\Pr\left(\Sigma_t=(0,0), X_t\leq x\right) \\
\geq\ & -2\lambda\Pr\left(\Sigma_t=(0,0), X_t\leq x\right)+\Pr\left(\Sigma_t=(1,0), X_t\leq x\right) \\
& +\Pr\left(\Sigma_t=(0,1), X_t\leq x\right).
\end{aligned}$$

Since the upper and lower bounds of the linear combination of derivatives are equal, we conclude

$$\begin{aligned}
& \frac{\partial}{\partial t}\Pr\left(\Sigma_t=(0,0), X_t\leq x\right)-c\frac{\partial}{\partial x}\Pr\left(\Sigma_t=(0,0), X_t\leq x\right) \\
=\ & -2\Pr\left(\Sigma_t=(0,0), X_t\leq x\right)+\lambda\Pr\left(\Sigma_t=(1,0), X_t\leq x\right) \\
& +\lambda\Pr\left(\Sigma_t=(0,1), X_t\leq x\right).
\end{aligned}$$

Since we are interested in the stationary distribution, we assume the probabilities do not change with respect to time. Thus, setting $\frac{\partial}{\partial t}\Pr\left(\Sigma_t=(0,0), X_t\leq x\right)$ equal to zero and multiplying by $-\frac{1}{c}$, we arrive at the result:

$$\begin{aligned}
& \frac{\partial}{\partial x}\Pr\left(\Sigma_t=(0,0)\,,X_t\le x\right) \\
= \; & \frac{2\lambda}{c}\Pr\left(\Sigma_t=(0,0)\,,X_t\le x\right)-\frac{1}{c}\Pr\left(\Sigma_t=(1,0)\,,X_t\le x\right) \\
& -\frac{1}{c}\Pr\left(\Sigma_t=(0,1)\,,X_t\le x\right).
\end{aligned}$$

The derivations of the other equations in (1) are analogous. The drifts Δx that we use in the substitution in inequality (7) are equal to $(1-c)\,\Delta t$ in the equations for $\pi^{(0)}_{1,0}(x)$ and $\pi^{(0)}_{0,1}(x)$, and $(2-c)\,\Delta t$ in the equation for $\pi^{(0)}_{1,1}(x)$.

In the case that there is congestion, as in (3.2), information from Source < 2 > is not buffered. Therefore, the drifts Δx in the equations for $\pi^{(1)}_{0,1}(x)$ and $\pi^{(1)}_{1,1}(x)$ are equal to $-c\Delta t$ and $(1-c)\,\Delta t$, respectively. The proof for (3.2) is otherwise identical to that of (3.1a-b).

The proof of (3) is identical to that of (1). □

Proof of Proposition 3.4. This proposition is explained in the text. □

Proof of Proposition 3.5. We begin by considering the case that $0<c<1$. In this situation, the high-priority source's information will be buffered whenever it is being sent and the buffer is not full to its capacity B. When the buffer is full, only a portion c of its information can be buffered. Since these are all disjoint possibilities, we add the individual rates weighted by their probabilities to get the total throughput rate:

$$
\begin{aligned}
T^{<1>} &= 1\left[\Pr\left(\Sigma_t=(1,0),X_t<B\right)|_{\frac{\partial}{\partial t}=0}+\Pr\left(\Sigma_t=(1,1),X_t<B\right)|_{\frac{\partial}{\partial t}=0}\right] \\
&\quad +c\left[\Pr\left(\Sigma_t=(1,0),X_t=B\right)|_{\frac{\partial}{\partial t}=0}+\Pr\left(\Sigma_t=(1,1),X_t=B\right)|_{\frac{\partial}{\partial t}=0}\right] \\
&= \pi^{(1)}_{<1,0>}(B)+\pi^{(1)}_{<1,1>}(B) \\
&\quad +c\left[\left(\frac{\lambda}{(1+\lambda)^2}-\pi^{(1)}_{<1,0>}(B)\right)+\left(\frac{\lambda^2}{(1+\lambda)^2}-\pi^{(1)}_{<1,1>}(B)\right)\right] \\
&= \pi^{(1)}_{<1,0>}(B)+\pi^{(1)}_{<1,1>}(B)+c\left[\frac{\lambda}{1+\lambda}-\pi^{(1)}_{<1,0>}(B)-\pi^{(1)}_{<1,1>}(B)\right].
\end{aligned}
$$

Meanwhile, the low-priority source's information will be buffered whenever the source is ON and the buffer's content is below the threshold level B_1. The buffer will also accept a portion c of its information when the content is exactly equal to B_1, the low-priority source is ON, and the high-priority source is OFF:

$$
\begin{aligned}
T^{<2>} &= 1\left[\Pr\left(\Sigma_t=(0,1),X_t<B_1\right)|_{\frac{\partial}{\partial t}=0}+\Pr\left(\Sigma_t=(1,1),X_t<B\right)|_{\frac{\partial}{\partial t}=0}\right] \\
&\quad +c\left[\Pr\left(\Sigma_t=(1,0),X_t=B\right)|_{\frac{\partial}{\partial t}=0}\right] \\
&= \pi^{(0)}_{<0,1>}(B_1)+\pi^{(1)}_{<1,1>}(B_1)+c\left[\pi^{(1)}_{<0,1>}(B_1)-\pi^{(0)}_{<0,1>}(B_1)\right] \\
&= \pi^{(0)}_{<1,0>}(B_1)+\pi^{(1)}_{<1,1>}(B_1)+c\left[\pi^{(1)}_{<1,0>}(B_1)-\pi^{(0)}_{<1,0>}(B_1)\right],
\end{aligned}
$$

where the third equality is due to the symmetric distributions of the two sources.

The high-priority source's information is lost only when it is ON and the buffer is completely full, and then only a portion $1-c$ ui/ut is lost, since the remaining c units will be accepted as the buffer dispatches information:

$$
\begin{aligned}
L^{<1>} &= [1-c]\left[\Pr\left(\Sigma_t=(1,0),X_t=B\right)|_{\frac{\partial}{\partial t}=0}+\Pr\left(\Sigma_t=(1,1),X_t=B\right)|_{\frac{\partial}{\partial t}=0}\right] \\
&= [1-c]\left[\frac{\lambda}{1+\lambda}-\pi^{(1)}_{<1,0>}(B)-\pi^{(1)}_{<1,1>}(B)\right].
\end{aligned}
$$

The low-priority source's information is lost whenever it is ON and the buffer is full beyond the threshold level B_1. In the case that the buffer content is exactly equal to B_1 and only the low-priority source is ON, the portion $1 - c$ ui/ut is lost:

$$\begin{aligned}
L^{<2>} &= 1\left[\Pr\left(\Sigma_t = (0,1), X_t > B_1\right)|_{\frac{\partial}{\partial t}=0} + \Pr\left(\Sigma_t = (1,1), X_t \geq B_1\right)|_{\frac{\partial}{\partial t}=0}\right] \\
&\quad + [1-c]\left[\Pr\left(\Sigma_t = (0,1), X_t = B_1\right)|_{\frac{\partial}{\partial t}=0}\right] \\
&= \frac{\lambda}{(1+\lambda)^2} - \pi^{(1)}_{<0,1>}(B_1) + \frac{\lambda^2}{(1+\lambda)^2} - \pi^{(1)}_{<1,1>}(B_1) \\
&\quad + [1-c]\left[\pi^{(1)}_{<0,1>}(B_1) - \pi^{(0)}_{<0,1>}(B_1)\right] \\
&= \frac{\lambda}{(1+\lambda)} - \pi^{(1)}_{<1,0>}(B_1) - \pi^{(1)}_{<1,1>}(B_1) + [1-c]\left[\pi^{(1)}_{<1,0>}(B_1) - \pi^{(0)}_{<1,0>}(B_1)\right],
\end{aligned}$$

again by symmetry of distributions of sources.

For the case that $1 < c < 2$, bandwidth is sufficient for all of the high-priority source's information to be routed. Thus, the high-priority source's throughput is equal the rate it sends information to the buffer (1 ui/ut) weighted by the probability that it is ON:

$$\begin{aligned}
T^{<1>} &= 1\left[\Pr\left(\Sigma_t = (1,0)\right)|_{\frac{\partial}{\partial t}=0} + \Pr\left(\Sigma_t = (1,1)\right)|_{\frac{\partial}{\partial t}=0}\right] \\
&= \frac{\lambda}{(1+\lambda)^2} + \frac{\lambda^2}{(1+\lambda)^2} \\
&= \frac{\lambda}{1+\lambda}.
\end{aligned}$$

The low-priority source's information can be routed whenever it is ON. However, in the case that the buffer is completely full to its capacity B and both sources are ON, only a portion $c - 1$ ui/ut of the information it passes can be buffered:

$$\begin{aligned}
T^{<2>} &= 1\left[\Pr\left(\Sigma_t = (1,0)\right)|_{\frac{\partial}{\partial t}=0}\right] + c\left[\Pr\left(\Sigma_t = (1,1), X_t = B\right)|_{\frac{\partial}{\partial t}=0}\right] \\
&= \frac{\lambda}{(1+\lambda)^2} + \pi_{<1,1>}(B) + [c-1]\left[\frac{\lambda^2}{(1+\lambda)^2} - \pi_{<1,1>}(B)\right].
\end{aligned}$$

Since the high-priority source's information is always routed, its loss rate is zero. The low-priority source loses information at the rate $2-c$ ui/ut, only when the buffer is full and both sources are ON:

$$L^{<2>} = [2-c]\left[\frac{\lambda^2}{(1+\lambda)^2} - \pi_{<1,1>}(B)\right].$$

$\square$

Proof of Proposition 3.6. To show weak dominance of the strategy $b_i = v_i$, we must show that

$$\begin{aligned} u_i(v_i, b_{-i}) &\geq u_i(b_i, b_{-i}) \text{ for all possible } b_i \\ u_i(v_i, b_{-i}) &> u_i(b_i, b_{-i}) \text{ for some } b_{-i}. \end{aligned}$$

By bidding v_i, player i gains $(v_i - b_{-i})T^A$ per unit time if $v_i > b_{-i}$, or zero if $v_i < b_{-i}$.

Consider an alternative strategy in which player i bids in excess of her valuation. That is, $b_i = v_i + \Delta v_i$, for some positive amount Δv_i.

We will review this possibility for three disjoint but exhaustive cases for the bid of player i's rival b_{-i}.

First, suppose b_{-i} is less than player i's valuation; that is, $b_{-i} < v_i$. In this case, player i's excess bid is as good as a bid of v_i, since

$$v_i + \Delta v_i > b_{-i},$$

so

$$u_i(v_i + \Delta v_i, b_{-i}) = (v_i - b_{-i})T^A = u_i(v_i, b_{-i}).$$

Second, suppose b_{-i} is between player i's valuation and her bid; that is, $v_i < b_{-i} < v_i + \Delta v_i$. In this case, player i does worse than bidding v_i, since

$$v_i + \Delta v_i > b_{-i},$$

so

$$u_i\left(v_i + \Delta v_i, b_{-i}\right) = \left(v_i - b_{-i}\right) T^A < 0 = u\left(v_i, b_{-i}\right).$$

Third, suppose b_{-i} is more than player i's bid. In this case, player i does as well as bidding v_i, since

$$v_i < v_i + \Delta v_i < b_{-i},$$

so

$$u\left(v_i + \Delta v_i, b_{-i}\right) = 0 = u\left(v_i, b_{-i}\right).$$

We now consider an underbidding strategy; that is, player i employs a strategy $b_i = v_i - \Delta v_i$, for some positive amount Δv_i. Our analysis is similar.

Suppose player i's bid is still greater than her rival's bid b_{-i}. In this case, she does as well as bidding v_i, since

$$b_{-i} < v_i - \Delta v_i < v_i,$$

so

$$u_i\left(v_i - \Delta v_i, b_{-i}\right) = \left(v_i - b_{-i}\right) T^A = u\left(v_i, b_{-i}\right).$$

Now suppose b_{-i} is between player i's bid and her valuation; that is, $v_i - \Delta v_i < b_{-i} < v_i$. In this case, she loses the auction when it would have been beneficial to win, since

$$v_i - \Delta v_i < b_{-i} \text{ but } b_{-i} < v_i,$$

so

$$u_i\left(v_i - \Delta v_i, b_{-i}\right) = 0 < \left(v_i - b_{-i}\right) T^A = u\left(v_i, b_{-i}\right).$$

Finally, suppose b_{-i} is greater than player i's valuation. In this case, she does as well as she would by bidding v_i, since

$$v_i - \Delta v_i < v_i < 0,$$

so

$$u_i\left(v_i - \Delta v_i, b_{-i}\right) = 0 = u_i\left(v_i, b_{-i}\right).$$

$\square$

Proof of Proposition 3.7. The total value of the auction is equal to the value of the auction to both players, plus the value to the recipient; i.e.

$$TV_A = v_{<1>}\left(T^{<2>} + T^{<A>}\right) + v_{<2>}T^{<2>}.$$

We calculate the value of allocation by inherited priority assuming that the incremental bandwidth $T^{<A>}$ is automatically awarded to the high-valuing user without charge. We consider the cases that (1) the high-valuing user is assigned high priority and the low-valuing user is assigned low priority; and (2) the high-valuing user is assigned low priority and the high-valuing user is assigned high priority.

In case (1), the incremental bandwidth $T^{<A>}$ is assigned to the high valuing user, since she also enjoys high-priority treatment by the buffer; i.e.

$$\begin{aligned} TV_{IP1} &= v_{<1>}\left(T^{<2>} + T^{<A>}\right) + v_{<2>}T^{<A>} \\ &= TV_{A.} \end{aligned}$$

In case (2), the incremental bandwidth is assigned to the low-valuing user, since she now enjoys high-priority treatment by the buffer; i.e.

$$TV_{IP2} = v_{<2>}\left(T^{<2>} + T^{<A>}\right) + v_{<1>}T^{<A>}.$$

Since $v_{<2>} < v_{<1>}$ by definition, we conclude that $TV_{IP2} < TV_A$. □

Proof of Proposition 3.8. In calculating the expected benefit of the router when allocation is by auction, User 1 considers the probability distributions of her and her rival's valuations. Since she would bid after learning her valuation, she can reasonably expect to bid her value, since it is a best response to any of her rival's actions, as noted in Proposition 3.6. Thus, her expected benefit from the auction is her weighted average benefit in Nash equilibrium, in which the weights are all possible distributions of her and her rivals' valuations:

$$\begin{aligned}
EV_A &= \int_{v_2=0}^{1}\left[\int_{v_1=0}^{v_2} u_1\left(v_1, v_2 | v_1 < v_2\right) dv_1 + \int_{v_1=v_2}^{1} u_1\left(v_1, v_2 | v_1 > v_2\right) dv_1\right] dv_2 \\
&= \int_{v_2=0}^{1}\left[\int_{v_1=0}^{v_2} v_1 T^{<2>} dv_1 + \int_{v_1=v_2}^{1}\left(v_1\left(T^{<2>}+T^{<A>}\right) - v_2 T^{<A>}\right) dv_1\right] dv_2 \\
&= \frac{1}{2}\int_{v_2=0}^{1}\left[T^{<2>} + T^{<A>}\left(1+v_2^2-2v_2\right)\right] dv_2 \\
&= \frac{1}{2}\left(T^{<2>} + \frac{1}{3}T^{<A>}\right).
\end{aligned}$$

To calculate the expected benefit of the router when allocation is by inherited priority, User 1 considers the distribution of her own valuation and the disjoint probabilities that she will inherit rights to the two priority classes. Her rival's valuations are not relevant in this case.

$$\begin{aligned}
EV_{IP} &= \int_{v_1=0}^{1}\left[\begin{array}{c} v_1\left(T^{<2>}+T^{<A>}\right)\Pr(\text{User 1 is high priority}) \\ +v_1 T^{<2>}\Pr(\text{User 1 is low priority}) \end{array}\right] dv_1 \\
&= \int_{v_1=0}^{1}\left[v_1\left(T^{<2>}+T^{<A>}\right)\varphi + v_1 T^{<2>}(1-\varphi)\right] dv_1 \\
&= \frac{1}{2}\left(T^{<2>} + \varphi T^{<A>}\right).
\end{aligned}$$

□

References

[1982] Anick, D., D. Mitra, and M. M. Sondhi, "Stochastic theory of a data-handling system with multiple sources," The Bell System Technical Journal 61 (8), 1871-1894.

[1976] Becker, Gary, Comment on "Toward a more general theory of regulation," Journal of Law and Economics 19: 245-248, 1976

[1980] Chamberlain, G., "Analysis of covariance with qualitative data," Review of Economic Studies 47: 225-238.

[1992] Conner, Kathleen, "Obtaining strategic advantage from being imitated: When can encouraging 'clones' pay?" July 1992 working paper, later published in Operations Research.

[1992] Elwalid, A. I., and D. Mitra, "Fluid models for the analysis and design of statistical multiplexing with loss priorities on multiple classes of bursty traffic," IEEE Infocom '92, 415-425.

[1981] Greene, W.A., "Sample selection bias as a specification error: comment," Econometrica 49: 795-798.

[1995] Greene, W.A., Econometric Analysis, Second Edition, New York: Macmillan.

[1982] Griffin, J.M., "The welfare implications of externalities and price elasticities for telecommunications pricing," Review of Economics & Statistics 64(1): 59-66.

[1994] Gupta, A., D.O. Stahl, and A.B. Whinston, "A stochastic equilibrium model of internet pricing," Journal of Economic Dynamics and Control 21, 697-722.

[1997] Gupta, A., D.O. Stahl, and A.B. Whinston, "An economic approach to networked computing with priority classes," available at http://crec.bus.utexas.edu/alok/nprice/netprice.html, 1997.

[1984] Hannan, T.H., and J.M. McDowell, "The determinants of technology adoption: the case of the banking firm," Rand Journal of Economics 15: 328-335.

[1995] Hassin, R., "Decentralized regulation of a queue," Management Science 41 (1), 163-173.

[1978] Hausman, J., "Specification tests in econometrics," Econometrica 46: 1251-1271.

[1979] Hausman, J.A., and D.A. Wise, "Attrition bias in experimental and panel data: the Gary income maintenance experiment," Econometrica 47: 455-473.

[1945] Hayek, Friedrich A. von, The price system as a mechanism for using knowledge, Homewood, IL: Irwin, 1985 (reprint), pp. 29-40.

[1979] Heckman, J., "Sample selection bias as a specification error," Econometrica 47: 153-161.

[1981] Heckman, J., "The incidental parameters problemand the problem of initial conditions in estimating a discrete time-discrete data stochastic process," Structural Analysis of Discrete Data with Econometric Applications, Manski and McFadden (eds.), Cambridge: MIT Press, pp. 179-185.

[1999] International Telecommunication Union, World Telecommunication Indicators Database.

[1998] International Monetary Fund, World Economic Outlook.

[1974] Johnson, N., and S. Kotz, Distributions in Statistics – Continuous Multivariate Distributions, New York: Wiley.

[1988] Keane, M., R. Moffitt, and D. Runkle, "Real wages over the business cycle: Estimating the impact of heterogeneity with micro data," Journal of Political Economy 96: 1232-1266.

[1997] Kelly, F., "Charging and accounting for bursty connections," from Internet Economics, L.W. McNight and J.P. Bailey (eds.), Cambridge: MIT Press.

[1997] Kubasik, J., "Competition in the Central European telecommunications networks," Global Networking 97, Vol. II, pp. 217-227.

[1987] Levin, S.G., S.L. Levin, and J.B. Meisel, "A dynamic analysis of the adoption of a new technology: The case of optical scanners," Review of Economics and Statistics , pp. 12-17.

[1994] MacKie-Mason, J.K., and H.R. Varian, "Pricing the Internet," working paper.

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

[1969] Naor, P., "The regulation of queue size by levying tolls," Econometrica 37 (1), 15-24.

[1987] Noam, Eli, "The public telecommunications network: a concept in transition," Journal of Communication 37: 30-48, 1987.

[1976] Peltzman, Sam, "Toward a more general theory of regulation," Journal of Law and Economics 19: 190-244, 1976.

[1971] Rawls, J.; A Theory of Justice;Cambridge: Harvard University Press.

[1990] Ridder, G., "Attrition in multi-wave panel data," from Panel Data and Labor Market Studies, Hartog, Ridder, and Theeuwes (eds.), Elsevier, pp. 45-67.

[1978] Rohlfs, J., "Economically efficient Bell System pricing," Unpublished Bell Laboratories Memorandum, attachment H, transmitted by R.L. McGuire of AT&T to Congressman Lionel VanDeerlin, October 31, 1978.

[1981] Spencer, D., and K. Berk, "A limited information specification test," Econometrica 49: 1079-1085.

[1990] Spiller, P. and C. Cardilli. "The Frontier of Telecommunications Deregulation: Small Countries Leading the Pack," Journal of Economic Perspectives 11, 127-138.

[1997] Sunarno, R, "Impact of privatization in telecommunications of Indonesia," Global Networking '97 Joint Conference, Enslow et al, Ed., IOS Press, Amsterdam, 1997, Vol. II, pp. 95-98.

[1994] Taylor, Lester D., Telecommunications demand in theory and practice, Dordrecht: Kluwer, 1993.

[1988] Tirole, Jean, The Theory of Industrial Organization, Cambridge: MIT Press, 1988, p. 392.

[V 1996] Varian, H., "Advanced Microeconomics," 1996.

[1990] Verbeek, M., "On the estimation of a fixed effects model with selectivity bias," Economics Letters 34: 267-270.

[1961] Vickrey, W., "Counterspeculation, auctions, and competitive sealed tenders," Journal of Finance 16 (1), 8-37.

[1997] Wang, Q., J.M. Peha, and M.A. Sirbu, "Optimal pricing for integrated services networks," from Internet Economics, L.W. McNight and J.P. Bailey (eds.), Cambridge: MIT Press.

[1980] White, H., "A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity," Econometrica 48: 817-838.

[1997] "Ruggiero congratulates governments on landmark telecommunications agreement," WTO press release, Feb. 15, 1997, http://www.wto.org/wto/press/press67.htm

[1998] Conversation with Edward E. Zajac, 1998.