6. INDUSTRIAL ORGANIZATION OF TRANSPORTATION PROVIDERS

So far we have discussed desirable investment and pricing policies, but not what institutional structures can best bring them about. The dominant organizational form for providing urban transportation services to individual users, especially in developed nations, is public ownership. This is supplemented by regulation of those firms allowed to operate privately. Observers increasingly question the efficacy of these arrangements.

A fundamental problem with private transportation markets is their tendency toward scale economies. This tendency has long been recognized in intercity transportation industries such as ocean shipping and railroads, whose fixed costs take tangible forms like terminals and rail track. But scale economies also affect industries, such as airlines and trucking, where users place a premium on fast and reliable transfers across various links in a large network, because then scale economies on individual links — which can exist due to efficiencies of operating large vehicles — become important as firms seek to use those links to provide convenient service for many origin-destination pairs.

Both urban roads and urban public transit services operate on networks that collect users with diverse origins and destinations onto high-capacity links in order to take advantage of link-specific scale economies. Furthermore, we saw in Section 3.2.4 that the nature of scheduled services creates a type of scale economy even on feeder links. Therefore, urban transportation faces the same underlying cost condition as inter-city transportation. Scale economies mean that marginal-cost prices do not cover average costs. In order for private firms to operate in such markets, they must either receive subsidies or forego marginal-cost pricing. If we choose to forego marginal-cost pricing, then we must permit some degree of market power, which has its own problems that depend on the nature of markets and the more general state of public control of business practices (Button, 2005).

This chapter, then, reviews issues related to private operators in urban transportation. Section 6.1 discusses profit-maximizing price and capacity choice for private highway operators, and compares these to welfare-maximizing choices. Section 6.2 discusses regulation and franchising of private highways, while section 6.3 does the same for transit services.

Of course, transportation is only one of many industries for which the question of private or public operation has been greatly discussed. Two valuable general assessments are those by Kay and Thompson (1986) and Vickers and Yarrow (1991). One conclusions is that a lot depends on details of the industry. Analyses of transportation industries have been important in coming to these more general conclusions as well as in applying more general results to cases with significant public-policy implications. For these reasons, the economic study of transportation, and of public transit in particular, continues to provide insights of widespread interest.

6.1 Private Highways

Some observers, such as Roth (1996), suggest that many of the recent difficulties with financing and pricing publicly owned highways could be overcome through a return to private ownership, which was common in past eras.[1] Given perfect competition among road suppliers, first-best congestion pricing would be the equilibrium outcome (DeVany and Saving, 1980). Even under the more realistic conditions of monopoly or oligopoly, private ownership would provide a type of congestion pricing, as we shall see – although not at first-best price levels. Naturally, the outcomes and hence the desirability of private ownership depend critically on market structure, which includes the nature of the highway network and the relationship among suppliers. In this section, we examine how one can predict analytically such outcomes.

6.1.1 Single Road with Static Congestion

We start with an unregulated private monopolist on a single road, using the multi-period setup introduced in equation (4.13). We can analyze this case by modifying the benefit-cost framework developed in Chapter 5. Instead of choosing capacity and tolls so as to maximize $W=B-C$, as in equation (5.1), we assume the monopolist maximizes profit Π, equal to its revenues minus its own costs:

[1] See World Bank (2006) for a review of recent experience.

$$\Pi = \sum_h q_h \cdot V_h \cdot \tau_h - \rho \cdot K(V_K) \tag{6.1}$$

where again h denotes a time period of duration q_h. The user-equilibrium condition from equation (4.1), equating marginal benefit $d(V)$ to the generalized price $p \equiv c(V) + \tau$, of course remains valid for each period h. It is convenient to directly substitute this condition into the objective function:

$$\Pi = \sum_h q_h V_h \cdot \left[d_h(\boldsymbol{V}) - c_h(V_h; V_K) \right] - \rho \cdot K(V_K) . \tag{6.2}$$

We maximize (6.2) with respect to capacity V_K and flows V_h.

Maximizing with respect to capacity produces, perhaps surprisingly, the first-best condition already encountered in (5.3):

$$\rho \cdot K'(V_K) = -\sum_h q_h \cdot V_h \cdot \frac{\partial c_h(\cdot)}{\partial V_K} . \tag{6.3}$$

For given flows V_h, the monopolist chooses capacity to minimize total social cost, including user cost. This is important as it shows that the monopolist is cost-conscious, even with those resources supplied by its customers. The intuition is that for any given flows V_h resulting from some generalized prices $p_h = c_h + \tau_h$, the monopolist would like to minimize user cost c_h and therefore maximize toll τ, while maintaining that generalized price. Every dollar reduction in total user cost can be turned into an extra dollar of toll revenues for given flow levels. The monopolist therefore faces the optimal incentive to minimize the sum of user cost and capital cost, just as in welfare maximization.

Maximizing (6.2) with respect to traffic volume, however, does not yield the first-best rule derived previously. Instead, we obtain the following first-order condition:

$$\tau_h = V_h \cdot \frac{\partial c_h}{\partial V_h} - V_h \cdot \frac{\partial d_h}{\partial V_h} - \sum_{i \neq h} \frac{q_i}{q_h} \cdot V_i \cdot \frac{\partial d_i}{\partial V_h} . \tag{6.4}$$

The first term on the right-hand side is equal to the first-best toll in (4.14) and (5.2). Thus, the profit-maximizing toll does at least partly internalize the congestion externality. However, two extra terms are added that take into account demand elasticities. With downward-sloping demand

functions and substitutability across time periods, both terms are positive, so the toll is higher than optimal — just what we would expect from a monopolist. When demands in different time periods are independent of each other, the last term disappears and (6.4) simplifies to:

$$\tau_h = V_h \cdot \frac{\partial c_h}{\partial V_h} - V_h \cdot \frac{\partial d_h}{\partial V_h} \quad \Leftrightarrow \quad p_h \cdot \left(1 - \frac{1}{|\varepsilon_h|}\right) = mc_h \quad \Leftrightarrow \quad \frac{p_h - mc_h}{p_h} = \frac{1}{|\varepsilon_h|}, \tag{6.5}$$

where ε_h is the own-period price-elasticity of demand w.r.t. generalized price (p_h) and mc_h is defined, as before, as $\partial(V_h c_h)/\partial V_h = c_h + \partial c_h/\partial V_h$, *i.e.* marginal social cost.[2] Equation (6.5) looks like the familiar monopoly rule equating marginal revenue to marginal cost; but here the price and marginal cost both include user costs c_h. As usual with monopoly solutions, it is valid only when demand is elastic ($|\varepsilon_h|>1$).

The monopolist internalizes the congestion externality because it has an interest in making its service attractive, so that users will pay more for it. When choosing the toll, just as when choosing capacity, the monopolist would like to reduce user cost in order to charge a higher toll. In doing so, however, it is constrained both by congestion technology and by users' demand elasticities. In fact, as we can see from the first of equations (6.5), the upward slope of the average user-cost function (representing congestion) affects the toll in exactly the same way as the downward slope of the inverse demand function; indeed, the inverse demand function $\tau_h(V_h)$ for the monopolist is given by $d_h(V_h) - c_h(V_h;V_K)$.

Because the monopolist takes marginal social cost into account in setting price, it may be said to practice a form of congestion pricing — but it then adds a markup represented by the bracketed term in the second equation in (6.5). This term multiplies the entire user-perceived price p_h, not just the toll τ_h. As a result, a substantial fee may be charged even during time periods when the optimal congestion fee is zero. An example of this in practice is the fact that the private operator of the express lanes on State Route 91 in southern California between 1995 and 2002 (see Section 4.3.3), in setting its time-varying fee, chose a non-zero fee even during nighttime hours.

[2] This result is equivalent to that of Mohring (1985) for a monopolist owner of a congested port (his equation 4, for one time period only). The equivalence involves converting his demand curve (stated as a function of fee τ) to one that is a function of generalized price p.

In the special case of perfectly elastic demand ($|\varepsilon_h|=\infty$), the demand-related monopolistic mark-up disappears and the monopolist undertakes socially optimal pricing and investment. This is the case in the classic analyses by Pigou (1920) and Knight (1924). Although important as a benchmark, it is of limited use in practice.

Note also from (6.5) that the fractional mark-up on marginal social cost, as given by the "Lerner index" $(p_h–mc_h)/p_h$, is simply the inverse of the absolute value of the demand elasticity $|1/\varepsilon_h|$. Again this is consistent with conventional microeconomics. A similar relationship holds in Ramsey pricing, in which social welfare is maximized subject to a minimum-profit constraint (Ramsey, 1927; Baumol and Bradford, 1970). We can derive the Ramsey result in this case by maximizing $B–C$ subject to a minimum constraint $\Pi^{\#}$ on profit Π, the latter defined in (6.1). Thus we maximize the Lagrangian function:

$$\Lambda = \sum_h q_h \cdot \int_0^{V_h} d_h(v)\mathrm{d}v - \sum_h q_h V_h \cdot c_h(V_h;V_K) - \rho \cdot K(V_K)$$

$$+ \lambda \cdot \left\{ \sum_h q_h V_h \cdot \left[d_h(V_h) - c_h(V_h;V_K) \right] - \rho \cdot K(V_K) - \Pi^{\#} \right\}$$

where λ is the Lagrange multiplier associated with the budget constraint. The first-order condition for capacity V_K is unchanged: once again, V_K is chosen to minimize total social cost. The first-order condition for volume V_h, however, changes. It may be solved to yield:

$$\tau_h = V_h \cdot \frac{\partial c_h}{\partial V_h} - \frac{\lambda}{1+\lambda} \cdot V_h \cdot \frac{\partial d_h}{\partial V_h} \quad \Leftrightarrow \quad p_h \cdot \left(1 - \frac{\lambda}{1+\lambda} \cdot \frac{1}{\varepsilon_h} \right) = mc_h$$

$$\Leftrightarrow \quad \frac{p_h - mc_h}{p_h} = \frac{\lambda}{1+\lambda} \cdot \frac{1}{|\varepsilon_h|}$$

where again *mc* denotes marginal social congestion cost. This equation is a well-known result for Ramsey pricing, but adapted here to incorporate a mutual externality and user-supplied costs.[3] When the constraint is not binding, substitution of $\lambda=0$ confirms that we are back at first-best

[3] It is derived in a more conventional context, for example, by Oum and Tretheway (1988), who go on to generalize it to handle externalities imposed by the price-setting firm on society in general.

pricing. When the constraint becomes increasingly hard to satisfy, so that $\lambda \to \infty$, the toll approaches the profit-maximizing one of (6.5). Likewise, note that the profit-maximizing toll (6.5) and capacity (6.3) are indeed also the same as those for a public operator who faces an infinite marginal cost of public funds, given (for the case of a single time period of duration normalized to one) by (5.16) and (5.17) when $\lambda_\tau = \lambda_K \to \infty$.

6.1.2 Single Road with Dynamic Congestion

In models with endogenous scheduling, demands at different times within the peak period are determined by individual travelers' tradeoffs between travel delay and schedule delay. Does this affect a monopolist differently from a welfare maximizer?

Arnott, de Palma, and Lindsey (1993) show that if the monopolist charges only a time-invariant fee, the problem is exactly like that just analyzed. However, if a time-varying fee is possible, one might wonder whether the travelers' tradeoffs across time periods set up varying time-specific elasticities to be exploited by a monopolistic road owner. If so, the monopolist would choose a *pattern* of time variation that differs from the optimal pattern.

We can solve the problem for the basic bottleneck model, defined in Chapter 3 and used in Section 4.1.2. We consider a downward sloping inverse demand function $d(Q)$. For convenience, we follow de Palma and Lindsey (2002) and decompose the time-varying toll $\tau(t')$ (for an exit at time t') into a time-independent "base toll" τ_0 and a purely time-varying component $\tau_v(t')$ that is zero for the first and last users to travel. This enables us to distinguish between the monopolist's choice of toll level t_0 and toll pattern $\tau_v(t')$. What we find is that the toll *pattern* is unaffected by monopoly, *i.e.* it is the same as the optimal toll pattern; whereas the toll *level* is higher for a monopolist by exactly the same amount (and for the same reasons) as in the static result (6.5).

The reasoning is as follows. First, consider the toll *pattern*. A revenue maximizer would set the toll pattern so as to eliminate any queuing, because queuing time can be replaced by toll revenues without affecting the generalized price p – exactly as we argued in the previous subsection when considering profit-maximizing investment. The toll schedule must therefore be at least as steep as the optimal one. But given the absence of queuing, the revenue maximizer would also not make the toll schedule steeper than the optimal one. If it did, this would create

periods within the peak where the bottleneck remains idle; but then it would be possible to extract some revenue from the earliest or the latest driver by shifting that driver to an empty slot within the peak, for which this driver is willing to pay because of lower scheduling cost. The purely time-varying toll component will consequently follow the same pattern as the first-best time-varying toll of (4.23).

Now consider the toll *level*. To see how the base toll τ_0 is chosen, recall from (4.25) that with this time-varying toll pattern, users adjust so that their average congestion-related user cost is $\bar{c}_g^1 = \frac{1}{2}\delta \cdot Q / V_K$, where Q is the total number of trips over the rush hour and δ is a composite measure of how costly it is to deviate from the desired schedule. They also pay an average of $\bar{c}_g^1$ in tolls. Thus we can write

$$\bar{c}_g^1(Q;V_K) = \bar{\tau}_v(Q;V_K) = \tfrac{1}{2}\delta \cdot \frac{Q}{V_K} \tag{6.6}$$

With the additional base toll added, generalized price, which must be equalized across users for them to be in equilibrium, is:

$$\begin{aligned} d(Q) \equiv p &= \tau_0 + \bar{\tau}_v(Q;V_K) + \bar{c}_g^1(Q;V_K) \\ &= \tau_0 + \delta \cdot Q / V_K . \end{aligned} \tag{6.7}$$

Profit is toll revenue less capital cost, or:

$$\Pi(Q,V_K) = Q \cdot \left[d(Q;V_K) - \bar{c}_g^1(Q;V_K) \right] - \rho \cdot K(V_K) . \tag{6.8}$$

This equation is just like (6.2) with one time period h, with $q_h V_h$ replaced by Q, and with c_h replaced by $\bar{c}_g^1$. So it leads to first-order conditions with the same properties as (6.3) and (6.4): namely, capital is chosen efficiently (given Q), and the toll level is set to account for marginal congestion cost (through $\bar{\tau}_v$) but with a monopoly markup (through τ_0). Writing out explicitly the first-order condition with respect to Q (and suppressing V_K as an argument in the functions), we have:

$$d(Q) - \bar{c}_g^1(Q) + Q \cdot d'(Q) - Q \frac{\partial \bar{c}_g^1}{\partial Q} = 0$$

where d' is the slope of the inverse demand curve. Equation (6.6) implies that the fourth term is equal to $-\bar{\tau}_v$. Because $d-\bar{c}_g^1-\bar{\tau}_v=\tau_0$, we find:

$$\tau_0 = -Q \cdot d'(Q) . \tag{6.9}$$

There are two ways that we can interpret the base toll in (6.9) as a mark-up over marginal cost. First, as argued in Section 4.1.2, the time-varying toll component $\tau_v(t')$ is equal to the time-varying marginal external congestion cost *mecc* for a user exiting at t'. With the base-toll τ_0 set according to (6.9), the total toll $\tau(t')$ is therefore equal to *mecc* (t') plus a time-independent demand-related mark-up.

Alternatively, we can rewrite (6.9) in a form like (6.5) by defining the marginal cost of adding a new user, given the optimal toll pattern:

$$mc = \partial(Q \cdot \bar{c}_g^1)/\partial Q = \bar{c}_g^1 + Q \cdot \partial \bar{c}_g^1 / \partial Q = 2\bar{c}_g^1 .$$

Then generalized price is

$$p = \tau_0 + \bar{\tau}_v + \bar{c}_g^1 = \tau_0 + 2\bar{c}_g^1 = \tau_0 + mc$$

so that (6.9) becomes (6.5) with the *h* subscripts removed and $1/|\varepsilon|$ defined as $(Q/p)\cdot -d'(Q)$.

The main insights from the static model on profit-maximizing tolling and capacity choice therefore survive in the basic bottleneck model.

6.1.3 Heterogeneous Users

When individuals' values of time differ, another deviation between conditions for optimality and those for profit maximization is created. This is because the non-discriminating monopolist considers the interests only of marginal travelers (those nearly indifferent to using the highway in question); whereas conditions for optimality include inframarginal travelers as well (Edelson, 1971).

We can elaborate using the analysis of David Mills (1981). Mills considers the situation when individuals with different reservation prices (hence accounting for different parts of the inverse demand curve) have different values of time. Then the monopolist may allow too much or too little congestion because revenues resulting from a marginal change in price depend on

just the marginal user, ignoring the benefits or costs for others. For example, suppose users with relatively high reservation prices (*i.e.* they are willing to pay a lot to travel) also have relatively high values of time. The existence of these users tends to increase the level of the first-best toll because they would benefit a lot from reduced congestion. But such users do not affect the marginal revenue of the operator — the left-hand side of the second of equations (6.5) — because they will take trips regardless of marginal changes in congestion. (Hence they are called inframarginal users.)

If the monopolist could price discriminate, *i.e.* charge different prices to users with different values of time, then the profit-maximizing and first-best congestion levels would coincide. However, distributional outcomes would differ because the price-discriminating monopolist would be able to extract consumer surplus otherwise enjoyed by users, in contrast to the toll authority in first-best pricing.

6.1.4 Private Toll Lanes: The Two-Route Problem Revisited

One way to allow a private road operator to implement pricing while limiting its market power is to maintain a free close substitute. This is in fact an arrangement of increasing practical interest, although usually it is combined with discounted or free travel for carpools.

We analyzed a similar situation under the rubric of second-best pricing and investment in sections 4.2.1 and 5.1.3, by positing two parallel links that are perfect substitutes for each other, one tolled (route *T*) and the other untolled (route *U*). There, we maximized welfare (benefits minus costs) subject to a user-equilibrium constraint on each link, as summarized in the Lagrangian problem of equation (5.9). Here, we assume that a private operator would maximize profit, subject to the same constraints. This means maximizing the Lagrangian function:

$$\begin{aligned}\Lambda = {} & V_T \cdot \tau_T - \rho \cdot K_T(V_{KT}) \\ & + \lambda_T \cdot \left[c_T(V_T; V_{KT}) + \tau_T - d(V_T + V_U)\right] + \lambda_U \cdot \left[c_U(V_U; V_{KU}) - d(V_T + V_U)\right]\end{aligned} \tag{6.10}$$

where $d(\cdot)$ is again the inverse demand curve for the entire corridor. The first-order conditions can be solved to yield:

$$\tau_T = V_T \cdot \frac{\partial c_T}{\partial V_T} - V_T \cdot d'(V) \cdot \left(\frac{\frac{\partial c_U}{\partial V_U}}{\frac{\partial c_U}{\partial V_U} - d'(V)} \right) \tag{6.11}$$

$$\rho \cdot K_T'(V_{KT}) = -V_T \cdot \frac{\partial c_T}{\partial V_{KT}} \tag{6.12}$$

where d' and K' denote derivatives.[4] The investment rule (6.12) has the familiar first-best structure indicating that conditional on travel volumes, capacity is chosen to minimize total social cost.

The toll formula in (6.11), however, is not second-best or even quasi first-best, as can be seen by comparing it with (4.35) and (4.6). Its first term shows that the profit-maximizer internalizes the congestion externality on the road under its control. Its second term gives the demand-related mark-up, which depends on how demand for the tolled link is affected both by congestion (on the untolled link) and by the overall corridor demand elasticity. (The markup is positive because d' is negative). This second term is a fraction, defined by the term in large brackets, of the mark-up that would apply if there were no free alternative — *i.e.* the markup that occurs in the profit-maximizing toll (6.5) for a single road. This fraction is zero when the competing route U is uncongested, since then demand for the toll road itself is perfectly elastic. The fraction rises to one as congestion on route U becomes highly sensitive to traffic ($\partial c_U/\partial V_U \to \infty$), since then traffic on the free route is effectively fixed and so inverse demand for the toll road has the same slope d' as total inverse demand.

It is illuminating to compare (6.11) term by term with the corresponding second-best toll of equation (4.35). In both cases, the congestion externality on the toll road itself is accounted for through the usual term reflecting marginal external congestion cost on the toll road, which we may denote $mecc_T$. Where they differ is in how they account for congestion on the competing road. In computing the second-best toll, a positive term is *subtracted* from $mecc_T$ to account for

[4] The toll formula (6.11) is derived for the case of fixed capacities by Verhoef, Nijkamp and Rietveld (1996). It could be derived alternatively by using the user-equilibrium conditions to translate total demand for the corridor into demand just for the toll road, then applying (6.5).

congestion spill-over to route *U*. But in computing the profit-maximizing toll, a positive term is *added* to $mecc_T$ to account for the additional revenue that can be extracted when congestion on the free road is heavy. Thus the profit-maximizing toll is higher than the second-best toll. It is no surprise, then, that the welfare gains from applying a profit-maximizing toll are below the already small gains from second-best tolling. Indeed they may well be negative when compared to the unpriced situation for the same capacity; this situation is in fact illustrated by the lowest curve in Figure 4.5 in Chapter 4. This is why Liu and McDonald (1998), in their study of partial pricing on the Californian SR-91, find a substantial efficiency *loss* in moving from no pricing to revenue-maximizing pricing on the express lanes; whereas they find a small but positive welfare *gain* from second-best pricing.

Why, then, is private ownership of express lanes receiving such favorable attention as a policy option? There are several reasons why it might be desirable in practice despite these theoretical results. First, although we have compared alternative regimes for a given amount of capacity, private ownership may in fact be the key to providing new capacity — as was true for SR-91 in California. In that case the relevant comparison is between a single free road and the same road augmented by a privately operated express road. Computing the welfare gain then involves knowing the capital cost of the new capacity. Nevertheless, if the private road is a financial success and there are no adverse spillovers elsewhere on the network, then the net benefits of adding and pricing the express road cannot be negative because the free road offers travel at least as fast as before, the toll road is used only voluntarily (hence its users must be at least as well off as they were on the free road), and the operator makes non-negative profits.

A second reason is user heterogeneity. Small and Yan (2001) and Verhoef and Small (2004) find that even holding total capacity fixed, the welfare losses from profit maximization (as compared to no pricing) become smaller, and may in some cases turn into gains, when users have heterogeneous values of time. The reason is the same as for public express-lane pricing, discussed in Section 4.2.1, and involves socially beneficial self-selection of users according to value of time.

A third reason could be favorable impacts of revenue-maximizing pricing on departure times. As we have seen using the basic bottleneck model, revenue-maximizing pricing leads to a

toll pattern over time that eliminates queuing. This remains true when an unpriced alternative exists and represents a potentially huge welfare gain (de Palma and Lindsey, 2000).

Finally, the two roads may be imperfect substitutes. An example is Route 407 in suburban Toronto, a privately operated road that parallels, several miles distant, the main east-west freeway through the city center. Imperfect substitutability could either increase or decrease the distortions from revenue-maximizing pricing, depending on whether it serves more strongly to undermine the operator's market power or to reduce the congestion spillovers. Viton (1995) analyzes such a model.

6.1.5 Competition in Networks

Private ownership has also been analyzed in various network configurations other than the classic two-route problem with a single, unpriced substitute. De Palma and Lindsey (2000), for example, consider various ownership regimes for a network of parallel links characterized as interacting bottlenecks. One result is that a duopoly of two private operators of parallel links achieves most of the potential efficiency gains from first-best pricing (over 90% in their base case with time-varying tolling). A mixed duopoly, with one public and one private operator, is even more efficient — consistent with more general results from oligopoly theory. (These results assume Nash-Bertrand competition, meaning that each operator takes the other's toll as given in choosing its own profit-maximizing toll.)

De Borger, Proost and Van Dender (2005) also consider a network of parallel links, but with static congestion and with two types of traffic: "regional" (able to choose the more favorable link) and "local" (forced by circumstances to use a particular link only). They study how two governments, each controlling one link, may engage in "tax competition," meaning they each try to attract revenue-producing travelers from the other's facility. They consider cases where the governments can and cannot distinguish between the two types of traffic in setting tolls, and also where they cannot toll the regional traffic at all. It turns out that social welfare is substantially enhanced by the ability to toll regional traffic. The ability to distinguish between regional and local traffic, by contrast, does not matter much for social welfare.

A different type of tax competition is studied empirically by Levinson (2001). He presents evidence that states in the US are more likely to apply tolls to their major through roads when their traffic contains a higher share of non-residents.

For more general networks, we can appeal to more general results of Economides and Salop (1992) involving substitutes and complements. These results suggest that competition among producers of goods that are substitutes (*e.g.*, operators of competing parallel roads) leads to lower prices than a combined monopolistic producer. By contrast, competition among producers of complements (*e.g.*, operators of serial links) leads to higher prices (but lower total profits) than a single monopoly.[5] If each good is produced by a separate (but otherwise monopolistic) firm, then each firm applies a conventional demand-related markup. The consumer needs all goods when these are perfectly complementary, so the final combined good (the trip, in a roads context) gets multiple mark-ups applied on top of each other. These results suggest that in a general network of roads with private ownership (or private franchised operation), the results of various degrees of competition depend on whether the private operators control parts of the network that are predominantly substitutable (parallel) or complementary (serial). Yang and Huang (2005) provide more specific results.

We can illustrate this idea formally by considering two extreme cases of dividing control of a single corridor with only regional traffic. We allow a number F of identical revenue-maximizing firms to control different parts of the corridor, with F varying between one (monopoly) and infinity (perfect competition). Each firm's capacity and costs are fixed. In one case, the corridor is divided into F equal-capacity parallel roads, each operated by a different (and otherwise unregulated) firm. Total corridor traffic V divides across the roads according to the Wardrop conditions, constrained by $\Sigma_f V_f = V$. In the other case, the corridor is divided serially into F segments, each controlled by a different firm. The total corridor traffic level V then

[5] This latter result is closely related to "double marginalization," which occurs with vertically organized monopolistic producers of intermediate and final goods. With double marginalization, however, the upstream firm faces the downstream firm's marginal revenue function as its inverse demand function. With pure complements as discussed in the main text, the firm faces the market inverse demand function, shifted downward by the prices charged by the other firms. Double marginalization in urban transport would occur when a monopolistic transit firm needs to travel on a private highway without substitutes, or when a transit operator cannot do without the services provided by a private station operator.

applies to each firm. In both cases, because the firms are identical, we consider only symmetric equilibria, *i.e.* outcomes for which all firms have identical tolls, traffic, and hence revenues.

First, we consider the case of parallel roads that are perfect substitutes, a case analyzed by Engel, Fischer, and Galetovic (2004). The equilibrium toll for firm f can be derived by maximizing the following Lagrangian:

$$\Lambda_f = V_f \cdot \tau_f + \lambda_f \cdot \left[c_f(V_f) + \tau_f - d(V_f + \sum_{g \neq f} V_g) \right] + \sum_{g \neq f} \lambda_f^g \cdot \left[c_g(V_g) + \tau_g - d(V_f + \sum_{h \neq f} V_h) \right] \tag{6.13}$$

where $d(\cdot)$ is the inverse demand function for the entire corridor and each term in square brackets represents a user-equilibrium constraint. The first-order conditions can be solved to yield the following toll formula:

$$\tau = \tau_f = V_f \cdot \left(c_f' - d'\right) + \frac{-d'(F-1)}{c_{-f}' - d'(F-1)} \cdot V_f d' , \tag{6.14}$$

where τ is the toll that each user will pay for a trip; τ_f is the toll for a specific firm (these are identical across firms and equal to τ because of symmetry); c_f is the user cost for the road controlled by firm f; and c_{-f} is the user cost for any other road (symmetry of course implies that $c_f = c_{-f}$). Equation (6.14) shows how the toll is equal to the monopolistic toll of (6.5) when $F=1$, while it approaches the first-best toll of (4.6) when $F \to \infty$. These results are intuitive, and suggest that the equilibrium toll level is closer to the first-best level when the number of firms is large so that each firm has little market power.

Now consider when individual firms occupy serial segments, each carrying the same traffic V as determined by the inverse demand curve.[6] The following Lagrangian applies to firm f:

[6] An example is the recent privatization of two separate US toll roads, the Chicago Skyway and the Indiana Toll Road. These roads, in adjacent states, cover parts of the same interstate highway route (I-90) and so carry a lot of through traffic. Perhaps the problem of excessive tolling, illustrated in this paragraph, is part of the reason why the consortia of firms that won the (separate) franchise auctions for the two roads include two large firms in common, Cintra Concesiones de Infraestructuras de Transporte (from Spain) and Macquarie Infrastructure Group (from

$$\Lambda_f = V\tau_f + \lambda \cdot \left\{ c_f(V) + \tau_f + \sum_{g \neq f} \left[c_g(V) + \tau_g \right] - d(V) \right\} \tag{6.15}$$

where $c_g(V)$ is the average user cost incurred just on the segment operated by firm g, so that the user cost for the entire trip is $c=\Sigma_g c_g$. The first-order conditions yield the following firm-specific toll:

$$\tau_f = V \cdot \left(c_f' + \sum_{g \neq f} c_g' - d' \right). \tag{6.16}$$

This firm thus internalizes not only the congestion on its own road segment ($V \cdot c_f'$), but also that on each other firm's segment ($V \cdot c_g'$). The reason is that congestion on the other firms' segments affects the firm's marginal revenues in exactly the same way as congestion on its own segment. Since every firm internalizes the congestion on the entire road, the combined toll facing the traveler over-internalizes congestion if $F>1$. Furthermore, each firm applies a demand-related markup ($-V \cdot d'$) that is identical to what a single monopolist would charge; thus this markup gets charged F times when considering the entire trip. Writing this formally, the total trip toll τ is:

$$\tau = \sum_f \tau_f = F \cdot V \cdot \left(c' - d' \right), \tag{6.17}$$

which is exactly F times the monopoly toll of (6.5) — which already exceeds the first-best toll. We therefore now find the opposite result to the parallel competition case: here, the lower the number of firms, the closer the overall toll approaches the efficient level — although even with one firm it will never reach that level unless demand is perfectly elastic, whereas a larger number of firms produces an ever-larger total trip toll and hence drives demand toward zero.[7]

These opposing results for the extreme cases are exactly in the direction predicted by the general analysis of Economides and Salop, as just discussed. Thus, the desirability of private

(..continued)
Australia). These firms have an interest in internalizing the combined revenue potential from the two roads and therefore could have influenced their respective consortia to bid more, on the expectation that such internalization would be realized yielding higher total revenues.

[7] De Borger, Dunkerley and Proost (2006) find that similar mechanisms are relevant when different governments control different parts of a corridor.

road ownership, and the ideal number of competitors, depend critically on the network configuration and the distribution of firms over that network. Generally, an increase in competition among substitute roads would bring equilibrium tolls closer to first-best levels, while the opposite applies for complementary roads. This suggests that private operators, if allowed on a network, should serve full-length corridors but should face competition when doing so.

Another option, of course, is to regulate the private firms, a topic to which we now turn.

6.2 Regulation and Franchising of Private Roads

The local monopoly power of a private road operator provides a potentially strong economic rationale for regulation. Moreover, it is impractical to allow unrestricted free entry of private road operators given the physical nature of road investment, including its network aspects, lumpiness, irreversibility, right-of-way requirements, and land-use implications. Under what institutional set-up, then, can private roads contribute most to social welfare? The question gains relevance with the growing interest in and importance of private involvement in road operations throughout the world (Estache, 2001; World Bank, 2006). This section explores various dimensions of such institutional arrangements, which are generally known as public-private partnerships (PPPs). Many of these considerations are part of a more general analysis of privatization in transportation, which is discussed at greater length by Nash (2005). Many of them also can be fit into a broader theory of contract design with limited information, such as that provided by Laffont and Tirole (1993) and discussed in Section 6.3.5.

Private involvement is often motivated by the desire to bring in private capital when public budgets are tight. A second motivation is the hope that private management will be more efficient than public management. Another motivation might be that the public would more readily accept road pricing from a private than from a public entity.

There are few if any truly standardized formats for PPPs, but several categories are generally recognized. The most basic is Build-Operate-Transfer (BOT). Under such a scheme, the concessionaire finances, builds, operates, and maintains the road — usually to predefined specifications — while collecting tolls for a certain period such as 30 years, after which control

is transferred to the government. A variant is Rehabilitate-Operate-Transfer (ROT), involving the rehabilitation of an existing road instead of the construction of a new one. The Design-Build-Finance-Operate (DBFO) or Design-Build-Operate-Transfer (DBOT) format is similar to BOT, but the private party is invited to propose how the road should be configured as well as how it will be built and operated. Another format is leasing, where the government sells to a private operator the right to operate and charge users for an existing road for a specified time period; prominent recent examples are the Chicago Skyway and the Indiana Toll Road, two adjacent portions of US Interstate Route 90 for which long-term leases were sold in 2005 and 2006, respectively.

As an alternative to actual tolls, systems of "shadow toll" have also been used in certain countries including the UK, Finland, and the Netherlands. In this case, users do not pay actual tolls, but the authority remunerates the concessionaire depending on the degree of utilization. A shadow toll may be better or worse, from the point of view of social welfare, than an actual toll depending on all the considerations discussed in Section 6.1. Shadow tolls are often used in conjunction with a DBFO system of private operation.

Experience with highway franchising has not always been positive. Engel, Fisher and Galetovic (1997) highlight two pitfalls: the frequent use of government guarantees and renegotiations in the face to financial trouble. The first reduces the incentives to control construction costs, while the second encourages bidders to submit overoptimistic bids ("lowballing") on the assumption that discrepancies will be made up later. Engel *et al*. attribute these problems mainly to the fact that most franchises are awarded for a fixed period. They therefore propose to use a variable-term contract instead, in which the franchise is awarded to the bidder that requires the *least present value of revenue* (LPVR) from tolling. In a LPVR auction, each possible revenue stream is converted to a present value through procedures defined in advance as part of the request for bids.[8] The bidder then specifies an amount for this present value that, once reached through the accumulation of toll revenues, ends the term of the franchise. Assuming there are multiple bidders, the smallest such bid wins the franchise. Such an approach is likely to limit the need for guarantees and the scope for contract renegotiations, and

[8] Present value is computed using a formula like (5.18).

therewith the distorting impacts that such practices exert on both the original franchise and the subsequent operations.

Alternative criteria for auctions that have been used in practice include the total capacity cost, duration of the construction period, the toll rate at opening, and the length of the concession (World Bank, 2006). Verhoef (2007) finds that toll rates and capacities may in fact depend strongly which criterion is used. His analysis considers static congestion, homogeneous travelers, neutral scale economies, competitive auctions, and allows for unpriced congestion elsewhere on the road network. Perhaps surprisingly, a criterion based on maximizing traffic flow is usually capable of reproducing the zero-profit second-best outcome. An exception is when that the road would produce something akin to a Braess paradox — in which case this criterion could lead to a minimum social surplus — but presumably this situation represents a planning failure at the very start of the process.

A key element of franchising is how the risk of uncertain future demand is shared between the government and the private operator. The parties to the agreement (including financial institutions providing capital to the franchisee) vary in their costs of bearing risk, in their information about contingencies, and in their ability to influence these contingencies. It is important to take account of these variations, in particular to provide incentives against any misuse of private information, yet while allowing enough flexibility in setting fares to enable pricing to achieve its welfare-improving allocative effects. As an illustration of how such considerations are sometimes ignored, Nash (2005) gives the example of the UK paying shadow tolls to private road franchisees, ostensibly to transfer to them the risk of inadequate traffic to justify the road; yet the public retained the power to build competing roads and restricted the franchisee's ability to develop new interchanges or other measures that would increase traffic on its road.

Because of the difficulties of foreseeing all contingencies, the franchising agreement may include some form of price and/or capacity regulation. Otherwise, competition to win the franchise would push bidders towards a profit-maximizing combination of capacity and toll which, as we have seen, may be far from socially optimal when the road is part of a network. But setting rigid toll rates in advance makes future changes highly political, and may discourage

pricing policies that are in the public interest. One solution to this is regulations that limit the rate of return on the project, as specified for example in the franchise that enabled the express lanes to be built on California's State Route 91 as described earlier. Private financial and consulting firms contain personnel with experience in designing contracts under conditions of uncertainty, and for this reason government agencies have often engaged a private firm as a financial advisor early in the process. Still, there is a basic dilemma: the more conditions are included in the franchise the better it can be designed to fulfill the public interest, but the more openings it creates for incompetence, political manipulation, or corruption on the part of the public authority — all of which tend to defeat the purpose of privatization and may discourage sufficient numbers of bidders for competition to have its desired effect.

Another dilemma is the need for non-compete provisions that protect a private toll-road operator against competing free roads that might be built by the public sector. Naturally, no private investor will want to put money at risk without some assurance against such competition, especially since responsible public officials and political parties turn over frequently. At the same time, the public tends to resent agreements that foreclose public options to solve public problems, and this can undermine support for the franchising operation. An example is State Route 91 in Orange County, California, where the original private operator of the 91 Express Lanes (under a very long-term lease) retained veto power over any capacity improvements on the free portion of the road. After it exercised this power in court, the provision became so controversial that it was part of the motivation for a public buyout of the express lanes. Similarly, citizens of Sidney, Australia, became incensed when competing roads were closed as part of a franchise agreement with the builder of the Cross City Tunnel. This resulted in a decision to reverse some of the road closings, presumably with compensation to the tunnel operator, and a public commitment by the prime minister not to permit such provisions in the future.[9]

A skeptic might well argue that, given such complexities, public road provision and tolling should remain the preferred option. A more pragmatic viewpoint is that private road operation should not be a goal in itself, but should be an option when conditions warrant. Such conditions

[9] "Sydney Tunnel Mired in Dispute," *Public Works Financing*, July-August 2006, p. 21.

could include competition from substitute road links, a lack of privately controlled complementary road links, an inefficient public road supplier, political limitations on the public sector's ability to implement pricing or to undertake desirable price discrimination, disadvantages for the public sector in capital markets, public budgeting constraints that lead to insufficient funds for investment, relatively small external effects other than congestion, and the availability of an efficient and effective auction mechanism. But libertarians might argue that some of these conditions are universally present, so that private provision should be the norm and only if it involves demonstrable and insurmountable problems should public provision be substituted (*e.g.*, Foldvary, 2006).

We conclude, then, that it is impossible to say in general whether private or public road provision is more desirable in the long run can. The extent to which the conditions just described apply, and therewith the relative desirability for private road supply, can be expected to vary strongly across nations, regions, and over time. Fortunately, many degrees and forms of privatization are possible and understanding these conditions is the key to choosing among them.

6.3 Privately Provided Transit Services

Following widespread socialization of public transit during the middle of the twentieth century, the world has witnessed since 1980 a wave of privatization and deregulation that has resulted in much experimentation in organizational forms for providing transit service. This has created opportunities for comparing results of alternatives ranging from public ownership and operation, through various public coordinating or regulatory roles, to fully deregulated private provision. Nash (2005) provides an insightful review of issues and experience.

Dissatisfaction with public control was fueled by huge budgetary commitments to urban public transit, which in many places grew enormously following consolidation of operators under public ownership. For example, Pickrell (1983) and Lave (1989) examine the sources of rising US transit operating deficits, tracing much of the cause to wage increases and inefficiently capital-intensive operations, which in turn were encouraged by incentives built into subsidy programs. Analysts distinguish several other possible disadvantages of public ownership

including political interference and inability of the public sector to finance timely investments. As noted by Nash (2005), these differing motivations for turning to the private sector do not necessarily lead to compatible directions for change: for example, cost control may require competition whereas raising funds for investment may require some monopoly power.

Of course, transportation is only one of many industries for which the question of private or public operation has been greatly discussed. Two valuable general assessments are those by Kay and Thompson (1986) and Vickers and Yarrow (1991). One conclusions is that a lot depends on details of the industry. Analyses of transportation industries have been important in coming to these more general conclusions as well as in applying more general results to cases with significant public-policy implications. For these reasons, the economic study of transportation, and of public transit in particular, continues to provide insights of widespread interest.

In this section, we review the main forms that private-sector involvement takes, followed by an analysis of the outcomes likely under some of these forms — particularly as influenced by the nature of competition or lack thereof. We then examine empirical evidence on the extent to which private operators display lower costs and/or higher labor productivity. Next, we review worldwide experience with privatization and deregulation of public transit, with special attention to the rich lessons revealed by the UK since 1985 and to special problems in developing nations. Finally we consider paratransit (a loose collection of transit-like services, often provided privately) and conventional taxi service.

6.3.1 Forms of Privatization

We have already seen, in discussing private highways, that there are many intermediate positions between the extremes of full public operation and unregulated private ownership. With transit, there are even more dimensions along which such intermediate positions can be defined, since we must consider not only infrastructure provision but also ongoing operation.

The least drastic form of privatization is *tendering*, also called *contracting* or *contracting out*. The public authority can retain full control over network design and services offered, but contract with private firms to carry out specific parts of this overall design such as operating

prespecified bus runs or maintaining rolling stock (*i.e.* vehicles). An example of tendering is the private bus operators with which London Transport contracted in the early phases of UK deregulation during the 1980s.

Going somewhat further, the public authority can *franchise* some of these services by licensing private firms to operate them under less specific guidelines.[10] The franchise is for a specified period of time, which in the case of regional rail services has ranged from as little as two or three years in Sweden to several decades in parts of South America (Nash, 2005). Performance goals may be mandated or encouraged through incentives, including the prospect for favorable consideration for later renewal of the franchise. Desired investments whose useful lifetimes would exceed the length of the franchise can be encouraged by contractual terms, government investment subsidies, or co-ownership arrangements. Needless to say, each of these options opens the possibility of contract disputes, strategic renegotiations, and outright abandonment of obligations by a financially failing firm.

Going further still, the market to provide certain services may be simply turned over to one or more private firms, as is common for example with telecommunications and electricity in many nations. This could be a *regulated monopoly*, a single firm allowed to provide services under tightly controlled terms of price and service quality. If freedom of entry is allowed, the regulations over price and service may be relaxed on the assumption that competition will produce a desirable result just as for other goods in a largely market economy. Depending on how completely such regulations are relaxed, the result is some degree of *privatization with deregulation*. In virtually all cases, government oversight is maintained over such things as safety, financial disclosure, and matters covered by general business policies.

Naturally, the relative advantages of different forms of privatization depend on the nature of the industry. A primary consideration is whether or not the market is a natural monopoly,

[10] Definitional lines vary. Preston (2005) defines tendering as "firms bidding for the right to operate services" (p. 65) and defines franchising as a particular type of tendering, involving "contracting out some of the tactical … as well as operational functions" with an emphasis on arrangements that "expose bidders to revenue risk" (p. 66). Halcrow Fox (2000), however, applies a narrower definition of franchising: an arrangement in which "the authority is in the lead in specifying the broad public transport product and is prepared to incur the costs of doing so;" whereas a "concession" is somewhat closer to a true free market: a situation in which "the authority imposes a few basic requirements and has no financial responsibility" (p. 3).

meaning that costs display scale economies that are strong enough to make it unacceptably inefficient to have more than one producer. Natural monopoly is often thought to characterize infrastructure (*e.g.* rail track, large bus terminals) but not operations. However, as we have seen, transit operations are also subject to scale economies when they require substantial access costs on the part of users and when the desire to reduce such access costs, by providing frequent and/or geographically dense service, is a limiting factor in choosing the size of transit vehicles. Thus transit operations may also be a natural monopoly and it is no coincidence that free entry into privatized transit markets has often led to consolidation of the market by one or at most a very few firms.

6.3.2 Market Structure and Competitive Practices

If markets are left partially or fully unregulated, what will happen? This question has been addressed specifically for public transport through theoretical, empirical, and simulation analysis.

A first question is whether private firms could operate profitably without subsidies. Several authors have found a range of conditions under which this is possible. Harker (1988) formulates a model in which several types of transit compete with each other and with (uncongested) auto and applies it to three Philadelphia-area corridors; he finds profitable bus service in one of them, which has relative high density and low incomes. Cervero (1990), reviewing individual transit routes in twenty-five U.S. cities, similarly finds that those operating at a profit serve mostly high-density areas with low-income people taking short trips. By contrast, case studies analyzed by Morlok and Viton (1980, 1985) find a niche for expensive, high-quality service by commuter rail (Chicago), rapid transit (Lindenwold line into Philadelphia), and bus (express service into Manhattan). It seems then that there are two potentially profitable markets for conventional urban transit: high-quality express service from affluent suburbs to large employment centers, and local bus service serving low-income people in high-density areas.

A second question is whether private operation produces desirable results. This is a much broader question and to analyze it, we have to consider the complexities of competition under various market conditions.

One line of inquiry is the nature of unregulated and imperfectly competitive equilibria in which two or more firms compete. Often these are modeled as some variation of a Bertrand equilibrium, in which each firm assumes that the price and quality of service offered by other firms are fixed. For example, Evans (1987) considers an unregulated non-cooperative oligopoly with free entry. The equilibrium exhibits higher fares and higher service frequency than would result either from unconstrained welfare maximization or from welfare maximization subject to a breakeven constraint. However, in Evans's simulations (p. 23), welfare in the oligopolistic case falls only slightly short of that resulting from constrained or unconstrained welfare maximization, whereas it far exceeds (at most demand levels) that resulting from monopoly. Hence Evans results are supportive of deregulation as a viable policy when the market is likely to accommodate two or more firms.

But will such an oligopolistic structure emerge in an unregulated market? Dodgson and Katsoulacos (1988b) examine entry conditions in order to address this question for local bus service. They find a wide range of market conditions under which just two firms share the market. However, the firms differentiate their products in order to increase their market power, so the results are not necessarily as desirable as in Evans' more symmetric solutions. Similarly, Viton (1981a) finds that two transit firms would significantly differentiate their products if they engaged in Cournot-like competition, in which each assumes the other will respond to its actions so as to maintain its customer base.

Despite the theoretical possibility of oligopoly, experience in Great Britain suggests that in most deregulated local bus markets, one firm becomes dominant through superior efficiency, predatory practices, mergers, or luck. Thus we need to ask whether it is necessary to regulate such a firm in order to prevent the high price and low ridership expected from a monopoly. A key question here is whether *potential* entry by competitors would serve to discipline a monopolist's decisions about price and service. At the extreme, we can ask whether the transit market is *contestable*: Is the prospect of hit-and-run entry sufficiently threatening to force a

monopolist to choose competitive fare and service policies? Contestability requires that the entrant have low barriers to entry and exit (the latter requiring an absence of sunk costs), and also that the incumbent be unable to change fares and service levels too quickly (Baumol, Panzar, and Willig, 1982).

Button (1988) argues that there is substantial though not perfect contestability in urban transit. Certain features favor low barriers to entry and exit: lack of significant economies of scale in providing vehicle-hours of service, low setup costs, and a good market in used bus equipment. On the other hand, substantial investments may be required to establish a reputation, build terminal facilities, or achieve efficiency through learning-by-doing. These investments by an entrant cannot be retrieved if the monopolist responds to entry by quickly lowering fare or increasing service. If the monopolist can credibly threaten to do so temporarily, in order to drive out the entrant, it is said to be capable of *predation*, which discourages entry. Dodgson and Katsoulacos (1988a) analyze when a rational monopolist would respond in this way, showing that informational asymmetries can lead to successful predation.

Indeed, the limited empirical evidence suggests that transit markets are not fully contestable. Evans (1988) describes the experience in Hereford, England, where transit service was deregulated beginning in 1981. Following a brief period of intense competition, the dominant firm drove out all its rivals except in one small segment of its market. Fares ultimately returned nearly to the levels that prevailed prior to the experiment, but service levels remained substantially higher. Evans suggests that potential entry constrains a monopolist's service levels, which cannot be quickly increased in response to entry, but not its fares. As we will see in Section 6.3.4, higher prices and more frequent service were consistent results of deregulation of local bus service in Great Britain (outside of London) starting in 1986.

Such experience is consistent with the theory of competition with differentiated products — in this case, service at different times of day (Schmalensee, 1978). In such a market, the threat of potential entry will typically cause a monopolist to offer an excessive number of products in order not to leave an open niche for a competitor. The reason is that the monopolist can protect its high profits on each product through price predation, giving it a strong incentive to maintain

dominance in each product; but it cannot so easily protect against new products because doing so would require immediately matching an entrant's product characteristics.

Van der Veer (2002) performs numerical simulations to compute the results of such behavior on a prototype bus line. He finds that, as expected, a profit-maximizing monopolist would like to offer service that is *less* frequent than would be optimal — even less than would be second-best optimal subject to a breakeven constraint. But if the monopolist wants to deter entry, it will instead offer service that is *more* frequent than optimal. Van der Veer also finds inefficiencies regarding other dimensions of service quality, which can be partly ameliorated if the government offers a per-rider subsidy. A greater improvement can be achieved by combining a per-rider subsidy with a lump-sum payment required of the firm, which is what could result from a competitive franchising system. The idea here is that the ridership-related subsidy encourages a lower price and better quality of service, while the lump-sum tax (assumed to apply equally to a potential entrant) reduces the incentive to oversupply service because it makes entry more difficult.[11]

The models described above mostly assume constant returns to scale in producing intermediate outputs. Furthermore, most assume implicitly that any economies of scale due to user-supplied time is at a system rather than a firm level: that is, the traveler cares only about total bus frequency on the route, not about the frequency provided by a given firm. This, however, raises troubling questions about the viability of a non-integrated system of urban transit. What if it is not feasible for each firm to use the same stops, for example because they use vehicles of different sizes or because major terminals are owned by one firm? What if consumers care about the reputations of firms whose vehicles they are about to enter? What if the unregulated equilibrium entails differentiated products, *e.g.*, high-fare express and low-fare local service, so that travelers with a strong preference for one cannot benefit from the extra service frequency offered by the other? In all these situations, scale economies are lost by

[11] To see why a per-rider subsidy is at least partly passed through in lower prices, consider an example with constant marginal production cost mc and linear demand curve $p=a-bq$, where p is price and q is quantity. Marginal revenue is then $mr=a-2bq$, and the monopolist chooses q where $mc=mr=a-2bq$, yielding $q^*=(a-mc)/2b$ which can be achieved by charging price $p^*=(a+mc)/2$. With a per-rider subsidy s, it will set $mc=s+a-2bq$, which is the new marginal revenue; this yields $q^{**}=(s+a-mc)/2b$, achieved by charging price $p^{**}=p^*-(s/2)$. Thus in this example, half of the subsidy is passed through in lower fares.

allowing multiple providers, because the waiting times of a given firm's riders are not diminished by an increase in service supplied by other firms. Nash (1988) emphasizes the importance of system integration in realizing these user-cost savings and also reminds us of several other sources of economies of scale and scope, such as through ticketing of passengers and scheduling of drivers, that occur in an integrated system. The implication is that efficiency may be lost unless a central authority takes a proactive role in coordinating service.

Klein, Moore, and Reja (1997) tackle yet another issue of entry conditions. Part of a firm's set-up requirements for entering the market for local bus service, they argue, is the need for customers to learn that if they show up at a particular bus stop, a bus will appear to take them where they want to go. This requires establishing a reputation for frequency of service, perhaps by providing initially a greater frequency than could be justified otherwise. But once the availability of service at that location is widely known, competing firms can pick up those same waiting passengers, depriving the initial entrant of the returns from its investment. If the potential entrant understands this in advance, entry may never occur. The solution suggested by Klein *et al.* is to establish "curb rights" that allocate a given curb location to a given firm, which would then be able to reap the advantages of its reputation either directly or by licensing the right to others.

The theoretical considerations described here are consistent with more general analyses of privatization of industries mentioned earlier. Those analyses stress the importance of competition and other institutional structures in providing incentives for good management. Neither private nor public nor private ownership guarantees a strong or a weak set of incentives; instead, much depends on specific rules and policies.

6.3.3 Efficiency of Public and Private Providers

Two types of studies have attempted to compare the costs or productivities of public and private transit operators. The first type compares firms across cities, often estimating cost functions to control for factors other than the type of ownership. The second examines the results in a given city when tendering or franchising of transit services is introduced. Karlaftis (2007) and Frick, Taylor, and Wachs (2007) review these studies carefully.

Comparisons Across Areas

Cross-sectional studies have reached varying conclusions about whether or not private operators are more efficient than public operators.[12] These studies are complicated by potential biases that may make private operators falsely appear more efficient. Public operators often experience sharper daily peaks, and a public authority may take over previously failing private firms or spin off its more successful operations, thereby leaving it with less inefficient operations at any point in time. After accounting for these factors, Iseki (2003), in one of the most careful analyses, finds modest cost savings from contracting of around 5 to 8 percent in the US.

Before-and-After Comparisons

Preston (2005) and Karlaftis (2006) review a number of cases where publicly operated bus and rail services switched to a tendering system. Again the evidence is mixed, but generally positive. Cost savings and/or productivity improvements have been reported for several cases in Sweden, Spain, Australia, New Zealand, and the US. For bus services, most cases reported have shown some immediate reductions in unit costs, averaging around 20 percent if services remained unchanged and more if services were restructured.

The relatively simple but apparently effective practice of *gross cost contracting* (the "Scandinavian model" in Preston's terminology), in which firms bid on the cost at which they will offer specified services, has been used in Sweden, Norway, Copenhagen, London, Helsinki, Rome, Auckland, and Las Vegas (Nevada), among other places. Some studies have suggested that savings are greater with *net cost contracting*, in which the firm collects and keeps fare revenue. Competitive bids then specify the amount of subsidy required or, perhaps, the amount of profits the firm is willing to return to the government. With this arrangement, the risk created by uncertain demand is shifted to the private firm — which has advantages and disadvantages. On the plus side, the firm is given an incentive to provide services that attract users. On the other hand, private firms may be less able to bear risk than a government, resulting in higher bids and less competition for the contract (Estache and Gómez-Lobo, 2005).

[12] For reviews, see Perry, Babinsky, and Gregersen (1988) and De Borger and Kerstens (2000).

The experience with tendering service is less favorable with rail than with bus. Rail service has higher fixed costs and involves a more complex relationship between infrastructure and operations. These traits create more scope for strategic bidding and predatory pricing as means for firms to attempt to control the market. Attempts to privatize the infrastructure itself have been the most problematic, as we will discuss in Section 6.3.4.

As with many economic policies, success depends in part on the particular mechanism used and how well it matches conditions of the local market. There are many dimensions for choosing a form of tendering or franchising: for example, contract duration, ownership of vehicles and terminals, types of bonuses and penalties, and location of responsibility for choosing fares and service levels. Thus it is not surprising to find a lot of variation in outcomes.

Furthermore, the results of any real privatization will depend greatly on the extent and nature of regulation of private firms. A comprehensive theory of regulation developed by Laffont and Tirole (1993) emphasizes information asymmetries between a regulator and the regulated firms — typically the firms know more about their own cost structures, which is information the regulator needs in order to set regulatory parameters. A central conclusion of the theory is that one can often design contracts that induce firms to implicitly reveal their information, and to voluntarily make desirable choices, by offering an appropriate menu of contractual options. Gagnepain and Ivaldi (2002) apply this theory by estimating cost functions that include regulatory variables, using a data set drawn from urban transit providers in France. They find significant departures from optimal contracting arrangements, with cost-plus contracts proving to be especially inefficient. There have been a few other attempts to apply this theory to public transit services, described by Estache and Gómez-Lobo (2005).

Conclusions

While results of private transit provision are promising, the evidence is not straightforward. Rather, it supports the conclusion from the theoretical literature, described earlier, that the most important factors are those that affect the nature of management incentives, especially the nature of competition and regulation, rather than the type of ownership itself. This observation provides

a useful background as we examine, in the next subsection, the practical experience with institutional changes in public transport.

6.3.4 Experience with Privatization and Deregulation

A great deal of experience is now available to help assess the implications of privatization and deregulation of public transit services. Nash (2005) is particularly helpful, covering not only urban transit but intercity modes as well. Here we focus on two special cases that have proven illuminating: the UK starting in 1985, and developing nations.

UK since 1985

One of the most far-reaching and varied experiments with privatization of transit services took place in Great Britain following the British Transport Act of 1985. Useful reviews include Glaister (1997), Small and Gómez-Ibáñez (1999, Section 5.5), Darbéra (2004), and Nash (2005, Section 4).[13]

We can distinguish three quite different experiments in British urban areas. Outside London, urban bus services were mostly privatized and free entry was permitted, with municipal operators required either to privatize or to operate on a commercial basis. (Subsidies were allowed but had to be made available on equal terms to all firms.) Within London, the public bus operator (London Transport) was retained but it was required to tender services through competitive contracts, while maintaining central control over schedules, routes, and fares. The London Underground, by contrast, was unaffected by the 1985 act, but starting in 2003 its infrastructure maintenance and investment activities were spun off through public-private partnerships (PPPs).

The main results outside London were large service increases (as measured by vehicle-kilometers), higher fares, lower patronage, and substantial cost savings. Real wages for drivers stabilized after a prior increase but have not substantially declined, implying that the cost savings represent improved productivity. Much of the service increase represented a switch to smaller

[13] These reviews in turn rely on many earlier studies, of which two of the most comprehensive are White (1995) and Mackie, Preston, and Nash (1995).

buses, called "minibuses." The higher fares resulted not from the new market structure but from a drastic reduction in government subsidies that was made simultaneously with deregulation.

The patronage decline was the biggest surprise. Several studies from the mid-1990s compare patronage with counter-factual scenarios to see how much of the decline was due to deregulation (Small and Gómez-Ibáñez, 1999). Some results suggest that fare increases alone cannot explain the decline, with authors suggesting that lack of integration of service among competing operators may have diminished the quality of service. Neither is the decline explained by transitional difficulties, as it continued throughout the 1990s and, at a slower rate, to the time of this writing.[14]

The nature of competition varied among metropolitan areas. In most cases, any serious competition was soon eliminated by aggressive increases in route frequency, predatory pricing, or mergers. Mackie, Preston, and Nash (1995) explain this consolidation, at least in part, as reflecting inherent advantages of incumbents such as local knowledge. Scale economies of the kind described in Chapter 3 may also help explain it.

The experience with London's buses is in some ways similar: more service, higher fares, and dramatic cost reductions. However, patronage did not fall, but instead has shown a steady increase.[15] This offers some support to the hypothesis that users benefit from the integrated planning of service offerings that continues in London, but it also reflects a much more moderate subsidy-cutting program in London than elsewhere.

The PPPs for the London Underground, forced through by the national government over the strenuous objections of the Mayor of London, consist of three extremely detailed contracts with private consortia of firms to maintain and improve the track, stations, rolling stock, and other infrastructure of specific groups of Underground lines.[16] These contracts were competitively bid, and two of them were awarded to the same consortium; thus, in fact there are two firms with responsibility for the Underground's infrastructure. London Underground, a

[14] Darbéra (2004, Fig. 10). The decline has continued at least through 2005/06, according to UK Department for Transport (2006, Table C).

[15] Darbéra (2004, Fig. 10). Also this increase has continued at least through 2005/06, according to UK Department for Transport (2006, Table C).

[16] See Transport for London (2001) and O'Connor (2002).

public agency, retains responsibility for train operations and fare collection, and is given somewhat circumscribed responsibility to monitor the contracts. The contracts call for frontloading of expenditures by the private firms, for the purpose of speeding up needed upgrades to what everyone agreed was a seriously deteriorated system. These expenditures were covered by private financing arranged by the consortia as part of their bids.

Performance of the two "infrastructure companies," called "infracos," inevitably involved some well publicized problems including a series of braking failures in 2005, which led London Underground to intervene in the responsible infraco's subcontracting arrangements under a safety clause in the PPP. (Financial penalties were subsequently imposed under the PPP's performance-based provisions.) In other aspects, London Underground (2006) reports mostly satisfactory results from the infracos as of March 2006. But only time will tell whether the Mayor's Commissioner for Transport was right in claiming that divided management will undermine the coordination between operations and infrastructure activities needed for some planned major construction projects.

Developing nations

A number of special characteristics of developing nations influence the performance of privately provided transit. First, gaps in managerial capacity in government agencies typically makes it much less likely that regulations will be consistently enforced. This means that a regulated monopoly or a regulated market of private firms may behave quite differently — generally more chaotically — than intended.

Second, the availability of much low-wage labor and the difficulty with which small businesses can raise capital create the possibility of very small companies. Thus many developing cities are characterized by hundreds of separate bus companies — estimated for example at 200 on average for a single minibus route in Lagos, Nigeria (Gwilliam, 2005, p. 7). At the same time, a third trait of developing cities is a high modal share for bus transit — estimated for example at 61 percent for Santiago, Chile, far greater than for any European city (Estache and Gómez-Lobo, 2005, p. 147). These factors combine to create the potential for huge

numbers of individually owned buses competing for passengers in an unregulated or under-regulated environment.

Fourth, transit riders in any poor country are likely to value their time at a far lower monetary amount than in richer countries. This factor tends to lower the optimal frequency, which depends on a tradeoff of value of time against operating costs, some of which are for capital goods and so are less correlated with value of time. As we have already seen, a free-entry equilibrium is likely to result in higher than optimal frequency anyhow, and so this tendency is even stronger in developing countries.

Finally, these tendencies toward oversupply of buses interact with a fifth trait: the prevalence of high levels of congestion, air pollution, and traffic accidents in large developing cities. Bus transit accounts for a high proportion of air emissions in many developing cities — for example nearly one-fourth of fine particulates and more than one-third of nitrogen oxides in Santiago in 2000 (Gwilliam, 2005, p. 16). Thus in such cities the tendency of free markets toward excess supply exacerbates congestion and pollution, in contrast to developed countries where excess transit service would tend to reduce congestion and pollution by diverting some people from car trips (Estache and Gómez-Lobo, 2005, p. 147). As for accidents, transit buses are heavily involved in developing cities and any market structure that encourages "on-the-road competition" for passengers, in the form of drivers racing to the next bus stop in order to collect fares from those passengers, makes the situation even worse.

Experience in two South American cities illustrate how policymakers have tried to alleviate for these problems. In Bogotá, Columbia, the TransMilenio project begun in 2000 establishes a single public company to design the bus network, to oversee tendering of routes to private operators, and to organize a centralized (and separately tendered) fare collection system (Estache and Gómez-Lobo, 2005, pp. 153-155). At the same time, several elements of Bus Rapid Transit were added, including exclusive bus lanes and enclosed bus stops. Operators work on gross-cost contracts so that they have no incentive to compete for passengers. In its first year, very favorable results were reported: a 32 percent reduction in average trip times, reductions in bus-related accidents and injuries by 89 and 74 percent, respectively, and 13 to 54 percent

reductions in air pollutant concentrations. These gains are despite the fact that 85 percent of bus trips are on parts of the system that are not part of TransMilenio.

A broadly similar project, Transantiago, began operating in Santiago, Chile, in 2006.[17] It has many of the same goals as TransMilenio: faster travel times, fewer operators, integrated fares (including transfers to and from the subway system), and elimination of drivers racing for passengers. It follows an earlier system of competitive tendering based on net-cost contracts, begun in 1991, which reportedly reduced the number of buses in central Santiago by 31 percent (Kain and Liu, 2002, p. 159). Transantiago hopes to further reduce bus proliferation and on-the-road competition by changing to gross-cost contracts, with pollution one of the factors in the criteria for awarding tenders. The system aims to require no public subsidies.

6.3.5 Paratransit

Private entrepreneurs and firms, in addition to providing conventional transit service, sometimes fill market niches with other services that, like public transit, involve strangers sharing a vehicle. Examples include subscription commuter buses, semi-scheduled jitney services by vans or minibuses, airport shuttle vans, demand-responsive services activated by telephone or hailing, shared-ride taxi, commuter vanpools, and rental cars. These services, known generally as "paratransit," are usually discouraged by competition from subsidized transit systems and are often strongly inhibited or prohibited outright by regulations. Thus one outcome of deregulation or privatization of transit may be the spontaneous emergence of paratransit.

What market characteristics can we expect of paratransit? Cervero (1997) provides a comprehensive review of experience. Some types, namely subscription vans and airport shuttles, have proven commercially viable in very specialized markets and seem not to provoke a lot of controversy. Others, such as commuter vanpools, are mostly arranged through large employers (often with government pressure to increase employees' average vehicle occupancy) and again thrive in very limited markets. Yet another type, demand-responsive transit or "dial-a-ride," in which specialized vehicles provide shared-ride service with advance reservations, is very

[17] See Gwilliam (2005), pp. 17-18, for a prospective description.

expensive (at least with current dispatching technology) and in the US is almost entirely limited to government-mandated service for elderly and physically handicapped riders.

Here, we concentrate on two other types of paratransit — jitneys and shared-ride taxis — that are most likely to arise spontaneously and that seem capable, in certain circumstances, of carrying substantial market shares of urban trips. Both modes have certain supply characteristics that keep costs down: low overhead expense, small general-purposes vehicles, ability to use part-time labor, and flexibility in adjusting to changing demand conditions.

Jitneys are vans or small buses that follow somewhat regular routes but generally not on a published schedule and often with *ad hoc* route deviations to accommodate passenger needs. From Cervero's observations, it appears that cities with long narrow corridors, limited parking, and a major trip generator (like a rail station or compact business district) offer a favorable environment for jitney service. Jitneys tend to be politically favored in emergency situations such as when a city recovers from hurricane or earthquake damage or during gasoline supply disruptions. They also perform better and improve their image when they regulate themselves concerning safety, driving practices, customer service, and the like through industry associations. However, jitneys are often eliminated by regulations instigated by hostile competitors, including a public transit system, and they are highly vulnerable to targeted competition from a subsidized transit system. Thus, while jitneys have thrived in many US cities for periods of a few years to several decades, they had almost entirely disappeared by the late 1990s.

In many developing nations, by contrast, jitneys are an important component of urban transportation — especially in Latin America, Southeast Asia, and Africa. Mexico City has an extensive system, accounting for one-third of all motorized trips in the metropolitan area in 1994, with government regulation of fares, routes, and certain performance standards (Cervero, 1997, p. 128). In Bangkok, Jakarta, and Manila, jitneys and shared-ride taxis together accounted for 18 to 30 percent of all motor vehicles in the early 1990s (Cervero, p. 134). In Africa, jitneys have mostly replaced conventional bus transit; and similar trends appear to be underway in Eastern Europe and central Asia (Gwilliam, 2003, p. 201).

Shared-ride taxi service means that drivers can combine passengers who are not necessarily traveling together but who have origins and destinations in compatible locations or

directions. The ability to do this is mostly determined by the nature of taxi regulation. Where permitted, such service is encouraged by a zone fare system (used in Washington, D.C.) and by a pay structure giving the driver all the incremental revenue resulting from carrying additional travelers. Like jitney service, it is vulnerable to competition from cheap bus transit and in the absence of self-regulation it can easily get a reputation for poor service.

6.3.6 Conventional Taxi Service

Exclusive-ride taxi is an important but somewhat neglected sector in urban transportation. Taxis handle a large number of passenger trips and provide an alternative to car ownership or rental for short occasional trips, including many by low-income people. Service is heavily regulated in most cities, for reasons that are complex and vary with local conditions.

A number of experiments with deregulation have been undertaken during the past three decades, providing evidence about the nature of the industry and, by comparison, about the effects of the prior regulatory regimes. The loosening of entry and price controls has consistently resulted in significant increases in the number of taxis operating: 18 to 127 percent in seven US cities deregulated in the decade prior to 1985; 15 percent in Sweden's two largest metropolitan areas (deregulation 1991); and approximately 100 percent in New Zealand (deregulated 1989) and in Ireland (2000). Fares, however, have not declined as predicted by many analysts who believed that the industry behaves like a monopoly under regulation (due to regulatory "capture" by industry leaders) and like a perfectly competitive market when unregulated. In fact, fares in cities that deregulated have risen as fast or faster than elsewhere. In New Zealand, deregulated fares in real terms appear to have fallen slightly, but in the US they rose about equally in regulated and deregulated cities, and in Sweden (Stockholm and Göteberg) they rose by around 30 percent.[18] Also in the Netherlands, deregulation in 1999 was in the next four years followed by a 50% increase in the numbers of taxis and firms, an 11% increase in real fares, and a 17% reduction in the number of rides (TNS/NIPO KPMG, 2004).

[18] The figures quoted in this paragraph are from Teal and Berglund (1987), Gärling *et al.* (1995), Morrison (1997), Barrett (2001). For the fare statistics, see especially Morrison (pp. 921-924), Teal and Berglund (Table 3), and Gärling *et al*. (Tables 2-4).

Service quality seems to have improved in some cases in terms of availability (*e.g.* an 18 percent decrease in average access time in Sweden); but on some cases service has also deteriorated in terms of refusals or no-shows. Productivity dramatically declined in several cases as taxi drivers spent more time cruising or simply queuing at cab stands. In the US at least, these results seem due to a deregulated market structure typically characterized as an oligopoly with a competitive fringe, with the fringe adding unnecessary excess service to dense markets (airports and other large pickup spots), and perhaps with less reliable oversight over the reliability of the drivers. There were also increased passenger complaints over drivers who did not know the city and/or who could not speak the local language well. On the whole, it seems safe to conclude that successful deregulation needs to take close account of the fact that even if the market appears to have many firms, the individual "products" being bought and sold depend on matching a particular provider with a particular origin-destination request, and this makes the actual market behavior far from competitive.

Indeed, taxicab service exhibits scale economies analogous to those on scheduled transit service, both in the cruising sector (where cabs are hailed by sight) and in the dispatch sector (where cabs are routed in real time over a large network). This is because the average waiting time for finding a cab declines with the density of available cabs, which in turn rises with demand in the medium term. Probably these scale economies help account for the tendency of the dispatch market, where passengers get to choose the firm they contact, to be dominated by a few large firms. Scale economies by themselves would imply that an unregulated equilibrium will have too little service; but countering this effect are congestion and pollution externalities.

The basic idea that scale economies arise from the dependence of waiting time on number of available vehicles underlies a number of formal models.[19] All such models produce the result that a first-best optimum involves negative profits. Some also consider oligopolies and entry barriers such as may be created by the need to be part of a radio-dispatch service. The models differ on the viability of a second-best optimum (one with profits constrained to be non-negative), the result depending on what degrees of freedom the firms are assumed to have. For

[19] Recent examples include Frankena and Pautler (1986), Häckner and Nyberg (1995), Cairns and Liston-Heyes (1996), and Yang and Wong (1998). Some of these authors cite Orr (1969) as an inspiration.

example, Frankena and Pautler (1986) assume that waiting time depends on the number of taxis, and suggest that the second-best optimum could be obtained just through price regulation; whereas Cairns and Liston-Heyes (1996) assume that each driver can determine hours of operation, arguing that even if price and number of vehicles are regulated at second-best optimum levels, drivers will choose to operate too many hours per day and therefore provide too much service in aggregate. It seems that to make more progress, it is crucial to accurately match modeling assumptions to characteristics of a specific regulatory environment, making it likely that quite different outcomes can be expected depending on fine details of the local situation.

One disappointment from past efforts to deregulate taxi service is that little innovation occurred — nothing like the enormous transformation in strategies that have characterized airlines, trucking, and inter-city railroads, for example. However, the onset of on-board guidance and global positioning systems could dramatically change the potential for larger taxi firms to manage their service quality and dispatching efficiency. Thus, in the future we may see significant changes in the industry when regulatory restrictions allow them, and we will need new modeling to help regulators know which such changes should be encouraged.

6.4 Conclusions

Dissatisfaction with publicly provided transportation infrastructure and services has sparked renewed interest in applying free-market principles to urban transportation. Current research suggests that although the transportation sector is far from meeting the conditions under which unregulated markets are fully efficient, selective use of private enterprise can improve incentives and bring about significant cost savings. We are learning a great deal about the effects of specific regulatory measures on market structure and performance, both from more fine-tuned theoretical models and from careful empirical examination of the many experiments being carried out around the world.