Economics of Transportation 1 (2012) 49–63

Contents lists available at SciVerse ScienceDirect

Economics of Transportation

journal homepage: www.elsevier.com/locate/ecotra

Road pricing and investment

Robin Lindsey*

Sauder School of Business, University of British Columbia, 2053 Main Mall, Vancouver, British Columbia, Canada V6T 1Z2

ARTICLE INFO

Article history:
Received 2 March 2012
Received in revised form
8 July 2012
Accepted 10 July 2012

Keywords:
Congestion pricing
Second-best pricing
Road capacity
Induced demand
Public transportation
Private roads

ABSTRACT

Traffic congestion is a bane of modern city life. Transportation economists have long supported road pricing as a tool for controlling congestion and the idea is slowly coming into practice. This paper reviews the theory of congestion pricing and the relationship between optimal congestion tolls and optimal road capacity. It is organized around four questions. Is congestion pricing according to marginal-social-cost principles consistent with covering the costs of road infrastructure? How does road pricing affect optimal road capacity? How does road pricing affect optimal public transportation fares and capacity? Do private toll road operators make socially efficient toll and capacity decisions? The paper concludes with an assessment of long-run trends in travel demand and technology that could alter the evolution of traffic congestion and priorities for road pricing and investment.

1. Introduction

Traffic congestion is a bane of modern city life. In its latest annual Urban Mobility Report (Schrank et al., 2011), the Texas Transportation Institute estimates that in 2010, congestion in the 439 major urban areas of the US caused approximately 4.8 billion hours of travel delay and 1.9 billion gallons of extra fuel consumption with an estimated total cost of $101 billion. The average cost per automobile commuter was $713, and in six of the largest urban areas it exceeded $1000. Broadly similar estimates are found in other developed countries (VTPI, 2011). These estimates understate the full costs of congestion because they exclude the costs related to pollution, vehicle wear and tear, and retiming trips to avoid peak congestion.

For decades the established approach to controlling congestion was to forecast traffic growth, and then build enough road capacity to accommodate it. This "Predict and provide" strategy was finally abandoned in the UK during the 1990s in the face of evidence that new capacity soon fills up with new traffic (Cervero, 2003). Similar policy shifts have occurred in the U.S. and elsewhere. Other policies to combat congestion have been tried around the world: land-use planning, improvements in traffic management, odd–even license plate restrictions on car use, and so on. But these policies are expensive to implement. They are also blunt instruments for targeting congestion, and to the extent that they make driving more attractive they share the weakness of road building in encouraging more driving.

Cities face other transportation problems as well: accidents, air pollution, and other road-use externalities; rising costs of road construction and maintenance; and shortage of funds for public transit. Economists have long argued that road pricing is the best single tool to address these problems because it serves three functions. First, it controls usage of roads by influencing all dimensions of travel behavior without banning any particular trips. Second, it generates revenues that can be used to fund road investment, maintenance, and operations as well as public transportation and other services. And third, toll revenues provide a signal whether capacity expansion is warranted.

The practice of tolling roads to cover their construction and maintenance costs dates back to Roman times. This funding role remains important today. However, most of the road pricing literature has focused on the demand management role of tolls in controlling congestion. Pigou (1920) planted the idea of treating congestion as an external cost and taxing it like pollution. Tolls are sometimes called taxes although user fees are a more appropriate term because toll payment is directly related to road usage.

Tolling schemes designed specifically for congestion pricing have been slow to develop. The main examples are Singapore's Electronic Road Pricing system, the London congestion charge, the Stockholm cordon charge, and High Occupancy Toll (HOT) lanes in the US. Technological barriers to road pricing have been largely overcome, and collecting tolls is becoming cheaper. The biggest remaining challenge is public opposition.

An economist new to transportation economics may wonder why road pricing has attracted a voluminous literature. Are the

* Tel.: +1 604 822 3323; fax: +1 604 822 9574.
E-mail address: Robin.Lindsey@sauder.ubc.ca

2212-0122/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ecotra.2012.07.001

principles of pricing and investing in roads not adequately covered by general microeconomic and public finance theory? One response to this question is that road pricing and investment affect travel behavior, and travel behavior has many interrelated dimensions that require specialized models to analyze. Another response is that although several other bodies of literature do provide insights, they have features that make them unsuitable for application to roads. Four will be mentioned here:

Peak-load pricing theory (Crew et al., 1995) was developed for public utilities such as electricity that are *loss systems*. Customers cannot be queued for service in a loss system, and each is either served immediately or not at all (e.g., during an electricity blackout). By contrast, roads are *delay systems* in which vehicles can be stored or queued up on road links, at intersections, and in parking lots. The economics of optimal pricing and capacity decisions differ for loss systems and delay systems.

Club theory (Scotchmer, 2002) has some common elements with roads. Clubs are congestion-prone, and access can be rationed by membership fees and user charges. Club theory is also concerned with optimal club size and whether members are efficiently allocated between competing clubs. However, clubs lack a network dimension comparable to roads. Time-of-use decisions are also rarely addressed, and the more recent clubs literature has focused on diversity of preferences and interactions between club members which are of limited relevance for roads.

Queuing theory was largely developed in operations research. While the theory is mainly concerned with delay systems, the M/M/1 and other workhorse queuing models assume that customer arrival times and service times are random. Such randomness is not an important characteristic of congested roads for which service times are relatively uniform (typical time headways between vehicles are 1.5–2 s). More importantly, most of the queuing literature focuses on steady-state conditions whereas road usage tends to be highly peaked and changes too rapidly for steady-state equations to be applicable (Hurdle, 1991).[1]

Telecommunications and computer networks are similar to road networks in that they sometimes operate as delay systems. However, as in queuing theory, user arrival rates and individual usage durations are highly variable. Moreover, much of the equilibrium analysis used in the telecommunications and computer network literature has been adopted, or adapted, from transportation (Altman et al., 2006), and in that respect has few lessons to offer for road applications.

The literature on road pricing is large and growing rapidly, and it is impossible to give a comprehensive review here.[2] Instead, this paper focuses on road pricing and road capacity investments for the purpose of relieving congestion. The literature on this subtopic is mainly theoretical and relatively small. Attention is mainly focused on tolling of road links because most studies of road pricing and investment have treated this form of road pricing.[3] Coverage is also limited to deterministic models of congestion and congestion pricing. Although demand and capacity uncertainty are practically important for roads, and a literature on dynamic (i.e., state-dependent) road pricing is developing, the implications of uncertainty for investment in roads have not been explored very far. Public acceptability and equity dimensions of road pricing are omitted as well.

A number of questions about road pricing and road investment have been addressed in the literature as well as debated by policy makers, politicians, and the public. The review is organized around the following four questions:

1. Is marginal-cost pricing consistent with cost recovery?The demand management role of road pricing calls for setting tolls at short-run marginal cost in order to support efficient usage. Depending on existing road capacity, short-run marginal cost pricing can yield surpluses or deficits, and resultant cross-subsidies between transportation user groups or between transportation and other sectors of the economy. This is inconsistent with cost recovery and the user pay principle. Under what conditions are demand management and cost recovery consistent if both tolls and capacities can be freely chosen? How likely are these conditions to be met?
2. How does road pricing affect optimal road capacity?An important question for long-run transportation planning is whether more or less road capacity should be built if road pricing is introduced. Investment is usually considered less valuable if tolling is implemented since, by curbing traffic, tolls reduce the number of trips that can benefit from higher capacity. What this argument overlooks is that without tolls, adding capacity induces more travel that has a private value less than its social costs so that some of the potential benefit from investment is "wasted". The answer to question 2 turns on the strength of this induced travel demand.
3. How does road pricing affect optimal public transportation fares and capacity?Public transit service is heavily subsidized around the world. One reason is that low fares encourage people to take transit rather than drive, which helps to alleviate traffic congestion and other externalities. This rationale falls away if road pricing is introduced. Thus, one might expect fares to increase and optimal transit capacity adjustments to be determined along similar lines as for roads as discussed in question 2. But there is an additional force at play because the introduction of tolls increases transit demand directly. The net effect of these various forces is unclear *a priori*.
4. Do private toll road operators make socially efficient toll and capacity decisions?Many intercity highways and some urban roads have been designed, financed, built, operated, maintained and/or tolled by the private sector. Private entities can be more cost efficient than public operators, and they can accelerate construction of new roads by contributing to financing. But the private sector has an incentive to exercise market power by setting high tolls and underinvesting (or possibly overinvesting) in capacity. How serious are these distortions, and how do they depend on the way in which control over road network links is assigned to private operators?

The paper is organized as follows. Section 2 describes the principles of road congestion pricing, beginning with first-best pricing and then delving into the messier but more relevant theory of second-best pricing. The principles described in this section are of interest not only as they relate to optimal road capacity but also by themselves since efficient usage of roads is a goal whether capacity is optimal or not. Section 3 considers road investment and the interplay between pricing and optimal capacity. It addresses the first three questions posed above. Section 4 reviews the economics of private toll roads and addresses the fourth question. Section 5 summarizes the main conclusions, and then considers long-run trends in travel demand and technology that could alter the evolution of traffic congestion and priorities for road pricing and investment.

[1] Deterministic queuing models with transient queues are better suited to road applications. The Vickrey (1969) bottleneck model, mentioned later, is a prominent example.

[2] Yang and Huang (2005) provide a detailed mathematical review of road pricing theory. Recent surveys of road pricing are found in Parry (2009), Anas and Lindsey (2011), and Santos and Verhoef (2011).

[3] Other forms of road pricing are toll rings, toll cordons, zonal schemes, and distance-based schemes. Road pricing methods and technologies are reviewed in Tsekeris and Voß (2009) and de Palma and Lindsey (2011).

2. Road pricing

2.1. First-best pricing

2.1.1. The basic model[4]

The economic principles of road congestion pricing were developed by Walters (1961) using a static, partial-equilibrium, supply-demand approach. It ignores externalities other than congestion, distortions in other markets, and toll collection costs. Individuals (hereafter "users") travel one per vehicle from a common origin to a common destination along a single road. Users are treated as a continuum rather than discrete entities, and they are identical except for the monetary value they place on a trip. The inverse demand curve for trips is $d(V)$ where V (volume) is the number of trips made per unit time. The private cost of a trip, $c(V)$, includes fuel consumption and other vehicle operating costs, tolls if any, and the opportunity cost of time. If the road is congested, travel time increases with volume and $c(V)$ is upward-sloping as shown in Fig. 1. Time cost is assumed to be linear in travel time and is written $c(V)=wT(V)$ where $T(V)$ is travel time and w is the opportunity cost or "value" of time.[5]

The total cost of V trips is $TC=c(V)V$, and the marginal social cost (MC) of a trip is

$$MC = dTC/dV = c(V)+c'(V)V, \tag{1}$$

where $c'(V)\equiv dc(V)/dV \geq 0$. If there is no toll the equilibrium number of trips, V_E, is defined by the condition that marginal willingness to pay equals private cost:

$$d(V_E) = c(V_E) \tag{2}$$

Equilibrium occurs at point E in Fig. 1. In contrast, the socially optimal number of trips, V_O, is defined by the condition that marginal willingness to pay equals MC:

$$d(V_O) = c(V_O)+c'(V_O)V_O, \tag{3}$$

which occurs at point O.

The social or first-best optimum can be supported by imposing a congestion toll equal to the gap between the demand and MC curves at volume V_O:

$$\tau = MC(V_O)-c(V_O) = c'(V_O)V_O \tag{4}$$

The toll in Eq. (4) is often called the Pigouvian toll in honor of Pigou (1920). It equals the marginal external cost imposed by an additional trip on each user, $c'(V_O)$, multiplied by the number of users affected, V_O. Toll revenue is $R=\tau V_O = c'(V_O)V_O^2$ which corresponds to the shaded rectangular area in Fig. 1. The efficiency gain from the toll (the increase in social surplus) equals the shaded area EOA. Inclusive of the toll, users incur a trip cost equal to $p=\tau+c(V_O)$ where p is the *full price* or *generalized cost* of a trip. The user cost is a real, social cost (sometimes called a "resource" cost) while the toll is a transfer with a zero net social cost.

The Pigouvian toll formula is simple and appealing, and it forms the basis of arguments for congestion pricing that transportation economists have been making for decades. However, even in the simple Walters' model some practical challenges are apparent. One is that the toll must be evaluated at traffic volume V_O rather than V_E which is observed in the pre-toll equilibrium. To deduce V_O it is necessary to estimate both the demand curve and the cost curve.[6] Second, toll collection is costly. If demand is relatively inelastic, the efficiency gain given by area EOA is a small fraction of the toll revenues so that tolling is not worthwhile unless collection costs per dollar are small. Third, while tolling reduces congestion delay it does not eliminate it because the travel cost at point B in Fig. 1 typically exceeds the cost under free-flow conditions, $c(0)$. Tolling therefore cannot be touted as a "cure" for congestion (although it may support near-free-flow conditions if the speed–flow curve is relatively flat for flows up to V_o). More worrying, tolling increases users' private costs by p_O-p_E. Users are thus likely to oppose tolling unless at least some of the toll revenues are used in ways that benefit them. As discussed in Section 3, toll revenues are often earmarked for this reason.

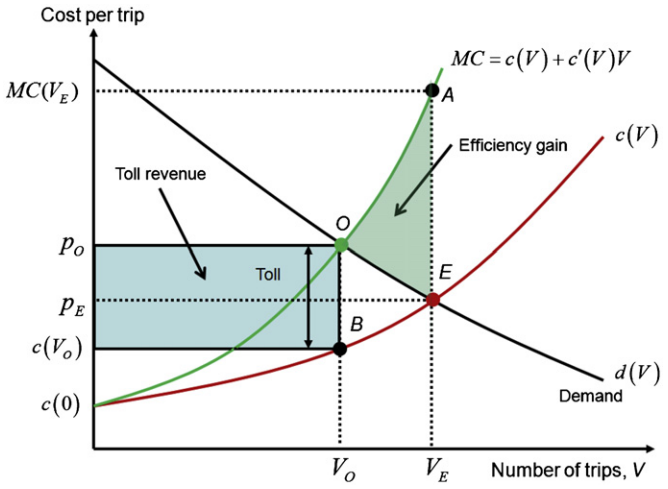

Fig. 1. Equilibrium and social optimum, congestion toll, toll revenue and efficiency gain.
Source: Author's construction.

Walters' model has been extended in various directions. Attention is limited here to dynamics and user heterogeneity. The important network dimension is introduced with second-best pricing in Section 2.2.

2.1.2. Dynamics

Walters' model is static. Yet congestion is inherently dynamic because travel demand varies strongly by time of day, day of week, and season whereas capacity is fixed. Both traffic engineering and dynamic travel demand models are necessary to describe the dynamics of congestion. Engineering models describe how given traffic flows evolve over time and space. Foremost is the hydrodynamic model in which vehicles are treated as a compressible fluid.[7] Dynamic demand models describe travelers' trip-timing preferences. The most popular model is due to Vickrey (1969) who assumed that users prefer to reach their destinations at a particular time, and incur *schedule delay costs* if they arrive earlier or later. During periods of high demand they face a choice between suffering long delays in order to arrive on time, and avoiding delays at the cost of traveling inconveniently early or late. Vickrey combined this demand-side specification with a simple supply side in which congestion delay takes the form of deterministic queuing at a bottleneck. Travel through the bottleneck is unimpeded if the arrival rate does not exceed capacity;

[4] The exposition and notation used here broadly follow Small and Verhoef (2007, Section 4.1).

[5] Small and Verhoef (2007, Section 2.6) and Hensher (2011) review the theory underlying the value of time as well as empirical estimates of it. Function $T(V)$ can be derived from an engineering speed–volume curve. Various functional forms are used. A common choice is the power function $T(V)=a+bV^c$ with V measured in vehicles per lane per hour. Estimated values of parameter c range from 2.5 to 10.

[6] In practice the toll could be determined iteratively by trial and error (Yang et al., 2004) which would bypass the need to estimate the demand curve. However, repeatedly changing the toll in a tâtonnement-like manner might upset users.

[7] Daganzo (1997) provides a detailed review of the hydrodynamic model.

otherwise a queue forms and the bottleneck operates at capacity until the queue drops back to zero.[8]

Three types of tolls have been analyzed using Vickrey's bottleneck model: flat tolls that do not change over time, step tolls that change in discrete increments and decrements, and a "fine" toll that varies continuously in order to eliminate queuing. An important conclusion from this analysis (Arnott et al., 1993) is that time-varying tolls yield much higher welfare gains than do flat tolls because they induce changes in departure time that flatten the peak and reduce the resource costs incurred for a given number of trips.

2.1.3. User heterogeneity

Users differ in their travel preferences as well as in the characteristics of their vehicles. This raises the question whether tolls need to be differentiated to support a first-best optimum, or whether they can be anonymous (i.e. varied by location and time, but not by user or vehicle characteristics). Arnott and Kraus (1998) show that anonymous tolling suffices as long as users impose the same external congestion costs. If so, the optimal toll given by Eq. (4) is replaced by $\tau = \sum_i c_i'(V_{iO})V_{iO}$ where V_{iO} is the optimal flow for user type i, and $c_i(\,)$ is the user cost function for type i. All users pay the same toll, but they differ in how much they gain from lower trip costs. Higher-income users tend to gain more because of their higher values of travel time.

Charging drivers different tolls is necessary even for first-best pricing if they differ in their driving behavior. In principle, differentiating tolls by age and sex (as is done for insurance premiums) as well as speed and driving style is feasible although it may be opposed on privacy and discrimination grounds. Vehicles differ in size, maneuverability, and visual intrusion as well as characteristics such as weight, engine size, and fuel type that affect other road-usage externalities. Traffic engineers account for vehicle heterogeneity using Passenger Car Equivalents (PCEs). If a truck has a PCE of 2, the Pigouvian toll for a truck is twice the toll for a car. Toll differentiation on the basis of PCEs is feasible as long as the relevant vehicle characteristics are observable or can be recorded on vehicle transponders.

2.2. Second-best pricing

The models described so far are based on first-best conditions that are invariably violated in practice. This section considers a few major distortions that are empirically important for road pricing and investment decisions.[9]

2.2.1. Road networks

The formulation of second-best toll policies on road networks is complex and notationally burdensome in its full generality. To keep the presentation manageable attention is limited to the static model with homogeneous users. The road network is represented by a directed graph comprising a set of nodes $\boldsymbol{N}$, and a set of links $\boldsymbol{L}$. There is a set $\boldsymbol{M}$ of "markets" for travel between some (possibly all) pairs of nodes. Let N_m denote the number of trips made in market m, and $d_m(N_m)$ the inverse demand curve for trips in market m.[10] Market m is connected by a set of routes $\boldsymbol{R}_m$. The correspondence between links and routes is described by indicator variables δ_{lr}, where $\delta_{lr}=1$ if route r uses link l, and $\delta_{lr}=0$ otherwise. Similarly, the correspondence between routes and markets is described by indicator variables δ_{rm}, where $\delta_{rm}=1$ if $r \in \boldsymbol{R}_m$, and $\delta_{rm}=0$ otherwise.

Let v_l denote the traffic flow on link l, and V_r the traffic flow on route r. The v_l, V_r, and N_m variables are related by the accounting identities

$$v_l = \sum_r \delta_{lr} V_r, \quad l \in \boldsymbol{L}, \tag{5}$$

$$N_m = \sum_r \delta_{rm} V_r, \quad m \in \boldsymbol{M} \tag{6}$$

User cost on link l, $c_l(v_l)$, is assumed to depend only on the flow on link l and not on flows on other links.[11] Finally, the toll (if any) on link l is τ_l.[12] User equilibrium on the network is described by Wardrop's First Principle (Wardrop, 1952):

For any market, all routes that are used have equal generalized costs, and all unused routes have higher (or possibly equal) generalized costs.

Application of Wardrop's First Principle results in two sets of equilibrium conditions:

$$\sum_l \delta_{lr}(c_l(v_l)+\tau_l) \geq d_m(N_m), \quad \text{for } r \in \boldsymbol{R}_m,\ m \in \boldsymbol{M} \tag{7}$$

$$\left(\sum_l \delta_{lr}(c_l(v_l)+\tau_l) - d_m(N_m)\right) V_r = 0 \quad \text{for } r \in \boldsymbol{R}_m,\ m \in \boldsymbol{M} \tag{8}$$

Condition (7) states that willingness to pay for a trip in any market cannot exceed the generalized cost incurred on any route. Condition (8) states that willingness to pay must equal generalized cost on a route that is used (i.e., with $V_r > 0$).[13] Both sets of conditions are intuitive. If Condition (7) was violated on route r serving market m, more travelers in market m would take route r. If Condition (8) was violated on route r, the number of trips in market m using route r would adjust up or down.

The social optimum is derived by maximizing the sum of net consumers' surplus in all markets:

$$B = \sum_m \int_{n=0}^{N_m} d_m(n)dn - \sum_l c_l(v_l)v_l \tag{9}$$

subject to accounting identities (5) and (6), equilibrium conditions (7) and (8), and non-negativity constraints on link flows and route flows. The solution depends on which links can be tolled and whether the levels of the tolls are constrained. If all links can be tolled freely a first-best optimum can be realized. Yang and Huang (1998) show that the analog of the Pigouvian toll in Eq. (4) is levied on each link:

$$\tau_l = c_l'(v_l)v_l, \quad l \in \boldsymbol{L} \tag{10}$$

Although varying the toll on one link affects flows on other links, the first-best toll is imposed as if other links do not exist. The reason for this is that if all links are efficiently priced, marginal changes in flows on other links are welfare-neutral and the envelope theorem applies.

In contrast to Eq. (10), formulas for second-best tolls are complicated and opaque, and solutions are difficult to compute numerically. Useful insights can nevertheless be drawn from simple networks. The simplest example is the two-links-in-parallel network considered by Pigou (1920) and many others

[8] Small and Verhoef (2007, §4.1.2) and de Palma and Fosgerau (2011) review Vickrey's bottleneck model and some of its extensions.

[9] More extensive reviews of second-best pricing are found in Small and Verhoef (2007, Section 4.2).

[10] Demands are assumed to be independent across markets. This assumption can be relaxed without changing the basic results.

[11] Flow on one link can affect costs on other links because of intersection delays and delays in passing created by opposing traffic on non-separated highways.

[12] In principle, tolls can be imposed on routes rather than links. However, route-based tolling requires tracking of vehicle paths and raises privacy concerns.

[13] The Wardrop equilibrium coincides with the Nash equilibrium if there is a continuum of users. In an atomic game with discrete users (e.g., as in laboratory experiments) a Nash equilibrium can exist without the costs of all used routes being equal so that the Wardrop equilibrium conditions are violated. Haurie and Marcotte (1985) describe the relationship between Nash and Wardrop equilibria.

since then. Each link serves as one route. Suppose the toll on link 2 is fixed at $\tau_2 \geq 0$. The second-best toll on link 1 works out to[14]

$$\tau_1 = c_1' v_1 + \frac{|d'|}{c_2' + |d'|}(\tau_2 - c_2' v_2) \tag{11}$$

where $d' < 0$ is the slope of the inverse demand curve. The second-best toll on link 1 equals the first-best toll, $c_1' v_1$, plus an adjustment factor that accounts for how τ_1 affects the net congestion externality on link 2. Eq. (11) reduces to the first-best toll if link 2 is efficiently priced. Otherwise, the toll on link 1 differs from the first-best toll. If link 2 is not tolled, τ_1 is set below the first-best level, and if the toll on link 2 is too high, τ_1 is set above it. Deviating from the first-best toll is optimal because small deviations cause only second-order losses of efficiency on link 1, but first-order gains on link 2.

Eq. (11) simplifies in two limiting cases. If demand is perfectly inelastic ($d' = -\infty$) then $\tau_1 = c_1' v_1 - (c_2' v_2 - \tau_2)$. The toll equals the difference between the externality on link 1 and the net externality on link 2. With total demand fixed, the only goal is to support an efficient division of traffic between links. One instrument (τ_1) suffices to achieve this goal and the first-best optimum can be attained. In the second limiting case of perfectly elastic demand ($d' = 0$), the adjustment factor in Eq. (11) vanishes and the second-best toll matches the first-best toll. The reason for this is that with perfectly elastic demand the generalized cost of travel on each link is fixed at the reservation price for travel. The number of trips on link 2 is therefore a given, independent of τ_1, and the network effectively consists of two isolated, unconnected links.

The two-links-in-parallel network example shows that second-best tolls are determined not only by travel conditions on the tolled link themselves, but also by travel conditions on other links as well as demand functions. It also shows how the potential benefits from tolling are weakened if drivers can escape payment by rerouting to untolled alternatives so that congestion (and other externalities) get displaced rather than reduced.

High Occupancy Toll (HOT) lanes in the US resemble the two-links-in-parallel network example. Drivers can choose whether to pay a toll to use a HOT lane, or take the adjacent (but generally slower) toll-free lanes. HOT lanes differ from the two-link example in that vehicles meeting a minimum occupancy requirement (usually two or three people) can also use the HOT lane without paying. Some HOT lane facilities have been converted from High Occupancy Vehicle (HOV) lanes which are often so underutilized that the benefits to users are outweighed by increased delays for other drivers squeezed onto the remaining lanes. HOT lane facilities achieve a better balance of traffic by adjusting the toll to regulate the volume of drivers who choose to pay to use the HOT lanes.

Another simple network case is a two-links-in-series network with link 1 upstream of link 2. Suppose the toll on link 2 is fixed at $\tau_2 \geq 0$. Because links 1 and 2 are perfect complements it is possible to support a first-best optimum by setting the toll on link 1 equal to the sum of the externality on link 1 and the net externality on link 2: $\tau_1 = c_1' v_1 + c_2' v_2 - \tau_2$. Links 1 and 2 are effectively one link, and inability to toll one of them freely does not impair efficiency. This would not be the case if trips were made to or from the node joining the two links because it would then not be possible to face users in each market with their marginal social costs.

A third network of interest is one with two links in parallel joining the origin to an intermediate node, and a third link connecting the intermediate node to the destination. Traffic flow on this network is described by the number of trips taken, and the division of trips between the two routes. Since two instruments suffice to control these two variables, one of the three links can be left untolled. Similar to the second example this shows that constraints on tolling are sometimes inconsequential.[15]

Real-world urban networks have many links and, unless some form of comprehensive distance-based pricing is introduced, it is practical to toll only a small fraction of them. In cities with heavily congested central cores area-based schemes are logical candidates. Such schemes can take two forms. One is a cordon in which a toll is paid for crossing the perimeter of a charge area either inwards, outwards, or in both directions. A cordon can be implemented by tolling the inbound and/or outbound links that form the perimeter. The other type of scheme is a charging zone for which a toll is paid not only for crossing the perimeter but also for traveling anywhere within the zone. The Stockholm congestion charge and the Milan EcoPass (which was introduced to reduce pollution) are cordons, whereas the London congestion charge is a zonal scheme. Area-based schemes are generally viewed as user-friendly due to their simplicity and clarity, but they have some drawbacks (May et al., 2002): (i) they fail to intercept journeys made wholly outside the charge area (and also wholly inside in the case of cordons), (ii) they induce rerouting of some journeys to escape payment, and (iii) they impose a single charge regardless of distance traveled. These drawbacks can be alleviated by creating multiple cordons or zones, but at the expense of extra cost and complexity for users. Another challenge is that in cities without natural boundaries it is necessary to choose the location of the charged area. This is a computationally difficult problem, but also important to solve accurately since the efficiency of such schemes is sensitive to their design (May et al., 2008).

2.2.2. Public transportation

The model in Section 2.2.1 can be extended to include public transportation with some modifications that account for the characteristics of public transportation trips. First, trips by car and transit are imperfect substitutes and Wardrop's Principle does not apply. Second, in addition to the fare and in-vehicle travel time the generalized cost of public transportation includes access time to and from bus stops and rail stations, waiting time, transfer time, crowding on platforms and in transit vehicles, and the schedule delay due to timetable-imposed constraints on departure and arrival times. Because public transportation service exhibits scale economies from service frequency, spacing of stops, and density of routes, the marginal social cost of a passenger trip (inclusive of system operating costs and user costs) is less than the average cost. In effect, new transit users create a positive externality on existing users once transit supply adjusts to the greater demand. Third, buses have a higher PCE than cars, but a passenger on a fully loaded bus has a much smaller congestion footprint than a person in a car. The congestion delay experienced by buses depends on whether they have a separate right of way or share the road with cars. In the latter case, tolling cars benefits bus travel by reducing both operator and passenger costs.

Given a multimodal transportation network, second-best prices can be derived for both road links and public transit links or routes. Efficiency of equilibrium is defined by three criteria: the total number of trips made in each market, link flows on each

[14] To economize on notation, the dependence of costs on flows is omitted. Verhoef et al. (1996) and Small and Verhoef (2007, §4.2.1) consider the special case where $\tau_2 = 0$.

[15] The three-link network is studied by Verhoef and Small (2004) using a model in which users differ in their values of time. They find that tolling just the downstream link yields most of the welfare gains obtained from tolling all three links.

transportation mode network, and the division of trips between modes or "modal split". Although driving and taking transit can be complements (e.g., for Park and Ride trips), in most cases they are substitutes. Hence if car travel is underpriced, fares should be set below marginal social cost. This will depress fares further below average costs and exacerbate deficits. Correspondingly, if transit fares are overpriced—perhaps because of cost recovery constraints—second-best tolls for cars exceed first-best levels to alleviate the distortion from overpricing transit. The implications for optimal transit capacity—the main element of the third question posed in the introduction—are less clear-cut, and are addressed in Section 3.

2.2.3. Labor market distortions[16]

Commuters account for a large fraction of peak-period trips when congestion is severe and Pigouvian tolls correspondingly high. Since work trips are complementary to labor supply which is heavily taxed in many countries, the second-best toll for commuters is below the first-best toll. Parry and Bento (2001) show that levying first-best tolls can be welfare-reducing if the revenues are distributed in lump-sum fashion back to drivers. However, if second-best tolls are applied and the revenues are used to reduce labor taxes in a revenue-neutral way, the welfare gains are larger than if there were no labor market distortions.

Van Dender (2003) generalizes the Parry–Bento model by treating work trips and other trips separately, and compares the welfare gains from a uniform toll with a toll that discriminates between the two trip purposes. De Borger (2009) uses a wage bargaining model to show how congestion charges can reduce employment levels by increasing union wage demands. These and other papers demonstrate the importance of extending consideration beyond transportation markets when formulating road pricing policies.

2.2.4. Revenue generation

Levying tolls to generate revenues is a priority for governments if the marginal cost of public funds from other revenue sources is high. It is also necessary to meet a self-financing constraint if other tolls such as vehicle registration fees are unavailable. In either case the second-best toll accounting for revenue needs is a weighted average of the Pigouvian toll and the revenue-maximizing toll where the weights depend on demand elasticities (Oum and Tretheway, 1988). The limiting case where the weight on revenue is one will be considered in Section 4 on private roads.

Public acceptability and equity concerns generally militate against high tolls. For this reason, governments may be interested in congestion pricing while attempting to minimize toll revenues as a secondary objective. There is a literature on *minimum-revenue tolling* (Bergendorff et al., 1996) which seeks to identify system-optimal tolls that generate the least amount of revenue. As shown by the examples in Section 2.2.1, it may be possible to support a first-best optimum without tolling every link.

2.2.5. Complexity of tolling schemes

The efficiency of a road pricing scheme depends, of course, on whether travelers know how much they will pay in tolls for a given trip as well as on how much effort they expend to obtain information about tolls and to process it. System complexity is a function of how much tolls vary by time of day and day of week; on how many links are tolled; and on characteristics of the payment system such as quantity discounts, minimum or maximum payments per day or within the accounting period, and how the toll varies with payment method (e.g., cash, transponder recorded, video recorded). If travelers are misinformed about tolls, they are prone to making mistakes that leave them worse off and reduce system efficiency as well. Drivers may oppose complex systems unless they can be justified by the objectives of the scheme (Bonsall et al., 2007).

Keeping tracking of tolls is all the more demanding for motorists if tolls are adjusted in real-time in response to accidents, bad weather, transit strikes, special events, and other shocks that affect road capacity and travel demand. State-dependent or responsive pricing has been implemented on some HOT lane facilities where tolls are adjusted every few minutes to maximize utilization of the toll lanes while maintaining high speeds. This goal is readily understood by motorists, and it is easy for them to avoid the tolls at the last minute by taking the toll-free lanes that run in parallel on the same corridor. But responsive tolling is not as well suited to area-based schemes because avoiding the toll is more difficult or impossible, and also because the optimal responsive toll may vary from entry point to entry point. A challenge for responsive pricing generally is to inform travelers of tolls sufficiently far in advance that they can modify their travel decisions at an opportune time. Smart phones and other portable devices that display real-time tolls may be useful in this regard. Travelers can also program in-vehicle navigational units to guide them onto routes with the lowest expected generalized cost inclusive of tolls.

3. Investment

There are three types of road infrastructure investment. One entails building new roads in order to enable driving in previously inaccessible regions. This is still a priority in some lesser-developed countries, but it is of limited relevance in large cities of the developed world. A second type of investment is intended to facilitate safe, high-speed travel by reducing grades, improving sight lines, upgrading signalized intersections, and so on. Finally, roads can be expanded to alleviate congestion. This is the highest priority in most large cities and it ties in with congestion pricing in Section 2. Attention is therefore focused here on capacity expansion for congestion relief.

To analyze investment decisions it is necessary to add capacity to the model. Let K_l denote capacity on link l, and $c_l(v_l,K_l)$ the user cost function on link l, where $\partial c_l/\partial v_l \geq 0$, $\partial c_l/\partial K_l \leq 0$, and $\partial^2 c_l/\partial v_l \partial K_l \leq 0$. The last assumption assures that capacity expansion has a greater impact on reducing user cost when traffic volumes are high. The annualized cost of capacity is assumed to be a strictly increasing function $F_l(K_l)$.[17] Welfare is given by consumers' surplus in Eq. (9) less capacity costs:

$$B = \sum_m \int_{n=0}^{N_m} d_m(n)dn - \sum_l c_l(v_l,K_l)v_l - \sum_l F_l(K_l) \tag{12}$$

3.1. First-best optimum[18]

The first-best optimum is derived by maximizing (12) with respect to τ_l and K_l, $l \in L$, subject to the constraints identified in Section 2. Optimal tolls are still given by Eq. (10):

$$\tau_l = \frac{\partial c_l(v_l,K_l)}{\partial v_l} v_l, \quad l \in \boldsymbol{L} \tag{13}$$

[16] See Small and Verhoef (2007, Section 4.2.5).

[17] Maintenance costs that depend on v_l and K_l can be added without affecting results of interest.

[18] See Small and Verhoef (2007, Section 5.1).

Given $\partial^2 c_l/\partial v_l \partial K_l \leq 0$, τ_l is a decreasing function of K_l for any given level of usage. By the envelope theorem, optimal capacity can be derived by differentiating (12) with respect to K_l holding traffic volumes fixed. The first-order condition is:

$$\underbrace{-\frac{\partial c_l(v_l,K_l)}{\partial K_l} v_l}_{\text{(a) User benefit}} - \underbrace{\frac{\partial F_l(K_l)}{\partial K_l}}_{\text{(b) Capacity cost}} = 0, \quad l \in \boldsymbol{L} \tag{14}$$

Term (a) in Eq. (14) is the marginal reduction in user costs. Term (b) is the marginal increase in capacity cost. The two quantities are equal at the optimum. As shown in Section 3.3, under second-best conditions such that Eq. (13) does not hold, Eq. (14) does not hold either because the envelope theorem does not apply.

3.2. Self-financing: The cost recovery theorem

Supporting efficient road usage calls for marginal-cost pricing. Paying for roads requires pricing at average cost. The potential inconsistency between these two goals of pricing troubled economists for a long time, and it continues to be raised by opponents of road pricing. Yet, in their celebrated Cost Recovery Theorem (CRT), Mohring and Harwitz (1962) showed that under certain assumptions the two goals are consistent. The CRT was extended to road networks by Yang and Meng (2002) and their version of the theorem is presented here. Let $\varepsilon_l \equiv (\partial F_l/\partial K_l)(K_l/F_l)$ denote the (local) elasticity of capacity cost, R_l the toll revenue on link l, and $\rho_l \equiv R_l/F_l(K_l)$ the fraction of costs that are paid for by tolls or the cost recovery ratio.

3.2.1. The cost recovery theorem for a network:

Assume: (i) The user cost functions, $c_l(v_l,K_l)$, $l \in L$, are homogeneous of degree zero, (ii) $\varepsilon_l = 1$, $l \in L$, and (iii) capacity on each link is perfectly divisible. Then, at the optimum, the toll on each link just pays for capacity on the link: $\rho_l = 1$, $l \in L$.[19]

Empirical evidence on the three assumptions underlying the CRT is limited and somewhat equivocal: (i) Homogeneity degree zero of the user cost function implies that there are constant returns to scale in road use and the user cost functions can be rewritten in the form $c_l(v_l/K_l)$. Increasing usage and capacity in equal proportions thus leaves user cost unchanged. Put another way, for a given traffic flow per lane, user costs are independent of the number of lanes. In practice, if a road is expanded from two lanes to four lanes, capacity more than doubles at low traffic volumes because passing is easier. However, passing is usually difficult under congested conditions for which the CRT is of interest. Moreover, approximately constant returns prevail for road with more than four lanes. (ii) Road infrastructure costs exhibit mild scale economies for individual road links, but these economies are at least partly offset in large cities by diseconomies from intersections and rising land prices. (iii) Capacity is clearly indivisible insofar as traffic lanes must be large enough to accommodate vehicles. However, road capacity can be adjusted by altering lane widths, horizontal and vertical alignments, grades, quality of pavement surface, lighting, and other design features.

Overall, Assumptions (i)–(iii) are likely to be satisfied approximately in heavily traveled corridors where many traffic lanes are built. Moreover, surpluses and deficits on individual links may average out so that total costs are approximately recovered on large networks.

When the assumptions of the CRT are satisfied, but capacity is not optimal, the cost recovery ratio provides a signal how capacity should be adjusted. If $\rho_l > 1$ (i.e., there is a surplus), capacity should be expanded. If $\rho_l < 1$ (i.e., there is a deficit), capacity is excessive and should not be replaced when it wears out. And if $\rho_l = 1$, capacity is optimal.[20] As Arnott and Kraus (2003) explain, the relationship between cost recovery and optimal capacity adjustment closely parallels the relationship between profits and entry/exit of firms in a perfectly competitive industry.

The Cost Recovery Theorem has been extended in various directions and found to be fairly robust.[21] It continues to hold: in dynamic models, with heterogeneous users, with other road usage externalities, with demand and capacity uncertainty, and with constraints on the structure of tolls. The main requirements are that assumptions (i)–(iii) continue to hold and tolls be sufficiently flexible to price usage at marginal cost.

3.3. Second-best capacity

This subsection addresses the second question posed in the introduction: how does road pricing affect optimal road capacity? Similar to second-best pricing, second-best investment rules are complex and general results are elusive. For this reason, attention will be mainly focused on a single road link.

3.3.1. One road link with non-optimal pricing[22]

Assume that the toll on link l is fixed at some level $\tau_l \geq 0$. The first-order condition for optimal capacity works out to:

$$\underbrace{-\frac{\partial c_l(v_l,K_l)}{\partial K_l} v_l}_{\text{(a) User benefit}} - \underbrace{\frac{\partial F_l(K_l)}{\partial K_l}}_{\text{(b) Capacity cost}} + \underbrace{\left(\tau_l - \frac{\partial c_l(v_l,K_l)}{\partial v_l} v_l\right)\frac{dv_l}{dK_l}}_{\text{(c) Induced demand}} = 0. \tag{15}$$

Terms (a) and (b) in Eq. (15) match the first-best capacity rule in Eq. (14). Term (c) is an *induced-demand effect* that arises because expanding capacity attracts additional flow. Suppose τ_l is set below the first-best toll while capacity is held fixed at its first-best level. Term (b) does not change. But usage increases so that term (a) is larger than in the first-best solution. This *usage effect* underlies the conventional wisdom that optimal capacity is larger when usage is underpriced. However, given $dv_l/dK_l > 0$ term (c) is negative so that the induced-demand effect works in the opposite direction to the usage effect.

Without specific assumptions about the user cost and demand functions it is not possible to determine in general whether the usage effect or the induced-demand effect dominates, and consequently whether second-best capacity is larger or smaller than first-best capacity. Two useful results have nevertheless been derived. First, Wheaton (1978) shows that if the toll is only slightly below the first-best toll, the usage effect outweighs the induced-demand effect so that second-best capacity exceeds first-best capacity. This is because the benefits from additional capacity to existing users (term (a) in Eq. (15)) increases by a

[19] A more general version of the CRT states that $\rho_l = \varepsilon_l$, $l \in \boldsymbol{L}$, without requiring $\varepsilon_l = 1$.

[20] Proof: Since $c_l(v_l,K_l)$ is homogeneous of degree zero, its first derivatives are homogeneous of degree-1. The toll in Eq. (13) can thus be written $\tau_l = c_l'(v_l/K_l)(v_l/K_l)$. Toll revenue is $R_l = \tau_l v_l = c_l'(v_l/K_l)(v_l/K_l)^2 K_l$, and $\rho_l = c_l'(v_l/K_l)(v_l/K_l)^2 K_l/F(K_l)$. With constant scale economies, $F(K_l) = kK_l$ for some constant k, and $\rho_l = c_l'(v_l/K_l)(v_l/K_l)^2 k^{-1}$. If capacity is expanded, and the toll is adjusted optimally, v_l/K_l decreases unless demand is perfectly elastic. The cost recovery ratio is therefore a strictly decreasing function of capacity. Capacity is optimal when $\rho_l = 1$. The conclusions for $\rho_l > 1$ and $\rho_l < 1$ follow.

[21] See de Palma and Lindsey (2007).

[22] The exposition in this subsection draws on de Palma and Lindsey (2004) and Lindsey (2009). The analysis applies either to an isolated link, or to one link of a network when all other links are priced efficiently.

first-order magnitude, whereas the deadweight loss from attracting new users is only second-order magnitude.

The second result, due to Arnott and Yan (2000), is that when user costs are homogeneous of degree zero, and usage is underpriced, the volume–capacity ratio v_l/K_l is always greater in the second-best optimum than the first-best optimum. Thus, if first-best congestion pricing was introduced, and capacity was adjusted from the second-best level given in Eq. (15) to the first-best level given in Eq. (14), congestion would be reduced regardless of how capacity is adjusted. Congestion pricing therefore relieves congestion in the long run as well as short run.

Broadly similar results have been derived using the bottleneck model. Arnott et al. (1987) show that if the elasticity of demand is constant and less than one in magnitude (which is almost certainly the case during peak hours), second-best capacity with no toll is larger than first-best capacity. Under the same demand elasticity assumption, Arnott et al. (1993) show that among the tolling regimes, optimal capacity is smallest for the fine toll, intermediate for a step toll, and largest for a flat toll. The intuition for this ranking is that more finely time-differentiated tolls reduce the resource cost of trips, which leaves less scope for further cost reductions from expanding capacity provided demand is not too elastic. A more general lesson is that optimal capacity depends not only on whether tolls are implemented, but also on their structure.

3.3.2. Non-optimal pricing of other road links

The problem of determining optimal link capacity when other links are inefficiently priced has not received much attention, but a few results are available. Consider first the two-links-in-parallel network and suppose that link 2 cannot be tolled. The second-best toll for link 1 is given by Eq. (11) with $\tau_2=0$. Optimal capacity for link 1 is given by the first-best rule in Eq. (14) because usage of link 1 is controlled by the toll, the envelope theorem applies, and capacity is chosen as in the first-best setting to minimize the sum of user and capacity costs.[23] The actual amount of capacity chosen depends on whether capacity of link 2 can be chosen as well. If it can, and link 2 is at least as long as link 1, link 2 should be eliminated and first-best policy can then be followed on link 1.[24]

Another possibility is that neither link 1 nor link 2 can be tolled. Expanding link 1 will draw users off link 2. If link 2 is congested, congestion will be relieved on both links so that the benefits from the investment will be underestimated if link 2 is ignored. However, if link 2 is wide enough to be congestion-free, expanding link 1 will provide no benefit on either link because travel time is fixed at the free-flow travel time on link 2. This unsettling result is called the *Pigou–Knight–Downs paradox*. It is an example of how an ostensibly reasonable strategy of expanding congested bottlenecks can go astray when usage is uncontrolled.[25] The paradox also illustrates the *Fundamental law of traffic congestion*, formalized by Downs (1962), that new capacity can quickly fill up with new traffic as travel patterns adjust. Later, Downs (1992) called the combined effects of changes in route, mode, and trip-timing *triple convergence*: a term that has also become part of the lexicon.

Paradoxical results also occur on more complex networks. The most famous is the *Braess Paradox* in which adding a new link to a network causes travel costs to increase for everyone. As explained in the next section, paradoxes also occur on networks with public transportation.

3.4. Public transportation

Section 2 briefly addressed the implications of mispricing car trips or public transportation trips for second-best pricing of the other mode. This subsection extends consideration to capacity decisions for the two modes when one or both are mispriced. The analytics depend on whether public transportation operates on a separate right of way.

3.4.1. Separate rights of way

Subway, light rail, and Bus Rapid Transit systems operate on separate rights of way from cars so that the two traffic streams do not interact. To begin the analysis suppose car trips are underpriced. As noted in Section 2.2.2, fares should be set below their first-best level. Second-best transit capacity is governed by several forces. Similar to roads, expanding transit capacity creates a usage effect and an induced-demand effect. If car trips are priced only slightly below the first-best level, fares will be too. By analogy, Wheaton's (1978) analysis of roads suggests that transit capacity should exceed the first-best level. This reasoning is reinforced by the fact that expanding transit capacity in tandem with lowering fares provides another way to reduce car travel. However, underpricing of roads also has a direct effect of reducing transit ridership.

The net result of these forces is unclear *a priori*. Kraus (2012) investigates it using the bottleneck model to describe road congestion and a simple model of rail transit service that has declining long-run marginal costs. Kraus assumes that driving creates an exogenous environmental externality as well as queuing congestion. First-best pricing then requires both a flat environmental charge and a fine (i.e., continuously time-varying) toll to eliminate queuing. For analytical purposes Kraus assumes that the environmental charge is introduced first, and the toll second. Introducing the environmental charge raises the cost of driving. The second-best fare increases, transit capacity decreases, and transit usage consequently declines as well—at least when transit usage is sufficiently high.[26]

Imposing the fine toll eliminates queuing and decreases the marginal social cost of driving without changing the private cost since the toll substitutes perfectly for queuing delay. Since driving becomes more efficient, it is optimal to increase driving. Provided the fare elasticity of transit demand is less than two in magnitude (a condition that is almost certainly satisfied in practice) the transit fare increases, and transit usage and capacity both decline. Implementing first-best pricing to internalize either environmental or congestion externalities therefore causes transit usage and capacity to decline.

This adjustment contrasts with the strategies of upgrading public transit that were adopted for the London and Stockholm congestion charges as well as the Milan EcoPass. One explanation for the discrepancy is that to overcome public opposition, the "stick" of tolls must be packaged with the "carrot" of better transit service. Another explanation is that much of the public transit service in these cities is provided by buses that share the road with cars (see below). A third explanation for London is that, as Kraus (2012) notes, transit fares in London were not changed when the congestion charge was introduced.

[23] Verhoef et al. (2010) show that this result extends to general networks when tolls can be set on some links, capacities chosen on other links, and both tolls and capacities chosen on a third set of links.

[24] As Small and Verhoef (2007, Section 5.1.3) point out, if users differ in their values of travel time it may be optimal to retain some capacity on link 2 for users with a low VOT while providing a higher service quality for other users on link 1.

[25] Arnott and Small (1994) describe the Pigou–Knight–Downs paradox as well as the Braess paradox and Downs–Thompson paradox mentioned below.

[26] Kraus derives this result using linear approximations of the demand curves for the two modes.

A parallel analysis can be undertaken when overpricing of transit, rather than underpricing of cars, is the primary distortion. Kidokoro (2010) shows that in this case tolls should be set above the first-best toll, and transit capacity expanded beyond the level indicated by the first-best transit investment rule. Overpricing of transit trips and underpricing of cars trips therefore both call for more transit service.

What about road capacity? The implications of underpricing car trips for second-best road capacity were investigated in Section 3.3. If transit is overpriced as well, road investment decisions can go seriously astray if pricing distortions are ignored. This is nicely illustrated by considering a polar case. Suppose that driving is untolled while transit is priced at average cost to satisfy a self-financing constraint. Due to scale economies, average total cost of transit decreases with ridership. If road capacity is expanded, some transit passengers will start driving. Transit ridership falls, average cost rises, and the fare has to be raised. In the new equilibrium the generalized cost of travel is higher for both modes. This phenomenon, known as the *Downs–Thompson paradox*, is more extreme than the Pigou–Knight–Downs paradox in which road capacity expansion yields zero benefits.[27] The Downs–Thompson paradox arises because there are two externalities at work: a negative road congestion externality, and a positive transit scale economy externality. Expanding the road induces a modal shift that exacerbates the negative externality and undermines the positive externality. The paradox can be avoided by pricing both modes efficiently, and in favorable circumstances expanding road capacity may still be worthwhile.

3.4.2. Shared rights of way

Some bus services operate on bus-only lanes, but many share the road with cars. Road pricing then has an additional benefit by allowing buses to circulate more freely as car traffic diminishes. As Small (2004) explains, road pricing sets off a virtuous circle as car use declines, bus ridership and service expand, car use drops further, and so on. Indeed, Ahn (2009) shows that travelers can be better off even without any toll-revenue recycling. The size of the modal shift and the benefits from congestion pricing depend on various factors: the severity of congestion, the initial share of trips taken by transit, the cross-price elasticity of demand between car and bus trips, whether or not buses are full, fare policy, and how much toll revenue is allocated to transit. Small (2005) illustrates some of these dependencies using numerical examples for London and a typical US city.

3.4.3. Practical limitations to expanding public transportation

The models just described suggest that public transportation investments can be complementary to road pricing. In practice, the scope to expand transit capacity is often limited. On bus systems without a separate right of way, expanding bus service contributes to road congestion and partially offsets the benefits from less car traffic. Subways and rail systems avoid this problem, but they often face space constraints, are much more costly to build, and are prone to cost overruns (Flyvbjerg et al., 2003). To facilitate access to transit stations, park-and-ride facilities may be required or bus feeder service that entails reconfiguration of bus lines and possible cancellations of longer-distance bus routes. Rail service may also cannibalize bus ridership and undermine bus scale economies (Pickrell, 1992) in a similar way to road expansion in the Downs–Thompson paradox.

[27] Arnott and Yan (2000) show that allowing for imperfect demand substitutability between car and transit trips does not eliminate the Downs–Thompson Paradox.

3.5. Scale of induced demand

Sections 3.3 and 3.4 illustrate the role of induced demand in determining the benefits of road capacity expansion when road usage is not efficiently priced. Induced demand is caused by changes in route, mode, trip-timing, and other dimensions of travel behavior. Over time, new demand can also materialize from residential and other land-use developments.

How important is induced demand in practice? Assessments vary widely. One view is that it is not significant enough to undermine the benefits of highway capacity additions or public transit service extensions.[28] The opposing view derives from a belief in the Fundamental law of road congestion according to which "you can't pave your way out of traffic congestion".[29] One measure of the strength of induced demand is the elasticity of traffic volume with respect to road capacity. This elasticity has been estimated for individual highway projects as well as for regions. If v denotes aggregate flow, and K denotes capacity measured in lane-km, the elasticity is defined as $\varepsilon \equiv (dv/dK)\times(K/v)$. A value of $\varepsilon=0$ corresponds to no induced demand, whereas $\varepsilon=1$ implies that capacity expansion attracts an equi-proportional increase in traffic so that travel speeds do not improve at all. Small and Verhoef (2007, §5.1.3) review empirical studies that obtained estimates of ε ranging from 0.2 to 0.8.

Elasticities tend to increase over time as residential and other land-use developments respond to improved mobility. Elasticities also tend to be larger for individual road projects because traffic diversion from alternative routes, or modal shifts from transit, are ready sources of induced demand. Elasticities are generally smaller at a regional level because there is less scope for traffic diversion. Two recent studies have estimated elasticities at a regional level using US data. Duranton and Turner (2011) use data on interstate highway kilometers and highway vehicle kilometers traveled for cities. Consistent with the *Fundamental law* they obtain estimates of ε close to one. They attribute the high level of induced demand mainly to increases in driving by current residents and increases in transportation-intensive production activity (e.g., trucking and warehousing). Migration and traffic diversion are of secondary importance. They also find that public transportation capacity has no statistically significant effect on total distance traveled—suggesting that any reductions in car trips caused by people shifting to transit are replaced by new car trips by other people.[30] Overall, they conclude that neither road capacity expansion nor public transit investment is effective in addressing traffic congestion which "leaves congestion pricing as the main candidate tool to curb traffic congestion." (p. 2646).

The second study by Hymel et al. (2010) uses vehicle miles traveled at the state, rather than city, level as a dependent variable, and total length of state roads as a measure of capacity. They obtain elasticity estimates of 0.037 in the short run and 0.186 in the long run, with about 60 percent of the induced demand attributable to increased accessibility and 40 percent to decreased congestion. Insofar as migration and traffic diversion are less important at the state than the city level, the fact that their estimates are smaller than those obtained by Duranton and Turner (2011) is understandable. Nevertheless, the differences in

[28] Duranton and Turner (2011) attribute these views to the American Road and Transport Builders Association and the American Public Transit Association respectively.

[29] This view was starkly expressed by David Begg, then chairman of the UK Commission for Integrated Transport, who is quoted by The Economist (2002) as saying: "A big road-building programme without pricing is as ludicrous as giving a heroin addict a last fix."

[30] By comparison, Winston and Langer (2006) find that an increase in rail transit service does reduce congestion costs, but bus service actually increases them.

estimates are so large that other factors are almost certainly at work that deserve investigation.

To sum up: the strength of induced demand is highly context specific. It depends on the geographical scale over which capacity investment is undertaken, on the time period considered, on the types of roads that are built or expanded (e.g., limited-access highways versus city streets), on the quality of public transit in the affected region, and so on. The consequences of implementing road pricing are therefore likely to be varied. Construction or expansion may be warranted for some links. Other links should be abandoned when they require major rehabilitation or reconstruction. Still other links should be built years later than they would be without road pricing.

3.6. Other benefits and costs of road infrastructure

As noted at the beginning of Section 3, road investment entails more than capacity expansion. Building roads with thicker pavement improves durability. Wider traffic lanes, wider shoulders, and better sight lines enhance safety and permit higher speed limits. Well-maintained roads also enhance safety and reduce vehicle wear and tear. Larsen (1993) describes road attributes that contribute to quality of travel as forming a road "standard" and argues that the benefits from improved road standards should be included in investment criteria. Some of these attributes have been studied in connection with road pricing. Newbery (1988) and Small et al. (1989) show how congestion tolls can be combined with infrastructure damage charges for heavy vehicles to pay for the construction and maintenance costs of roads designed to optimal capacity and durability. Ng and Small (2012) argue that unless congestion is relieved by road pricing or other means, at least some urban roads should be (re)designed for low-speed travel. For example, capacity can be increased by building roads with narrower lanes and allowing shoulders to be used as running lanes during peak periods. This illustrates the important point that the optimal *form* of investment could depend on whether road pricing is implemented.

An implicit assumption thus far is that road infrastructure provides benefits only to users while they are traveling. This is too restrictive. For example, wider roads facilitate access for fire engines and other rescue vehicles, and impede the spread of fires. More generally, road networks improve accessibility. These benefits may be significant if the road network is sparse, and alternative transportation modes are lacking, as is the case in some rural areas and developing countries. The benefits are unlikely to be as important in cities. However, by increasing the geographical scale of markets, good roads may increase competition and reduce distortions due to market power. Workers also have a wider choice of employment opportunities, and consumers have a more varied choice of goods and services. Metz (2008) argues that these benefits are more important than the travel time savings from road investments that are still at the core of cost–benefit analysis in transportation.

Yet another benefit of improved accessibility is that it facilitates agglomeration economies which arise when the spatial concentration of economic activity creates increasing returns. These economies derive in large part through labor markets from enhanced opportunities for matching of workers and employers, and from exchange of ideas. The geographical range of agglomeration economies varies, but it is believed to extend at least across urban areas.[31]

A final point is that road infrastructure also imposes external costs. Roads create a "barrier effect" by impeding walking and bicycling. Elevated roads block views. Roads and parking lots contribute to the urban heat island effect as well as susceptibility to flooding by increasing impermeable surface area. These effects are considered in the environmental review process for road projects, but they are usually disregarded in studies of road pricing and investment. External costs due to the presence of road infrastructure itself do not warrant pricing usage of the infrastructure.

3.7. Road pricing and investment in practice

The pricing and investment rules reviewed so far apply to a centralized, far-sighted, and benevolent planner. In practice, pricing and investment decisions are made by multiple levels of government with different jurisdictions and susceptibility to different interest groups. The results can be uncoordinated, myopic, and self-interested (Levinson, 2005; Vickerman, 2005). Local authorities may discriminate against through-traffic by tax-exporting, or try to shift congestion to other jurisdictions (De Borger et al., 2005). And local authorities that control links in series tend to set tolls that are too high, and invest in too little capacity from the perspective of the region.[32]

The danger of myopic investment policies were apparent to Vickrey (1969) when he wrote (pp. 252–253):

> "…the construction of facilities to accommodate additional traffic on the one mode or route will not only encounter increased construction costs by reason of other existing facilities that cross its path, but such construction will at the same time be increasing the cost of constructing any other transportation facilities across its path in the future…. It is very rarely that any account is taken, in the estimating of the cost of constructing facilities for a given transportation route and mode, of the increased costs of such future crossings by other links."

Planners have since become more aware of the network-wide implications of investment, but some investments are still aimed at regional development goals. Moreover, pricing has yet to be fully integrated with investment decisions, and fear remains that user charges will be used to generate tax revenues rather than to manage demand or recover costs (Vickerman, 2005).

Partly due to such fears, toll revenues for most existing and proposed road-pricing schemes are earmarked for specific transportation purposes.[33] Public finance theory disapproves of earmarking because it may not represent the most efficient use of money, reduces flexibility in the face of changing priorities, and hampers effective budget control. Nevertheless, earmarking revenues to roads is consistent with the beneficiary principle that those who pay tolls should receive the benefits, and it is consistent with public utility pricing models that have been applied to other services. Earmarking also provides a way to compensate non-users for the external costs of roads. For example, use of revenues from the toll rings in Norway was originally restricted to roads, but can now also be used for local public transportation, environmental quality, and safety.

[31] Estimates of the elasticity of average labor productivity with respect to employment density range from 2 to 10 percent for manufacturing industries, and up to 20 percent for some service industries (Mackie et al., 2011).

[32] Serial-link configurations are studied by De Borger et al. (2007) and Ubbels and Verhoef (2008). As discussed in Section 4, similar problems can arise with private roads.

[33] Arguments for and against earmarking are reviewed in Newbery and Santos (1999), de Palma and Lindsey (2007), and Small (2010).

4. Private roads

Private toll roads have been gaining favor as a supplement or alternative to public, toll-free roads. In part, support derives from the same concerns that motivate road pricing generally: shortages of public funds, dwindling revenues from fuel taxes, and growing acceptance of the user-pay principle. Private firms may be more cost efficient than public operators due to stronger financial incentives, and greater freedom from procurement rules and political interference. The public may also accept innovative pricing mechanisms such as peak-period tolls more readily from private firms because they are common in airline, hotel, and other private markets.

Views differ widely on private roads. Proponents emphasize the potential advantages just mentioned. Indeed, Knight (1924) criticized Pigou (1920) on the grounds that private toll-road operators implement efficient congestion tolls so that government control or oversight is unnecessary. Opponents worry about monopoly power and lack of coordination in toll and investment decisions if control of road networks is devolved to multiple private firms. To assess the pros and cons of private toll roads, this section first examines the analytics of toll and capacity choice decisions by private operators and then briefly addresses some practical concerns.

4.1. One firm controlling a single link

Consider an unregulated profit-maximizing firm that controls link l on a road network. As in the analysis of public roads in Section 2, toll choice is considered first and then capacity.

4.1.1. Toll choice

The firm sets τ_l to maximize revenue, $\tau_l v_l$, given an inverse link-usage demand curve $d_l(v_l)=p_l$ and the constraint $p_l=c_l(v_l)+\tau_l$ where p_l is the generalized cost on link l. Function $d_l(v_l)$ is determined by the distribution of market demands on the network, and tolls on other links which the firm treats as given. The profit-maximizing toll works out to:

$$\tau_l = c_l'(v_l)v_l - d_l'v_l = (c_l'(v_l) - d_l')v_l \tag{16}$$

The toll equals the first-best congestion toll in Eq. (10) plus a markup due to the firm's market power. The functional form of Eq. (16) reflects that the firm faces an inverse demand function net of user cost of $d_l(v_l)-c_l(v_l)$ so that the slopes of the inverse demand and user cost functions affect willingness to pay in the same way. In effect, the firm incurs two costs to attract an additional user. One is the marginal external congestion cost the user imposes on other users which reduces their willingness to pay. The firm accounts for this cost in the same way as does a public operator. The second cost is that the firm has to decrease the toll in order to attract the new user. This causes a loss of revenue from inframarginal users equal to $-d_l'v_l$. The firm adds this loss to the toll as a markup. The markup causes a deadweight loss since the lost revenue is a transfer to users with no social cost.

Eq. (16) can be rearranged as $\tau_l + d_l'v_l = c_l'(v_l)v_l$ which shows that the firm balances the marginal revenue from attracting an additional user with the marginal cost the user creates in congestion. Eq. (16) can also be rewritten in the form of a conventional markup rule:

$$\frac{\tau_l - c_l'(v_l)v_l}{p_l} = \frac{p_l - (c_l(v_l)+c_l'(v_l)v_l)}{p_l} = \frac{1}{|\varepsilon_l|}, \tag{16a}$$

where ε_l is the elasticity of demand for usage of link l.[34] If demand is perfectly elastic, perhaps because there is a congestion-free parallel link, the markup is zero and the firm sets the first-best toll. Otherwise the toll is too high. If demand is relatively inelastic, and congestion is not severe, the benefits from the toll in congestion relief are outweighed by the loss from tolling off too many users and welfare is lower than if the link remained untolled.

The principles of profit-maximizing tolls carry over, with some modifications, to dynamic models. A firm has an incentive to reduce peak-period congestion because any reduction in user costs can be recovered in greater toll revenue without affecting demand. Thus, in the bottleneck model a firm will use a fine toll to eliminate queuing. The firm will also add a flat or base component to the toll as a markup. But because the firm gains revenue from higher tolls within the peak period it has an incentive to limit the size of the base toll in order to attract more users. Time-variation of the toll therefore increases productive efficiency by eliminating queuing, as well as allocative efficiency by reducing the markup (de Palma and Lindsey, 2000).

The conclusion that private firms internalize congestion costs efficiently breaks down if users are heterogeneous. A firm still accounts for differences in the external costs that users impose, and if toll differentiation is feasible it will charge tolls based on user and vehicle characteristics. But unlike a public operator, a firm cares only about the costs borne by marginal users. If inframarginal users value travel time more highly than do marginal users, the firm will include too small a congestion externality component into the toll. In theory, the toll could be smaller than the first-best toll although this seems unlikely.

4.1.2. Capacity choice

With homogeneous users, a firm's capacity choice rule is the same as Eq. (14) for a public operator. This is an example of Spence's (1975) general result that if users value product quality equally, and outputs are the same, profit-maximizing and socially-optimal quality choices coincide. Since a firm sets a toll above the first-best optimal toll, output (i.e., link volume) is too low, and the firm chooses a capacity below first-best capacity. However, if the assumptions underlying the Cost Recovery Theorem hold, the volume–capacity ratio is the same as in the first-best optimum.[35] The firm therefore provides optimal quality, but too little quantity.

4.2. Competition

4.2.1. Toll competition

Competition between toll-road firms on general road networks is analytically intractable, and most economic studies have focused on simple networks with links in parallel or series. As expected, with parallel links toll competition is most effective in the symmetric case when firms control links with equal capacities and free-flow travel times. The allocative efficiency of equilibrium improves with the number of firms, and in the limit attains the first-best optimum because firms lose all their market power (Engel et al., 2004).

By contrast, when firms control links in series, increasing the number of firms has the opposite effect because links are perfect complements and each firm effectively has monopoly control over usage. Each firm adds a monopoly markup, and the end result is an equilibrium generalized trip cost far higher than in the first-best outcome. This suggests that the most efficient market

[34] Eqs. (16) and (16a) appear as Eq. (6.5) in Small and Verhoef (2007).

[35] The proof parallels that of the Cost Recovery Theorem. The user cost function $c_l(v_l,K_l)$ can be written $c_l(v_l/K_l)$, and Eq. (14) simplifies to $c_l'(v_l/K_l)(v_l/K_l)^2 - k = 0$ which defines a unique value of v_l/K_l independent of the operator's choice of K_l.

structure is one with multiple competing routes with a single firm in control of the links that comprise each route.[36]

4.2.2. Toll and capacity competition

When firms compete in both tolls and capacity, equilibrium depends on the timing of decisions. Three games have been studied for the case where firms operate links in parallel. In the simplest game all firms choose their capacities and tolls simultaneously. This game is covered in Section 4.1. Each firm chooses the socially optimal volume–capacity ratio, but adds a markup that depends on other firms' decisions.[37]

A second type of game involves two stages. In stage 1, firms simultaneously and independently choose their capacities, and in stage 2 they simultaneously and independently choose their tolls. Consistent with general industrial organization theory (Fudenberg and Tirole, 1984) in this game firms behave strategically and hold back on capacity in stage 1 in order to soften toll competition in stage 2. As a result, firms choose a higher volume–capacity ratio, and a correspondingly lower service quality, than in the first-best optimum. De Borger and Van Dender (2006) study a duopoly version of this model with linear demand and constant marginal capacity costs, and show that capacities can be either strategic substitutes or strategic complements.[38]

The third game, studied by Van den Berg and Verhoef (2011), is a Stackelberg game in which firms choose their capacities in sequence, and then choose tolls simultaneously once all links have been built.[39] In this game firms face conflicting incentives: they gain from restricting capacity to limit toll competition (as in the two-stage game), but they also gain from building a high capacity to induce subsequent firms to build less capacity. Van den Berg and Verhoef show that the first firm to move chooses a capacity that is larger than in the corresponding two-stage game, whereas the last firm chooses a capacity that is smaller. They find that welfare can be higher or lower in the Stackelberg game than the two-stage game.

To sum up: analytical studies of private toll roads draw several conclusions consistent with standard economic and industrial organization theory. If users are homogeneous, firms internalize congestion externalities efficiently in their choices of both toll and capacity, but they add a distortionary markup to the toll. Decentralization of control over the network to independent private firms enhances allocative efficiency relative to private monopoly when substitute links are controlled by different firms, but sharply reduces efficiency if the links are complements. If firms act strategically, and limit capacity to soften toll competition, efficiency is harmed, whereas if they increase capacity to deter other firms from investing, the effect on efficiency is ambiguous a priori.

4.3. Practical considerations

Although private toll roads are experiencing a resurgence in popularity, resistance continues to be strong in some countries. Traditional arguments for public provision of roads (as well as other transportation infrastructure) still carry weight (Vickerman, 2005). To a degree, roads are natural monopolies and have the opportunity to set high toll markups. Road capacity is also rigid and lumpy. During the "ramp-up period" when demand is growing on a new link, a firm may run a large deficit. Links in regions with low traffic volumes may never be profitable. Moreover, profitability is neither a necessary nor a sufficient condition for construction of a new link to be welfare-improving (Mills, 1995). There are two opposing biases: new links create social surplus that firms cannot fully expropriate, but some of their profits may come at the expense of profits on other links. These same biases apply generally to entry decisions in differentiated-products markets (Mankiw and Whinston, 1986).

A further difficulty is that toll roads are immobile and have few, if any, alternative uses. Asset specificity creates significant risks, especially in currently underdeveloped areas where demand depends on future development decisions, land-use regulations, and so on. All this suggests that some form of public-sector involvement is inevitable. There is a growing literature on regulation, contract design, risk sharing, competition for the market, and public private partnerships.[40]

One stream of literature addresses what form of regulation should be applied to a design-build-operate-and-toll concession when distortions elsewhere on the road network are not a concern. Toll regulation alone is insufficient because the first-best toll is below the profit-maximizing toll so that the firm will choose too small a capacity. Capacity regulation alone is also inadequate because the firm will set an excessive toll. One option is to regulate both the capacity choice and the toll. A simpler alternative is to regulate usage by imposing a minimum volume constraint (Guo and Yang, 2009). The firm will then choose the optimal capacity that minimizes the sum of construction and lifetime congestion delay costs because it can recoup any reductions in congestion costs by raising the toll. It will also set the optimal toll in order to meet the volume constraint. As explained in Section 4.1.1, the firm has an incentive to employ a time-varying toll to reduce peak-period congestion. Since this is socially beneficial, it should be encouraged to do so. Indeed, if toll regulation is adopted it should be imposed on the average level of the toll rather than on the time structure. In the bottleneck model, this would entail a ceiling on the base toll while allowing the firm to impose the fine toll that eliminates queuing.

5. Conclusions

5.1. Summary

Urban traffic congestion has persisted in the face of various policies to combat it including major road capacity investments. Many transportation economists and a growing number of other transportation professionals support road pricing as the most promising approach to address congestion and its collateral effects. This paper has selectively reviewed the theoretical literature on congestion pricing with a focus on the interface between congestion pricing and investment. Partial answers to the four questions posed in the introduction can be drawn.

1. *Is marginal-cost pricing consistent with cost recovery*? Efficient usage of roads calls for short-run marginal cost pricing. Cost recovery and the user pay principle require tolls set at average costs. The two goals generally diverge if capacity is determined arbitrarily. However, if the costs of building and using roads both exhibit constant returns to scale, and capacity is perfectly divisible, a road designed to optimal capacity just pays for

[36] See Small and Verhoef (2007, p. 201).

[37] Wu et al. (2011) prove this result for a general network, and show that it also holds if the generalized cost of travel is regulated.

[38] Capacities are strategic substitutes if an (anticipated) increase in capacity by one firm induces another firm to reduce its capacity so that its reaction curve is negatively-sloped. Capacities are strategic complements if the opposite is true.

[39] This setting is plausible for a toll-road industry. Entry is protracted because of the time required for environmental approval of each road and then to build it. In contrast, long-term toll contracts that prevent a firm from changing its toll may be difficult to write or enforce.

[40] See Small (2010) and Estache et al. (2011).

itself. These assumptions are likely to be satisfied approximately in heavily traveled corridors where the indivisibility of lane capacity is not a major factor. Moreover, surpluses and deficits on individual links may average out on large networks.

2. *How does road pricing affect optimal road capacity?* Tolls are usually seen as a substitute for investment since, by curbing traffic, tolls reduce the number of trips that can benefit from congestion relief. However, when usage is underpriced, adding capacity induces new demand with a private value less than its social costs so that the investment is partly "wasted". Empirical studies at both the individual facility and regional level find evidence of induced demand. Depending on the time span considered, the type of road investment, and other factors such as quality of public transit service, induced demand may be quite strong. If so, the potential travel-time savings from expanding capacity are largely dissipated if usage is underpriced. By filling the gap, road pricing enhances the benefits of investments and strengthens the case for adding capacity. Road pricing and investment may also be complementary in terms of public acceptability. Three common objections to road pricing are: paying for something that was previously free, double taxation, and inequity. Each of these objections applies with less force to tolls that are imposed on new roads—particularly if toll revenues are used to fund capacity expansion, operations, and maintenance.
3. *How does road pricing affect optimal public transportation fares and capacity?* Public transit service is heavily subsidized in many cities. One reason is that car travel is underpriced, and low fares encourage people to take transit which alleviates traffic congestion and other externalities. If road pricing is introduced, this rationale falls away and fares should increase. The implications for transit capacity are less obvious because while road pricing encourages transit use, higher fares discourage it. The effects of road pricing also depend on whether transit operates on a separate right of way or shares the road with cars. In the latter case, road pricing enhances transit service directly by allowing transit vehicles (usually buses) to circulate more freely as car traffic diminishes. Road pricing then sets off a virtuous circle as car use declines, transit ridership and service expand, car use drops further, and so on. Expanding transit service also helps to overcome public opposition by reducing or offsetting the out-of-pocket costs incurred from tolls.
4. *Do private toll road operators make socially efficient toll and capacity decisions?* When users are homogeneous and value trip quality equally, a private firm is cost efficient. It correctly accounts for the congestion costs borne by users, and chooses a road capacity that minimizes the sum of road construction and user costs. But a private firm is allocatively inefficient because it exploits its market power by adding a markup to the toll. The markup is inversely proportional to the elasticity of demand. If firms operate competing parallel roads, the market power of each firm is limited and decreases with the number of competitors. By contrast, if firms operate road links in series, the outcome can be grossly inefficient because each firm adds a monopoly markup to the price of a trip. This suggests that the most efficient market structure is one with multiple competing routes, with a single firm in control of the links that comprise each route.

Competition "in the market" between private road operators is still rare, and there is little experience to judge how it will play out. But as support increases for road pricing as a way to deal with traffic congestion, pollution, and declining fuel tax revenues, organizational and regulatory issues with private toll road markets are likely to become more pressing. Similar issues will arise if control over toll-road networks ends up divided within the public sector as municipal, regional, state/provincial, and national governments compete for business and residents while also trying to generate revenues for their respective budgets.

5.2. Future developments

This review has focused on the role of road pricing and investments to relieve congestion. Building new roads is very expensive and the costs are sunk. Road pricing is expensive to implement as well—especially if it is done comprehensively using either roadside infrastructure or satellite technology. Whether these investments are worthwhile depends on how road travel demand evolves, and whether traffic congestion remains a top policy priority.

There is little doubt that automobile ownership and usage will continue to grow in China, India, and other developing countries for many years. Freight transportation in developed countries may also continue an upward path. Yet evidence is mounting that annual per-capita distance traveled has saturated, and even started to decline, in many developed countries (Metz, 2010; Millard-Ball and Schipper, 2011). Various reasons for saturation have been suggested including economic stagnation, rising fuel prices, population aging, increasing urbanization, and growing health and environmental concerns (Litman, 2011). Another, more intriguing and encompassing, explanation is the *constant travel time budget hypothesis*.[41] According to this hypothesis, the fraction of time spent traveling per day has remained roughly constant over the centuries despite huge improvements in travel speed and convenience; growth of income, commerce, and tourist opportunities; and so on (Metz, 2008). Combined with a flattening in travel speeds the hypothesis provides an explanation for saturation in distance traveled. Unfortunately, the hypothesis has not been thoroughly tested and the data requirements are formidable (Axhausen, 2010).

Regardless of the reasons for stagnation in personal travel, if the change is permanent the case for comprehensive congestion pricing and building new roads loses some force. Priorities for road pricing will shift towards cost recovery, while investment priorities will shift towards road maintenance and rehabilitation, or reconstruction of aging infrastructure. Kahn and Levinson (2011) propose a new scheme for highway infrastructure funding in the US that reflects these priority shifts.

Technological change is another force that could alter the economics of road pricing and investment as well as other features of travel such as safety. Advanced Traveler Information Systems (ATIS) technology has been in use for some time. Like road pricing, ATIS have the potential to influence all travel-related decisions. Travelers benefit from learning about new choices, greater travel time reliability, and so on. Traffic controllers can respond more quickly to incidents. Still, the scope to control traffic flows using ATIS is limited because drivers will comply with advice or instructions only if it is in their self-interest to do so (Bonsall, 2008). Moreover, like other supply-side measures ATIS technology is susceptible to induced demand unless road pricing is implemented in tandem.

Information and Telecommunications Technology offers a substitute for personal and business travel as well as courier services and postal mail. But it also complements travel by helping people establish new contacts, and learn about new products and places to visit. Electronic devices such as laptops, wireless Internet, and cell phones are also complementary during travel by allowing people to make more productive use of travel time and therefore reduce the

[41] The hypothesis is also called *Zahavi's Law* after Yacov Zahavi (Zahavi, 1977).

opportunity costs of travel time. Whether telecommunications are a substitute or a complement for travel overall remains unclear although empirical evidence for complementarity is accumulating (Mokhtarian, 2002; Choo et al., 2012).

The most rapid technological advances currently under way are in electric and other alternative-fueled vehicles. These vehicles will provide local and global environmental benefits, but they also require large investments in fuel distribution infrastructure and they hasten the shift away from fuel-tax revenues as the main source of funding for roads. The implications for traffic congestion and road capacity requirements are less obvious. In the longer term, automated highways and self-driving vehicles could provide huge benefits in congestion relief, time savings, and safety, although they also present substantial legal and other challenges (Markoff, 2012). The technology would require massive investments in vehicles and roadside infrastructure, and might operate on a user-pay basis. This would spawn a new line of research on road pricing and investment.

Role of the funding source

Financial support from the UBC Sauder School of Business new faculty start-up grant is gratefully acknowledged. The School was not involved in any specific aspects of this review or in the decision to submit the paper for publication.

Acknowledgments

The author is grateful to Erik Verhoef and two anonymous referees for very helpful comments.

References

Ahn, K.-J., 2009. Road pricing and bus service policies. Journal of Transport Economics and Policy 43 (1), 25–53.
Altman, E., Boulognea, T., El-Azouzi, R., Jiménez, T., Wynter, L., 2006. A survey on networking games in telecommunications. Computers and Operations Research 33, 286–311.
Anas, A., Lindsey, R., 2011. Reducing urban road transportation externalities: road pricing in theory and in practice. Review of Environmental Economics and Policy 5 (1), 66–88. (2011).
Arnott, R., de Palma, A., Lindsey, R., 1987. Bottleneck congestion with elastic demand. Discussion Paper 690. Institute for Economic Research, Queen's University.
Arnott, R., de Palma, A., Lindsey, R., 1993. A structural model of peak-period congestion: a traffic bottleneck with elastic demand. American Economic Review 83 (1), 161–179.
Arnott, R., Kraus, M., 1998. When are anonymous congestion charges consistent with marginal cost pricing? Journal of Public Economics 67 (1), 45–64.
Arnott, R., Kraus, M., 2003. Transport economics. 2nd ed. In: Hall, R.W. (Ed.), Handbook of Transportation Science, vol. 56. Kluwer, Dordrecht, The Netherlands, pp. 689–726. (International Series in Operations Research and Management Science).
Arnott, R., Small, K.A., 1994. The economics of traffic congestion. American Scientist 82, 446–455. (Sept.–Oct.).
Arnott, R., Yan, A., 2000. The two-mode problem: second-best pricing and capacity. Review of Urban and Regional Development Studies 12 (3), 170–199.
Axhausen, K.W., 2010. The limits to travel: how far will you go? Transport Reviews 30 (2), 271–273.
Bergendorff, P., Hearn, D.W., Ramana, M.V., 1996. Congestion toll pricing in traffic networks. In: Hager, W.W., Hearn, D.W., Paardalos, P.M. (Eds.), Lecture Notes in Economics and Mathematical Systems. Springer Verlag, Berlin. (Network Optimization, pp. 52–71).
Bonsall, P., 2008. Information systems and other intelligent transport system innovations. In: Hensher, D.A., Button, K.J (Eds.), Handbook of Transport Modelling. Elsevier Science, Oxford, pp. 559–574.
Bonsall, P., Shires, J., Maule, J., Matthews, B., Beale, J., 2007. Responses to complex pricing signals: theory, evidence and implications for road pricing. Transportation Research Part A 41 (7), 672–683.
Cervero, R., 2003. Road expansion, urban growth, and induced travel: a path analysis. Journal of the American Planning Association 69 (2), 145–163.
Choo, S., Chang, Y., Mokhtarian, P.L., Hwang, K., 2012. Are the complementary relationships between transportation and communications for industrial uses dominant? A case study for Asian countries. In: 91st Annual Meeting of the Transportation Research Board. Washington, D.C. Conference CD Paper no. 12-2985.
Crew, M.A., Fernando, C.S., Kleindorfer, P.R., 1995. The theory of peak-load pricing: a survey. Journal of Regulatory Economics 8, 215–248.
Daganzo, C.F., 1997. Fundamentals of Transportation and Traffic Operations. Elsevier Science, New York.
De Borger, B., 2009. Commuting, congestion tolls and the structure of the labour market: optimal congestion pricing in a wage bargaining model. Regional Science and Urban Economics 39 (4), 434–448.
De Borger, B., Proost, S., Van Dender, K., 2005. Congestion and tax competition on a parallel network. European Economic Review 49, 2013–2040.
De Borger, B., Dunkerley, F., Proost, S., 2007. Strategic investment and pricing decisions in a congested transport corridor. Journal of Urban Economics 62, 294–316.
De Borger, B., Van Dender, K., 2006. Prices, capacities and service quality in a congestible Bertrand duopoly. Journal of Urban Economics 60 (2), 264–283.
de Palma, A., Fosgerau, M., 2011. Dynamic traffic modeling. In: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), Handbook in Transport Economics. Edward Elgar, Cheltenham, UK and Northampton, Mass, USA, pp. 188–212.
de Palma, A., Lindsey, R., 2000. Private toll roads: competition under various ownership regimes. Annals of Regional Science 34 (1), 13–35.
de Palma, A., Lindsey, R., 2004. Basic economic concepts for pricing and financing transport systems. In: de Palma, A., Quinet, E. (Eds.), La Tarification des Transports: Enjeux et Défis. Economica, Paris, pp. 37–64.
de Palma, A., Lindsey, R., 2007. Transport user charges and cost recovery. In: de Palma, A., Lindsey, R., Proost, S. (Eds.), Investment and the Use of Tax and Toll Revenues in the Transport Sector, Research in Transportation Economics, vol. 19. Elsevier, Amsterdam, pp. 29–58.
de Palma, A., Lindsey, R., 2011. Traffic congestion pricing methodologies and technologies. Transportation Research Part C 19 (6), 1377–1399.
Downs, A., 1962. The law of peak-hour expressway congestion. Traffic Quarterly 16, 393–409.
Downs, A., 1992. Stuck in Traffic: Coping with Peak-Hour Traffic Congestion. The Brookings Institution, Washington, DC.
Duranton, G., Turner, M.A., 2011. The fundamental law of road congestion: evidence from US cities. American Economic Review 101 (6), 2616–2652.
Engel, E., Fischer, R., Galetovic, A., 2004. Toll competition among congested roads. The B.E. Journal of Economics Analysis & Policy 4 (1). (Article 4).
Estache, A., Ellis, J., Trujillo, L., 2011. Public–private partnerships in transport. In: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), Handbook in Transport Economics. Edward Elgar, Cheltenham, UK and Northampton, Mass, USA, pp. 708–725.
Flyvbjerg, B., Skamris Holm, M.K., Buhl, S.L., 2003. How common and how large are cost overruns in transport infrastructure projects? Transport Reviews 23 (1), 71–88.
Fudenberg, D., Tirole, J., 1984. The fat cat effect, the puppy-dog ploy and the lean and hungry look. American Economic Review: Papers and Proceedings 74 (2), 361–366.
Guo, X., Yang, H., 2009. Analysis of a Build–Operate–Transfer scheme for road franchising. International Journal of Sustainable Transportation 3 (5), 312–338.
Haurie, A., Marcotte, P., 1985. On the relationship between Nash–Cournot and Wardrop equilibria. Networks 15, 295–308.
Hensher, D., 2011. Valuation of travel time savings. In: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), Handbook in Transport Economics. Edward Elgar, Cheltenham, UK and Northampton, Mass, USA, pp. 135–159.
Hurdle, V., 1991. Queuing theory applications. In: Papageorgiou, M. (Ed.), Concise Encyclopedia of Traffic and Transportation. Pergamon Press, Oxford, pp. 337–341.
Hymel, K.M., Small, K.A., Van Dender, K., 2010. Induced demand and rebound effects in road transport. Transportation Research Part B 44 (10), 1220–1241.
Kahn, M.E., Levinson, D.M., 2011. Fix it First, Expand it Second, Reward it Third: A New Strategy for America's Highways. The Hamilton Project. Washington, DC, Discussion Paper 2011-03, February.
Kidokoro, Y., 2010. Revenue recycling within transport networks. Journal of Urban Economics 68, 46–55.
Knight, F., 1924. Some fallacies in the interpretation of social costs. Quarterly Journal of Economics 38 (4), 582–606.
Kraus, M., 2012. Road pricing with optimal mass transit. Journal of Urban Economics 72 (2–3), 81–86.
Larsen, O.I., 1993. Road investment with road pricing—investment criteria and the revenue/cost issue. In: Talvitie, A., Hensher, D., Beesley, M.E. (Eds.), Privatization and Deregulation in Passenger Transportation, Second International Conference on Privatization and Deregulation in Passenger Transportation. c/o Viatek Ltd., Espoo, Finland, pp. 273–281.
Levinson, D.M., 2005. The evolution of transport networks. In: Button, K.J., Hensher, D.A. (Eds.), Handbook of Transport Strategy, Policy and Institutions 6. Elsevier, Amsterdam, pp. 175–190.
Lindsey, R., 2009. Tolls, earmarking and optimal road capacity. International Journal of Sustainable Transportation 3 (5), 385–411.
Litman, T., 2011. The Future Isn't What it Used to be: Changing Trends and Their Implications for Transport Planning. November 6 (〈http://vtpi.org/future.pdf〉 [February 18, 2012]).
Mackie, P., Graham, D., Laird, J., 2011. The direct and wider impacts of transport projects—a review. In: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), Handbook in Transport Economics. Edward Elgar, Cheltenham, UK and Northampton, Mass, USA, pp. 501–526.

Mankiw, N.G., Whinston, M.D., 1986. Free entry and social inefficiency. Rand Journal of Economics 17 (1), 48–58.
Markoff, J., 2012. Collision in the Making Between Self-Driving Cars and How the World Works. The New York Times. January 23 (〈http://nytimes.com/2012/01/24/technology/googles-autonomous-vehicles-draw-skepticism-at-legal-symposium.html?_r=1&nl=todaysheadlines&emc=tha26〉 [February 12, 2012]).
May, A.D., Liu, R., Shepherd, S.P., Sumalee, A., 2002. The impact of cordon design on the performance of road pricing schemes. Transport Policy 9 (3), 209–220.
May, A.D., Shepherd, S., Sumalee, A., Koh, A., 2008. Design tools for road pricing cordons. In: Richardson, H., Bae, C. (Eds.), Road Congestion Pricing in Europe: Implications for the United States. Edward Elgar, Cheltenham, UK and Northampton, Mass, USA, pp. 138–155.
Metz, D., 2008. The Limits to Travel: How Far Will You Go? Earthscan, London.
Metz, D., 2010. Saturation of demand for daily travel. Transport Reviews 30 (5), 659–674.
Millard-Ball, A., Schipper, L., 2011. Are we reaching peak travel? Trends in passenger transport in eight industrialized countries. Transport Reviews 31 (3), 357–378.
Mills, G., 1995. Welfare and profit divergence for a tolled link in a road network. Journal of Transport Economics and Policy 29 (2), 137–146.
Mohring, H., Harwitz, M., 1962. Highway Benefits: An Analytical FrameworkNorthwestern University Press, Evanston Illinois.
Mokhtarian, P.L., 2002. Telecommunications and travel: the case for complementarity. Journal of Industrial Ecology 6 (2), 43–57.
Newbery, D.M.G., 1988. Road damage externalities and road user charges. Econometrica 56 (2), 295–316.
Newbery, D.M.G., Santos, G., 1999. Road taxes, road user charges and earmarking. Fiscal Studies 20 (2), 103–132.
Ng, C.-F., Small, K.A., 2012. Tradeoffs among free-flow speed, capacity, cost, and environmental footprint in highway design. Transportation, http://dx.doi.org/10.1007/s11116-012-9395-8.
Oum, T.H., Tretheway, M.W., 1988. Ramsey pricing in the presence of external costs. Journal of Transport Economics and Policy 22 (3), 307–317.
Parry, I.W.H., 2009. Pricing urban congestion. Annual Review of Resource Economics 1 (1), 461–484.
Parry, I.W.H., Bento, A., 2001. Revenue recycling and the welfare effects of road pricing. Scandinavian Journal of Economics 103 (4), 645–671.
Pickrell, D., 1992. A desire named street car. Journal of American Planning Association 58 (2), 158–173.
Pigou, A.C., 1920. The Economics of Welfare. Macmillan, London.
Santos, G., Verhoef, E.T., 2011. Road congestion pricing. In: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), Handbook in Transport Economics. Edward Elgar, Cheltenham, UK and Northampton, Mass, USA, pp. 561–585.
Schrank, D., Lomax, T., Eisele, B., 2011. The 2011 Urban Mobility Report Powered by INRIX Traffic Data. College Station, Texas Transportation Institute. Texas A&M University, September. (〈http://tti.tamu.edu/documents/mobility-report-2011-wappx.pdf〉 [October 4, 2011]).
Scotchmer, S., 2002. Local public goods and clubs. In: Auerbach, A., Feldstein, M. (Eds.), Handbook of Public Economics, vol. 4. North-Holland, Amsterdam, pp. 1997–2042.
Small, K.A., 2004. Road pricing and public transport. In: Santos, G. (Ed.), Road Pricing: Theory and Evidence, Research in Transportation Economics, vol. 9. Elsevier Science, pp. 133–158.
Small, K.A., 2005. Unnoticed lessons from London: road pricing and public transit. Access 26, 10–15.
Small, K.A., 2010. Private provision of highways: economic issues. Transport Reviews 30, 11–31.
Small, K.A., Verhoef, E.T., 2007. The Economics of Urban Transportation. Routledge, London.
Small, K.A., Winston, C., Evans, C.A., 1989. Road Work. Brookings, Washington, DC.
Spence, M., 1975. Monopoly, quality and regulation. Bell Journal of Economics and Management Science 6, 417–429.
The Economist, 2002. Transport: Tolled You So. April 27 (〈http://economist.com/node/1099057〉 [January 27, 2012]), p. 56.
Tsekeris, T., Voß, S., 2009. Design and evaluation of road pricing: state-of-the-art and methodological advances. Netnomics 10 (1), 5–52.
Ubbels, B., Verhoef, E.T., 2008. Governmental competition in road charging. Regional Science and Urban Economics 38 (3), 174–190.
Van den Berg, V.A.C., Verhoef, E.T., 2011. Is the service quality of private roads too low, too high, or just right when firms compete Stackelberg in capacity? Tinbergen Institute Discussion Paper no. 2011-079/3, May 16 (〈http://tinbergen.nl/discussionpapers/11079.pdf〉 [May 25, 2011]).
Van Dender, K., 2003. Transport taxes with multiple trip purposes. Scandinavian Journal of Economics 105, 295–310.
Verhoef, E.T., Koh, A., Shepherd, S., 2010. Pricing, capacity and long-run cost functions for first-best and second-best network problems. Transportation Research Part B 44 (7), 870–885.
Verhoef, E.T., Nijkamp, P., Rietveld, P., 1996. Second-best congestion pricing: the case of an untolled alternative. Journal of Urban Economics 40 (3), 279–302.
Verhoef, E.T., Small, K.A., 2004. Product differentiation on roads: constrained congestion pricing with heterogeneous users. Journal of Transport Economics and Policy 38 (1), 127–156.
Vickerman, R., 2005. Infrastructure policy. In: Button, K.J., Hensher, D.A. (Eds.), Handbook of Transport Strategy, Policy and Institutions, vol. 6. Elsevier, Amsterdam, pp. 225–235.
Vickrey, W.S., 1969. Congestion theory and transport investment. American Economic Review (Papers and Proceedings) 59 (2), 251–260.
Victoria Transport Policy Institute, 2011. Transportation Cost and Benefit Analysis II—Congestion Costs. 19 August (〈http://vtpi.org/tca/tca0505.pdf〉 [February 18, 2012]).
Walters, A.A., 1961. The theory and measurement of private and social cost of highway congestion. Econometrica 29 (4), 676–699.
Wardrop, J., 1952. Some theoretical aspects of road traffic research. Proceedings of the Institute of Civil Engineers 1 (2), 325–378.
Wheaton, W., 1978. Price-induced distortions in urban highway investment. Bell Journal of Economics and Management Science 9, 622–632.
Winston, C., Langer, A., 2006. The effect of government highway spending on road users' congestion costs. Journal of Urban Economics 60, 463–483.
Wu, D., Yin, Y., Yang, H., 2011. The independence of volume–capacity ratio of private toll roads in general networks. Transportation Research Part B 45 (1), 96–101.
Yang, H., Huang, H.-J., 2005. Mathematical and Economic Theory of Road Pricing. Elsevier Science, New York.
Yang, H., Huang, H.-J., 1998. Principle of marginal-cost pricing: how does it work in a general road network? Transportation Research Part A 32 (1), 45–54.
Yang, H., Meng, Q., 2002. A note on 'Highway pricing and capacity choice in a road network under a build-operate-transfer scheme'. Transportation Research Part A 36, 659–663.
Yang, H., Meng, Q., Lee, D.-H., 2004. Trial-and-error implementation of marginal-cost pricing on networks in the absence of demand functions. Transportation Research Part B 38 (6), 477–493.
Zahavi, Y., 1977. Equilibrium between travel, demand, system supply and urban structure. In: Visser, E.J. (Ed.), Transport Decisions in an Age of Uncertainty. Martinus Nijhoff Publishers, The Hague, pp. 194–199.