The Economics of Traffic Congestion

Rush-hour driving strategies that maximize an individual driver's convenience may contribute to overall congestion

Richard Arnott and Kenneth Small

Traffic congestion has become one of the plagues of modern life in a big city. Time spent ensnarled in traffic is not simply time wasted; for most of us, it is time miserably wasted.

The dimension of the problem can be gauged from a simple back-of-the-envelope calculation. In the 39 metropolitan areas in the United States with a population of one million or more, roughly one-third of all vehicular travel takes place under congested conditions in which speed averages half of its free-flow value. About half of this congested driving is on expressways, causing a delay of about six-tenths of a minute per kilometer of travel; the remaining half is on other arterials, causing about 1.2 minutes delay per kilometer of travel. With some 75 million licensed drivers in heavily populated areas, each averaging roughly 16,000 kilometers per year within those areas, there are approximately 1,200 billion kilometers driven annually in metropolitan areas, bringing the total delay to 6 billion vehicle-hours each year.

Richard Arnott is a professor of economics at Boston College. Born in London, he obtained his undergraduate degree from MIT in urban studies and his Ph. D. from Yale in economics. His visiting positions include ones at Oxford, Princeton, Stanford and Munich, and he has served on the editorial boards of journals in urban economics and public finance. His current research is on the economic theory of urban housing markets and of rush-hour traffic congestion. Kenneth A. Small is a professor and Chair of Economics at the University of California, Irvine, and specializes in urban and transportation economics. He earned an M.S. in physics and a Ph. D. in economics from the University of California, Berkeley. He is coeditor of Urban Studies *and serves on the editorial boards of three other journals. He has served on advisory panels for local, national and international organizations involved with transportation, land use and environmental issues. Address for Small: Department of Economics, University of California, Irvine, CA 92727.*

Research has shown that the cost of driving is quantifiable. Through their actual choices, drivers have demonstrated a willingness to pay, on average, about $1.33 to save 10 minutes travel time, or $8.00 per hour. This figure does not include the costs of disruption from the unpredictability of traffic delays, the costs of inconvenient schedules caused by attempts to avoid delays, nor the costs of extra fuel, accidents and air pollution. Even without taking all of these additional factors into account, the annual cost of driving delays comes to $48 billion, or $640 per driver.

Such congestion has policy-making itself in gridlock. Every policy considered either is too unpopular, is too expensive or has proven ineffective. Why is congestion so intractable, and what can be done?

Answering these questions turns out to require a sophisticated understanding of the behavioral interactions that determine when and where congestion occurs. Transportation researchers have identified three paradoxes in which the usual remedy for congestion—expanding the road system—is ineffective or even counterproductive. The resolution of these paradoxes employs the economic concept of externalities to identify and account for the difference between personal and social costs of using a particular roadway. This not only clarifies the economics of traffic congestion, but it also points to ways in which the congestion problem can be solved with clever applications of the standard pricing tools of economics.

Intractable Congestion

The standard remedy to traffic congestion is to "build our way out." Building our way out of the current jam, however, would be prohibitively expensive. A few years ago, the Southern California Association of Governments estimated the cost of accommodating expected 25-year growth in the Los Angeles region through expansion of highways and new rapid-transit lines. Their cost estimate was $111 billion, a figure that now seems conservative in light of cost escalations in some recent projects. As urban areas grow more dense around existing facilities, planning and building new capacity becomes extraordinarily complex, expensive and politically controversial.

There is, of course, the alternative solution of building new capacity in the form of mass transit. Experience shows that such an approach is unable by itself to attract more than a tiny fraction of the peak demand for highway facilities. Don Pickrell of the Transportation Systems Center in Cambridge, Massachusetts, meticulously documented the cost of each trip diverted from cars to public transit for eight major rail transit projects. This was done by dividing the total annualized cost of the system (including interest and depreciation on capital) by the annual number of transit riders who formerly used automobiles.

For three of the projects, there was no diversion because the number of bus patrons who shifted to personal cars overshadowed the rise in rail patronage. For the others, the cost (at today's prices) ranged from about $12 to $43 for each new transit trip. Achieving significant reductions in automobile congestion through subsidies of this magnitude is financially infeasible. Furthermore, the advantages of the car are simply too great: Not only does it provide considerably more comfort, privacy and convenience than mass transit, but it is also much better suited to the decentralized American city. The

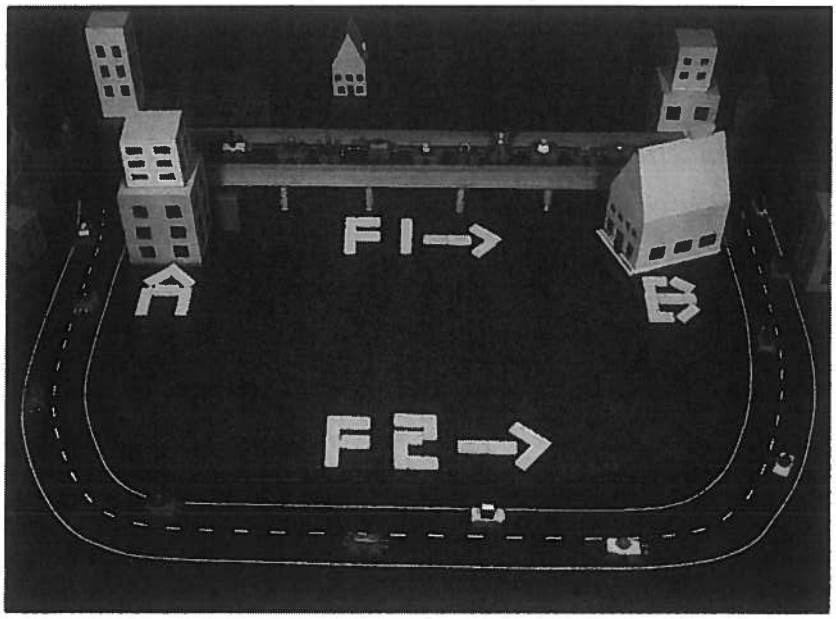


Figure 2. Expanding road capacity creates its own demand, a phenomenon known as the Pigou-Knight-Downs paradox. Because route 1, over the bridge, is the most direct route from point A to point B, more people want to use it, and the resulting congestion makes route 1 take as long as the more circuitous route 2. Travel time on each route is 15 minutes. Expanding the capacity of the bridge over route 1 only attracts more users, and the travel time remains unchanged. The paradox disappears only if the bridge capacity exceeds twice the total travel flow.

$T_1 = 10 + 10(F_1/C_1)$, $T_2 = 15$, $F_1 + F_2 = 1{,}000$

Scenario A: Suppose $C_1 < 2{,}000$

And $T_1 = 10 + 10\,(F_1/C_1) = 15 = T_2$

Then: $F_1 = 1/2\ C_1$, $T_1 = T_2 = 15$

Scenario B: When $C_1 > 2{,}000$

$F_1 = 1{,}000$ and $F_2 = 0$, $T_1 = 10 + 10{,}000/C_1$

Example: Suppose $C_1 = 2{,}500$

Then: $F_1 = 1{,}000$, $F_2 = 0$, $T_1 = 10 + 10{,}000/2{,}500 = 14$

Everyone uses the bridge.

Figure 3. Mathematical expression of the Pigou-Knight-Downs paradox shows that increasing the bridge capacity to any value less than twice the traffic flow has no effect on travel time (T_1). Suppose that route 1 in Figure 2, the route over the bridge, takes 10 minutes with no traffic, but travel time rises linearly with the ratio of traffic flow (F_1) to bridge capacity (C_1). Route 2 always takes 15 minutes (T_2). There are 1,000 travelers faced with the choice of route 1 or route 2. In scenario A, the bridge's capacity is defined as less than 2,000. The traffic flow over the bridge adjusts to $1/2C_1$, so that travel time on routes 1 and 2 are equal at 15 minutes. In scenario B, the bridge capacity is increased to exceed 2,000. In that case, everyone uses the route with the bridge, but the travel time decreases, as can be seen in the example where bridge capacity equals 2,500.

urban sprawl that was encouraged by massive subsidies to automobile travel in past decades cannot be reversed.

Even if new highway construction and new mass transit were cheaper, building our way out of the problem would be much harder than it might appear at first glance. One reason for this is a phenomenon called latent demand. The traffic we see does not represent the full demand for peak travel at the prevailing monetary cost, since congestion itself causes many potential rush-hour vehicle trips to be canceled, diverted (for example, to mass transit, to car pools and to less-congested routes and destinations) or rescheduled. Any reduction in congestion resulting from capacity expansion encourages others to drive during hours or on routes they ordinarily would not use. So measures to relieve congestion are at least partially undone by latent demand.

The other reason capacity expansion alone does not work is that congestion is mispriced. Because drivers do not pay for the time loss they impose on others, they make socially inefficient choices concerning how much to travel, when to travel, where to travel and what route to take. As the paradoxes will show, the combination of latent demand and mispriced congestion may be so perverse that an expansion of capacity brings about no change in congestion, or even makes it worse.

Traffic Paradoxes

The first of the paradoxes, the Pigou-Knight-Downs paradox, helps to explain why expanding road capacity can elicit new demand with no improvement in congestion. Suppose 1,000 peak-hour travelers between two cities can choose between a direct route containing a narrow bridge and a more circuitous, but wider, road, as illustrated in Figure 2. The first route takes 10 minutes with no traffic, but travel time rises linearly with the ratio of traffic flow (which we will call F_1) to bridge capacity (C_1). In the example, capacity is defined as the traffic flow at which the speed drops to half of the free-flow speed. Travel time (T_1) therefore can be described as the 10 minutes it takes without traffic, plus the extra time it takes if the road is congested, as in the following equation:

$$T_1 = 10 + 10\,(F_1/C_1)$$

The second route always takes 15 minutes. Each traveler chooses the road with the lower travel time. As long as bridge capacity exceeds 2,000, the first route can accommodate all of the 1,000 travelers and still takes less than the 15 minutes of the second route. If bridge capacity is set at 2,500, for example, travel time is 14 minutes. Under these circumstances, everyone takes the shorter route, and there is no paradox.

The paradox occurs when the bridge capacity, C_1, is less than 2,000. In this case, travelers divide themselves across

the two routes, such that travel time on each route is 15 minutes, which implies that traffic flow over the bridge is exactly half its capacity. Therefore, expanding the bridge's capacity to anywhere in the range from 0 to 2,000 has absolutely no effect on anyone's travel time. Instead, it diverts more people from the route with spare capacity to the route crossing the bridge. In other words, the new bridge capacity generates its own demand.

Attempts to reduce congestion on the bridge by instead encouraging car pooling, expanding mass transit or improving telecommunication facilities would likewise be frustrated unless total vehicular traffic were reduced to below half of the bridge's capacity. So long as any traffic remained on the second route, latent demand for the bridge would undermine these attempts to relieve its congestion.

The crux of the paradox lies in the distinction between the private and the social costs of a trip. The private cost is the cost the driver incurs. The social cost equals the private cost plus the external cost, which is the cost the driver imposes on other drivers by slowing them down. In the example, the social cost of traveling on the bridge exceeds the private cost because it is congested. Typically, drivers choose the route with the lower cost to them—the lower private cost. This results in an equilibrium in which private costs on the two routes are equalized. If, instead, drivers were distributed across the two routes so as to equalize the social cost, the paradox would disappear; bridge expansion would relieve congestion. This suggests that conventional policies to relieve congestion would work better if each driver faced the social cost of his or her trip.

The second paradox, called the Downs-Thomson paradox, is even more perverse. The example of this paradox is like that of the previous paradox, except the alternative to taking the congested route is now a privately operated train line. The train operator breaks even financially by ensuring that all of the train cars are full. If more people take the train, then trains run more frequently, saving people some waiting time at the station. In this case, let us say that the maximum travel time by train is 20 minutes, and that 10 minutes will be cut from the trip for every 3,000 travelers. We can describe the effect of the actual number of people using the train (F_2) on travel

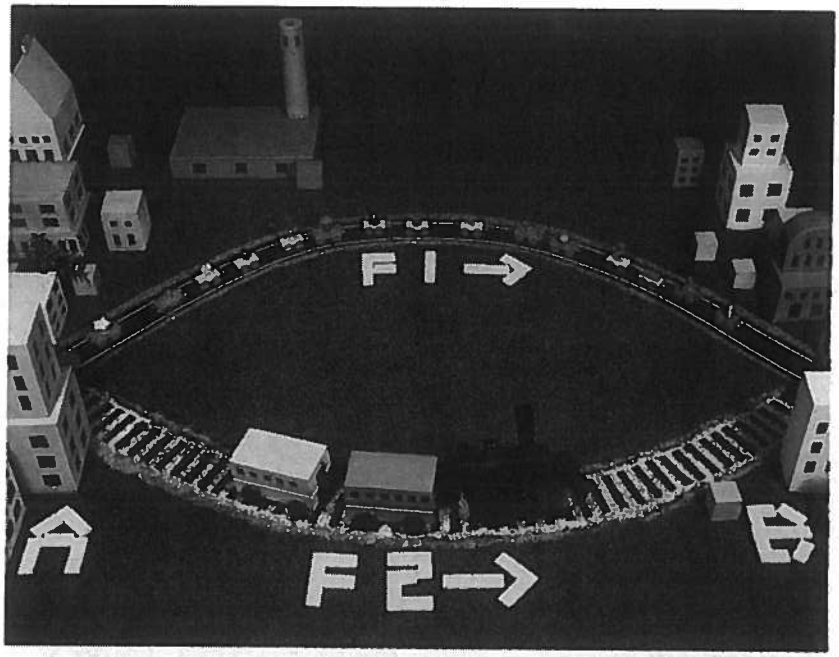


Figure 4. Increased capacity leads to more, rather than less, congestion in the Downs-Thomson paradox. Here the second route, a passenger train, shows increasing returns with added flow because service quality improves as more travelers use it. Expanding road capacity draws people off the train, worsening train service. Equilibrium between the two routes dictates that road travel becomes worse as well, such that increasing road capacity actually increases travel time on both routes.

$T_1 = 10 + 10\,(F_1/C_1)$, $T_2 = 20 - (F_2/300)$, $F_1 + F_2 = 1{,}000$

Scenario A: When $C_1 < 1{,}000$

$T_1 = 10 + 10\,(F_1/C_1) = 20 - [(1{,}000 - F_1)/300]$

So: $F_1 = C_1/[1.5 - (C_1/2{,}000)]$; $T_1 = 10 + 10/[1.5 - (C_1/2{,}000)] = T_2$

Examples of equilibrium solutions:

If $C_1 = 250$, then $F_1 = 182$ and $T_1 = T_2 = 17.27$

If $C_1 = 750$, then $F_1 = 667$ and $T_1 = T_2 = 18.89$

If $C_1 \rightarrow 1{,}000$, then $F_1 \rightarrow 1{,}000$ and $T_1 = T_2 \rightarrow 20$

Scenario B: When $C_1 > 1{,}000$

$F_1 = 1{,}000$, $F_2 = 0$, $T_1 = 10 + (10{,}000/C_1)$

Example: Suppose $C_1 = 2{,}000$

Then $T_1 = 15$

Figure 5. Downs-Thomson paradox, expressed mathematically, shows how increasing road capacity in the situation in Figure 4 actually raises travel time, as long as the road capacity (C_1) is smaller than the number of travelers. Suppose that the equation for travel time by the congested highway route (T_1) is the same as in Figure 3; that the maximum travel time by train (T_2) is 20 minutes; and that 10 minutes will be cut from the train trip for every 3,000 travelers. Since there are 1,000 total travelers, the number using the road (F_1), plus the number using the train (F_2) will equal 1,000. In scenario A, at equilibrium, some of the travelers use the road, and others use the train. Under these conditions, as the road capacity approaches 1,000, travel time for each route, T_1 and T_2, approaches 20 minutes. In scenario B, the road capacity is expanded to exceed 1,000. All of the travelers use the road, so that the traffic flow over the road is 1,000, whereas the traffic on the train is zero. Expanding the capacity of the road to 2,000, for example, lowers travel time to 15 minutes.

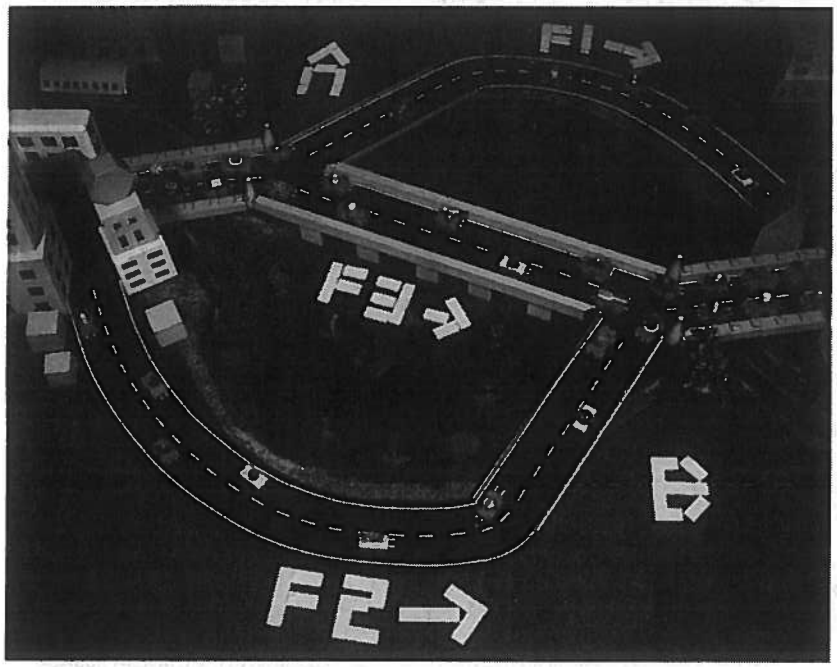


Figure 6. Adding a link to a congested road network can cause everyone's travel time to go up. The Braess paradox shows how too much traffic may become attracted to the most congested route segments. Here the old routes each made use of only one bridge apiece. In contrast, the new causeway, route 3, entices half the travelers to choose a shorter route that takes them across both bridge A and bridge B, increasing congestion on both bridges and slowing traffic.

Traffic on bridge A: $F_A = F_1 + F_3$, $T_A = F_A/100$

Traffic on bridge B: $F_B = F_2 + F_3$, $T_B = F_B/100$

$T_1 = 15 + T_A$, $T_2 = 15 + T_B$, $T_3 = 7.5 + T_A + T_B$, $F_1 + F_2 + F_3 = 1{,}000$

Scenario A: Equilibrium with no causeway

$F_1 + F_2 = 1{,}000$, $F_3 = 0$

$T_1 = T_2 = 15 + (F_1/100) = 15 + (1{,}000 - F_1)/100$

So: $F_1 = F_2 = 500$, $T_1 = T_2 = 20$

Scenario B: Equilibrium with causeway

$F_1 + F_2 + F_3 = 1{,}000$

$T_1 = T_2 = T_3 = 15 + (F_1 + F_3)/100 = 15 + (F_2 + F_3)/100 = 7.5 + (F_1 + F_2 + 2F_3)/100$

So: $F_1 = F_2 = 250$, $F_3 = 500$, $T_1 = T_2 = T_3 = 22.5$

Figure 7. Mathematical expression of the Braess paradox shows how travel time increases when the causeway is added. Both bridges are the congested points. (The time it takes to travel bridge A is expressed as T_A, bridge B as T_B; traffic flow on the bridges is expressed as F_A and F_B.) In scenario A, before the additional link, equilibrium is reached when the total time (T_1) to travel on route 1, over bridge A, is equal to the time (T_2) to travel over route 2, which uses bridge B. Under these circumstances, the traffic flow on route 1 (F_1) and the traffic flow on route 2 (F_2) are each equal to 500 (half of the 1,000 travelers); the total travel time on each route is 20 minutes. In scenario B, the causeway has been added, so travelers have the choice of taking this additional route, which we call route 3. Route 3 takes traffic over both bridges A and B, so the bridges become even more congested than before. Equilibrium is reached when travel times on all three routes, T_1, T_2 and T_3, are equal. When this happens, traffic flow over route 3 (F_3) is 500 vehicles, and travel time on all three routes is 22.5 minutes.

time (T_2) on the train, illustrated in Figure 4, with the following equation:

$$T_2 = 20 - 10\,(F_2/3{,}000)$$

Each traveler chooses the faster mode, so that travel times are equalized when both modes are used. (The time it takes to travel on the road is described by the same equation used in the first paradox.) The equalized travel time is calculated in Figure 5.

The intriguing feature of this situation is that now travel time increases with any increase in bridge capacity within the range from 0 to 1,000. The reason is that, just as in the earlier example, capacity expansion diverts people to the congested road. But now the diversion causes train service to get worse, so equilibrium can occur only when congestion is worse also. Here, new capacity generates more than its own demand.

The reason this paradox is even more perverse than the previous one is that there is not only an external cost imposed by each automobile user, as before, but there is now an external benefit created by each user of the train as well. This is because using the train causes the frequency of service to increase and hence reduces other users' waiting times. This is a technological property of all types of mass transit, including bus and even taxicab service, as was demonstrated in 1972 by Herbert Mohring of the University of Minnesota, who formulated a detailed model of a bus line, taking into account road speed, frequency of service and the extra time required for each passenger to get on or off at a bus stop.

The same perverse result can be obtained if instead of expanding the road, well-intentioned planners entice some fraction of travelers away from both routes by providing some third alternative such as subsidized vanpools, telecommuting centers—or even a new train service. Nor is this unrealistic: The cases studied by Pickrell included some where initiating a new train service diverted so much traffic from the existing transit system (in this case, bus transit) that the overall quality of transit service deteriorated, causing a net diversion to automobiles and, presumably, a worsening of road traffic. Had the existing transit service been improved instead, the improvements might have reinforced, rather than thwarted, the external benefits inherent in transit service.

If, in our example, the bridge capacity is expanded to equal or exceed 1,000, a very different thing happens. Capacity exceeds demand, and everyone starts using the road. The number of train users, F_2, drops to zero. Now, further increasing the bridge capacity does reduce the travel time. For example, increasing capacity to 1,500 decreases the travel time to 16.67 minutes, the same time a train trip would take if all commuters traveled by train. Further increasing the road capacity lowers road travel time more, so the paradox disappears.

The final paradox is the Braess paradox, named for a German operations researcher who in 1968 described an abstract road network in which adding a new link causes total travel time to increase. Our version involves 1,000 people traveling from district A of a city to district B, where the districts are separated by marshland, as pictured in Figure 6. District A lies south of a river at the west end of a marsh, and district B is north of the river at the east end of the marsh. Two routes connect A to B. Route 1, carrying traffic F_1, crosses the river at bridge A and circles north of the marsh to B. Route 2, carrying traffic F_2, circles south of the marsh and crosses the river at bridge B. Travel on both routes is uncongested except at the bridges. Travel time on either route is 15 minutes under uncongested conditions. Ten minutes travel time is added on either route for every 1,000 drivers going over its bridge, so that the total time for either route can be described as follows

$$T_{1\,or\,2} = 15 + 10\,(F_{1\,or\,2}/1{,}000)$$

At the point of equilibrium, where travel times on the two routes are equal, equal numbers of people use each route. Since there are 1,000 total travelers, this means that 500 are on each route. In that case, the time to travel on each route is 20 minutes.

A causeway is then constructed across the marsh from the north end of bridge A to the south end of bridge B. The causeway can be traversed in 7.5 minutes, regardless of traffic volume. There is now a third route from A to B—across bridge A, along the causeway and then across bridge B. Its traffic is shown as F_3 in Figure 6. What happens when the causeway is opened? Each bridge now carries the traffic for two distinct routes, the previous one plus the route over the causeway. Now there are two ways to approach bridge B, either by route 2 or by the causeway, route 3. The total traffic flow over bridge B becomes the sum of the two inflowing routes. Likewise the total flow over bridge A becomes the sum of the total flows for both routes 1 and 3. Travel time for bridges A and B can be rewritten as follows:

$$T_A = 10\,(F_A/1{,}000) = 10\,(F_1 + F_3)/1{,}000$$
$$T_B = 10\,(F_B/1{,}000) = 10\,(F_2 + F_3)/1{,}000$$

Traffic will be in equilibrium when travel times on all three routes are equal, a condition that gives two equations. The three traffic volumes add up to 1,000 vehicles, as we have defined the total peak-hour usage, so we can readily solve for all three traffic flows as follows:

$$\begin{aligned} T_1 = T_2 = T_3 &= 15 + 10\,(F_1 + F_3)/1{,}000 \\ &= 15 + 10\,(F_2 + F_3)/1{,}000 \\ &= 7.5 + 10\,(F_1 + F_2 + 2F_3)/1{,}000, \end{aligned}$$

so that:

$$F_1 = F_2 = 250;\ F_3 = 500;$$
$$T_1 = T_2 = T_3 = 22.5 \text{ minutes.}$$

The result shows that half the traffic takes the causeway route, and the other half divides evenly between the two previous routes. Hence each bridge carries 750 travelers, 50 percent more than before, producing a travel time on each route of 22.5 minutes, as opposed to the 20 minutes it took before the causeway was constructed. Adding the causeway has made everyone's trip longer.

The paradox is explained by congestion externalities on the bridges; that is, because each traveler ignores the external cost he or she imposes by crossing a bridge, too many people choose the causeway route, which crosses both bridges. The faster the causeway, the more people are enticed to take it, and the worse is their trip. If causeway-traversal time were only 5 minutes, all 1,000 would choose that route, and travel time would rise to 25 minutes. Only if the causeway speed were infinite would equilibrium travel time return to its original 20 minutes.

Are these paradoxes more than intellectual curiosities? It has been claimed that the Braess paradox explains some traffic problems observed in Stuttgart, Manhattan and Oslo. Martin Mogridge of University College, London, has forcefully, if controversially, asserted that the Downs-Thomson paradox explains the deterioration of road speeds over 20 years or so in central London. As for the Pigou-Knight-Downs paradox, it is so enshrined in transportation planning that it is often called "the fundamental law of traffic congestion."

Externalities and Pricing

The concept of externalities provides a powerful tool for analyzing congestion in a more general context. An externality is brought about when a person does not face the true social cost of an action. By modeling congestion systematically, it is possible to define the social cost of driving on a congested road by observing how aggregate travel delays are related to the number of travelers. Combining this with a model of demand for the road, one can determine both equilibrium travel patterns (as in the above examples) and optimal travel patterns under some defined objective such as minimizing aggregate travel time.

In order to apply the concept of externalities, we need to convert travel time to a cost. For simplicity, let us ignore the out-of-pocket costs of travel. Assume also that everyone places an identical monetary value on each minute of travel time. Multiplying travel time (T) by this monetary value then gives the private cost of a trip.

The existence of congestion implies that this travel time depends on the traffic flow, F. Thus we obtain the curve relating private cost to traffic flow shown in Figure 8. The lower (solid) part of the curve, marked $pc(F)$, where private cost is increasing as flow increases, corresponds to situations of modest congestion and lends itself to analyzing the congestion externalities discussed earlier. At any level of flow, one can calculate total cost as the flow F multiplied by each driver's private cost $pc(F)$. Then it is possible to calculate how much total cost increases when flow is increased by one unit, which is referred to as the social cost of a trip. This quantity, known as the social cost of a trip, is plotted as $sc(F)$ in Figure 8. By definition, the social cost of a trip equals the private cost plus the external cost—the cost the added driver imposes on others by slowing them down. Thus the external cost equals the vertical distance between the social cost and private cost curves. (Because drivers impose external costs on each other, multiplying the social cost curve by flow does not lead to a meaningful total cost.) A more complete analysis would also consider

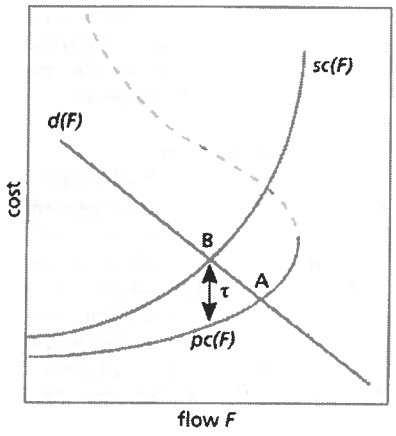


Figure 8. Maximum efficiency and equilibrium do not occur at the same point; they are separated by the difference between the cost of road usage to an individual (the private cost) and the social cost. The private cost of using a road, *pc(F)*, rises with the traffic flow (F) because of congestion. This implies that each one-unit increment to *F* raises total cost *F* x *[pc(F)]* by an amount called the social (or marginal) cost of a trip, which exceeds the private cost, as shown. The social cost is written mathematically as *sc(F) = pc(F) + F* x *[d(pc)/dF]*, where *d(pc)* and *dF* are the change in private cost and the change in traffic flow, respectively. The social cost exceeds the private cost by *F* x *[d(pc)/dF]*, which is called the external social cost of a trip because it represents the cost that is imposed by a traveler on others. If the demand curve is *D(F)*, equilibrium occurs at point A. But the efficient solution would be at point B, where the marginal trip is just worth its social cost. At the efficient solution, the external cost of a trip is τ. This is the toll that, if charged, would shift the equilibrium from A to B.

the social costs of noise, air pollution and so forth.

The demand for using a road is generally some flat or downward-sloping function of the private cost. (In the paradoxes it was flat, an extreme case in which we assume that each route is a perfect substitute for the alternatives.) Such a relationship is shown in Figure 8 as *D(F)*. Equilibrium occurs at point A, where the demand curve intersects the private-cost curve; at this level of flow the benefit of an extra trip equals its private cost. Efficiency reaches its maximum at point B, where the benefit of an extra trip equals its social cost. In equilibrium, travel is underpriced because drivers do not pay for the congestion they cause. Consequently, too many trips are taken.

The trick for planners is to create conditions under which the system operates at point B instead of at point A. The solution is simply to charge a payment, known as a congestion toll, equal to the external cost. In Figure 8, the optimal congestion toll τ is measured by the vertical distance between the social and the private cost curves at point B. By thus bringing the private cost faced by the traveler up to the level of the social cost, privately-made decisions will lead to the social optimum (point B).

We can see how this approach might work to relieve the congestion predicted by the Downs-Thomson paradox in Figure 4. The simplicity of demand for the two alternative modes of travel in this paradox permits a diagrammatic analysis of the private and social costs on both routes. In Figure 9 one can visualize both the paradox itself (an equilibrium, also called a user optimum) and the cost-minimizing traffic pattern (a social optimum or system optimum). It is drawn for the case that the capacity of route 1 is equal to 750. The private cost pc_1 and social cost sc_1 of travel by car on the congested route are plotted as functions of flow F_1, starting with an original flow of 0, so that we designate this initial point as the first origin O_1. The corresponding costs of train travel are plotted backwards, as a function of passenger flow F_2 on the train with O_2 as its origin. The distance between the origins is 1,000, ensuring that $F_1 + F_2 = 1{,}000$. Note that the social cost of car travel exceeds its private cost. But the social cost of train travel is less than its private cost, reflecting the external benefit that each train user confers on other train users by causing the frequency of service to increase and waiting time to decrease. The equilibrium, at point A, occurs where the number of travelers using each mode of transportation (car or train) equalizes the private costs—in this example, two-thirds travel by car.

The social optimum, at point B, occurs when travelers are divided between the two modes in such a way as to equalize the social costs—in this case one-sixth of the travelers go by car. Under these circumstances, switching one traveler from car to train, or vice versa, neither increases nor decreases the social cost of accommodating that traveler. At this division, the private costs by car and train (measured by the heights of lines pc_1 and pc_2) are both lower than they were at point A. But because they are not equal, this point is not an equilibrium. However, by imposing a road toll equal to the difference, people will be led to choose the social optimum.

It can be shown that the total private cost associated with point B decreases as road capacity C_1 is expanded. Hence when social costs, rather than private costs, are equalized, the paradox disappears. Note that in this example, the road toll is interchangeable with a train subsidy. In real situations, however, there are so many substitutes for peak-hour car and train travel that this equivalence breaks down. In such cases, the theory calls for a road toll of $sc_1 - pc_1$ and a train subsidy of $pc_2 - sc_2$, both measured at point B.

These examples illustrate a policy known as congestion pricing. Versions of it have been implemented or are being considered in Europe, Asia and in the United States. Congestion pricing is also an example of marginal-cost pricing, a term with much broader meaning. Briefly, marginal-cost pricing refers to setting the price of a unit of a commodity equal to the incremental social cost of producing one more unit of the commodity. Mathematically, marginal cost is the derivative of the total-cost function. In the traffic context, the social cost of a trip is the increment in total cost to all travelers caused by adding one more trip; by facing the trip maker with this social cost, society effectively sets the full price of the trip (including both money and time) equal to its marginal cost.

One concern that crops up with congestion pricing is the overall welfare of the travelers. If one regards people only in their roles as travelers, everyone is made worse off by being forced to pay a toll that raises the cost of using the road, even with a reduction in congestion. (Our paradoxes illustrate extreme cases where travelers are not made worse off.) But travelers are also citizens, so one must consider what happens to the toll revenues. Paying a toll, after all, does not use up resources; it is only a paper transaction—or, more likely, an electronic one. If the toll revenues are used to benefit citizens generally, the gains people receive as citizens more than offset their losses as travelers. In fact, the more formal statement of "efficiency" is precisely this: There is some way of redistributing the toll revenues that leaves everyone as well, or better, off.

To show how this could happen, we return to the first example, the Pigou-

Knight-Downs paradox. If all 1,000 travelers value their time equally, the efficient traffic allocation across routes minimizes aggregate travel time. The efficient allocation occurs at the point where the number of vehicles using the bridge equals one-quarter the bridge's capacity, with the remainder of vehicles taking the longer route. According to the formula for congestion on the bridge, the travel time on the bridge route is 12.5 minutes. The other travelers use the second route, which has a travel time of 15 minutes. The aggregate travel time is described as follows:

$$1/4C_1 \times 12.5 + (1{,}000 - 1/4C_1) \times 15 = 15{,}000 - 0.625C_1 \text{ minutes}$$

Suppose, for example, that capacity C_1 were exactly 1,600. In the efficient allocation, 400 travelers take the bridge with travel time 12.5 minutes, while the other 600 travelers take the longer route with a travel time of 15 minutes. Aggregate travel time is 14,000 minutes, the lowest possible with this capacity. For the sake of simplicity, suppose time is valued at 10 cents per minute; this traffic allocation can then be achieved by charging a toll of 25 cents for crossing the bridge, since then, everyone's trip cost is $1.50, either in time (for users of the longer route) or in time plus toll (for bridge users). This is the same trip cost that prevailed in the unpriced equilibrium; thus it is easy to allocate all of the $100 in toll revenue so that everyone is better off.

In more realistic examples, it would probably not be possible to target the redistribution of toll revenues so carefully that everyone was made better off by a toll. Hence in practice, congestion pricing (or any policy change) can be justified only if it is acceptable to make some people worse off when the overall gains are enough. Some argue that this is justified because the existing system of highway finance subsidizes drivers, so the proposed change actually corrects an existing inequity.

We asserted earlier that all the paradoxes disappear if every driver pays the social cost of his or her travel. Thus with (optimal) congestion pricing, expansion of capacity always creates benefits (which, of course, must be weighed against the cost of expansion). We have just shown this in the example of the Pigou-Knight-Downs paradox: Application of the efficient congestion toll results in aggregate travel time equal to $15{,}000 - 0.625C_1$, which falls with expansion of the bridge. We invite the reader to check our assertion for the other two paradoxes.

Congestion pricing has the added advantage that it makes transportation planning easier. Whether or not congestion pricing is employed, the merits of a proposed expansion of a transportation link can be evaluated by comparing the cost of expansion with the total cost savings it produces. In the absence of congestion pricing, calculation of these cost savings requires knowing how the expansion will alter traffic flows and travel times over the entire network. But with congestion pricing, the savings can be evaluated knowing only the traffic flow on that link.

A final, and very important, point concerning congestion is that some traffic congestion is usually optimal. Congestion could be eliminated entirely by prohibiting travel or by spending vast sums on transportation systems. And it could probably be reduced to negligible levels by requiring that trips be evenly spread over the 24 hours of the day. But any of these solutions would generate social costs far exceeding the current costs of congestion. Huge benefits from concentrating economic activity within a geographical location derive from the reduction of transport costs (even with congestion). There are also great advantages from schedule coordination—having people work or play at common times. Congestion is simply a cost that goes hand in hand with these benefits. Congestion pricing ensures that a given level of benefit is achieved at minimum congestion cost.

Practical Pricing?

Economists have advocated congestion pricing for at least three decades, since the pioneering work of William Vickrey of Columbia University. They have failed, however, to overcome a number of counterarguments, including costly and inconvenient toll collection, especially on downtown streets; regressive distributional impact, since lower-income people spend a larger proportion of their income on commuting and have less work-schedule flexibility; lack of trust in government to dispose of toll revenues wisely; and benefits that in some cases are so small as to be insignificant.

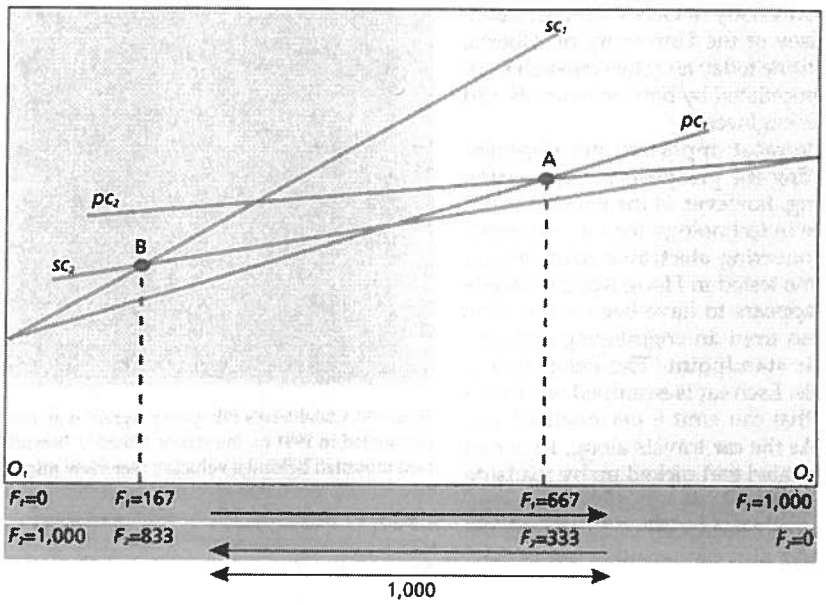
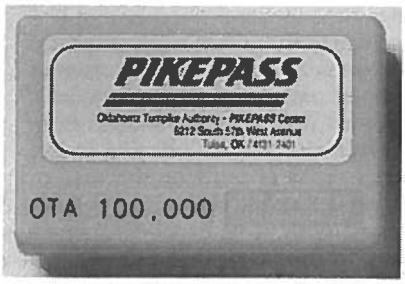


Figure 9. Downs-Thomson paradox can be resolved by bringing private costs in line with external social costs. The paradox depicted in Figure 4 disappears when social costs, rather than private costs, are equalized for the two travel alternatives, the road versus the train. Here, private cost pc_1 and social cost sc_1 of travel by car on the congested route are plotted as a function of flow F_1, with O_1 as the origin. The corresponding costs of train travel are plotted backwards, as a function of passenger flow F_2 on the train, with O_2 as its origin. Equilibrium occurs at point A, and the social optimum is at point B. At point B, switching one traveler from car to train, or vice versa, neither increases nor decreases the social cost of accommodating that traveler. Imposing a road toll equal to $(sc_1 - pc_1)$ and subsidizing the train fare by $(sc_2 - pc_2)$ will lead people to choose the social optimum.



The case for congestion pricing is significantly stronger today. Because of worsening congestion and financial constraints, the public is more receptive to pricing solutions. A recent survey in London, for example, found that a majority of automobile commuters would favor congestion pricing if the revenues were used to upgrade the transport system. Other proposals for using the toll revenues address the impacts on income distribution. As for the benefits of pricing, a new generation of models that take into account trip rescheduling produce estimates of benefits many times larger than earlier work based on a rush hour of fixed duration. These models, introduced by Vickrey and further developed by one of us (Arnott) in collaboration with André de Palma of the University of Geneva and C. Robin Lindsey of the University of Alberta, constitute today an active research frontier populated by both economists and traffic engineers.

The most important development affecting the prospects for congestion pricing, however, is the enormous advance in technology for toll collection. A pioneering electronic road-pricing scheme tested in Hong Kong a decade ago appears to have been a complete success from an engineering and economic standpoint. The basic idea is simple. Each car is equipped with a device that can emit a personalized signal. As the car travels along, its signal is activated and picked up by roadside receptors at designated charging points. A central computer records the charges and periodically sends each car owner a bill based on that car's travel history. Enforcement is based on photographing license plates of cars failing to emit the signal when queried electronically. Another variation is to have prepayments coded on a "smart card" mounted in the vehicle, thereby eliminating the need to record the vehicle's location.

Commercial equipment is readily available, and electronic toll collection is now in operation on toll highways or bridges in many places, including Oklahoma, Texas, Florida, France, Italy and Norway. These have proven that existing technology can handle road-pricing transactions quickly and efficiently without appreciably slowing traffic. A sophisticated system of congestion pricing using such technology will be implemented on a new privately operated roadway in the median strip of the Riverside Freeway in Southern California that is slated for completion in 1995.

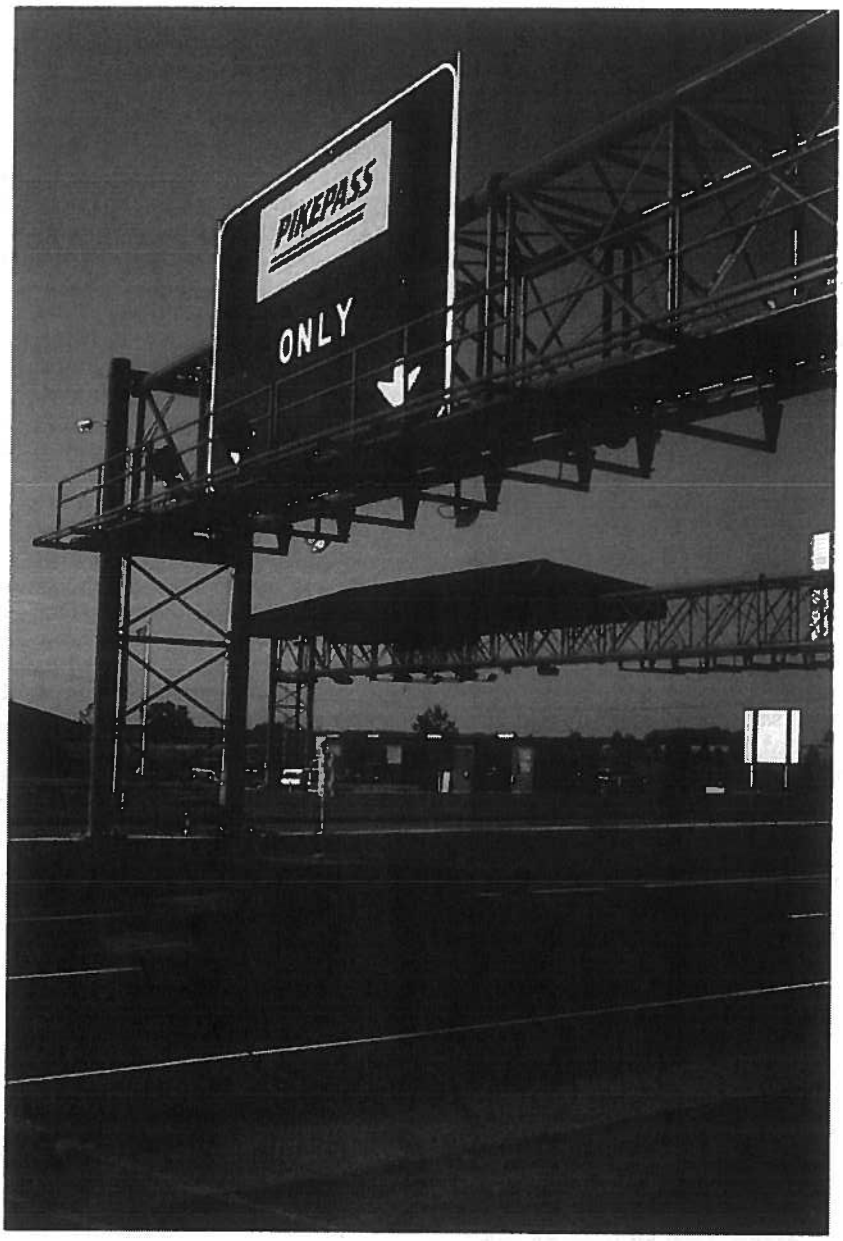


Figure 10. Oklahoma's Pikepass program is an example of electronic toll collection. Pikepass, implemented in 1991 on the state's 478-mile Turnpike system, uses a palm-sized integrated-circuit card mounted behind a vehicle's rear-view mirror. The card responds to a radio beacon emitted by a transmitter beside special Pikepass entry lanes, which are identified by an overhead sign. A charge is automatically deducted from the user's prepaid account. Users need not stop or slow down. Video cameras record license plates of violators. About 100,000 vehicles currently participate, and the revenues collected from Pikepass participants account for almost one-third of Oklahoma's annual toll revenues. Such systems can easily be adapted to vary the toll depending on time of day, as is now done in Norway and will be done in 1995 on privately built express lanes on the Riverside Freeway in California. (Photographs courtesy of Amtech Inc.)

Today, the remaining practical problems of implementation are much narrower and appear amenable to solution by the usual kinds of development efforts implemented for any public policy. For example, every system needs to specify an option for occasional travelers who lack an electronic device. Protection of privacy (a major factor in Hong Kong's decision not to implement the system it tested) is quite feasible, but conflicts to some degree with the need for tracing to correct mistakes. How finely tuned the pricing system should be is a question involving trade-offs between efficiency and simplicity. At one extreme, the city of Cambridge, England, considered a pricing system that would depend on actual congestion encountered moment by moment.

Political acceptability, however, remains the key. A well-designed and credible plan for spending the toll revenues is essential. Only with such a plan can the public be assured that a proposed pricing scheme would provide needed financing for transportation improvements, offset at least some of the regressive distributional impact of the tolls and protect against misappropriation of the revenues for wasteful purposes.

Whatever the prospects for congestion pricing, it is clear that congestion is a more complex phenomenon than some of our current policy analyses assume. It is also clear that some of the common-sense solutions do not solve the problem. Only by understanding the full nature of people's travel decisions and how they interact can sensible policies be formulated.

Bibliography

Arnott, R., A. de Palma and R. Lindsey. 1990. Economics of a bottleneck. *Journal of Urban Economics* 27:111–130.

Cameron, M. 1991. *Transportation Efficiency: Tackling Southern California's Air Pollution and congestion*. Oakland, Calif.: Environmental Defense Fund and Regional Institute of Southern California.

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*. Washington: Brookings Institution.

Hau, T. D. 1992. *Congestion Charging Mechanisms for Roads*. World Bank Working Paper WPS-1071. Washington: World Bank.

Holden, D. J. 1989. Wardrop's third principle: Urban traffic congestion and traffic policy. *Journal of Transport Economics and Policy* 23:239–262.

Lewis, N. C. 1993. *Road Pricing Theory and Practice*. London: Thomas Telford.

May, A. D. 1992. Road pricing: An international perspective. *Transportation* 19:313–333.

Mohring, H, ed. 1994. *The Economics of Transport*. Brookfield, Vt.: Edward Elgar.

Murchland, J. D. 1970. Braess's paradox of traffic flow. *Transportation Research* 4:391–394.

National Research Council Committee for Study on Urban Transportation Congestion Pricing. 1994. *Curbing Gridlock: Peak-Period Fees to Relieve Traffic Congestion*. I: Committee Report and Special Recommendations; II: Commissioned Papers. Transportation Research Board Special Report 242. Washington: National Academy Press.

Pickrell, D. H. 1989. Urban Rail Transit Projects: forecast versus Actual Ridership and Costs. Cambridge, Massachusetts: U. S. Department of Transportation, Transportation Systems Center.

Small, K. A. 1992. Using the revenues from congestion pricing. *Transportation* 19:359–381.

Small, K. A. 1992. Urban transportation economics. *Fundamentals of Pure and Applied Economics* 51. Chur, Switzerland: Harwood Academic Publishers.

Vickery, W. S. 1963. Pricing in urban and suburban transport. *American Economic Review Papers and Proceedings* 53:452–465.

Walters, A. A. 1961. The theory and measurement of private and social cost of highway congestion. *Econometrica* 29:676–699.