Contents

Lecture 8

Models of Highway Congestion

The Bottleneck Model

The model considered here was developed by William Vickrey [3]. My discussion is largely a restatement of results found in an excellent paper by Arnott, De Palma and Lindsey [1].

Every morning N identical commuters must travel from home to work. Highway traffic is slowed by a single bottleneck, a bridge located between the commuters' homes and their workplace. Cars are able to pass over this bridge at the rate s per minute. All commuters would prefer to arrive at work at the same time t^*, but the bottleneck makes this impossible. Instead there will be a "rush hour", lasting for N/s minutes, the time that it takes the N commuters to pass through the bottleneck. We assume that the time required to drive from home to the bridge is negligible, as is the time required to drive to work after crossing the bridge. Commuters cross the bridge in the same order in which they leave home. In equilibrium, a rush-hour queue will form as commuters leave home more rapidly than the bridge can serve them.

Everyone's first choice would be to pop out of bed at t^* and arrive at work immediately, without meeting any traffic. But this is not possible for everyone. Commuters face a tradeoff between rising early to beat the traffic, arriving at work late, or spending more time in the traffic queue in order to arrive at work at a preferred time. Let us assume that the cost of the inconvenience of arriving at work early, at time $t^* - t$, is βt and the cost of

arriving at work late, at $t^* + t$ is also βt.[1] Time spent waiting in a traffic queue is even more costly. The cost per minute of sitting in the traffic queue is α, where $\alpha > \beta$.

In equilibrium, commuters must be indifferent about which time during the rush hour they leave for work. The first commuter to leave for work will face no queue and will be first to arrive at work. The last commuter to leave for work will also face no queue and will be the last to arrive. In equilibrium the first and last commuter must be equally well off, and the time elapsed between their departures will be N/s. Therefore the first commuter must depart at $t^* - \frac{N}{2s}$ and the last at $t^* + \frac{N}{2s}$. Since they spend no time in traffic queues, each of these commuters has total travel costs of $\beta N/2s$.

Let us consider commuters who arrive at work before t^*. For these workers, arriving at work later is better than arriving earlier. In equilibrium, we expect that those arrive at work closer to their preferred time must spend time in the traffic queue. Let $D(t)$ be the length of the traffic queue at time t. Since cars cross the bridge at the rate s, a commuter who joins the queue at time t must spend $D(t)/s$ minutes in the queue and arrive at work at time $t + D(t)/s$, which is $t^* - t - (D(t)/s)$ minutes before t^*. In equilibrium, all commuters have the same utility as the first commuter to leave for work. This implies that for commuters who leave for work at time t and arrive at work before t^*,

$$\beta \frac{N}{2s} = \alpha \frac{D(t)}{s} + \beta(t^* - t - \frac{D(t)}{s}). \tag{8.1}$$

Solving this equation for $D(t)$, we have:

$$D(t) = \frac{\beta N}{2(\alpha - \beta)} + \left(\frac{s\beta}{\alpha - \beta}\right)(t - t^*). \tag{8.2}$$

For any time t during the rush hour, let $r(t)$ to be the rate at which people leave home to go to work. The rate of growth of the queue at time t is then

$$\dot{D}(t) = r(t) - s \tag{8.3}$$

From equations 8.2 and 8.3 it follows that

$$\dot{D}(t) = \frac{s\beta}{\alpha - \beta} = r(t) - s \tag{8.4}$$

---

[1]Arnott et al show that it not hard to solve this problem when the cost per minute of being late is different from that of being early. I will leave this case (and still more general cases for you to do as homework.

and hence

$$r(t) = \frac{s\alpha}{\alpha - \beta} > s, \tag{8.5}$$

and

$$\dot{D}(t) = r(t) - s = \frac{s\beta}{\alpha - \beta}, \tag{8.6}$$

From Equations 8.5 and 8.6 we see that, starting at time $t^* - N/2s$, people leave home at a constant rate $r > s$ and the traffic queue grows at the rate $s\beta/(\alpha - \beta)$ so long as those who join the queue get to work before t^*. Let $\tilde{t}$ be the departure time that would allow a commuter to arrive at work exactly at t^*. Then it must be that $\tilde{t} + D(\tilde{t})/s = t^*$, and therefore it follows from equation 8.2 that

$$\tilde{t} = t^* - \frac{\beta}{\alpha}\frac{N}{2s}. \tag{8.7}$$

Commuters who leave home after time $\tilde{t}$ will get to work later than the preferred time t^*. The queue will get shorter after $\tilde{t}$ as later departures mean an ever more inconvenient arrival time at work. The maximum length that the queue reaches is

$$D(\tilde{t}) = \frac{\beta}{2\alpha}N \tag{8.8}$$

It is interesting to notice that this maximum length does not depend on the capacity of the bridge. Thus a highway engineer who expects to reduce the length of queues by increasing capacity is going to be disappointed. Of course if capacity is increased, the amount of time that commuters spend in a queue of a given length is diminished proportionately.

In equilibrium, commuters who leave home after $\tilde{t}$ have the same costs as those who leave before. This implies that where $\tilde{t} < t < t^* + N/2s$,

$$\beta\frac{N}{2s} = \alpha\frac{D(t)}{s} + \beta\left(\left(t + \frac{D(t)}{s}\right) - t^*\right) \tag{8.9}$$

and therefore

$$D(t) = \frac{\beta N}{2(\alpha + \beta)} - \left(\frac{s\beta}{\alpha + \beta}\right)(t - t^*). \tag{8.10}$$

We see from equation 8.10 that after $\tilde{t}$, the queue diminishes at the constant rate

$$\dot{D}(t) = -\frac{s\beta}{\alpha + \beta} \tag{8.11}$$

until it disappears entirely at time $t^* + N/2s$, when the last commuter leaves for work. Since $\dot{D}(t) = r(t) - s$, it follows that from $\tilde{t}$ until the end of the rush hour at $t^* + N/2s$, commuters leave for work at the rate

$$r(t) = s\frac{\alpha}{\alpha + \beta} < s. \qquad (8.12)$$
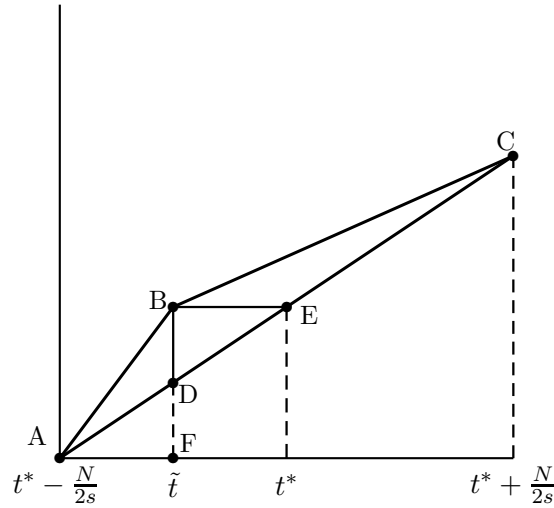
Figure 8.1: Rush Hour

Figure 8.1 shows the situation graphically. The line AC is drawn with constant slope s. The height of the line at any time t represents the cumulative number of commuters who have crossed the bridge by time t. The line AB is drawn with slope $r(t) = s\alpha/(\alpha - \beta)$, which is the departure rate who arrive at work early. The height of this line at time t denotes the number of commuters who have left for work before t. The commuter who leaves for work at time $\tilde{t}$ will arrive at work at exactly t^*. After time $\tilde{t}$, the rate of departures from home is $r(t) = s\alpha/(\alpha + \beta)$ and the height of the line BC shows the cumulative number of persons who have left for work by the corresponding time. Since at any time, the number of commuters who have left home is the height of the broken line ABC and the number who have arrived at work is that of the line AC, the vertical distance between these two lines is the length of the traffic queue at the corresponding time. For any time t, the horizontal distance from the corresponding point on ABC to the line AC is the length of time that a commuter who leaves home at

t will spend in the traffic queue. For example, at time $\tilde{t}$, BF commuters have left for work, while DF of them have passed over the bridge. Thus the length of the queue is BD. The number of persons who will have passed over the bridge at time t^* is equal to BF which is the same as the number who will have left home by $\tilde{t}$. Thus the commuter who leaves home at time $\tilde{t}$ will pass over the bridge at t^*. The amount of time that this commuter spends in the queue is the horizontal distance BE.

Welfare Analysis and the Use of Tolls

In equilibrium, each of the N commuters incurs the same total commuting cost, which we have seen to be $\beta N / 2s$. Therefore the total commuting costs of all N individuals are $\beta N^2 / 2s$. Let us call the part of a commuter's costs that is due to arriving at work at a less than ideal time the *displacement cost* and the part that is due to waiting in line, the *queueing cost*. Since commuters arrive at work at a uniform rate of s from time $t^* - N/2s$ until $t^* + N/2s$, it is easy to see that the average displacement cost for those who arrive at work before t^* is $\beta N / 4s$ and the average displacement cost for those who arrive at work after t^* is also $\beta N / 4s$. Thus total displacement cost summed over all commuters is $\beta N^2 / 4s$, which is exactly half of total costs $\beta N^2 / 2s$ borne by commuters.

The other half of costs incurred by commuters are queueing costs. These costs are entirely "wasted" in the sense that they could be avoided if commuters would coordinate their departures from home appropriately. Suppose that commuters could be persuaded to leave their homes at a constant rate s starting at time $t^* - N/2s$ and ending at time $t^* + N/2s$. Since they would leave home at the same rate as they cross the bridge, no queue is formed and no commuter has to spend time in a queue. In fact, their pattern of arrival at work and their total displacement costs would be exactly the same as that we found in equilibrium. But since they would have no congestion costs, their total costs would be exactly half of the costs found in equilibrium. The trouble with the coordinated solution is that it is hard to see how it could be maintained as a decentralized equilibrium. In the absence of queues, those commuters who leave for work at times closer to t^* are better off than those who arrive either near the beginning or near the end of the rush hour. How can we persuade some people to accept a less convenient departure time than is allowed to others?

Time-varying tolls

A toll that changes over by time of use could be designed to achieve full efficiency. Suppose that the toll is set at zero for those who commute at the earliest time, $t^* - N/2s$, then rises at the rate β per minute until time t^* and finally, after time t^* falls at the rate β per minute. In the absence of congestion, this tax scheme would leave commuters indifferent about the time they leave for work, since the higher costs of arriving at work at a less desired time are exactly compensated by the lower taxes collected at this time. With this toll in place, there would be an equilibrium with no congestion. Total costs of each commuter, including taxes plus displacement costs are exactly $\beta N/2s$, which is the same as total cost in the absence of a toll. Thus we see that nobody is made worse off by the imposition of the toll. The difference is that the queueing costs that were wasted with no tolls are collected by the government in the form of taxes and are available for other purposes. Total revenue collected from the tax is

$$\frac{\beta N^2}{4s} \tag{8.13}$$

Unless the toll revenue is entirely wasted, this toll is strictly Pareto improving.

The time varying toll is remarkably efficient relative to most taxes. In fact, this tax imposes literally no burden on commuters relative to the no-tax outcome. This stands in remarkable contrast to the standard result that ordinary excise taxes in the absence of congestion impose an "excess burden greater even than the amount of revenue collected.

Uniform Tolls

With electronic technology, it seems to be technically feasible to charge tolls that vary essentially continuously over the rush hour period and to collect these tolls without disrupting traffic. But this does require investment and may not be an acceptable solution in many circumstances. Thus it is useful to consider the possibility of charging a single toll to all who use the bridge. Although such a toll would not affect the pattern of traffic queuing directly, we can show that if the number of commuters is responsive to the cost of commuting, then a uniform toll will in general reduce the number of commuters and in so doing improve efficiency.

Let us suppose that the number of commuters who demand to cross the bridge is determined by a downward-sloping "demand function $\Phi(p)$ where

p is the cost to a consumer of commuting, including displacement costs, queueing costs as well as any toll that is charged.

In equilibrium, without a toll, if there are N commuters, we must have $\Phi(p) = N$. From our earlier analysis, we also know that all individuals face commuting costs of $\beta N/2s$. We can find equilibrium with familiar-looking "supply and demand curves as in Figure 8.2. Equilibrium occurs where the curve $\Phi(p) = N$ cuts the line $p = \beta N/2s$, with the quantity N_0 shown in the figure.

When a uniform toll of τ is applied, the congestion patterns are the same as those without a toll. The only effect is to change the number of commuters. Thus when the number of commuters is N and the toll is τ, the cost of commuting is $p = \tau + \beta N/2s$. The equilibrium number of commuters is then the solution to the simultaneous equations $N = \Phi(p)$ and $p = \tau + \beta N/2s$. This occurs with the quantity $N(\tau)$ shown in the figure.

Figure 8.2: Demand and Supply for Commuting
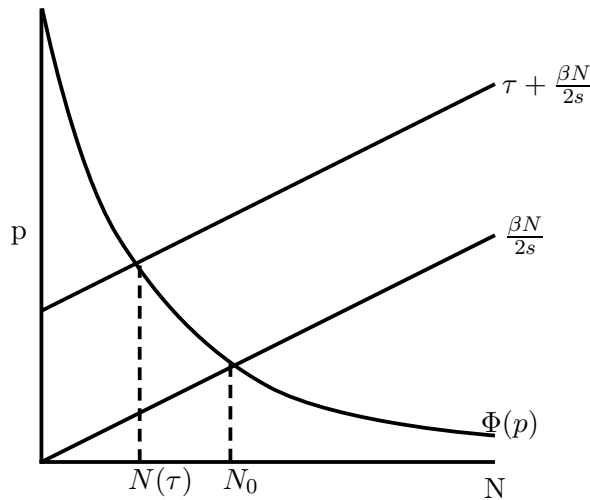
It is easy to show that if $\Phi(p)$ is a decreasing function, and if also $\Phi(0) > \tau$ and $\Phi(p)$ approaches zero for large p, then for each tax rate τ, there is exactly one solution for the equilibrium quantity amount of commuting. Let us denote this quantity by $N(\tau)$. It is also easy to show that $N'(\tau) < 0$ whenever the demand curve $\Phi(p)$ is downward sloping.

When the toll is set at τ, total commuting costs excluding taxes are

$\beta N(\tau)/2s$ for each commuter. Thus the sum of commuting costs for all commuters equals $\beta N(\tau)^2/2s$. The marginal increment to total costs caused by an extra commuter is therefore $\beta N(\tau)/s$. With tax rate τ, the private cost of commuting for any consumer is $\tau + \beta N(\tau)/2s$. The Samuelson efficiency conditions require that commuters travel only if they value the trip at least as highly as the total congestion costs it imposes. These conditions are satisfied if the toll is set so that after-tax cost to a commuter of making a trip equals its social cost. This condition can be stated as $\tau + \beta N(\tau)/2s = \beta N(\tau)/s$, or equivalently

$$\tau = \beta N(\tau)/2s. \tag{8.14}$$

This condition gives us a good idea of the appropriate magnitude for congestion tolls if the only tax instrument available is a uniform toll. From equation 8.14 it follows that the optimal toll rate is about as large as the displacement cost incurred by the first individuals to leave for work, which is also equivalent to the queuing cost borne by those who arrive at the preferred time t^*.

We should notice that the optimal uniform toll, unlike the optimal time-varying toll is not automatically a Pareto improvement over the no-toll outcome. In fact, if the revenue from the uniform toll were simply wasted, all commuters would be made worse off by the existence of this toll. This is apparent because in equilibrium with a positive uniform toll, the total cost of commuting is higher for every commuter than it is with no toll. (See Figure 8.2.) With the optimal toll in place, it is possible to work out a scheme of rebates of the tax revenue that makes every commuter better off than with no rebates, but this will not be the case in general for arbitrary use of the tax revenue.

Optimal Capacity, With and Without Tolls

The total cost of congestion at a bottleneck depends on the number of users, the pricing mechanism, and on the capacity of the bottleneck. Let's look into the optimal amount of capacity under alternative pricing institutions.

Let $C(s)$ be the cost of providing capacity s at the bottleneck. For a concrete example, we might assume that $C(s) = Ks^a$. This function is said to be homogeneous of degree a. There are increasing, decreasing, or constant returns to scale, depending on whether $a < 1$, $a > 1$, or $a = 1$. The amount of capacity will be efficient if the sum of costs is minimized where costs include the cost of congestion plus the cost of constructing capacity. Thus at optimal capacity, the marginal cost of capacity should be equal to minus the derivative of congestion costs with respect to capacity.

Optimal capacity with an optimal time-varying toll

We first consider the condition for optimal capacity that applies when an optimal time-varying toll is used. With the optimal toll, queuing costs are eliminated and if tax revenue is not wasted, the net social cost of congestion is $\beta N^2/4s$. Thus the efficiency condition becomes:

$$-\frac{d}{ds}\frac{\beta N^2}{4s} = \frac{\beta N^2}{4s^2} = C'(s). \tag{8.15}$$

which for the case of a cost function homogeneous of degree a implies:

$$\frac{\beta N^2}{4s} = aKs^a = aC(s) \tag{8.16}$$

Recall from 8.13 that the expression on the left side of 8.16 is equal to total revenue from the optimal time-varying tax. Thus we see that if there are constant returns to scale in the production of capacity, then the optimal time-varying tax would exactly pay for a bridge of optimal capacity. We also see that the optimal tax would collect more or less than the cost of optimal capacity depending on whether there are decreasing returns to scale ($a > 1$) or increasing returns to scale ($a < 1$).

Optimal Capacity with a uniform toll

In our analysis of the effects of a uniform toll, we showed that for a fixed bridge capacity if the number of commuters depends on the cost of commuting according to a demand function $\Phi(p)$, then the equilibrium number of commuters at toll τ is given by a decreasing function $N(\tau)$. If capacity is also variable, then we can show that the number of commuters will depend both on the capacity, s as well as tax rate τ. Thus we could define the equilibrium number of commuters as $N(s, \tau)$.

If the number of commuters is $N(s, \tau)$ then the equilibrium total commuting cost for each consumer is $\beta N(s, \tau)/2s$. If the number of commuters were to remain unchanged as capacity changes, then the marginal reduction in commuting costs for each current commuter resulting per small increase in capacity would be $\beta N(s, \tau)/2s^2$. This is the amount that each current commuter would be willing to pay for a marginal increase of capacity. Summed over all commuters we would have a total willingness to pay of $\beta N^2(s, \tau)/2s^2$.

Suppose that the cost of constructing capacity s is $C(s)$. Then the first-order condition for an efficient capacity (assuming that the number of commuters is held constant would be:

$$\frac{\beta N(s,\tau)^2}{2s^2} = C'(s).\tag{8.17}$$

With a homogeneous cost functions, it follows that Equation 8.17 is equivalent to

$$\frac{\beta N(s,\tau)^2}{2s} = aKs^a = aC(s)\tag{8.18}$$

This condition implies the interesting result that efficient expenditure on capacity on bridge capacity is greater than, equal to, or smaller than the total of displacement costs and queuing costs incurred by commuters depending on whether there are decreasing, constant, or increasing returns to scale in providing capacity. .

Since $N(s,\tau)$ is a decreasing function of τ, the solution of 8.17 for optimal s is greater if there is no toll than if there is a toll. But in this calculation, we have assumed that the number of commuters does not change with capacity. If there is no toll either before or after new construction, this assumption is not appropriate. Greater capacity attracts more users and this effect reduces the gains to each current user accordingly. Thus the optimal capacity with no tolls would be lower than that found by solving equation 8.17.

In the case where an optimal toll is charged, it is also true that increased capacity increases the number of users. But in this case, the extra congestion caused by these extra users is exactly balanced by the toll revenue collected from them. Thus the condition in equation 8.17 is the appropriate first order condition.

Transportation Network Problems

This section discusses the Downs-Knight-Pigou traffic paradox, drawing on an article written for a general audience by Arnott and Small [2].

It is important to recognize that the demand for travel on one roadway is strongly influenced by conditions on alternate roads. This sometimes leads to surprising results.

For example, suppose that there is a population of N workers who live on the opposite side of town from the factory where they work. As in our previous example, workers prefer to arrive at work at some time t^* and arriving earlier or later is costly to them There are two ways for workers to get from their homes to the factory. There is a direct route directly through the town and there is also a much longer, indirect route that circles the town. The direct route passes over a bridge with limited capacity s. If N_1

workers use the bridge, then as in our previous discussion, each of these workers bears a cost of $\beta N_1/2s$

The indirect route is a large, uncrowded freeway and the amount of time that it takes to get to work by this route is T_2 minutes, regardless of how many commuters use this route. Commuters who take this route can arrive at work just on time, but for them, the cost of spending T_2 minutes on the road is αT_2.

In equilibrium, commuters will be indifferent between the two routes. Thus the number of commuters who cross over the crowded bridge will be N_1 such that $\beta N_1/2s = \alpha T_2$ and hence

$$N_1 = 2s\frac{\alpha}{\beta}T_2.$$

Thus, in equilibrium nobody is better off than the people who use the ring road and indeed the existence of the shortcut over the bridge provides no net benefits to anyone. Moreover, so long as there are no restrictions on access, small increases in the capacity of the bridge will typically result in no net benefits. The newly expanded bridge will simply attract more users until using the bridge is no better than using the ring road. Of course if the capacity of the bridge is made large enough so that $\beta N/2s < \alpha T_2$, then nobody would use the ring road and further increases in bridge capacity actually reduce equilibrium travel costs.

If bridge access were managed to minimize total commuting costs, then it would have to be that the number N_1 of bridge users would minimize

$$N_1\frac{\beta N_1}{2s} + (N - N_1)\alpha T_2.$$

Applying calculus to this expression, we see that costs are minimized when

$$N_1 = s\frac{\alpha}{\beta}T_2.$$

Thus efficiency requires that the shortcut be used by only half as many commuters as the number who use it in equilibrium. This outcome could be enforced by an appropriate toll.

Exercises

8.1 Every morning, 6,000 commuters must travel from East Potato to West Potato. There are two ways to make the trip. One way is to drive straight across town, through the heart of Middle Potato. The other way is to take

the Beltline Freeway that circles the Potatoes. The Beltline Freeway is entirely uncongested, but the drive is roundabout and it takes 45 minutes to get from East Potato to West Potato by this means. The road through Middle Potato is much shorter and if it were uncongested it would take only 20 minutes to make the trip. Given the current size of the road through Middle Potato, if the number of commuters using the road is N_1, the time it takes to drive from East Potato to West Potato by this road is $20 + N_1/100$.

a). With the road at its current size and with no tolls, how many commuters will use the road through Middle Potato? What will be the total number of minutes per morning spent by all commuters going from East Potato to West Potato?

b). Suppose that a social planner controlled access to the road through Middle Potato and chose the number of commuters who were allowed to use the Middle Potato road in such a way as to minimize the sum of the number of minutes spent by commuters traveling from East Potato to West Potato. How many commuters would the social planner permit to use the Middle Potato road each morning? How long would it then take the commuters who used this road to make their morning commute?

c). Suppose that commuters value time saved from commuting at $\$w$ per minute. Suppose that the government charges a toll for driving on the road through Middle Potato and divides the revenue from the toll equally among all 6,000 commuters. If the government's objective is to minimize the total amount of time that people spend commuting, how high should it set the toll? How much revenue will it collect? With this policy, how much better off (evaluated in dollars) is each commuter than they were in the equilibrium with no tolls?

8.2 Suppose that a monopolist controls the road through Middle Potato and sets the toll that maximizes total revenue. What price will the monopolist set and how many people will use the road at this price? How does this price compare with the optimal toll?

    Hint: *First find the "demand curve" relating number of users to price. Then find the revenue-maximizing quantity and price.*

8.3 Suppose that the road through Middle Potato can be widened at a cost of $\$p$ per inch per day. If the road is widened by H inches and if N_1

people use the road, the time that it would take to drive to from East Potato to West Potato by this road would be $20 + N_1/(100 + H)$. Suppose that the government's objective function is to maximize the total value of time saved from commuting minus the sum of money spent on road building. Let w is the value per minute of time saved from commuting. Assuming that the government can not charge a toll for the road, or what values of the parameters p and w would it pay to widen the road at all. (Assume that there is no congestion for commuters coming home from work, only going to work.) If it pays to widen the road, how many commuters will be using the road when it is optimally widened? How long will it take them to make the trip, given that the road is optimally widened?

8.4 Suppose that as in the previous problem, the road through Middle Potato can be widened at a cost of $\$p$ per inch per day. Suppose also that the government can charge an efficient toll for using this road. What is the best combination of toll and highway expenditures? How does total toll revenue compare with the cost of building the highway? How does the optimal amount of highway expenditures compare with the optimal amount if the government cannot charge a toll? Hint: Think about two possible corner solutions.

8.5 Generalize the bottleneck problem so that the cost per minute of being late can take a different value γ from the cost per minute of being early. Show what happens in the limit as γ gets large.

8.6 Generalize the bottleneck problem so that the everyone has a utility $u(t)$ for arriving at work at time t where $u(t)$ is a fairly arbitrary single-peaked (or single-plateaued) function. (Assume that $u'(t) > -s\alpha$ for all relevant t.) Which results from our earlier discussion still apply and which do not?

8.7 What happens in the above generalization if $u'(t) < -s\alpha$

8.8 Generalize the model to allow two types of commuters with different preferred times of arrival.

Bibliography

[1] Richard Arnott, Andre De Palma, and Robin Lindsey. A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *American Economic Review*, 83(1):161–179, Maarch 1993.

[2] Richard Arnott and Kenneth Small. The economics of traffic congestion. *American Scientist*, 82(446-455), 1994.

[3] William Vickrey. Congestion theory and transport investment. *American Economic Review*, 59(2):251–260, May 1969.