the only strategy for each player were Not-Confess (an admittedly very simple game), then both players would be better off compared with the game where Confess is added as an option. (Indeed, that's why the police offer Confess as an option in the first place.)

Still, it's reasonable to view the analogous phenomenon at the heart of the Braess Paradox as more paradoxical, at an intuitive level. We all have an informal sense that "upgrading" a network has to be a good thing, and so it is surprising when it turns out to make things worse.

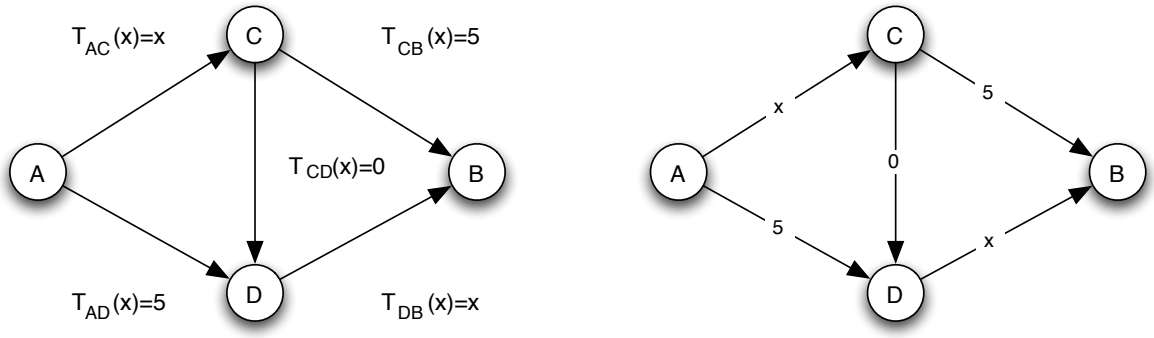
The example in this section is actually the starting point for a large body of work on game-theoretic analysis of network traffic. For example, we could ask how bad Braess's Paradox can be for networks in general: how much larger can the equilibrium travel time be after the addition of an edge, relative to what it was before? Suppose in particular that we allow the graph to be arbitrary, and we assume that the travel time on each edge depends in a linear way on the number of cars traversing it — that is, all travel times across edges have the form $ax + b$, where each of a and b is either 0 or a positive number. In this case, elegant results of Tim Roughgarden and Éva Tardos can be used to show that if we add edges to a network with an equilibrium pattern of traffic, there is always an equilibrium in the new network whose travel time is no more than 4/3 times as large [18, 353]. Moreover, 4/3 is the factor increase that we'd get in the example from Figures 8.1 and 8.2, if we replace the two travel times of 45 with 40. (In that case, the travel time at equilibrium would jump from 60 to 80 when we add the edge from C to D.) So the Roughgarden-Tardos result shows that this simple example is as bad as the Braess Paradox can get, in a quantitative sense, when edges respond linearly to traffic. (When edges can respond non-linearly, things can be much worse.)

There are many other types of questions that can be pursued as well. For example, we could think about ways of designing networks to prevent bad equilibria from arising, or to avoid bad equilibria through the judicious use of tolls on certain parts of the network. Many of these extensions, as well as others, are discussed by Tim Roughgarden in his book on game-theoretic models of network traffic [352].

8.3 Advanced Material: The Social Cost of Traffic at Equilibrium

The Braess Paradox is one aspect of a larger phenomenon, which is that network traffic at equilibrium may not be socially optimal. In this section, we try to quantify how *far* from optimal traffic can be at equilibrium.

We would like our analysis to apply to any network, and so we introduce the following general definitions. The network can be any directed graph. There is a set of drivers, and different drivers may have different starting points and destinations. Now, each edge e has



(a) *Travel times written as explicit functions of x.*

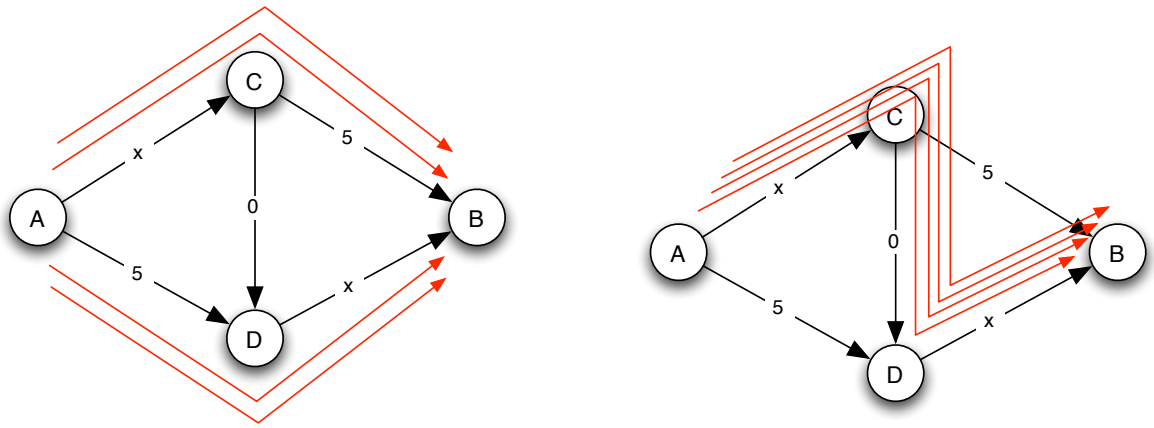
(b) *Travel times written as annotations on the edges.*

Figure 8.3: A network annotated with the travel-time function on each edge.

a *travel-time function* $T_e(x)$, which gives the time it takes all drivers to cross the edge when there are x drivers using it. These travel times are simply the functions that we drew as labels inside the edges in the figures in Section 8.1. We will assume that all travel-time functions are linear in the amount of traffic, so that $T_e(x) = a_e x + b_e$ for some choice of numbers a_e and b_e that are either positive or zero. For example, in Figure 8.3 we draw another network on which Braess's Paradox arises, with the travel-time functions scaled down to involve smaller numbers. The version of the drawing in Figure 8.3(a) has the travel-time functions explicitly written out, while the version of the drawing in Figure 8.3(b) has the travel-time functions written as labels inside the edges.

Finally, we say that a *traffic pattern* is simply a choice of a path by each driver, and the *social cost* of a given traffic pattern is the sum of the travel times incurred by all drivers when they use this traffic pattern. For example, Figure 8.4 shows two different traffic patterns on the network from Figure 8.3, when there are four drivers, each with starting node A and destination node B. The first of these traffic patterns, in Figure 8.4(a), achieves the minimum possible social cost — each driver requires 7 units of time to get to their destination, and so the social cost is 28. We will refer to such a traffic pattern, which achieves the minimum possible social cost, as *socially optimal*. (There are other traffic patterns on this network that also achieve a social cost of 28; that is, there are multiple traffic patterns for this network that are socially optimal.) Note that socially optimal traffic patterns are simply the social welfare maximizers of this traffic game, since the sum of the drivers' payoffs is the negative of the social cost. The second traffic pattern, Figure 8.4(b), is the unique Nash equilibrium, and it has a larger social cost of 32.

The main two questions we consider in the remainder of this chapter are the following. First, in any network (with linear travel-time functions), is there always an equilibrium traffic



(a) *The social optimum.* (b) *The Nash equilibrium.*

Figure 8.4: A version of Braess's Paradox: In the socially optimal traffic pattern (on the left), the social cost is 28, while in the unique Nash equilibrium (on the right), the social cost is 32.

pattern? We have seen examples in Chapter 6 of games where equilibria do not exist using pure strategies, and it is not *a priori* clear that they should always exist for the traffic game we've defined here. However, we will find in fact that equilibria always do exist. The second main question is whether there always exists an equilibrium traffic pattern whose social cost is not much more than the social optimum. We will find that this is in fact the case: we will show a result due to Roughgarden and Tardos that there is always an equilibrium whose social cost is at most *twice* that of the optimum [353].[2]

A. How to Find a Traffic Pattern at Equilibrium

We will prove that an equilibrium exists by analyzing the following procedure that explicitly searches for one. The procedure starts from any traffic pattern. If it is an equilibrium, we are done. Otherwise, there is at least one driver whose best response, given what everyone else is doing, is some alternate path providing a strictly lower travel time. We pick one such driver and have him switch to this alternate path. We now have a new traffic pattern and we again check whether it is an equilibrium — if it isn't, then we have some driver switch to his best response, and we continue in this fashion.

This procedure is called *best-response dynamics*, since it dynamically reconfigures the

[2]In fact, stronger results of Roughgarden and Tardos, supplemented by subsequent results of Anshelevich et al. [18], establish that in fact every equilibrium traffic pattern has social cost at most 4/3 times the optimum. (One can show that this implies their result on the Braess Paradox cited in the previous section — that with linear travel times, adding edges can't make things worse by a factor of more than 4/3.) However, since it is harder to prove the bound of 4/3, we limit ourselves here to proving the easier but weaker factor of 2 between the social optimum and some equilibrium traffic pattern.

players' strategies by constantly having some player perform his or her best response to the current situation. If the procedure ever stops, in a state where everyone is in fact playing their best response to the current situation, then we have an equilibrium. So the key is to show that in any instance of our traffic game, best-response dynamics must eventually stop at an equilibrium.

But why should it? Certainly for games that lack an equilibrium, best-response dynamics will run forever: for example, in the Matching Pennies game from Chapter 6, when only pure strategies are allowed, best-response dynamics will simply consist of the two players endlessly switching their strategies between H and T. It seems plausible that for some network, this could happen in the traffic game as well: one at a time, drivers shift their routes to ones that are better for them, thus increasing the delay for another driver who then switches and continues the cascade.

In fact, however, this cannot happen in the traffic game. We now show that best-response dynamics must always terminate in an equilibrium, thus proving not only that equilibria exist but also that they can be reached by a simple process in which drivers constantly update what they're doing according to best responses.

Analyzing Best-Response Dynamics Via Potential Energy. How should we go about proving that best-response dynamics must come to a halt? When you have a process that runs according to some set of instructions like, "Do the following ten things and then stop," it's generally obvious that it will eventually come to an end: the process essentially comes with its own guarantee of termination. But we have a process that runs according to a different kind of rule, one that says, "Keep doing something until a particular condition happens to hold." In this case, there is no *a priori* reason to believe it will ever stop.

In such cases, a useful analysis technique is to define some kind of *progress measure* that tracks the process as it operates, and to show that eventually enough "progress" will be made that the process must stop. For the traffic game, it's natural to think of the social cost of the current traffic pattern as a possible progress measure, but in fact the social cost is not so useful for this purpose. Some best-response updates by drivers can make the social cost better (for example, if a driver leaves a congested road for a relatively empty one), but others can make it worse (as in the sequence of best-response updates that shifts the traffic pattern from the social optimum to the inferior equilibrium in the Braess Paradox). So in general, as best-response dynamics runs, the social cost of the current traffic pattern can oscillate between going up and going down, and it's not clear how this is related to our progress toward an equilibrium.

Instead, we're going to define an alternate quantity that initially seems a bit mysterious. However, we will see that it has the property that it strictly decreases with each best-response update, and so it can be used to track the progress of best-response dynamics [303]. We will

refer to this quantity as the *potential energy* of a traffic pattern.

The potential energy of a traffic pattern is defined edge-by-edge, as follows. If an edge e currently has x drivers on it, then we define the potential energy of this edge to be

$$\text{Energy}(e) = T_e(1) + T_e(2) + \cdots + T_e(x).$$

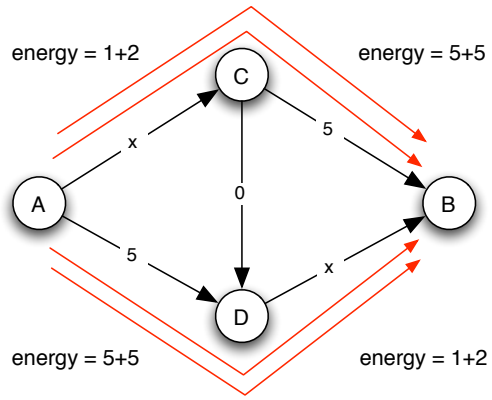
If an edge has no drivers on it, its potential energy is defined to be 0. The potential energy of a traffic pattern is then simply the sum of the potential energies of all the edges, with their current number of drivers in this traffic pattern. In Figure 8.5, we show the potential energy of each edge for the five traffic patterns that best-response dynamics produces as it moves from the social optimum to the unique equilibrium in the Braess-Paradox network from Figure 8.4.

Notice that the potential energy of an edge e with x drivers is not the total travel time experienced by the drivers that cross it. Since there are x drivers each experiencing a travel time of $T_e(x)$, their total travel time is $xT_e(x)$, which is a different number. The potential energy, instead, is a sort of "cumulative" quantity in which we imagine drivers crossing the edge one by one, and each driver only "feels" the delay caused by himself and the drivers crossing the edge in front of him.

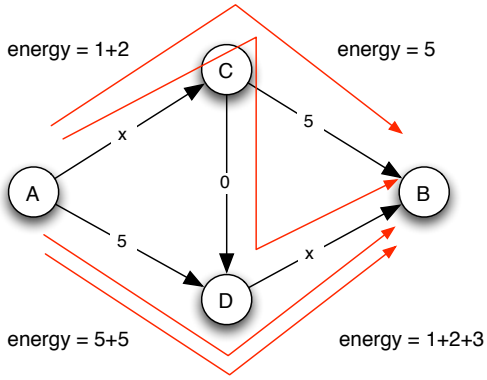
Of course, the potential energy is only useful for our purpose if it lets us analyze the progress of best-response dynamics. We show how to do this next.

Proving that Best-Response Dynamics Comes to an End. Our main claim is the following: each step of best-response dynamics causes the potential energy of the current traffic pattern to strictly decrease. Proving this will be enough to show that best-response dynamics must come to an end, for the following reason. The potential energy can only take a finite number of possible values — one for each possible traffic pattern. If it is strictly decreasing with each step of best-response dynamics, this means that it is "consuming" this finite supply of possible values, since it can never revisit a value once it drops below it. So best-response dynamics must come to a stop by the time the potential energy reaches its minimum possible value (if not sooner). And once best-response dynamics comes to a stop, we must be at an equilibrium — for otherwise, the dynamics would have a way to continue. Thus, showing that the potential energy strictly decreases in every step of best-response dynamics is enough to show the existence of an equilibrium traffic pattern.

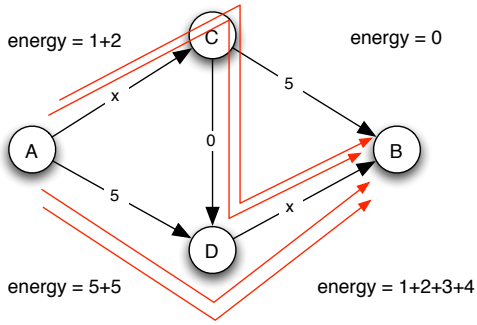
As an example, let's return to the sequence of best-response steps from Figure 8.5. Although the social cost is rising through the five traffic patterns (increasing from 28 to 32), the potential energy decreases strictly in each step (in the sequence 26, 24, 23, 21, 20). In fact, it is easy to track the change in potential energy through this sequence as follows. From one traffic pattern to the next, the only change is that one driver abandons his current path and switches to a new one. Suppose we really view this switch as a two-step process: first the



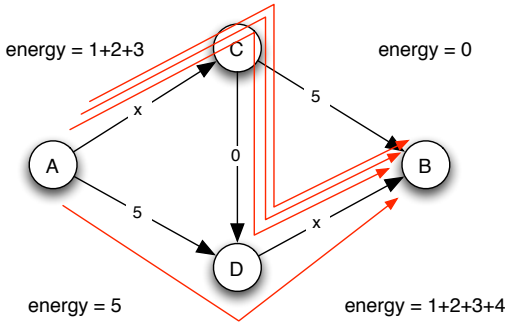
(a) *The initial traffic pattern. (Potential energy is 26.)*



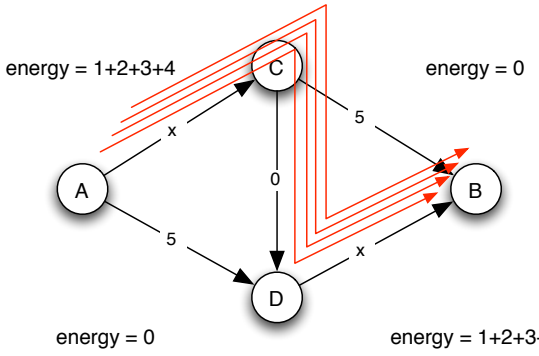
(b) *After one step of best-response dynamics. (Potential energy is 24.)*



(c) *After two steps. (Potential energy is 23.)*

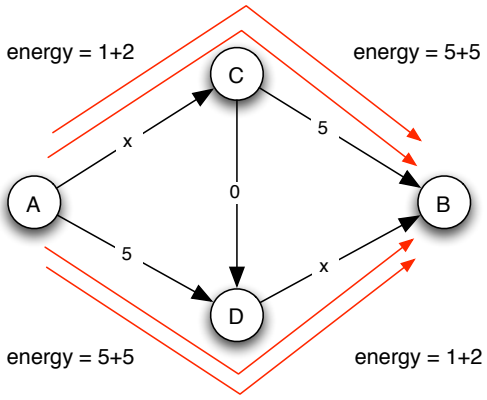


(d) *After three steps. (Potential energy is 21.)*

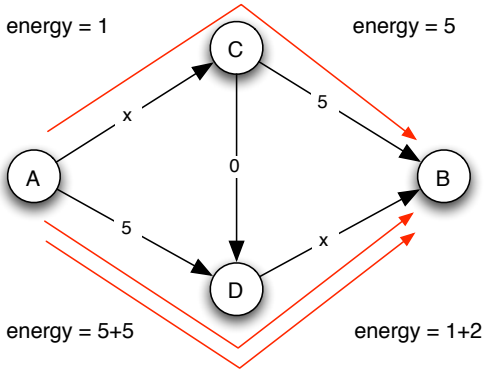


(e) *After four steps: Equilibrium is reached. (Potential energy is 20.)*

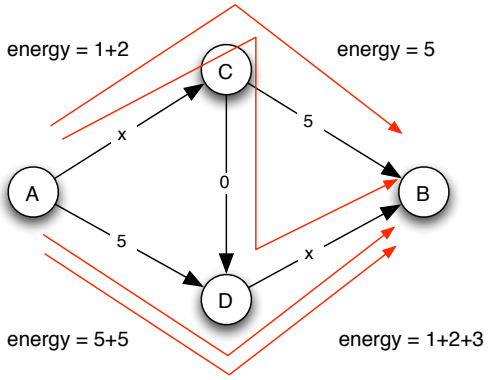
Figure 8.5: We can track the progress of best-response dynamics in the traffic game by watching how the potential energy changes.



(a) *The potential energy of a traffic pattern not in equilibrium.*



(b) *Potential energy is released when a driver abandons their current path.*



(c) *Potential energy is put back into the system when the driver chooses a new path.*

Figure 8.6: When a driver abandons one path in favor of another, the change in potential energy is exactly the improvement in the driver's travel time.

drivers abandons his current path, temporarily leaving the system; then, the driver returns to the system by adopting a new path. This first step releases potential energy as the driver leaves the system, and the second step adds potential energy as he re-joins. What's the net change?

For example, the transition from Figure 8.5(a) to 8.5(b) occurs because one driver abandons the upper path and adopts the zigzag path. As shown in Figure 8.6, abandoning the upper path releases $2 + 5 = 7$ units of potential energy, while adopting the zigzag path puts $2 + 0 + 3$ units of potential energy back into the system. The resulting change is a decrease of 2.

Notice that the decrease of 7 is simply the travel time the driver was experiencing on the path he abandoned, and the subsequent increase of 5 is the travel time the driver now experiences on the path he has adopted. This relationship is in fact true for any network and

any best response by a driver, and it holds for a simple reason. Specifically, the potential energy of edge e with x drivers is

$$T_e(1) + T_e(2) + \cdots + T_e(x-1) + T_e(x),$$

and when one of these drivers leaves it drops to

$$T_e(1) + T_e(2) + \cdots + T_e(x-1).$$

Hence the change in potential energy on edge e is $T_e(x)$, exactly the travel time that the driver was experiencing on e. Summing this over all edges used by the driver, we see that *the potential energy released when a driver abandons his current path is exactly equal to the travel time the driver was experiencing.* By the same reasoning, when a driver adopts a new path, the potential energy on each edge e he joins increases from

$$T_e(1) + T_e(2) + \cdots + T_e(x)$$

to

$$T_e(1) + T_e(2) + \cdots + T_e(x) + T_e(x+1),$$

and the increase of $T_e(x+1)$ is exactly the new travel time the driver experiences on this edge. Hence, *the potential energy added to the system when a driver adopts a new path is exactly equal to the travel time the driver now experiences.*

It follows when a driver switches paths, the net change in potential energy is simply his new travel time minus his old travel time. But in best-response dynamics, a driver only changes paths when it causes his travel time to decrease — so the change in potential energy is negative for any best-response move. This establishes what we wanted to show: that the potential energy in the system strictly decreases throughout best-response dynamics. As argued above, since the potential energy cannot decrease forever, best-response dynamics must therefore eventually come to an end, at a traffic pattern in equilibrium.

B. Comparing Equilibrium Traffic to the Social Optimum

Having shown that an equilibrium traffic pattern always exists, we now consider how its travel time compares to that of a socially optimal traffic pattern. We will see that the potential energy we've defined is very useful for making this comparison. The basic idea is to establish a relationship between the potential energy of an edge and the total travel time of all drivers crossing the edge. Once we do this, we will sum these two quantities over all the edges to compare travel times at equiibrium and at social optimality.

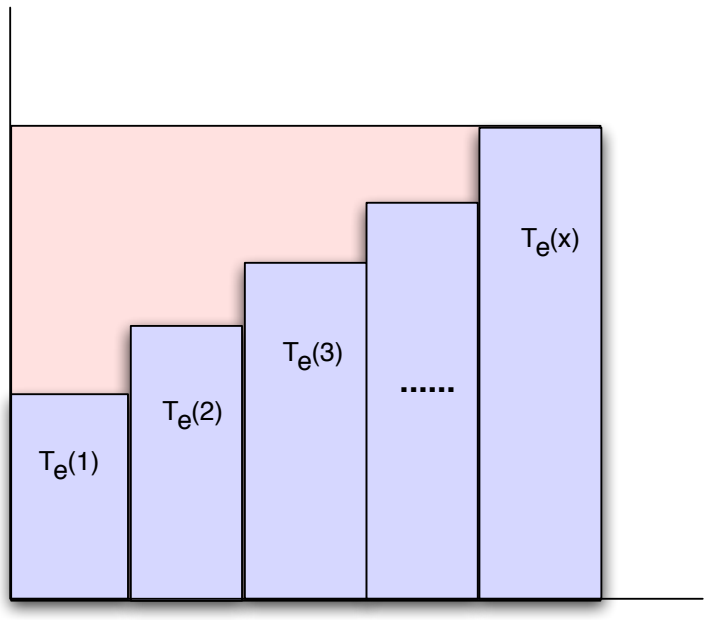


Figure 8.7: The potential energy is the area under the shaded rectangles; it is always at least half the total travel time, which is the area inside the enclosing rectangle.

Relating Potential Energy to Travel Time for a Single Edge. We denote the potential energy of an edge by Energy(e), and we recall that when there are x drivers, this potential energy is defined by

$$\text{Energy}(e) = T_e(1) + T_e(2) + \cdots + T_e(x).$$

On the other hand, each of the x drivers experiences a travel time of $T_e(x)$, and so the total travel time experienced by all drivers on the edge is

$$\text{Total-Travel-Time}(e) = xT_e(x).$$

For purposes of comparison with the potential energy, it is useful to write this as follows:

$$\text{Total-Travel-Time}(e) = \underbrace{T_e(x) + T_e(x) + \cdots + T_e(x)}_{x \text{ terms}}.$$

Since the potential energy and the total travel time each have x terms, but the terms in the latter expression are at least as large as the terms in the former, we have

$$\text{Energy}(e) \leq \text{Total-Travel-Time}(e).$$

Figure 8.7 shows how the potential energy and the total travel time compare when T_e is a linear function: the total travel time is the shaded area under the horizontal line with

y-value $T_e(x)$, while the potential energy is the total area under all the unit-width rectangles of heights $T_e(1), T_e(2), \ldots, T_e(x)$. As this figure makes clear geometrically, since T_e is a linear function, we have

$$T_e(1) + T_e(2) + \cdots + T_e(x) \geq \frac{1}{2} x T_e(x).$$

Alternately, we can see this by a bit of simple algebra, recalling that $T_e(x) = a_e x + b_e$:

$$\begin{aligned} T_e(1) + T_e(2) + \cdots + T_e(x) &= a_e(1 + 2 + \cdots + x) + b_e x \\ &= \frac{a_e x(x+1)}{2} + b_e x \\ &= x\left(\frac{a_e(x+1)}{2} + b_e\right) \\ &\geq \frac{1}{2} x(a_e x + b_e) \\ &= \frac{1}{2} x T_e(x). \end{aligned}$$

In terms of energies and total travel times, this says

$$\text{Energy}(e) \geq \frac{1}{2} \cdot \text{Total-Travel-Time}(e).$$

So the conclusion is that the potential energy of an edge is never far from the total travel time: it is sandwiched between the total travel time and half the total travel time.

Relating the Travel Time at Equilibrium and Social Optimality. We now use this relationship between potential energy and total travel to relate the equilibrium and socially optimal traffic patterns.

Let Z be a traffic pattern; we define Energy(Z) to be the total potential energy of all edges when drivers follow the traffic pattern Z. We write Social-Cost(Z) to denote the social cost of the traffic pattern; recall that this is the sum of the travel times experienced by all drivers. Equivalently, summing the social cost edge-by-edge, Social-Cost(Z) is the sum of the total travel times on all the edges. So applying our relationships between potential energy and travel time on an edge-by-edge basis, we see that the same relationships govern the potential energy and social cost of a traffic pattern:

$$\frac{1}{2} \cdot \text{Social-Cost}(Z) \leq \text{Energy}(Z) \leq \text{Social-Cost}(Z).$$

Now, suppose that we start from a socially optimal traffic pattern Z, and we then allow best-response dynamics to run until they stop at an equilibrium traffic pattern Z'. The social cost may start increasing as we run best-response dynamics, but the potential energy can only go down — and since the social cost can never be more than twice the potential energy, this shrinking potential energy keeps the social cost from ever getting more than twice as

high as where it started. This shows that the social cost of the equilibrium we reach is at most twice the cost of the social optimum we started with — hence there is an equilibrium with at most twice the socially optimal cost, as we wanted to show.

Let's write this argument out in terms of the inequalities on energies and social costs. First, we saw in the previous section that the potential energy decreases as best-response dynamics moves from Z to Z', and so

$$\text{Energy}(Z') \leq \text{Energy}(Z).$$

Second, the quantitative relationships between energies and social cost say that

$$\text{Social-Cost}(Z') \leq 2 \cdot \text{Energy}(Z')$$

and

$$\text{Energy}(Z) \leq \text{Social-Cost}(Z).$$

Now we just chain these inequalities together, concluding that

$$\text{Social-Cost}(Z') \leq 2 \cdot \text{Energy}(Z') \leq 2 \cdot \text{Energy}(Z) \leq 2 \cdot \text{Social-Cost}(Z).$$

Note that this really is the same argument that we made in words in the previous paragraph: the potential energy decreases during best-response dynamics, and this decrease prevents the social cost from every increasing by more than a factor of two.

Thus, tracking potential energy is not only useful for showing that best-response dynamics must reach an equilibrium; by relating this potential energy to the social cost, we can use it to put a bound on the social cost of the equilibrium that is reached.

8.4 Exercises

1. There are 1000 cars which must travel from town A to town B. There are two possible routes that each car can take: the upper route through town C or the lower route through town D. Let x be the number of cars traveling on the edge AC and let y be the number of cars traveling on the edge DB. The directed graph in Figure 8.8 indicates that travel time per car on edge AC is $x/100$ if x cars use edge AC, and similarly the travel time per car on edge DB is $y/100$ if y cars use edge DB. The travel time per car on each of edges CB and AD is 12 regardless of the number of cars on these edges. Each driver wants to select a route to minimize his travel time. The drivers make simultaneous choices.

   (a) Find Nash equilibrium values of x and y.

   (b) Now the government builds a new (one-way) road from town C to town D. The new road adds the path ACDB to the network. This new road from C to D has a travel