

Congestion Pricing Theory

Congestion pricing refers to pricing mechanisms designed to induce the economically-efficient use of congestible facilities. Examples of congestible facilities include highways, airport runways, shipping terminals, and the Internet. They are congestible in the sense that the costs faced by *each* user tend to increase with the number of users. As such, marginal usage costs can exceed average usage costs, resulting in *external* costs called *congestion externalities*. If users do not bear the external costs they generate then, in equilibrium, the facility's use will be inefficiently high and a deadweight loss will ensue. Congestion pricing prescribes fees that force users to internalize these externalities to some extent, thereby reducing or eliminating the deadweight loss.

Congestion-pricing principles are best illustrated by a common application: the use of *tolls* to manage highway traffic congestion. For example, consider an urban highway traveled by solo rush-hour commuters – a market for trips, v , both demanded and supplied by travelers. The *generalized price* of each trip is p , which includes travel-time and other non-pecuniary costs, and gives rise to an *inverse demand function* for trips, $p(v)$. The cost of *each* trip is $c = c(v/k)$, where k is the highway's capacity and (v/k) is the *volume-to-capacity* ratio; $\partial c / \partial v > 0$ indicates that the highway is congestible. The *total cost* of all trips is then $C = c \cdot v$. Accordingly, c is an *average cost* function, but it can be interpreted as *marginal private cost* because it gives the cost faced by each traveler when considering a trip. However, the *marginal social cost* of an additional trip is

$$MC = \frac{\partial C}{\partial v} = c + \frac{\partial c}{\partial v} \cdot v \quad (1)$$

comprising the private cost faced by the entering traveler (c), plus the increased cost imposed on all existing travelers ($\frac{\partial c}{\partial v} \cdot v$). This later cost is *external* to the entry decision and is thus called a *congestion externality*, measuring a gap between the private and social costs of travel. In equilibrium, entry will occur until

$$p(v) = c \quad (2)$$

but the *efficient* level of traffic is the solution to the net-benefit maximization problem

$$\text{Max}_v \int_0^v p(v') dv' - c \cdot v \quad (3)$$

yielding the first-order condition

$$p(v) = c + \frac{\partial c}{\partial v} \cdot v \quad (4)$$

such that the marginal benefit of the last trip taken equals its marginal social cost – including the external congestion cost it generates. As such, the equilibrium traffic level indicated by (2) will be inefficiently high. The efficient traffic level can be induced, however, by introducing a *congestion toll*, τ , levied on each traveler. The new equilibrium condition is then

$$p(v) = c + \tau \quad (5)$$

and it follows immediately from (4) and (5) that the optimal toll is

$$\tau = \frac{\partial c}{\partial v} \cdot v \quad (6)$$

which equals the congestion externality generated at the efficient traffic level (c.f. a *Pigouvian Tax*). This result is generalizable to a network of highways and multiple travel periods, and also to a variety of congestible facilities beyond highways. Note, however, that such tolls will not generally eliminate congestion externalities; they will instead reduce them to efficient levels.

Congestion-pricing policies often meet public opposition, partially due to the welfare losses they initially impose on travelers (including those priced off of the highway). The tolls levied on each traveler typically exceed the travel-cost savings they yield, implying that the bulk of the policy's net welfare gains take the form of toll revenues. As such, congestion pricing is only *Pareto Optimal* if these revenues are somehow returned to those it targets. One way to accomplish this is to invest in additional highway capacity; it can be shown that for travel-cost functions like $c = c(v/k)$ that tolls like (6) will generate just enough revenue to cover the cost of optimal capacity expansion. Another way is to use the revenues to reduce distortionary taxes in another market, thereby killing two deadweight losses with one toll (a “double dividend”).

The toll in (6) is a *first-best* toll because it achieves maximum welfare gains, which is only possible because it is derived without constraints. But such constraints exist in reality, such as an inability to toll some portion of a highway network. In such cases the goal is to derive a *second-best* toll by maximizing objective functions like (2) subject to these constraints. This second-best approach is particularly useful for analyzing real-world applications of congestion pricing, such as *cordon pricing* in Singapore, London, and proposed in New York, where tolls are charged only to enter a central business district, and *value pricing* such as tolled highway lanes in Southern California that run adjacent to non-tolled lanes.

Note that the rush-hour example above illustrates *static* congestion pricing, where travelers' departure times are treated as exogenous, implying a fixed commute-period duration. Endogenizing departure times gives rise to *dynamic* congestion pricing, which considers how tolls can influence departure times and, thus, the duration of the commute period. In a dynamic pricing framework it has been shown that optimal tolls can eliminate congestion (due to departure-time adjustments), and that the tolls equal the travel-cost savings they provide – an encouraging result in light of public opposition to congestion pricing.

References

Small, Kenneth, and Erik Verhoef (2007). *The Economics of Urban Transportation*. London and New York: Routledge.

Verhoef, Erik, Peter Nijkamp, and Piet Rietveld (1996). “Second-Best Congestion Pricing: the Case of an Untolled Alternative”, *Journal of Urban Economics* **40**, 279-302.

Author

Seiji S.C. Steimetz, Assistant Professor, California State University at Long Beach
 ssteimet@csulb.edu (WORDS: 795)