Journal of Public Economics 158 (2018) 113–125

Contents lists available at ScienceDirect

Journal of Public Economics

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

Pareto improvements from Lexus Lanes: The effects of pricing a portion of the lanes on congested highways☆

Jonathan D. Hall

University of Toronto, 150 St. George Street, Toronto, Ontario M5S 3G7, Canada

ARTICLE INFO

JEL classification:
D62
H41
R41
R48

Keywords:
Congestion pricing
Pareto improvement
Value pricing
Externality
Hypercongestion
Sufficient condition

ABSTRACT

Though economists have long advocated road pricing as an efficiency-enhancing solution to traffic congestion, it has rarely been implemented, primarily because it is thought to create losers as well as winners. This paper shows that a judiciously designed toll applied to a portion of the lanes of a highway can generate a Pareto improvement before using the revenue, a sufficient condition being that drivers with a high value of time travel at the peak of rush hour. I obtain these new theoretical results by extending a standard dynamic congestion model to reflect an important additional traffic externality: extra traffic does not simply increase travel times, but also introduces frictions that reduce throughput. The analysis draws attention to a practical policy that may help overcome the widespread opposition to road pricing.

1. Introduction

Traffic congestion is a major problem facing large cities worldwide. In the United States, for example, congestion consumes 42 hours per urban commuter annually—nearly an entire work week (Schrank et al., 2015), as well as imposing a host of other social costs.[1] At least since Pigou (1920), economists have advocated solving traffic congestion using tolls. Adding tolls would help drivers internalize the externalities they impose on others and would greatly increase social welfare. Yet tolling is rarely used in practice, in large part because of the received wisdom that such tolls impose losses on many, if not most, road users.[2] As Lindsey and Verhoef (2008) state, “most likely, these losses are the root of the longstanding opposition to congestion tolling in road transport,” a view echoed widely.[3]

A long literature has been concerned with the distributional consequences from tolling. Prior research has identified specific situations where it is possible for tolling to generate a Pareto improvement with homogeneous agents,[4] by using the toll revenue,[5] or charging negative tolls off-peak.[6] In practice, however, road users are not homogeneous, it is difficult to target the spending precisely enough to actually generate a Pareto improvement,[7] and charging negative tolls is impractical. The

☆ I am especially grateful for the guidance and support that Gary Becker and Eric Budish have given me. I am also grateful for helpful feedback from Richard Arnott, Jan Brueckner, Ian Fillmore, Mogens Fosgerau, Edward Glaeser, Brent Hickman, William Hubbard, Kory Kroft, Ethan Lieber, Robin Lindsey, Robert McMillan, Peter Morrow, John Panzar, Devin Pope, Mark Phillips, Allen Sanderson, Ken Small, Chad Syverson, George Tolley, Matt Turner, Vincent Van den Berg, Jos Van Ommeren, Clifford Winston, and Glen Weyl, as well as seminar audiences at the University of Chicago, Northwestern University, University of Toronto, Brigham Young University, Clemson University, Technical University of Denmark, Tinbergen Institute, RSAI, Kumho-Nectar, World Bank Conference on Transport and ICT, and the NBER Summer Institute. All remaining errors are my own.

E-mail address: jonathan.hall@utoronto.ca.

[1] Congestion also wasted 3.1 billion gallons of fuel in 2014 (Schrank et al., 2015), releasing an additional 28 million metric tons of carbon dioxide into the atmosphere. This additional pollution amounts to more than six times the annual emissions saved by the current fleet of hybrid and electric vehicles, and is responsible for up to 8600 preterm births a year (Currie and Walker, 2011). Congestion also retards economic growth; cutting congestion delay in half would raise employment growth by an estimated 1 % per year (Hymel, 2009).

[2] See Small and Verhoef's (2007) classic textbook for an explanation of this standard result (pp. 120–127) and see Appendix A for a brief discussion of other barriers to congestion pricing.

[3] For further examples, see Beesley (1973), Starkie (1986), Cohen (1987), Giuliano (1992), Arnott et al. (1994), Lave (1994), Small et al. (2005), and Small and Verhoef (2007).

[4] See Johnson (1964), De Meza and Gould (1987), Arnott and Inci (2010), Arnott (2013), and Fosgerau and Small (2013).

[5] See Foster (1975) and Arnott et al. (1994).

[6] See Braid (1996) and van den Berg and Verhoef (2011).

[7] Foster (1975) first noted the difficulty in targeting the spending. Small (1983, 1992) makes practical proposals regarding how to use the revenue to improve the distributional effects of congestion pricing but is careful to state that it is very unlikely that following his proposals would generate a Pareto improvement. In addition, even if we can design transfers that make a policy Pareto-improving, they can still be difficult to implement. As Stiglitz (1998) points out, the transfers are transparent and thus harder to defend than the implicit transfers the status quo entails; further, the government cannot commit to maintaining the transfers.

https://doi.org/10.1016/j.jpubeco.2018.01.003
Received 14 September 2017; Received in revised form 8 January 2018; Accepted 9 January 2018
0047-2727/

literature has also shown that pricing a portion of the lanes (a practice generally called "value pricing") reduces, but does not eliminate, the harm done by congestion pricing.[8] Additionally, Arnott et al. (1994) find that if only agents with the highest value of time (i.e., the rich) travel at the peak of rush hour, then it is possible to generate a Pareto improvement with heterogeneous agents. Nevertheless, as one of these authors argues, "economists came to appreciate…the practical impossibility of designing a tolling scheme that leaves everyone better off" (Lindsey, 2006, p. 332).

In this paper, I show it is possible to design such a tolling scheme. The main result of this paper is that a carefully designed, time-varying toll on a portion of the lanes of a highway can generate a Pareto improvement, even before the toll revenue is spent and even with realistic heterogeneity. I obtain this new result by extending the bottleneck congestion model of Vickrey (1969) and Arnott et al. (1993) to reflect an important additional traffic externality that has been identified by the transportation engineering literature but that has largely been ignored in the economics literature: not only does each additional vehicle slow others down, but in heavy enough traffic, additional vehicles can create frictions that reduce throughput (the number of trips per unit time).

By combining this empirically relevant additional externality with pricing a portion of the lanes we can obtain a Pareto improvement. The second externality pushes traffic off the production possibilities frontier (PPF). By using a time-varying toll to prevent the second externality from occurring, we can move back to the PPF, increasing both speeds and throughput (this is different than is typically assumed, and I explain how it is possible in Section 5). Were agents homogeneous, this would be enough to conclude everyone is better off.[9] However, road users are heterogeneous, and so pricing all the lanes likely hurts some of them. Adding tolls increases the financial costs of traveling while reducing the time costs. As not all road users value their time equally, this probably leaves some worse off.

Pricing a portion of the lanes can overcome these negative distributional effects. Adding tolls increases speeds and throughput in the priced lanes, allowing them to carry more traffic than they did before. This means the free lanes carry less traffic than before, improving travel times in the free lanes. Since travel times in the free lanes are lower, those who continue to use these lanes are better off. Those in the priced lanes could have stayed in the free lanes, and so by revealed preference are better off. We have generated a Pareto improvement, even without using the revenue.

Obtaining a Pareto improvement, however, often comes at a cost. By only pricing a portion of the lanes, we leave the other lanes congested, with all the resulting social costs. That said, inasmuch as generating a Pareto improvement makes it politically feasible to implement tolling, then we are trading potential, but unrealized, welfare gains for actual welfare gains.

My model focuses on two groups of agents: one rich, with high value of time, and the other poor, with low value of time. The groups also differ in the flexibility of their schedules and, within each group, desired arrival times are continuously distributed. Agents choose when to arrive at work and which route to take to minimize their total cost of traveling.

Using my model, I characterize the set of parameter values for which it is possible to generate a Pareto improvement without using the revenue. I find that pricing is more likely to generate a Pareto improvement the larger the reduction in throughput due to queuing, the greater the correlation between income and schedule inflexibility, and as income inequality decreases. Furthermore, I provide an intuitive sufficient condition for pricing a portion of the lanes to yield a Pareto improvement that holds even when there is an arbitrary number of groups: we simply need some rich drivers to be using the highway at the peak of rush hour. I then use the 2009 National Household Travel Survey to show that this sufficient condition is satisfied empirically—drivers with household incomes above $100,000 make up over 25% of those traveling at the peak. Finally, I characterize the trade-off between efficiency and equity by deriving the differences in social welfare gains and maximum harm done when pricing all the lanes or generating a Pareto improvement. I find that it is often the case that for parameter values such that pricing all of the lanes does the most harm, generating a Pareto improvement requires the least sacrifice of potential social welfare gains.

Central to the results of this paper is the notion that too many vehicles on the road reduces throughput. Over fifty years ago, Walters (1961) conjectured this additional externality existed, which Vickrey (1987) named "hypercongestion." As the theoretical arguments for hypercongestion did not carry over to dynamic models of highway congestion, the economics literature cast doubt on hypercongestion's existence.[10,11] Since then, the transportation engineering literature has identified the causal mechanisms and provided extensive empirical evidence that hypercongestion is an important real-world phenomenon—one that should be incorporated into models of congestion. I review this evidence in Section 2.

I build most closely on two important papers. First, in a valuable contribution, Arnott et al. (1994) extend the bottleneck model to allow agents to be heterogeneous. They work through several possible cases of heterogeneity, showing the distributional effects both before and after the revenue is either rebated lump sum or used to build new capacity. Most relevant to this paper, they show that when there are two groups with different values of time and schedule flexibility but homogeneous desired arrival time, pricing all of the lanes can help all agents before using the revenue if only those with a high value of time (i.e., the rich) travel at the peak of rush hour.[12]

I augment Arnott et al. (1994) by adding hypercongestion, value pricing, and heterogeneity in desired arrival times. I build on their results by showing how adding tolls can generate a Pareto improvement even when the poor travel at the peak of rush hour.

Second, an innovative paper by van den Berg and Verhoef (2011) extends Arnott et al. (1994) to allow agent preferences to vary continuously on two dimensions: value of time and schedule inflexibility. They show numerically that for intuitively reasonable parameter values, pricing all the lanes does not always hurt the majority of agents and that it is possible to generate a Pareto improvement by pricing a third of the lanes and forgoing revenue by charging a negative toll off-peak.

My model differs from van den Berg and Verhoef (2011) by adding hypercongestion and by allowing agent preferences to vary discretely along two dimensions (value of time and schedule inflexibility) and

[8] For example, see Braid (1996), Small et al. (2006), Light (2009), and van den Berg and Verhoef (2011). Light (2009) finds using the revenue to pay for new capacity generates a Pareto improvement while Braid (1996) and van den Berg and Verhoef (2011) find that charging a negative toll off-peak generates a Pareto improvement.

[9] Johnson (1964) and De Meza and Gould (1987) show this result in a static model, and Arnott and Inci (2010), Arnott (2013), and Fosgerau and Small (2013) derive it in dynamic models of downtown, rather than highway, congestion.

[10] The argument in favor of hypercongestion was that throughput (vehicles/hour) is the product of speed (miles/hour) times density (vehicles/mile), $T = S \times D$, and since speed is decreasing in density, $dS/dD < 0$, then it is possible for throughput to be decreasing in density, $dT/dD = D \cdot dS/dD + S \gtrless 0$ (e.g., Walters, 1961, Johnson, 1964, De Meza and Gould, 1987). The literature correctly responded that hypercongestion is a dynamic phenomenon, and showed that in dynamic models, the mathematical relationships above were not enough to generate hypercongestion (e.g., Newell, 1988; Evans, 1992; Verhoef, 1999, 2001, 2005; May et al., 2000; Small and Chu, 2003).

[11] There is some research arguing hypercongestion is possible for urban centers. See Small and Chu (2003), Arnott and Inci (2010) and Fosgerau and Small (2013).

[12] Vickrey (1973) derives a similar result also using the bottleneck model. He shows that if everyone is indifferent between all chosen arrival times (so we could assign arrival times so that only the rich are traveling at the peak), then adding tolls generates a Pareto improvement before using the revenue. The intuition for these results is that if the poor are willing to arrive at the worst equilibrium arrival time before a toll is added, then they can avoid any harm from tolling by choosing to do so once a toll is added.