Title
Distance-dependent Congestion Pricing for Downtown Zones

Authors
Daganzo, Carlos F
Lehe, Lewis J

Publication Date
2014-10-06

Distance-dependent congestion pricing for downtown zones *

by Carlos F. Daganzo and Lewis J. Lehe

October 6, 2014

Abstract

A growing literature exploits macroscopic theories of traffic to model congestion pricing policies in downtown zones. This study introduces trip length heterogeneity into this analysis and proposes a usage-based, time-varying congestion toll that alleviates congestion while prioritizing shorter trips. Unlike conventional trip-based tolls the scheme is intended to align the fees paid by drivers with the actual congestion damage they do, and to increase the toll's benefits as a result.

The scheme is intended to maximize the number of people that finish their trips close to their desired times. The usage-based toll is compared to a traditional, trip-based toll which neglects trip length. It is found that, like trip-based tolls, properly designed usage-based tolls alleviate congestion. But they reduce schedule delay more than trip-based tolls and do so with much smaller user fees. As a result usage-based tolls always leave those who pay with a large welfare gain. This may increase the tolls' political acceptability.

1 Introduction

The bulk of the theory of congestion pricing treats individual links of a road network. The advantage of a toll has historically been framed as either accounting for the externalities that users impose on each other in links and networks (Pigou, 1920; Beckman et al., 1956) or precluding wasteful queueing behind a bottleneck of fixed capacity (Vickrey, 1969). Reviews of both approaches can be found in Lindsey and Verhoef (2001). However, two developments have demanded a theory of congestion pricing as applied to entire zones of a road network.

The first is the practical difficulty of pricing individual links in dense urban networks. While dynamically-priced HOT lanes have been tried on particular American freeways, every city that has tried to price away downtown congestion has been forced to use cordon-

*Working Paper UCB-ITS-WP-2014-3. Publications in the Working Paper series are issued for discussion and are not considered final reports.

or area-based schemes[1] that price all trips equally. An exception might soon be found in ERP II, a GPS-based toll scheme under development in Singapore.

The second development is a macroscopic theory of traffic flow giving stable relationships between average traffic variables for street networks meeting certain conditions (Daganzo, 2005, 2007). The basic relationship underlying this theory, which has been called the "Macroscopic Fundamental Diagram" (MFD) in Daganzo and Geroliminis (2008), governs *flow:* under certain conditions, the average flow over all the network's links is a function of the average density in the network. The second macroscopic relationship in the theory, which in Gonzales and Daganzo (2012) is called the Network Exit Function (NEF), governs the *exit rate* (the rate at which trips are completed): if the average trip length is time-invariant, then by Little's formula, a rescaled version of the MFD yields the network's exit rate as a function of its vehicular accumulation. Daganzo (2007) points out that the relationship can be used for dynamic analyses if conditions change in time slowly compared with the relaxation time (the maximum duration of a commuter's trip). Geroliminis and Daganzo (2008) demonstrates the existence of a dynamic NEF in Yokohama, Japan using taxi and traffic data.

Since there are no routes in a macroscopic model—only arrival and exit times—a network governed by an MFD and NEF may be modeled as a simple, aspatial queueing system with familiar cumulative arrival and exit curves. Thus, a number of recent studies (Geroliminis and Levinson, 2009; Arnott, 2013; Fosgerau and Small, 2013) have used this zonal approach to model commuting choices and pricing in a downtown area.

So far, both the practice and the theory of zonal pricing have been somewhat coarse—notably neglecting heterogeneity in trip length among travelers. In practice, today's cordon schemes charge the same price to a vehicle which parks immediately upon crossing the cordon as to one which traverses the whole network. This lack of refinement is widely understood to create inequity as well as inefficiency: for example, a stated-preference survey in Holguín-Veras (2011) reveals that a proposed Manhattan cordon toll would affect delivery firms more as a lump-sum tax than a form of travel discouragement. Because trip-based tolls ("T-tolls"), rather than usage-based tolls ("U-tolls"), have prevailed in practice for technological reasons, modelers have also lacked an incentive to study the latter. This is no longer the case, however.

Accordingly, the goals of this paper are to design an efficient U-toll and to systematically compare it to an optimal T-toll, recognizing that user trip lengths vary. Tolling rules are sought that not only ease congestion, but also reorder trips so as to maximize the number of travelers that reach their destinations near their desired times. Section 2 introduces the problem and some modeling considerations. Section 3 derives the formula for the toll using a simple model of the network, and Sec. 4 evaluates it qualitatively and with simulations. To enhance realism, the simulation's physical parameters are chosen to roughly

[1] Although there are differences between cordon and area-based schemes, in this study we will use them interchangeably. See de Palma and Lindsey (2011) for a description of the types of road pricing now in practice.

approximate those of Yokohama, Japan.[2] It is found that, just like T-tolls, U-tolls reduce congestion. However, they do so with considerable savings in schedule penalties and much smaller toll payments.

2 Modeling Considerations

In Sec. 3 of this paper a traditional single-channel FIFO queuing model with a fixed service rate is used for design. It is later argued in Sec. 4 that the toll developed with this simple model also performs well with a slightly more realistic single-channel FIFO model in which the service rate is allowed to vary, depending on queue length as per an MFD. Furthermore, since it is known that this variable-rate FIFO model is a good approximation of reality if conditions change slowly with time, it is finally argued that the toll should also perform well in reality. This is verified with simulations using a realistic multi-channel non-FIFO queuing model. Section 2.1, below, discusses the three models that were just mentioned, and Sec. 2.2 the basic assumptions governing demand.

2.1 Queuing models

From now on it will be convenient to express the MFD as a function $V(n)$ giving the sum of the velocities of all vehicles in the network (veh-km-hr) for any given accumulation, n (veh). The variable V will be called from now on the "network total speed." Additionally, the term "exit," rather than the equivalent term"depart" from queueing theory, will be used for the act of completing a trip.

2.1.1 Multi-channel (MC) model

An MFD network is most realistically modeled as a non-FIFO multichannel $D/D/\infty$ queuing system with travelers as (discrete) customers with specific workloads, time-dependent arrivals and variable service rates. More specifically:

1. Customers' workloads, w, are the commuters' trip lengths, given in km.
2. There is an infinite number of servers, so that an arriving customer is immediately assigned a server—just as a commuter arriving in a zone immediately begins travel. Thus, the number of servers in action at once is the instantaneous vehicular accumulaton n.
3. Servers process customers' work at a rate equal to the average network speed, $v(n)$, which depends on the number of servers in action and is related to the MFD by: $v(n) = V(n)/n$.

[2] Geroliminis and Levinson (2009) suggested testing a zonal pricing scheme with an agent-based model, and Zheng et al. (2012) did it for a problem different from ours.

Given a list of commuters' arrival times and trips, the exit time of every commuter can be found with this model by evaluating, moment-by-moment, the accumulation n, the service rate $v(n)$ of the network and remaining workload of each commuter. Unfortunately, in this model a commuter's exit time depends on the history of antecedent arrivals, exits and trip lengths. Thus, the optimal feasible arrangement of commuters to arrival and exit times is complicated to find. Such an arrangement would also be impractical because it could not generally be achieved with a smooth toll. These concerns motivate a search for a simple yet efficient and practical toll. This will be done by working with the less realistic but much simpler model described below.

2.1.2 Single-channel, FIFO models with variable and fixed service rates (SCF-v and SCF-f)

The SCF-v model is rather conventional. It differs from the MC model in that it is single-channel, FIFO and treats customers as a fluid. More specifically, this model has the following properties:

1. Customers bring their workloads w as before, but now a single, very rapid, state-dependent server processes customers one at a time at rate equal to the total network speed, $V(n)$.

2. All customers in the system except the one being served are in a queue of length n with a First-In, First-Out (FIFO) discipline.

3. A commuter's trip time is his/her time in the system, including queuing for service, and not the commuter's service time.

This model behaves as if the servers of the MC model combined efforts to process customer workloads one at a time rather than in parallel. Although generally unrealistic, this approximation works relatively well if as stated in Daganzo (2007) "conditions change slowly compared with a trip time." These conditions refer to the accumulation $n(t)$ as well as the rates at which customers arrive and their trip lengths change over time.

The SCF-f model is identical in all respects to the SCF-v model, except the the service rate is fixed at some arbitrary value, V_r.

2.2 Demand

To design a toll scheme requires, of course, some picture of a "demand side." It is assumed that the system processes N customers with preassigned workloads. Customers individually and selfishly choose their arrival times to maximize utility.

The given workloads are modeled with distribution functions. Customers are numbered in order of increasing workload by an index i that ranges from 0 to N. Thus, a customer's

workload is related to its index by a non-decreasing function, $w(i)$, which is known. Associated with this distribution there are two cumulative distributions: (i) the cumulative workload of the first i customers, $W(i)$, and (ii) the cumulative workload of all customers with workloads less than or equal to w, denoted $\tilde{W}(w)$. Also used will be the area Ω under curve $W(i)$: $\Omega \doteq \int_0^N W(i)di$.

Arrival time choices are modeled assuming that customers have identical prefernces. Each customer is assumed to arrive at a time t_a of his/her choice and to exit at a time t_e dictated by the queuing mechanism. Customers share a common most desired time to exit the system (arrive at work), t_w. They choose t_a individually so as to minimize their user cost, which is the sum of the toll and a time-cost component c that includes queuing and schedule delay. The expression for this component, $c(t_a, t_e)$, is assumed to be the same for all commuters and be of a form that makes sense for a morning commute. The functional form in Vickrey (1969) is chosen:

$$c(t_a, t_e) = \alpha \cdot |t_a - t_e| + \max\left\{-\beta\left(t_e - t_w\right), \gamma\left(t_e - t_w\right)\right\}, \tag{1}$$

where the parameters α, β, and γ are, respectively, the per-time-unit cost of travel/queueing, earliness, and lateness. The second term is the schedule penalty function $P\left(t_e - t_w\right)$, which has a bilinear form. (For an evening commute, the penalty function would have as its argument $(t_a - t_w)$.)

From now on, without loss of generality, the origin of coordinates for the time variable is placed at $t = t_w$, so that $t_w = 0$, and we choose the monetary units so that $\alpha = 1$. Thus, (1) reduces to

$$c(t_a, t_e) = |t_a - t_e| + P(t_e), \quad \text{where} \quad P(t_e) \doteq \max\left\{-\beta t_e, \gamma t_e\right\}. \tag{2}$$

3 Toll design

The toll is designed with the SCF-f model. Section 3.1 defines the optimum travel pattern, and Section 3.2 presents the toll.

3.1 Least-Cost Arrangement (LCA)

The toll should maintain, as a Nash equilibrium, any least cost arrangement (LCA) of customer arrival and exit times. The only arrangements which are candidates to be the LCA are those that (i) occur in a single busy period and (ii) have no queuing, because queueing and idle periods can always be eliminated without adding any cost.[3] Arrangements of

[3]Idle periods between busy periods can be eliminated by time-shifting all the arrival and exit times on either or both sides by fixed amounts toward each other. This does not change the queuing time and always

this type are called "allowable," and denoted $\mathcal{A}$. Without queueing cost, the LCA is an allowable arrangement that minimizes the total schedule penalty.

Because the duration of the busy period for all allowable arrangement is $W(N)/V_r$, which is invariant, every allowable arrangement can be characterized by the beginning time of its busy period and the ordering of commuters within the period. It will be convenient to define said ordering in terms of increasing penalty rather than increasing time; i.e., with the convention that if the schedule penalty of customer i is greater than that of customer j, then $j > i$. The concept of ordering is defined recognizing that the indices i, j are real variables in $[0, N]$. Accordingly, an ordering $\mathcal{O}$ is a bijective, measure-conserving transformation of $[0, N]$ onto itself, and $\mathcal{O}(i)$ shall be the reordered index of customer i.[4]

The workload of the j-th customer in an ordering $\mathcal{O}$ shall be denoted $w(j|O)$. This function is related to the original ordering by $w(\mathcal{O}(i)|O) = w(i)$. The cumulative work distributions associated with the ordering $\mathcal{O}$ are denoted $W(j|\mathcal{O})$ and $\tilde{W}(w|\mathcal{O})$. Note that $W(j|\mathcal{O}) = \int_0^j w(i|\mathcal{O})di$, and $\tilde{W}(w|\mathcal{O}) = \tilde{W}(w)$.

An allowable arrangement $\mathcal{A}$ is also said to be "balanced" if the schedule penalties at the beginning and end of its busy period are the same. Since these busy periods have the same duration, $W(N)/V_r$, it follows that they must all begin at the same time and end at the same time. Thus balanced arrangements are completely characterized by their orderings. Furthermore, only balanced arrangements can be optimal.[5] In view of this, only balanced arrangements shall be considered from now on.

Now, let the schedule penalty of the i^{th} commuter in a balanced arrangement with ordering $\mathcal{O}$ be denoted $p\,(i|\mathcal{O})$. The following preliminary result can be stated.

Proposition 1. *In a balanced arrangement with ordering $\mathcal{O}$, customer i exits either at time $(W\,(i|\mathcal{O})\,/V_r)(-\beta/(\beta+\gamma))$, or at time $(W\,(i|\mathcal{O})\,/V_r)(\gamma/(\beta+\gamma))$. Furthermore, the schedule penalty associated with such customer is:*

$$p\,(i|\mathcal{O}) = (\theta/V_r)\cdot W(i|\mathcal{O}), \quad \text{where} \quad \theta \doteq \beta\gamma/\,(\beta+\gamma)\,. \tag{3}$$

can be done in a way that reduces schedule delay for every shifted customer. Thus, arrangements with idle times cannot be optimal. Arrangements with queuing cannot be optimal either because one can always hold the exit times constant and shift each arrival time to coincide with the prior customer's exit. This would reduce cost by eliminating all queuing without changing the schedule penalties. Thus, arrangements that include queuing cannot be optimal either.

[4] One way to construct some of these transformations is to partition $[0, N]$ into a finite (countable) number of subintervals. These intervals are then permutated without changing the ordering within the sets, and repositioned within $[0, N]$ in the new order so as to create a new partition. All the customer indices in each subinterval increase/decrease by the same amount. Thus, the transformation is defined by the original subintervals and the index-shift associated with each subinterval. The reader can verify that any measurable subset (e.g., interval) of $[0, N]$ is transformed into one with the same length measure (number of customers).

[5] If $\mathcal{A}$ is unbalanced then it is possible to move all the customers in an infinitessimal interval at the busy period's end that exhibits the largest penalty to the period's opposite end, and this would reduce total cost.

Proof. All the customers with equal of less work than i require $W(i|\mathcal{O})$ units of work, which must be processed in an interval of duration $W(i|\mathcal{O})/V_r$, positioned so that it includes the origin and such that the schedule penalty is the same at both its ends. Consideration shows that the times corresponding to these ends are $(W(i|\mathcal{O})/V_r)(-\beta/(\beta+\gamma))$ for the early end, and $(W(i|\mathcal{O})/V_r)(\gamma/(\beta+\gamma))$ for the late end. The reader can verify that the interval bracketed by these ends has the required duration and the same penalty (3) at both ends. □

Now let $\Omega(\mathcal{O})$ denote the area under the cumulative work curve for ordering $\mathcal{O}$: $\Omega(\mathcal{O}) \doteq \int_0^N W(i|\mathcal{O})\, di$. Note from (3) that the sum $S(\mathcal{O})$ of the schedule penalty across all customers in a balanced arrangement with ordering $\mathcal{O}$, $S(\mathcal{O}) \doteq \int_0^N p(i|\mathcal{O})\, di$, reduces to:

$$S(\mathcal{O}) \doteq (\theta/V_r) \cdot \Omega(\mathcal{O}) \tag{4}$$

This expression shows that any balanced arrangement whose ordering minimizes the area under its cumulative work curve, $\Omega(\mathcal{O})$, is an LCA. More specifically, the following can be stated.

Theorem 1. *An arrangement $\mathcal{A}$ is an LCA if and only if it is balanced and its ordering $\mathcal{O}$ is such that $W(i|\mathcal{O}) = W(i)$; i.e., such that commuters with less work suffer less schedule penalty. Furthermore, in an LCA customers with work w exit either at time $t_E(w) \doteq (\tilde{W}(w)/V_r)(-\beta/(\beta+\gamma))$, or at time $t_L(w) \doteq (\tilde{W}(w)/V_r)(\gamma/(\beta+\gamma))$.*

Proof. Since a balanced arrangement whose ordering has minimal area is an LCA, it suffices to show for the first part of the proof that the area, Ω, under $W(i)$ is minimal. This is true because $W(i)$ is by definition the sum of the work of the i customers with the least work. In other words, for any arbitrary commuter i: $W(i) \leq W(i|\mathcal{O}')$, $\forall \mathcal{O}'$, and this implies that $\int_0^N W(i)di \leq \int_0^N W(i|\mathcal{O}')di$, $\forall \mathcal{O}'$; i.e., that the area under $W(i)$ is minimal.

To prove the second part, consider a customer i with work $w = w(i)$; i.e., such that $W(i) = \tilde{W}(w)$. We have just seen that $W(i|\mathcal{O}) = W(i)$ for an optimum ordering. Thus, for this customer, and an optimum ordering: $W(i|\mathcal{O}) = W(i) = \tilde{W}(w)$. Consideration of these equalities and Proposition 1 reveals that if an ordering is optimum the customer must exit either at time $(\tilde{W}(w)/V_r)(-\beta/(\beta+\gamma))$, or at time $(\tilde{W}(w)/V_r)(\gamma/(\beta+\gamma))$, as claimed. □

The theorem establishes that the same orderings are optimal for all V_r. Furthermore, in view of (4), it also establishes that the optimal schedule delay is:

$$(\theta/V_r) \cdot \Omega. \tag{5}$$

3.2 Proposed toll

Let us now look for a toll scheme that achieves an LCA as a Nash equilibrium. As in the previous subsection, the service rate V_r is exogenously fixed.

The toll, τ, paid by a customer should depend on V_r and be adjusted across customers depending on their workloads w, and exit times, t_e. For this reason it will be denoted $\tau(t_e, w|V_r)$. The following formula is proposed:

$$\tau(t_e, w|V_r) = \max\left\{(\theta/V_r)\cdot\tilde{W}(w) - P\left(t_e\right), 0\right\}. \tag{6}$$

Recall that $\tilde{W}(w)$ is the combined workload of all vehicles with trip lengths less than w.

Figure 1 illustrates the expression. The first term of the non-zero entry in (6) is the height of the triangle in the figure. It gives the highest possible toll charged to a commuter traveling w distance units. This happens if the commuter exits at the ideal time $t_e = 0$. This maximum toll is a multiple of $\tilde{W}\left(w\right)/V_r$; i.e. of the time it takes to serve every customer with workload equal to or less than w. The toll drops linearly on both sides of the apex. Note that it reaches zero at the times $t_E\left(w\right)$ and $t_L\left(w\right)$ defined in Theorem 1.

It is now shown that this toll maintains the LCA as a Nash equilibrium. Note from the figure that the duration of the time interval with non-zero toll for customers with workload w is $\tilde{W}\left(w\right)/V_r$; i.e., the time it takes to process all the trips with length up to w. Note as well that the schedule penalty at both $t_E\left(w\right)$ and $t_L\left(w\right)$ is $(\theta/V_r)\tilde{W}(w)$, which is also the customer's total cost.

Theorem 2. *The toll given by (6) maintains the LCA as a Nash equilibrium.*

Proof. Consider an LCA and a customer with workload w. This person exits at either time $t_E(w)$ or $t_L(w)$, as per Theorem 1. As we have just seen such person suffers a total cost equal to the schedule penalty $(\theta/V_r)\tilde{W}(w)$. Thus, to show that the toll given by (6) preserves the LCA, it suffices to show that the cost experienced at any other exit time, $\$\left(t_e, w\right)$, is equal or greater than $(\theta/V_r)\tilde{W}(w)$. This is true, however, because:

$$\$\left(t_e, w\right) = P\left(t_e\right) + \tau\left(t_e, w|V_r\right) = \max\left\{(\theta/V_r)\tilde{W}\left(w\right), P\left(t_e\right)\right\} \geq (\theta/V_r)\tilde{W}\left(w\right).$$

□

Curiously, by exiting at either $t_E\left(w\right)$ or $t_L\left(w\right)$ a commuter with workload w avoids the toll. Since this is true for every w it follows that under the proposed tolling scheme no money changes hands and the agency collects no revenue. The proposed U-toll merely acts as a deterrent that encourages people to sort themselves out in the optimum way. This is of course true only in the fictitious SCF world, but as Section 4 below shows it is approximately true in the more realistic MC world.

Curiously, too, the reader can verify that the smallest possible Vickrey T-toll, which does not tax the travelers exiting at the beginning and end of the rush, happens to be equal to $\tau\left(t_e, w_{\max}|V_r\right)$ – i.e., to the U-toll presented to the person with the longest trip.

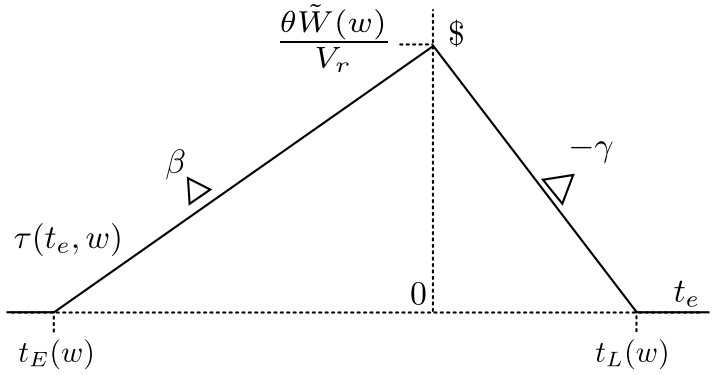


Figure 1: Toll structure.

This means that the proposed U-toll is smaller than the T-toll for all other trips, and that the shorter the trip the larger the discount; i.e. that the U-toll can be advertised as a conventional T-toll with a distance-based discount for short trips.

4 Toll performance

Section 3 established that if users are prearranged in an LCA form and then the proposed toll preserves the pattern as a Nash equilibrium in the SCF-f world. Qualitative arguments in Sec. 4.1 suggest that, at least in some circumstances, the toll continues to work well in the SCF-v world, and under tighter conditions even in the MC world. An agent-based microsimulation, described in Lehe and Daganzo (2014), confirms that this is true in Sec. 4.2. The simulation is also used to compare the performance of the U- and T-tolls.

4.1 System behavior under the toll in more realistic settings

Of practical interest is the impact of the toll not just on the customer arrangements with a fixed service rate but also on the rate itself when the rate is allowed to vary endogenously – depending on both the MFD relation, $V_r = V(n_r)$ and the customer choices. It is first argued below, that the proposed toll can still approximately preserve an LCA equilibrium in the SCF-v world. In this equilibrium $n(t) \approx n_r$ for most of the rush, and $V(t) = V(n(t)) \approx V_r$. It will then be argued that the tollit should also work well in the MC world.

One key assumption in the arguments is that $N >> n_r$. This is needed to ensure that the transient periods at the beginning and end of the rush, when $n(t)$ is not close to n_r, do not take up a significant part of the rush, and that $w(i)$ changes little over any interval of length n_r.

To see how the equilibrium comes to be in the SCF-v world, consider a standard FIFO queuing diagram of cumulative customer number vs. time with the LCA exit curve, $E(t)$,

and an arrival curve that satisfies $A(t) = n_r + E(t)$ during the middle of the rush. Since $w(i)$ changes little over ranges of i compared with n_r the workloads of the customers in queue at every t are very close to some value $w(t)$, which itself changes slowly with t. That is, the horizontal separation between the cumulative curves changes slowly with time. This means that the system is in a quasi-steady state in which any customer arriving or exiting at any time t in the middle of the rush experiences a time in the system that is close to the slow-varying quantity $w(t)/V_r$.

Thus, if the customer exiting at t considers other nearby times to exit (s)he will find that his/her time in the system would barely change. Thus, his/her choice will be mostly dictated by the schedule and toll costs. Since these costs are identical to those in the idealized case of Sec. 3.2 because the exit curve has not changed, the customer would find that his/her best choice is very close to his/her current exit time. Customers might move, but so little that the arrival curve should still satisfy $n(t) \approx n_r$ and the resulting customer arrangement after the moves should now be much closer to equilibrium. The final equilibrium pattern after a few adjustments should be very similar to the starting pattern. This suggests that the proposed toll can probably maintain the identified equlibrium LCA pattern without exogenously controlling for V_r.

The resulting total cost at the tolled equilibrium, $\$(n_r)$, is approximately the sum of the queueing delay, which is the area between the customer arrival and exit curves, and the optimual schedule delay (5). Since the arrival and exit curves are vertically separated by $n(t) \approx n_r$ customers most of the time, and since the rush lasts for approximately $W(N)/V(n_r)$ time units, the area between the curves is approximately $W(N)n_r/V(n_r)$. Thus the total cost is,

$$\$\left(n_r\right) \approx \frac{W(N)n_r}{V\left(n_r\right)} + \frac{\theta\Omega}{V\left(n_r\right)}. \tag{7}$$

Since V_r can be freely chosen in the recipe for the toll (6), one should choose it to be consistent with the value of n_r, denoted n^*, that minimizes $\$(n_r)$. Equation (7) shows that $\$(n_r)$ increases if $n_r \geq n_0$ (the value that maximizes $V\left(n\right)$). Thus the least cost is achieved for $n^* \leq n_0$ and $V^* \doteq V\left(n^*\right)$. Hence, if V^* is used in (6) to define the toll then an equilibrium queue with the optimum n^* can be expected for most of the rush.[6]

This qualitative discussion is no proof, of course, but it suggests that the proposed U-toll may, under some conditions, achieve approximately the dual purpose of metering the exit rate at an optimal level V^* and also reordering customers optimally. The proposed toll may even work in the more realistic MC world, in those cases where conditions change so slowly with time that the SCF model is a good approximation for the MC model.

Unfortunately, in many real cases the duration of people's trips may not be negligible compared with the duration of the rush, so it may not be reasonable to expect $N >> n^*$

[6] Although there is an $n_r \geq n_0$ that would also satisfy $V^* = V(n_r)$, this value cannot emerge spontaneusly as an equilibrium because the queue dynamics for $n_r \geq n_0$ are unstable – small perturbations from the steady state would grow and destroy the steady state.

and some of the other requirements to be satisfied. Furthermore, the arguments above only speak to the preservation of a tolled equilibrium: they do not suggest that the toll could drive the system toward an equilibrium starting from an arbitrary arrangement of trips. For these reasons, the following section tests the proposed U-toll in a MC scenario that closely resembles the conditions of a real city, starting from disequilibrium.

4.2 Experimental verification

The tests here will mimic to the extent possible the situation in Yokohama (Japan) circa 2000 because this city exhibits a reproducible MFD; see Geroliminis and Daganzo (2008). An approximate formula for the NEF in this city has been given in Daganzo et al. (2012). The formula used in this section for the MFD, $V(n)$, is a scaled up version of the NEF formula – scaled up by the 2.3 km average trip length observed in Yokohama. The result, which is linearly extrapolated for the range of vehicle accumulations not observed in Yokohama ($n > 14,000$) is:

$$V(n)\ \text{(veh-km/hr)} = 2.3 \cdot \begin{cases} 2.28 \cdot 10^{-8} n^3 - 8.62 \cdot 10^{-4} n^2 + 9.58n & \text{for } n < 14,000\ \text{(veh)} \\ 27,731 - 1.4n & \text{for } n \geq 14,000\ \text{(veh)} \end{cases} \tag{8}$$

Figure 2 shows the graphs for both $V(n)$ and $v(n) = V(n)/n$.

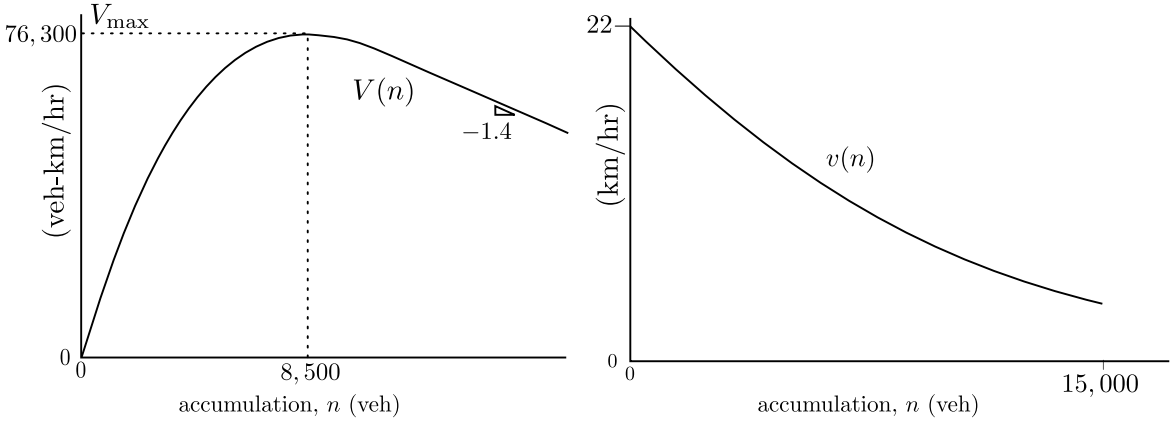


Figure 2: Macroscopic relations for simulation

As suggested in Small (1982), the earliness and lateness penalties are set to $\beta = 0.5$ and $\gamma = 2$. Four demand scenarios were considered with two demand magnitudes ($N = 65,000$ and $135,000$) and two levels of trip length variability around the observed mean (2.3 km). Trip lengths were taken to be uniformly distributed with a 2 km range for the low variability case and 3 km for the high variability case.

Each simulation started with a random distribution of trip arrival times for all commuters and was iterated over many "days." At the end of each day, a random, small set of drivers was chosen and they were allowed to switch arrival times so as to reduce their cost. The ensuing morning rush was then simulated with the changed arrival pattern using the MC framework.

It was found that in all cases tested if one sets $V_r = V_{\max}$ in (6) then the U-toll encourages users to arrive so close together and trip lengths to vary so rapidly that accumulations above the "optimum" n_0 occur in a way that invariably triggers catastrophic gridlock. So the U-toll cannot even eliminate queueing. Thus, the maximum network productivity, $V_{\max}$, is unattainable with the proposed U-toll.

For this reason values of $V_r = V(n_r) < V_{\max}$ were also tested. It was found that in all four scenarios the system tends toward a stable pattern resembling the LCA if $V_r/V_{\max} < 0.85$. In all cases tested, the stable pattern exhibited accumulations on the uncongested side of the MFD, $n_r < n_0$, in agreement with the theoretical arguments of Sec. 3.3. Details of the dynamics before convergence can be found in Lehe and Daganzo (2014).

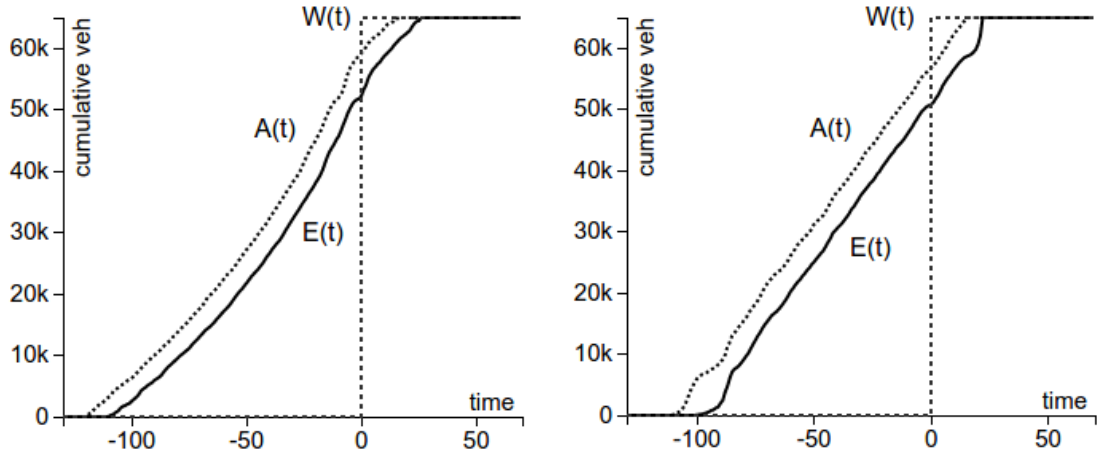
Figure 3 shows the final quasi-equilibrium patterns achieved with the two toll types for the low magnitude/low variability scenario. In order to compare apples to apples, the figure uses separate, experimentally determined optimum values, $n_r = n^*$ and $V_r = V^*$, for each toll.

The figure depicts the cumulative curves of (non-FIFO) arrivals and exits over time. The area between these curves represents the total time in the system for all commuters. The areas between the exit curve and the vertical axis are the total earliness and lateness experienced. Note how the exit curve of the U-toll is concave toward this axis, as this reduces the penalty cost. The maximum exit rate with the U-toll occurs for $t \approx 0$, since these are the times when travelers with the shortest trips are served. The savings afforded by the U-toll are considerable: it reduces schedule cost by about 13% and the total cost by about 11%. These reductions are more pronounced when the demand is high.

Figure 4 compares the two toll types on the four scenarios along four different dimensions: the social cost including travel and schedule delay, the schedule penalty, the travel cost and the toll paid. All costs are expressed in equivalent minutes of travel time per commuter. The results consistently favor the U-toll and are largely consistent with what was expected. Note in particular the remarkably low user fees collected with the U-toll. This suggests that the SCF-f model was a reasonable tool to develop the toll.

5 Conclusion

This study compared usage-based tolls to trip-based tolls, which charge travelers of all trip lengths the same price, for a morning commute problem with commuters whose trip lengths vary. Two aspatial models of network performance were used: a realistic but complex multi-channel model and a rough but more tractable single-channel model. The



(a) U-Toll (b) T-toll

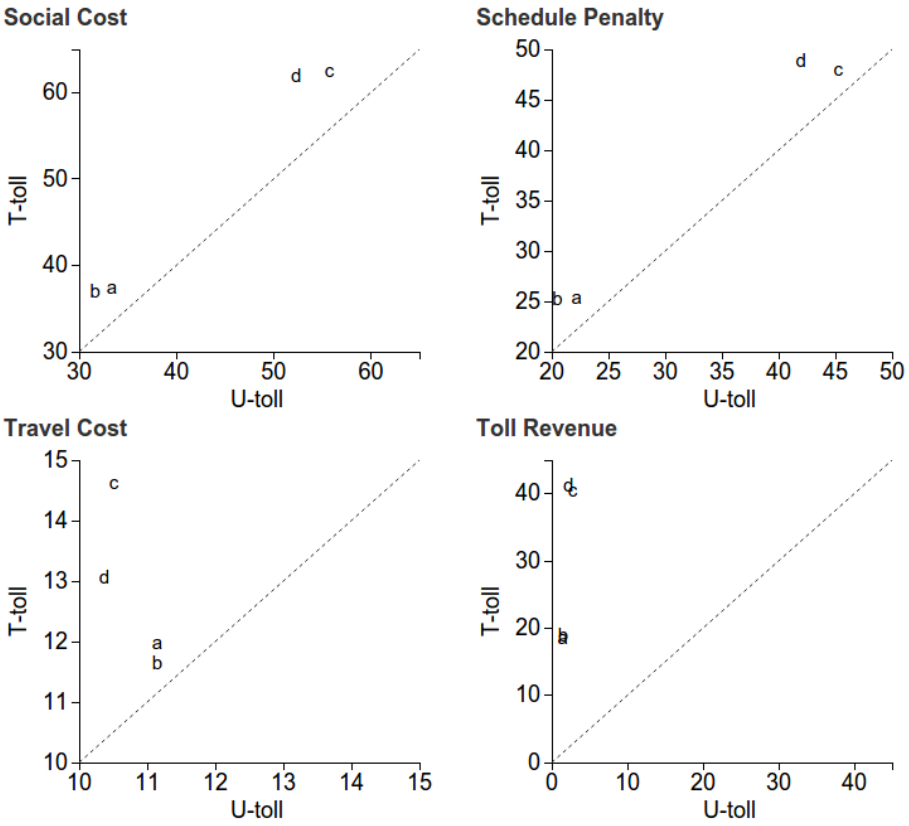
Figure 3: Exit curves from simulation

latter was used to design a usage-based toll, while the former was used to test both tolls in a realistic environment. It was found that, relative to the trip-based toll, the usage-based toll lowered social cost modestly and user cost significantly. Furthermore, it raised relatively little revenue and involved much smaller fees. Thus itcan be expected to increase user welfare considerably more.

The results of this paper can be extended to the evening commute. The optimum U-toll with a controlled exit rate in the SCF world would still be given by (6) of Section 3. This happens because in the optimum LCA with controlled exit rates there are no queues so that $t_a = t_e$. The only difference is that in the application with endogenous exit rates (and queues) t_a would have to be substituted for t_e, and this would matter in this case. Much of the qualitative argumentation in Sec. 3.3, however, can be repeated to suggest that this type of toll would also work well in the evening commute, and continue to be superior to T-tolls.

Whether the suggested usage-based tolls are of purely theoretical interest depends on their political feasibility. A chief difficulty in this regard is the tolls' non-intuitive dependence on trip length. Thus, an interesting avenue of research would be to evaluate more intuitive but sub-optimal U-tolls. One remedy, currently under investigation, is to charge drivers a fee per veh-km traveled that only depends on the time of day, and to use the same toll schedule for the morning and the evening.

On the other hand, the proposed U-tolls offset this problem with two political advantages. First, it allows those who are tolled to experience a considerable gain in welfare, since the toll only involves tiny fees. And second, it preserves privacy, since the toll is aspatial and only requires knowing the times at which each trip begins and ends.



Symbol	Magnitude	Variability
a	low	low
b	low	high
c	high	low
d	high	high

Figure 4: Cost comparisons.

Acknowledgement

Rsearch supported by a research grant from UC-Connect.

References

Arnott, R. (2013). A bathtub model of downtown traffic congestion. *Journal of Urban Economics*, 76:110–121.

Beckman, M., McGuire, C., and Winsten, C. B. (1956). *Studies in the Economics of Transportation*. Yale, New Yaven.

Daganzo, C. F. (2005). Improving city mobility through gridlock control: an approach and some ideas.

Daganzo, C. F. (2007). Urban gridlock: Macroscopic modeling and mitigation approaches. *Transportation Research Part B: Methodological*, 41(1):49–62.

Daganzo, C. F., Gayah, V. V., and Gonzales, E. J. (2012). The potential of parsimonious models for understanding large scale transportation systems and answering big picture questions. *EURO Journal on Transportation and Logistics*, 1(1-2):47–65.

Daganzo, C. F. and Geroliminis, N. (2008). An analytical approximation for the macroscopic fundamental diagram of urban traffic.

de Palma, A. and Lindsey, R. (2011). Traffic congestion pricing methodologies and technologies. *Transportation Research Part C: Emerging Technologies*, 19(6):1377–1399.

Fosgerau, M. and Small, K. a. (2013). Hypercongestion in downtown metropolis. *Journal of Urban Economics*, 76:122–134.

Geroliminis, N. and Daganzo, C. F. (2008). Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological*, 42(9):759–770.

Geroliminis, N. and Levinson, D. M. (2009). Cordon Pricing Consistent with the Physics of Overcrowding. *Transportation and Traffic Theory 2009: Golden Jubilee*, pages 219–240\r740.

Gonzales, E. J. and Daganzo, C. F. (2012). Morning commute with competing modes and distributed demand: User equilibrium, system optimum, and pricing. *Transportation Research Part B: Methodological*, 46(10):1519–1534.

Holguín-Veras, J. (2011). Urban delivery industry response to cordon pricing, time/distance pricing, and carrier/receiver policies in competitive markets. *Transportation Research Part A: Policy and Practice*, 45(8):802–824.

Lehe, L. and Daganzo (2014). Aspatial Simulation of Trip-timing in a Downtown Network (draft).

Lindsey, R. and Verhoef, E. T. (2001). Traffic congestion and congestion pricing. In Button, K. J. and Henscher, D. A., editors, *Handbook of Transport Systems and Traffic Control*, pages 77–105. Elsevier Science, Oxford.

Pigou, A. (1920). *The Economics of Welfare*. Macmillan, London.

Small, K. (1982). The scheduling of consumer activities: work trips. *The American Economic Review*, 72(3):467–479.

Vickrey, W. S. (1969). Congestion Theory and Transport Investment. *The American Economic Review*, 59(2):pp. 251–260.

Zheng, N., Waraich, R. A., Axhausen, K. W., and Geroliminis, N. (2012). A dynamic cordon pricing scheme combining the Macroscopic Fundamental Diagram and an agent-based traffic model.