Coarse tolling with heterogeneous preferences

Version of 21 August 2013

Vincent A.C. van den Berg[#]
Department of Spatial Economics
VU University Amsterdam
De Boelelaan 1105
1081 HV, Amsterdam, The Netherlands
v.a.c.vanden.berg@vu.nl

Abstract

This paper considers coarse tolling of congestion under heterogeneous preferences, and especially the distributional effects of such tolls. With coarse tolling, the toll equals a fixed value during the centre of the peak; outside this period, it is zero. This paper investigates three dimensions of heterogeneity. With the first, all values of time and schedule delay vary in fixed proportions, and this heterogeneity may stem from income differences. The second has differences in flexibility of users when to arrive. The third captures differences in willingness to arrive before or after the preferred arrival time. The paper uses three models of coarse tolling: the Laih, ADL, and Braking model.

All three dimensions affect the welfare gain of coarse tolling. In the Laih model, the generalised price with coarse tolling is in between the no-toll and first-best one. In the other models this is not so, and distributional effects may be non-monotonic and very different from the first-best toll's effects. In the Braking model, the capacity goes unused for some time during the tolled period. Compared with in the Laih model, this raises total cost, and it is most harmful for users with low values and difficulty to arrive late: e.g. low-income users with a strict work start time or a trip to the doctor.

Keywords: Coarse tolling, heterogeneous preferences, distributional effects, bottleneck model, proportional heterogeneity
JEL codes: D62, H23, R41, R48

[#] Affiliated to the Tinbergen Institute, Gustav Mahlerplein 117, 1082 MS Amsterdam

1. **Introduction**

This paper concerns step tolling in the bottleneck model with heterogeneous preferences, and especially the distributional effects. There is a large literature on fully-time-variant tolling, but in practice tolls are either constant over the day (e.g. London) or at most have a few steps in them (e.g. Singapore, Stockholm, the SR91 and Bay Bridge in California, and the SR520 and SR16 bridges in Washington State). We focus on single-step "coarse tolling" under three separate dimensions of preference heterogeneity. We only consider heterogeneity in values of time (α), schedule delay early (β) and/or schedule delay late (γ); all other user characteristics are homogeneous and demand is fixed. Preference heterogeneity is certainly present in reality (e.g. Small et al., 2005), and affects the welfare gain of tolling and leads to distributional effects (e.g. Arnott et al., 1988, 1994).

Although the economic case for congestion pricing is strong, applications remain scarce. A reason for resistance seems to be the redistributive effects. When tolling reduces travel time and increases monetary cost, a user is better off if her value of time is higher. This is often seen as meaning that the poor lose and the rich gain (e.g. Layard, 1977). Yet, things need not be so simple when considering a more realistic case with multiple dimensions of heterogeneity and dynamic congestion. Van den Berg and Verhoef (2011a) find that, in the bottleneck model under proportional and α heterogeneity, first-best tolling can *lower* the "generalised price" (i.e. toll plus travel costs, and henceforth "price" for brevity) for most users. Moreover, it is not the users with the lowest values who lose most, but those with intermediate values and strong inflexibility in when to arrive. We find that the distribution effects with coarse tolling may differ strongly from those with first-best tolling for each dimension of heterogeneity.

"Proportional heterogeneity" was introduced by Vickrey (1973). Under his definition, all three values vary proportionally: $\alpha_i=\mu\cdot\beta_i$ and $\gamma_i=\eta\cdot\beta_i$, where μ and η are homogeneous. Proportional heterogeneity could stem from income differences where a higher income increases all values proportionally, as all three values depend on the inverse of the marginal utility of income, which decreases with income. As Van den Berg and Verhoef (2011a) show, all users are *better off* with first-best (FB) tolling than without tolling, except those with lowest values who are unaffected. This also means that the gain of FB tolling increases with the degree of proportional heterogeneity.

Our second form of heterogeneity measures differences in the importance of travel time versus schedule delays, or alternatively how flexible users are in when to arrive. We denote such heterogeneity as "α heterogeneity", as the value of time (α) varies relative to the values of schedule delay. It could result from differences in type of job, family status, or trip purpose. Travel time costs are reduced by α heterogeneity, and thus the gain of FB tolling decreases with the degree of α heterogeneity. Now, all users—except for those with the highest value—lose due to

FB tolling. Arnott et al. (1988, 1994), Lindsey (2004a), and Van den Berg and Verhoef (2011ab) used this heterogeneity among others.

Our final form of heterogeneity captures differences in the willingness to arrive before or after the preferred arrival time, and could stem from job or trip type differences. We refer to it as "γ heterogeneity" as the value of schedule delay late (γ) varies relative to value of schedule delay early. With or without tolling, low-γ users arrive after the preferred arrival time, and a user arrives later the lower her value of schedule delay is. This self-selection lowers the gain from first-best tolling, but leads to no distributional effects from FB tolling (Arnott et al., 1988, 1994).

We consider three models of coarse tolling that differ in how the price is equalised before and after the toll is turned off. The "Laih model" of Laih (1994, 2004) has separate queues for tolled and untolled users. The "ADL model" of Arnott et al. (1990, 1993) has a mass departure. The "Braking model" of Lindsey et al. (2012) and Xiao et al. (2012) takes into account that users who would pass the tolling point just before the toll is lifted have an incentive to "brake" and delay passage until the toll is turned off. The braking means that capacity goes unused for some time during the peak, and this raises costs and lowers the gain of tolling.

Whereas with FB tolling distributional effects are monotonic, with coarse tolling this is not always so and the effects can very different from the first-best ones. We find that with proportional heterogeneity in the braking model, coarse tolling raises the price most for the users with intermediate values who are indifferent between the tolled and untolled periods. Untolled users are better off, the lower their values are. Tolled users are better off the higher their values are, and those with the higher values may gain. Xiao et al. (2011) study coarse tolling in the ADL model under proportional heterogeneity, and find that it lowers the price for all, and more so the higher a user's values are. With γ heterogeneity and the ADL model, coarse tolling lowers all prices; but most for mass users and especially for those with an intermediate γ. With α heterogeneity, coarse tolling raises the price more the higher the value of time is: in the Laih model the price increases are half that of FB tolling, the braking model has higher increases than the Laih model, and the ADL model lower ones. If braking can be prevented, coarse tolling lowers the price for most users if the degree of α heterogeneity is low enough compared to the other dimensions of heterogeneity.

The gain from tolling decreases with the degree of α and γ heterogeneity. Proportional heterogeneity raises the gain from tolling, and tends to make coarse tolling fare better compared with the FB toll. In the ADL model, the gain is higher than in the Laih model, in the braking model, it is lower; but both differences decrease with the degree of proportional heterogeneity.

The next section introduces the notation and the no-toll and first-best equilibria. Section 3 turns to coarse tolling under homogeneity. The three following sections separately study coarse tolling under our three dimension of heterogeneity. Section 7 discusses some caveats and directions for research. Section 8 concludes.

Table 1: Parameter description and (mean) value in the numerical example

Symbol	Description	(Mean) value in the example
	Possibly heterogeneous preferences	
α	Value of time	8
β	Value of schedule delay early	4
y	Value of schedule delay late	15.6
$\mu\equiv \alpha/\beta$	Relative size of the value of time to value of schedule delay early	2
$\eta\equiv \eta/\beta$	Relative size of the values of schedule delay late to early	3.9
	Parameters describing the distribution of the heterogeneous value *x*={*α, β, y, η, μ*}	
$f[x]$	PDF of the heterogeneous value x={$\alpha, \beta, y, \eta, \mu$}	$1/(\overline{x}-\underline{x})$
$F[x]$	CDF of the heterogeneous value x={$\alpha, \beta, y, \eta, \mu$}	$(x_i-\underline{x})/(\overline{x}-\underline{x})$
$n_j\equiv f[x_j]\cdot N$	Density of users with a x_j	-
n_j^k	Density of users with a x_j in arrival period k, the sum over all periods is n_i	-
$\underline{x}$	Minimum of the heterogeneous value x={$\alpha, \beta, y, \eta, \mu$}	-
$\overline{x}$	Maximum of the heterogeneous value x={$\alpha, \beta, y, \eta, \mu$}	-
$E[x]$	Mean of heterogeneous preference x={$\alpha, \beta, y, \eta, \mu$}	-
	Other parameters	
t^*	Preferred arrival time (which is normalised to 0)	0 (also in the analytical models)
N	Number of users	9000
s	Capacity of the bottleneck	3600

Table 2: Variables

Symbol	Description
	Timings
t	Arrival time
t_s	Start of the peak
t_e	End of the peak
t^+	Start tolled period
t^-	End tolled period
t^b	Moment the braking starts
$\Delta t\equiv t^- - t^b$	Time-span when the capacity goes unused in the braking model
	Prices and tolls
ρ	Level of the coarse toll
$\tau[t]$	The level of the toll (either first-best or coarse) for an arrival at t
P_i	(Generalised) price for users with a value of x_i
	Aggregate measures and indicators
V	Number of untolled users
M	Number of users in the mass
T	Indicator for the Tolled period (i.e. from t^+ to t^-)
UE	Indicator for the Untolled period before t^* (Untolled Early)
UL	Indicator for the Untolled period after t^* (Untolled Late)
$E[P]=N\int_{\underline{x}}^{\overline{x}} P_j \cdot f[x_j]dx_j$	Average generalised price
$TP=N\cdot E[P]$	Total price
$TR=\rho(N-V)$	Toll revenue
$TC=TP-TR$	Total cost

2. Set-up and the no-toll and first-best equilibria

This section introduces the set-up and notation. It also shortly reintroduces the no-toll and first-best equilibria, as these equilibria are well covered by earlier works such as Vickrey (1969, 1973) and Arnott et al. (1988, 1993, 1994), and textbooks such as Verhoef and Small (2007). Table 1 summarises the parameters, and Table 2 the variables. We use the point-queue bottleneck model of Vickrey (1969). Without a queue and as long as the arrival rate of users at the bottleneck is not above capacity, s, travel time is zero (and thus free-flow travel time is zero). Otherwise, travel time equals the length (in vehicles) of the queue when joining it divided by capacity.

A driver of type i faces two travel costs. The first, travel time cost, equals travel time multiplied by i's value of time, α_i. The second, schedule delay cost, equals the absolute difference between the arrival time, t, and the preferred arrival time, t^*, multiplied by the value of schedule delay early (β_i) or late (γ_i) depending on if she arrives before or after t^*.[1] For notational ease t^* is normalises to zero. When a user of type i arrives at t, she has a generalised price (P_i and hereafter referred to as price) that is the sum of the travel time and schedule delay costs and possible toll, τ:

$$P_i[t] \equiv \alpha_i \cdot T[t] + \tau[t] + \begin{cases} \beta_i \cdot \left(t^* - t\right) & \text{if } t \le t^* \\ \gamma_i \cdot \left(t - t^*\right) & \text{if } t > t^* \end{cases}. \tag{1}$$

This toll can be the first-best toll or the coarse toll. Square brackets indicate that something is a function of what is listed inside. Round brackets are used for arithmetic. The peak starts at t_s with a zero queue length, and ends at t_e when the queue has fully dissipated.

2.1. Homogeneous preferences

In user-equilibrium with homogeneous preferences, the price is constant over time during the peak, as otherwise some drivers would want to change their arrival time. Without tolling, this is achieved by a travel time that grows linearly over arrival time by β/α for arrivals before t^* and thereafter shrinks by $-\gamma/\alpha$, such that the sum of travel time and schedule delay costs is constant over arrival time. The no-toll (NT) equilibrium price is:

$$P = \beta \frac{\eta}{1+\eta} \frac{N}{s} = \delta \frac{N}{s}; \tag{2}$$

where $\delta \equiv \beta \cdot \eta/(1+\eta)$ and $\eta \equiv \gamma/\beta$. Further, N is the total number of uses and s the capacity of the bottleneck. The NT total cost is N times this price: $TC = \delta \cdot N^2/s.$

Travel time due to queuing is a pure loss: all queuing could be removed without increasing schedule delays by having the departure rate equal capacity. This is attained by a toll that varies over time such that sum of toll and schedule delay cost is constant, which implies that the first-best (FB) toll equals the NT travel time costs. As the FB toll exactly replaces the NT travel-time cost at all t, prices remain the same, but average travel cost and thus total cost are halved.

[1] The value of time is the marginal utility of travel time savings divided by the marginal utility of income; values of schedule delay early and late are defined analogously. It is assumed that $\alpha_i > \beta_i > 0$ and $\gamma_i > 0$, as otherwise the standard NT equilibrium of the bottleneck model does not hold. This is an assumption used in the entire bottleneck literature and is also needed in other congestion models.

2.2. Proportional heterogeneity

We now turn to the “proportional heterogeneity” of Vickrey (1973), which varies the values in fixed proportions: $\alpha_i=\mu\cdot\beta_i$ and $\gamma_i=\eta\cdot\beta_i$. We will refer to a type as having a certain β_i, where a type indicates all users with a certain set of values. The values follow a distribution function of $f[\beta_i]$, CDF of $F[\beta_i]$, and minimum and maximum of respectively $\underline{\beta}$ and $\bar{\beta}$. As discussed, this heterogeneity could stem from income differences. Although, income differences might also cause heterogeneity in the ratios of the values: e.g. rich people might be more flexible, and thus have low values of schedule delay relatively to their value of time (see also Koster and Koster, 2013).

Without tolling, travel times follow the same pattern as with homogeneity. This is because the ratios α/β and α/γ are the same for all and these ratios determine the arrival order of users, as the ratios measure how willing you are to reduce travel times by increasing schedule delays. With our other forms of heterogeneity, these ratios differ over types, and there is separation over time.

Queuing is again a pure loss, and the first-best toll eliminates it. Types now arrive ordered on their β_i. The type with the highest values arrives at t^*, as it is most willing to pay a toll to attain a lower schedule delay. The lowest-β type arrives the furthest from t^* at the start (t_s) and end (t_e) of the peak. The first-best (FB) toll thus fully separates the types, and not only removes the queuing but also reduces total schedule delay cost.

The NT-equilibrium price is similar to with homogeneity:

$$P_i=\beta_i\frac{\eta}{1+\eta}\frac{N}{s}\cdot \tag{3}$$

FB tolling removes all queuing, and now users self-select to an arrival time and this lowers total scheduling costs. The gain from this self-ordering increases with the degree of heterogeneity (where we define such an increase throughout as an increase of the variance for a given mean and shape of the distribution). The FB price is (see Van den Berg and Verhoef, 2011a):

$$p_i^{FB}=\frac{\beta_i}{s}\cdot\frac{\eta}{1+\eta}\left(\int_{\beta_i}^{\bar{\beta}}n_j\mathrm{d}\beta_j+\int_{\underline{\beta}}^{\beta_i}n_j\cdot\frac{\beta_j}{\beta_i}\mathrm{d}\beta_j\right)\equiv\frac{\beta_i}{s}\cdot\frac{\eta}{1+\eta}N\left(\int_{\beta_i}^{\bar{\beta}}f_j\,\mathrm{d}\beta_j+\int_{\underline{\beta}}^{\beta_i}f_j\cdot\frac{\beta_j}{\beta_i}\mathrm{d}\beta_j\right); \tag{4}$$

where $n_j\equiv f_j\cdot N$ is the density of users with β_j. FB tolling lowers the price for all but the lowest-β users, and more so the more heterogeneity there is. Accordingly, the gain of FB tolling increases with the degree of proportional heterogeneity.

2.3. α heterogeneity

Now the value of time, α_i, varies while the other values are fixed. However, what really matters is that the implied ratio $\alpha_i/\beta\equiv\mu_i$ varies. Users with a high ratio are less willing to queue or alternatively more flexible when to arrive, as a higher travel time is relatively more costly for them than a lower schedule delays. This heterogeneity was studied by Arnott et al. (1988, 1994) and

Van den Berg and Verhoef (2011b); Arnott and Kraus (1995) and Van den Berg and Verhoef (2011a) combine proportional and α heterogeneity; Newell (1987), Lindsey (2004a), de Palma and Lindsey (2002) and Hall (2013) look at more general heterogeneity.

The higher a user's α_i is relative to the other values, the less queuing she causes and the lower her congestion externality. The NT price concavely increases with α_i, and a highest-α user faces the same price as with homogeneity (Van den Berg and Verhoef, 2011b):

$$P_i^{NT} = C_{SD} + C_{TT} = \frac{\delta}{s} N\left(\int_{\underline{\alpha}}^{\alpha_i} n_j d\alpha_j + \alpha_i \int_{\alpha_i}^{\bar{\alpha}} (n_j / \alpha_j)\, d\alpha_j \right). \qquad (5)$$

The α heterogeneity has no effect on the first-best (FB) equilibrium, as queuing is still eliminated. The FB price is the sum of the schedule delay cost and toll:

$$P_i^{FB} = C_{SD} + toll = \frac{\delta}{s} N. \qquad (6)$$

This price is for all types the same as the price for the highest-α type in the NT situation, and hence the price increases for all (but the highest-α type), and more so the lower α_i is. Accordingly, the FB gain decreases with the degree of α heterogeneity.

2.4. γ heterogeneity

With our third form of heterogeneity, users differ in their value of schedule delay late γ_i, while the other values are the same for all. Again, what matters is not γ_i itself, but the ratio $\gamma_i/\beta=\eta_i$. Users with a high η_i (i.e. with a η_i above the indifferent η_1^*) arrive before t^*. As the values of time and schedule delay early are the same for all, these high-η types travel jointly and the price is the same for all these types. Users with a low η_i arrive after t^*, and the lower a user's η_i is, the further she arrive from t^*. The price of late users increases concavely with $\gamma_i=\eta_i{\cdot}\beta$. The self-ordering lowers total scheduling cost, and does so equally in the NT and FB cases; it also lowers travel time cost. The FB toll exactly the replaces travel time cost, and thus prices are unaffected by FB tolling.

It can be shown that NT and FB prices follow:

$$\begin{aligned} &P_i = \beta \cdot \frac{N}{s}(1-F[\eta^*]), && \eta_i \geq \eta_1^*, \\ &P_i = \beta \cdot \frac{N}{s}\left(1-F[\eta^*] - \int_{\eta_i}^{\eta^*} f[\eta_j]\left(\eta_j - \eta_i\right) d\eta_j\right), && \eta_i < \eta_1^*. \end{aligned} \qquad (7)$$

For a general distribution, there is no closed-form solution for η_1^*. But it is known that all types gain from γ heterogeneity: for each type *i* the self-ordering lowers the price compared to when all users would have the same value as *i*. Total cost is also lower (see Arnott et al., 1988).

FB tolling removes all queuing, but does not affect the prices. Thus γ heterogeneity does not lead to distributional effects. It does tend to lower a FB gain by lowering total travel time, as this FB gain equals the total NT travel time cost. The average NT travel time cost tends to decreases with the degree of heterogeneity by making the travel time function after t^* more convex.[2]

$$E[c_{TT}]=\frac{\beta}{2}\cdot\frac{N}{s}\left(1-F[\eta^*]\right)^2+F[\eta^*]\beta\frac{N}{s}\left(\int_{\underline{\eta}}^{\eta^*}\left(\int_{\underline{\eta}}^{\eta_i}f[\eta_j]\cdot\eta_j d\eta_j\right)f[\eta_i]d\eta_j\right); \qquad (8)$$

3. Coarse tolling under homogeneous preferences

With coarse tolling, the toll is on during the centre peak between t^+ and t^-, and equals the fixed ρ. Outside this period the toll is zero. We focus on coarse tolling under heterogeneous preferences, and will only summarise the results under homogeneity here, and Appendix A gives a short derivation (for details see Arnott et al. (1990, 1993), Laih (1994, 2004), and Lindsey et al. (2012)).

The coarse models only differ on what happens during the period after the toll is lifted at t^-. In the Laih (1994, 2004) model,[3] there are separate queues for toll payers and untolled-late users who will arrive after the tolled period. By assumption, these queues do not to interact. The first untolled-late user waits before the tolling point for such a time that her waiting-time cost equals the toll paid by the last tolled user. In the Laih model, coarse tolling removes half of the total travel time that occurs without tolling, and thus has a relative efficiency of a half. Relative efficiency equals the total cost reduction of a policy from the NT case divided by the FB reduction.

The ADL model of Arnott et al. (1990, 1993) has a mass departure at the moment the toll is lifted. The equilibrium mass size is such that the expected extra travel cost for a mass user equals the toll. The peak starts and ends later (i.e. is shifted to later) than in the Laih model, otherwise the expected price in the mass would be below that for early untolled users. This in turn implies that the coarse tolling actually lowers the equilibrium price and has a relative efficiency above a half.[4] To increase the benefit from the mass departure, the number of untolled users will be above that in the Laih model, which is achieved by setting a higher toll.

Finally, in the braking model of Lindsey et al. (2012) and Xiao et al. (2012), users stop passing the tolling point a time Δt before the toll is lifted. The equilibrium Δt is such that the extra travel cost of $(\alpha+\gamma)\cdot\Delta t$ for the first braker equals the toll. The other two models only have the described equilibrium if this braking is impossible. Braking increases total cost, and thus the relative efficiency is now below a half. To limit the cost from braking, the number of untolled users will be below that in the Laih model, which is attained by setting a lower toll.

[2] There are changes to the shape of the distribution that keep the same mean and increase the variance by only changing the distribution for high values $\eta_i>\eta^*$, this would then have no effect on the NT travel times and the FB gain. What is needed is that the distribution changes for low values. For a given shape and mean γ, an increase in the variance (i.e. an increase in the degree of heterogeneity) will lower the average travel time cost.

[3] Fosgerau (2011) uses the Laih model under general scheduling preferences instead of the time-invariant values used here.

[4] All this requires that $\alpha<\gamma$. With $\alpha>\gamma$, a different outcome results where there is no shift in the peak and the outcome of the model is the same as in the Laih model. Only the ADL model needs this $\alpha<\gamma$ assumption (see Lindsey et al., 2012).

4. Proportional heterogeneity and coarse tolling

4.1. The coarse tolling under proportional heterogeneity

We first present results that hold for any coarse toll model under proportional heterogeneity that varies all three values in fixed proportions. Total cost will be minimised w.r.t. the number of untolled users, *V*, and there are $N-V$ tolled users. The level of the coarse toll, ρ, is such that at the start, t^+, and end, t^-, of the tolled period the queue is zero.[5]

Proposition 1: With coarse tolling and proportional heterogeneity, the $N-V$ highest-values users travelled tolled. The *V* lowest-values users travel before and after the tolled period. The type that is indifferent between travelling tolled or untolled has a value of schedule delay early of $\beta^*[V]$. This self-selection lowers total cost, but less than with first-best tolling where there is the fully-separated self-selection.

Proposition 2: The price for a type is the same for any arrival time within a period (i.e. early and untolled, tolled or untolled and late), as long as this period does not have a mass departure. Therefore, types travel jointly within such a period.

Proofs: Appendix B.1 gives the proofs.

In all three models, the equations for the timings of the tolled period are unaffected by the heterogeneity: $t^+=-\eta\cdot t^-$ and $t^-=(N-V)/((1+\eta)s)$. The start of the peak, t_s, and end of the peak, t_e, follow different formulas.

As without a mass the price is constant throughout a period, we only need to calculate it for one moment of a period. The price for an untolled user equals the schedule delay cost at t_s of $-\beta_i\cdot t_s$:

$$P_i^U=-\beta_i\cdot t^s. \tag{9}$$

With a mass departure, its expected price equals the above price. The price in the tolled period is:

$$P_i^T=\rho-\beta_i\cdot t^+. \tag{10}$$

Hence, prices are piece-wise linear in β_i, and the function is kinked at β^*.

4.2. Laih model

We start with the simplest coarse toll model: the Laih model. Here, the start and end of the peak are independent of *V*, and are the same as in the NT and FB cases. Appendix B.3 shows that the total cost equals (where *TP* is the total price or average price times *N*, and *TR* is the toll revenue):

[5] A non-zero queue length would only raise costs (see also Xiao et al., 2011). This holds even if *V* is set suboptimally.

$$\begin{aligned} TC[V] &= TP[V]-TR[V] = -t^{s}\cdot\beta^{L}[V]\cdot V - t^{+}[V]\cdot\beta^{H}[V]\cdot\left(N-V\right) \\ &= \frac{\eta}{(1+\eta)s}\left(E[B]\cdot N^{2}-\beta^{H}[V]\cdot V(N-V)\right). \end{aligned} \tag{11}$$

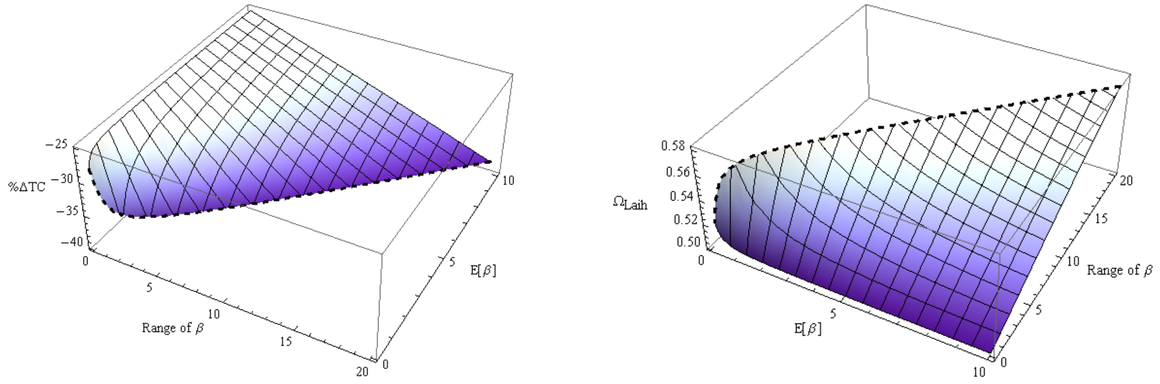
The E[β] is the average of β. The β^L is the average for the low-β users who travel untolled, β^H is the average for the tolled period. In the last line of (11), the first term between brackets measures the NT total cost, the second the travel cost reduction for the tolled users. Total cost tends to decrease with the degree of heterogeneity, as this tends increase the average β of the tolled period, which makes the travel cost saving of the tolled period more valuable.[6]

Appendix B.3 derives the f.o.c. for V for a general distribution, and shows that with a uniform distribution the relative efficiency depends only on the mean E[β] and the range d of the distribution (i.e. maximum minus minimum):

$$\begin{aligned} \Omega_{\text{Laih}} = &\frac{2\ (8\ E[\beta]+\ \sqrt{d^{2}\ +\ 2\cdot d\cdot E[\beta]\ +\ 4\cdot E[\beta]^{2}})}{9\ d\cdot d\ \ (d\ +\ 6\ E[\beta])} \\ &+\frac{2\left(d-2\ E[\beta]\ +\sqrt{d^{2}+2\ d\ \ E[\beta]+4\ E[\beta]^{2}}\right)\left(d^{2}-2\ E[\beta]-2\ E[\beta]+\sqrt{d^{2}+2\ d\ \ E[\beta]+4\ E[\beta]^{2}}\,]\right)}{9\ d\cdot d\ \ (d+6\ E[\beta])}. \end{aligned} \tag{12}$$

With a uniform distribution, the degree of heterogeneity only depends on the range d, and the relative efficiency increases with the range. When d is zero, the relative efficiency is ½; as d approaches 2·E[β],[7] the relative efficiency approaches $1/\sqrt{3}\approx 0.58$. Hence, the gain of coarse tolling increases faster with the range than the first-best gain. Fig. 1 illustrates the effect of the range and mean of β using the numerical example of Table 1.

Fig. 1: The Laih toll's percentage change in total cost from the NT case (left) and relative efficiency) over the mean and range of β



[6] An increase in the degree of heterogeneity is defined as an increase of the variance for a given mean and *distribution shape*. There are changes in the shape that increase the variance by only changing the distribution for low values, this would have no effect. What is needed is that the β^H increases.

[7] The d must be below 2·E[β] as otherwise for some users the value of time would equal the value of schedule delay late

4.3. The ADL model of coarse tolling with proportional heterogeneity

Xiao et al. (2011) introduced proportional heterogeneity to the ADL model. The ADL model needs the assumption $\alpha_i<\gamma_i$ for all i, which is not needed in the other models. Total cost is (see Appendix B.2):

$$\begin{aligned} TC[V] &= TP[V]-TR[V] = -t^s[V]\cdot\beta^L[V]\cdot V - t^+[V]\cdot\beta^H[V]\cdot\left(N-V\right) \\ &= \frac{\eta}{1+\eta}\frac{1}{s}E[B]\cdot N^2 - \beta^H[V]\frac{\eta+\mu}{2+\eta+\mu}(N-V)\cdot V - E[B]\frac{1}{1+\eta}\frac{V^2}{s}\frac{\eta-\mu}{2+\eta+\mu}. \end{aligned} \qquad (13)$$

The difference with the Laih model is that in the ADL model the total cost in the last line of (13) contains a third term measuring the gain from the mass. As in the Laih model, more heterogeneity tends to increase the gain from the lowered travel times and schedule delays in the tolled period. Yet, now there is also a second effect that more heterogeneity tends to lower the mean β of the untolled users (i.e. β^L), and this lessens the gain from the lowered schedule delays due to the mass.

In the ADL model, even with a uniform distribution, the formula for total cost with the equation for V inserted is extremely complex and hence omitted. Still, we do know from (13) that total cost decreases with the range. Compared with the Laih model, the gain in the ADL model decreases with d: a larger range not only increases the mean values during the tolled period making its travel time and schedule delay savings more valuable, it also lowers the gain from the mass by lowering the mean β of the untolled users.

4.4. Braking model

We now turn to the braking model, which takes into account that drivers that would pass the tolling point just before the toll is lifted have an incentive to wait passing the tolling point until the toll is turned off. In equilibrium, the bottleneck capacity goes unused for a time Δt during the peak. Therefore, the peak is Δt longer than in the other models, and this inefficiency raises costs.

Appendix B.3 shows that total cost equals:

$$\begin{aligned} TC[V] &= = -t^s[V]\cdot\int_0^V \beta_i[x]dx - t^+[V]\cdot\int_V^N \beta_i[x]dx = -t^s[V]\cdot\beta^L[V]\cdot V - t^+[V]\cdot\beta^H[V]\cdot\left(N-V\right) \\ &= \frac{\eta}{1+\eta}\left(\frac{E[B]}{s}\cdot N^2 - \beta^H[V]\cdot\frac{V}{s}(N-V) + \Delta t[v]\cdot\beta^L[V]\cdot V\right). \end{aligned} \qquad (14)$$

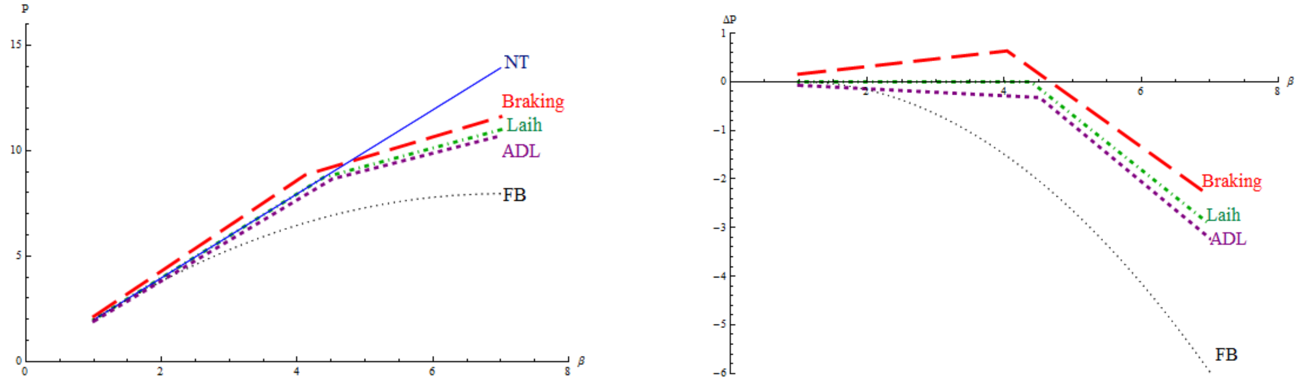
In the last line, the first and second terms between brackets are the same as in the previous models. The new third term measures the extra costs due to braking. This loss is lower than without heterogeneity as β^L (the mean value for untolled users) is below the value with homogeneity: the extra schedule delays imposed on untolled users is less costly as their values of schedule delay are lower. Coarse tolling with braking tends to become more attractive compared to in the Laih model as the degree of proportional heterogeneity increases, as the cost of braking tends to decrease.

It can be shown that, with a uniform distribution, the gain of coarse tolling increases with the range: this 1) the makes the travel time and schedule delay saving of the tolled period more valuable, and 2) makes the extra schedule delays due to braking less costly. Hence, the coarse toll in the braking model performs better against in the other models as the range increases.

4.4. Distributional effects

Proportional heterogeneity has important implications for the overall effects of coarse tolling Yet, for policy, these effects might be less important as coarse tolling always improves welfare and has a relative efficiency around 0.45 to 0.6. For policy, the distributional effects might be more relevant.[8] Fig. 2 compares the prices (left panel) and the prices changes for the tolling schemes from the NT case (right panel). It does so for the example as defined in Table 1, when β ranges between 1 and 7 and has mean of 4.

Fig. 2: Generalised prices (left) and changes in generalised prices from the NT equilibrium (right)



Proportional heterogeneity leads to interesting distributional effects. All users gain from First-Best (FB) tolling, and more so the higher their values are (except the lowest-β users, who are unaffected). In the Laih model, coarse tolling has no effect on untolled users. Yet, all types who strictly use the tolled period gain, and more so the higher their values are: the tolled period offers them a decrease in schedule delay and/or travel time for which they have to have a fixed toll, and higher the values makes these savings more valuable.

Conversely, in the ADL model, coarse tolling lowers the price for all, as untolled users also gain due to the mass departure. Still, FB tolling generally decreases prices more than the ADL model's coarse toll, only for drivers with very low values is coarse tolling better.

In the braking model, distributional effects are very different. There coarse tolling raises the price for the untolled drivers, and more so the higher their values are, and thus the indifferent type

[8] Although if toll collections costs are high enough a relative efficiency of 0.47 instead of 0.53, might mean that coarse tolling is not welfare improving.

faces the highest price increase. For the tolled period, coarse tolling with braking is better for a user the higher her β_i is: for low-β drivers the price increases, for high-β drivers it decreases. Unlike in the other models, now coarse tolling is not a Pareto improvement. The distributional effects in the braking model are similar as with second-best pricing with an untolled alternative with static congestion and a heterogeneous value of time (see Verhoef and Small, 2004).

The distributional effects are qualitatively robust to the used distribution of β. All tolling schemes are always Pareto improvements except in the braking model. In the Laih model, coarse tolling never affects the untolled users. In the braking model, most users are always worse off.

5. Coarse tolling and α heterogeneity

We now turn to coarse tolling under α heterogeneity where the value of time fares and the values of schedule delay and fixed. It is not α heterogeneity in itself that matters, but the implied heterogeneity in $\mu_i \equiv \alpha_i/\beta$ (and $\mu_i/\eta = \alpha_i/\gamma$). It is this heterogeneous ratio μ_i that also matters if there are multiple dimensions of heterogeneity (see Van den Berg and Verhoef, 2011a). The value of time is distributed with a maximum $\bar{\alpha}$, minimum $\underline{\alpha}$, distribution function $f[\alpha]$, and CDF $F[\alpha]$. The numerical example will follow Table 1 plus an α uniformly distributed between 5 and 11.

5.1. Generals of coarse tolling under α heterogeneity

We can divide the peak in 3 periods: (1) the Tolled (T), (2) the Untolled Early (UE) which is before t^*, and (3) the Untolled Late period (UL) which is after the tolled period. Without a mass departure, users arrive ordered on α_i within a period, and arrive closer to t^* the higher α_i is. The travel time when type i users travel has a slope α_i/β before t^* and α_i/γ thereafter. Hence, the travel time curve becomes steeper as one approaches t^*, and lowest-α users of a period face the longest travel time. With continuous heterogeneity, a type can only use one arrival time during a period, but may use multiple periods.

Prices are concavely increasing in α outside the mass, and linearly for mass users. The prices for non-mass users follow the same pattern as without tolling, as there still is queuing. The price equations are derived in Appendix C.1. The following propositions follow from the shapes of the price functions and that if a type uses multiple periods, its prices in them must be the same.

Proposition 3. Mass users travel fully separated: for there to be a user-equilibrium, all types in the mass (but for those with an indifferent value) cannot also use another period.

Proposition 4. If a type i uses multiple periods without a mass departure, then a type j with a higher value (i.e. $\alpha_i<\alpha_j$) either has: 1) no users in these periods or 2) j's users are shared in fixed proportions. It can occur that type j uses all the periods i uses, but type k with a α_k between i and j is absent from these periods.

Example of proposition 4: Suppose that there are no mass departures, that types *i* and *j* use the tolled and both untolled periods and $\alpha_i<\alpha_j$. Then, type *j*'s drivers use in equal numbers in the early-tolled and early-untolled periods; the late-tolled and late-untolled periods have a fraction $1/\eta$ of the users in the corresponding early period. Thus, periods need not have the same number of users of a type: if a period is twice as long as another, then the types *j* that use both should have twice as many users in the longer period.

Proposition 5. If the highest-α type uses a period without a mass departure, it uses all periods without a mass.

Proofs: Appendix C.1 gives the proofs to these three propositions.

The above discussion does not imply that low-α types have to use all periods. Indeed, in the ADL and Braking models they do not. For now, we will ignore mass departures. The higher-α types have to travel in all periods, and of each type half of the users travel untolled and the remained tolled. Then, if the tolled and untolled periods are not equally long, there is no room for the lowest-α users in the shorter period, and they only use the longer period.

Proposition 6. If the early-untolled period is longer than the early-tolled, the users with the lowest values of time travel only untolled, and vice versa. Without a mass departure, the same holds for late arrivals.

5.2. Laih model of coarse tolling

Using the pervious section, we need no further information to describe results in the Laih model. Just as with homogeneity, the optimal *V* equals *N*/2, and thus the tolled and untolled periods are equally long. Setting a different *V* would not affect prices for the high-α types that continue to travel tolled and untolled, but would lower toll revenue. It would also increase prices for low-α types that switch to using only one period, as they then use a period with relatively many low-α users, and these impose higher congestion externalities than high-α users (see Lindsey (2004a) for this effect without tolling).

However, the α heterogeneity does mean that prices increase due to coarse tolling for all but the users with the highest α_i. The equilibrium price is:

$$P_i^{Laih} = \frac{\delta}{2}\frac{N}{s} + \frac{\delta}{2}\frac{1}{s}\left(\int_{\underline{\alpha}}^{\alpha_i} n_j d\alpha_j + \alpha_i \int_{\alpha_i}^{\bar{\alpha}} (n_j/\alpha_j)\, d\alpha_j \right). \qquad (15)$$

In the Laih model, coarse tolling increases the price for all types by exactly half that of the first-best toll. The toll revenue is also half of the first-best revenue. Hence, the relative efficiency of the coarse toll is ½, and is independent of the degree of α heterogeneity.

5.3. ADL model of coarse tolling

In the optimum of the ADL model, the tolled period is shorter than the untolled, as having more users in the mass lowers costs (up to a point). Different from the Laih model, the ADL model does not have a closed-form solution for a general distribution, but there is with a uniform distribution.

Proposition 7: With a uniform distribution, the lowest-α cannot be in the mass. For them it would always be attractive to move to the untolled early period.

Appendix C.3 proofs proposition 7. The idea is that users with a low value gain from the α heterogeneity when there is normal queuing (i.e. no mass), and then the price concavely increases with α_i. In a mass, they lose this advantage, and thus using a mass is not interesting.

Using a uniform distribution, it can also be shown that users with an intermediate α have most to gain from being in the mass. The question is whether the users with the highest values are also in the mass. This is an empirical question, where the answer depends on the exact distribution. In the numerical example, and indeed for all tried parameterisations with a uniform distribution, the highest-α users will not be in the mass; but there may be alternative distributions were they will.

Proposition 8: In the equilibrium of the ADL model (under a uniform distribution of α), the continuum of types is separated in 4 groups, which are characterised by the periods they travel in. Group 1 with users with the lowest values of time (i.e. $\alpha_i < \alpha_1^*$) only travel in the untolled early period. The types in Group 2 with $\alpha_1^* \le \alpha_i < \alpha_2^*$ use the early untolled and the tolled period. An intermediate mass group travels only in the mass and has $\alpha_2^* < \alpha_i < \alpha_3^*$). Finally, Group 3 with $\alpha_3^* \le \alpha_i$ again travels early untolled as well as tolled. Groups 1 and 3 may be of zero size.

5.4. Braking model of coarse tolling

In the Braking model, users start braking at t^b and do not pass the tolling point until the toll is lifted at t^-. Since braking raises total cost and the braking time increases with the number of untolled users, it is optimal to have a longer tolled period than untolled period. Following proposition 6, this means that low-α users only travel tolled, as there is only room for them in the tolled period. The highest-α users travel in all periods and arrive at t_s, t^+, t^- and t_e.

Proposition 9. In the braking model, the continuum of types is separated in three groups. Group 1 with the lowest values ($\underline{\alpha} < \alpha_i < \alpha_1^*$) travels only tolled. Group 2 with intermediate values ($\alpha_1^* \leq \alpha_i < \alpha_2^*$) travels tolled and early untolled. Group 3 with the highest values ($\alpha_2^* \leq \alpha_i \leq \bar{\alpha}$) uses all periods. Type α_2^* users are the first to brake, the other brakers have higher values.

Proof: See Appendix C.4.

We can again calculate total costs by subtracting toll revenue from the total price. There is no closed-form solution for the optimal *V* even for a uniform distribution, although we do know that the optimal *V* will be below *N*/2 to limit the costs of braking.

5.5. Comparison of the coarse-toll regimes under α heterogeneity

Having established the equilibria of the coarse toll models, we now compare the effects in the numerical example. Fig. 3 depicts the prices (left panel) and the price changes from the no-toll case (right panel) when α ranges between 5 and 11. In Laih model, coarse tolling causes exactly half the price increase as in the first-best (FB) case, and does so for any parameterisation. Braking raises prices by ensuring that the capacity goes unused during the peak. Still, users with the lowest values of time (i.e. with α_i<7.2 or 38% of all users) are better off with the Braking model's coarse toll than with FB pricing: as the later remove the entire price advantage they have without tolling due to *α* heterogeneity. With α heterogeneity, the braking harms all types of users in a similar way. In the ADL model, the mass departure lowers costs. This makes the users with α_i>8.57 (about 40%) better off than without tolling; the others lose but less than in the other coarse-toll models.

Fig. 3: Equilibrium prices (right) and price changes from the no-toll equilibrium (left)

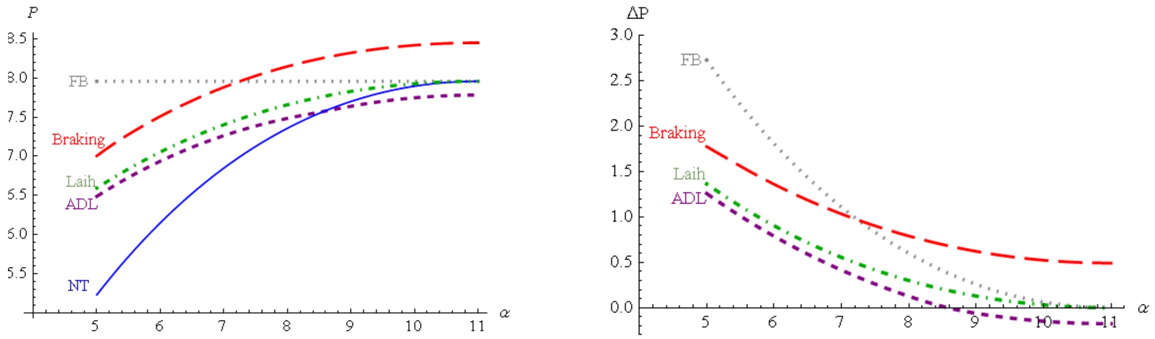
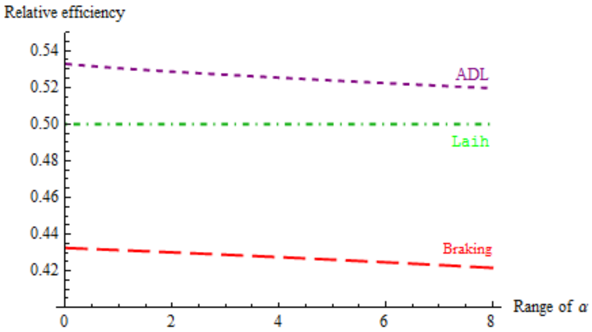


Fig. 4 looks at the effects of the range of α (i.e. the degree of heterogeneity), while the other parameters are at the levels of Table 1. The range needs to be below 8, otherwise some users would have a α_i equal to or below *β*. Consistent with the earlier discussion, a larger range of *α* lowers all average prices except the FB one. NT total cost decreases with the range, as this lowers congestion externalities. This also means that the gain of first-best pricing decreases with the range. The relative efficiency of coarse tolling in the Laih model is always a half. In the other models, the relative efficiency decreases with the range, but the effect is small. The relative efficiency in the ADL model approaches a half as the range increases, but never reaches it. With Braking, coarse tolling always raises the average price, and more so the larger the range of α.

Fig. 4: The range of α and the relative efficiencies



6. Coarse tolling under γ heterogeneity

6.1. The generals of coarse tolling under γ heterogeneity

Now we turn to heterogeneity in the value of schedule late (γ_i) with fixed values of time (α) and schedule delay early (β). The γ heterogeneity in itself is not important, the implied heterogeneity in the ratio $\gamma_i/\beta \equiv \eta_i$ is. This ratio η_i determines the choice whether to arrive early or late.

As in the NT and FB case, with coarse tolling all users with a η_i below the indifferent ratio η_1^* arrive late, as for them schedule delays late are less costly. In the Laih model, this ratio is the same as in the NT and FB cases, in the other two models it is not. The types that arrive early before t^* travel jointly, as γ_i only affects the price when arriving late. With coarse tolling, there is also a second indifferent ratio η_2^*, which separates the tolled-late and untolled-late users.

Proposition 10. Users with $\eta_i \equiv \gamma_i/\beta < \eta_1^*$ arrive after the preferred arrival time (t^*), and arrive ordered on γ_i with the lowest value arriving the furthest from t^*.

Proposition 11. Types with $\eta_i \geq \eta_1^*$ arrive on or before t^* and travel jointly. Their price is independent of their value of schedule delay late, and constant over arrival time.

The price for *the early-arriving* users will be the same in all models given η_1^*, although this ratio differs between the models. The price equation of early-arriving users still follows the same eq. (7) as for the NT and FB cases, and is independent η_i. The price for late-arriving toll payers also follows the same formula as before and concavely increases in η_i:

$$P_i = \beta \cdot \frac{N}{s}\left(1 - F[\eta_1^*] - \int_{\eta_i}^{\eta_1^*} f[\eta_j]\left(\eta_j - \eta_i\right) d\eta_j\right), \qquad \eta_1^* \geq \eta_i \geq \eta_2^*. \tag{16}$$

The price formula for the late-arriving no-toll payers differs over the three coarse toll models. There is no-closed form solution of η_1^* and η_2^* for a general distribution of $\eta_i \equiv \gamma_i / \beta$. Appendix E derives the conditions that determine these ratios.

6.2. Laih model of coarse tolling and γ heterogeneity

The Laih price for users who arrive after the tolled period follows the same equation as in the NT and FB equilibria, and the price concavely increases with η_i:

$$P_i = \beta \cdot \frac{N}{s}\left(1 - F[\eta_1^*] - \int_{\eta_i}^{\eta_1^*} f[\eta_j]\left(\eta_j - \eta_i\right) d\eta_j\right), \qquad \eta_2^* \geq \eta_i. \tag{17}$$

However, the γ heterogeneity does affect the choice of ρ, and thereby the choice of the number of tolled users and the indifferent ratios, and thus increases the actual levels of the prices. More γ heterogeneity reduces travel times when arriving late, and thus, for a given coarse toll, fewer users want to travel tolled.

6.3. ADL model of coarse tolling and γ heterogeneity

The mass departure ensures that for the indifferent type η_2^* the price in the tolled period is the same as the expected mass price.[9] The expected price is linear in η_i:

$$\begin{aligned} E[P_i] &= \beta \cdot \left(\eta_i \cdot t^- + \left(\mu + \eta_i\right)\frac{M}{2 \cdot s}\right) \\ &= \beta \cdot \frac{N}{s}\left(\eta_i \cdot \left(F[\eta_1^*] - F[\eta_2^*]\right) + \left(\mu + \eta_i\right) F[\eta_2^*]\right) \end{aligned}, \qquad \eta_2^* \geq \eta_i. \tag{18}$$

6.4. Braking model of coarse tolling and γ heterogeneity

In the equilibrium of the braking model, the bottleneck capacity goes unused for a period $\Delta t = \rho / ((\mu + \eta_2^*)\beta)$, where $\eta_2^* = \gamma_2^* / \beta$ is the ratio of the first braker. All users with a γ_i lower than γ_2^* also arrive during the late untolled period, and arrive later the lower their γ_i is. This self-ordering of users lowers the costs due to braking, as the extra schedule delays during the late period are imposed on users with a low γ_i. This then also implies that more γ heterogeneity tends to mean that the coarse toll in the braking model fares better compared to in the Laih model.

[9] For ADL tolling we assume that $\gamma_i > \alpha$ for all i which limits the degree of heterogeneity more than in the other models where we only need $\gamma_i > 0$. Note that this can be a very strict assumption, that may not hold in empirical studies. If $\gamma_i < \alpha$ for some types, then there will be normal queuing after the last mass user arrives. Still different from with homogeneity, the ADL model the need not have the same outcome as the Laih model, as the mass affects the prices of mass users, and thus the solution of η_1^*.

For users who arrive after the tolled period, the price equation is the same as in the Laih model. But prices will be higher with braking due to the Δt period when capacity is idle.

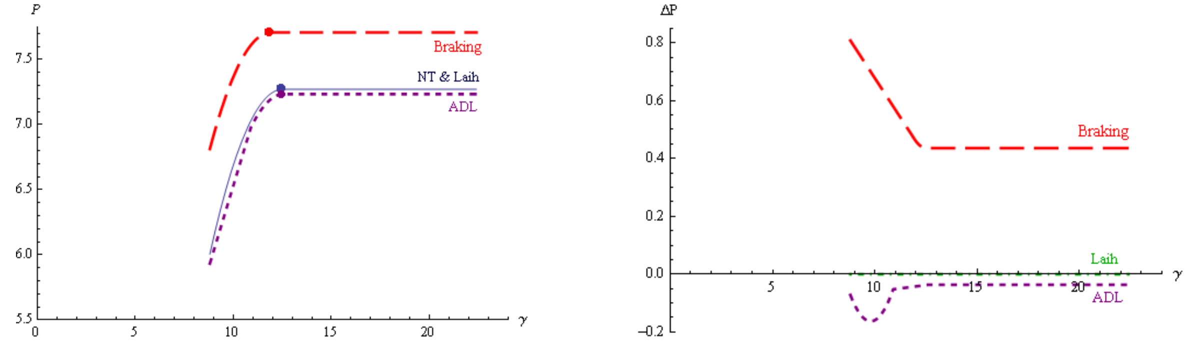
6.5. Numerical example

There are no closed-form solutions for a general distribution of γ, but with the numerical example of Table 1 it is possible to gain further insight. Fig. 5 depicts the equilibrium prices (left panel) and the price changes from the NT equilibrium (right panel) when γ ranges uniformly between 8.8 and 22.2.. For users who arrive early with $\eta_i \geq \eta_1^*$, the price functions are flat. In the Laih and Braking models, the price is concavely increasing for late-arriving low-γ users; in the ADL model, the price is linear for mass users. As discussed, prices are substantially higher in the braking model than in the Laih model.

With the Laih model's coarse toll and FB tolling, prices are the same as without tolling, and thus there are no distributional effects. With braking, coarse tolling increases the price for all, and for a late-arriving user the price increase is higher the lower her γ_i is. In the ADL model, the price decreases for all, but the distributional effect is non-monotonic: the mass users gain most, and the intermediate mass users gain most of all. Similar distributional patterns occur if we increase the range of γ or use a different distribution shape.

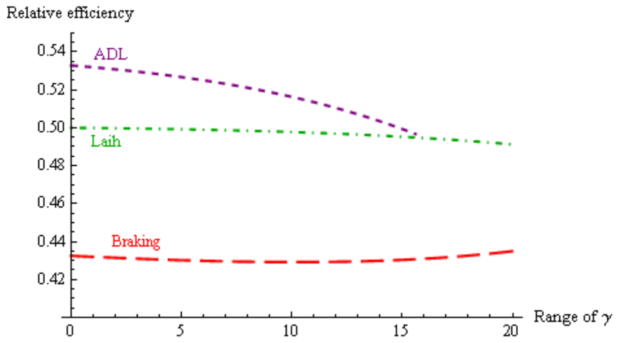
Now consider the distributional effects of coarse tolling when all three forms of heterogeneity. With braking, coarse tolling raises the price for all but perhaps those with very high values and low ratios α_i/β_i and β_i/γ_i (e.g. high-income users who have flexible schedules and can easily arrive late). If we can prevent braking, coarse tolling may actually lower the price for all or most users if the degree of α heterogeneity is low enough compared to the other dimensions of heterogeneity.

Fig. 5: Generalised prices (left) and percentage changes in generalised price from the NT case (right)



As Fig. 6 shows, the relative efficiency in the ADL model decreases with the range of γ, as a larger range lessens the beneficial effect from the mass departure. In the Laih model, there is almost no effect on the relative efficiency. In the braking model, the range has hardly an effect on the efficiency, and the effect is non-monotonic. Similarly, in Zhang et al. (2008), the gain of mixed private and public supply is non-monotonic in the degree of heterogeneity in the value of time in a large network with static congestion.

Fig. 6: The effect of the range of γ on the relative efficiency



7. Discussion

This section discusses some caveats and directions for future research. An important point is that we ignore price-sensitivity of demand, which is a common assumption in the literature on step-tolling and/or heterogeneous preferences.

With first-best tolling and *α* heterogeneity, Van den Berg and Verhoef (2011b) find that low-α users lose due to FB tolling. However, as these users therefore demand less travel, overall congestion decreases and therefore high-*α* users gain from FB tolling. With proportional heterogeneity, FB tolling increases the number of users as for most types the price decreases. This in turn increases congestion and thereby makes low-*β* users worse off.

Price-sensitive demand also affects coarse tolling. With homogeneity, Arnott et al. (1993) show that the coarse part of the toll minimises social cost for a given number of users, while a time-invariant addition to the toll optimises the number of users. Moreover, Van den Berg (2012) finds that step tolling raises the price, but less the more steps the toll has. With our forms of heterogeneity, the same set-up would be optimal, and step tolling would also tend to increase prices.

Another interesting extension would be heterogeneity in multiple dimensions at the same time. However, analytical analysis of this seems difficult. Hence, more numerical analysis may be more promising. Lindsey (2004b) analyses tolling in a network of concentric ring roads using METROPOLIS. There are four types of drivers that differ in their values of time and schedule delay; for each type, the preferred arrival time is uniformly distributed. Börjesson and Kristoffersson (2012) study—using the “Silvester” model—the Stockholm step-toll system and network with proportional, *α* and *γ* heterogeneity. De Palma and Lindsey (2004) and Van den Berg and Verhoef (2011) considered first-best tolling under proportional and α heterogeneity, where the former also studied heterogeneity in the preferred arrival time.

8. Conclusion

This paper derived the equilibria, as well as total costs and toll revenues, for coarse tolling with preference heterogeneity. It also examined the distributional effects of such tolls and the effects of the degree of heterogeneity. It used three models of coarse tolling.

In the Laih model, the (generalised) price with coarse tolling is in between that in the no-toll equilibrium and with first-best fully-time-variant tolling. In the other two models, this is not the case and the distributional effects may be non-monotonic. In the braking model with proportional heterogeneity, the higher a untolled user's values are, the more coarse tolling increases the price from the no-toll case, while a tolled user is better off the higher her values are. With the ADL model and γ heterogeneity, coarse tolling lowers all prices; but most for mass users and especially for those with an intermediate value. Users are always better off than in the Laih model as the mass departure lowers costs. Braking makes all users worse off than in the Laih model.

Braking not only lowers the gain from tolling, but also affects the distributional effects. Compared with the Laih model, with braking, coarse tolling is most harmful for low-values users and those with a relatively large value of schedule delay late: this could, for instance, be low-income users with strict work start times or on a trip to the doctor. With α heterogeneity, all types of users are harmed by the braking in a similar way. If braking can be prevented, coarse tolling may lower the price for all or most users if the degree of α heterogeneity is low enough compared to the other dimensions of heterogeneity. Nevertheless, even with braking, coarse tolling is a potential Pareto improvement, as welfare increases and thus the toll revenue could be used to compensate losers.

Proportional heterogeneity raises the welfare gain from tolling, and coarse tolling tends to fare better against the first-best toll the more heterogeneity there is. In the ADL model, the gain is higher than in the Laih model, in the braking model the gain is lower, and both differences decrease with the degree of heterogeneity. The gain from tolling decreases with the degree of α and γ heterogeneity. With α heterogeneity, the Laih model's coarse toll has half the gain of the FB toll; with γ heterogeneity, this "relative efficiency" is very close to a half in the numerical model.

Acknowledgements

I thank Erik Verhoef, Robin Lindsey, Paul Koster, and Alex Dimitropoulos for their helpful comments. Financial support from the ERC (AdG Grant #246969 OPTION) is gratefully acknowledged. The usual disclaimer applies.

Appendix A: Coarse tolling under homogeneous preferences

A.1. Laih model of coarse tolling

The coarse toll is turned on at t^+ and off at t^-. Just as without tolling, the slope of the travel time over t is β/α before t^* and $-\gamma/\alpha$ thereafter. This ensures that the price is constant over t within a period (i.e. early and untolled, tolled or untolled and late). The first toll-payer to arrive at t^+ has a zero travel time. The travel time of the last no-toll-payer to arrive before t^* is the highest of all and equals ρ/α, such that her travel time costs equals the coarse toll. The last toll-payer to arrive at t^- has a zero travel time, while the first no-toll payer to arrive thereafter has the highest travel time. Hence, users who arrive after the toll is lifted start queuing during the tolled period, and only pass the tolling point, and then the bottleneck, after t^-.

Just as in the NT and FB cases, the start and end of the peak are such that the schedule delay costs then are the same, and thus these timings are the same as before. The number of no-toll payers is V. There are $N-V$ toll-payers. The start and end times of the tolled period are found by equating the schedule delay costs then and by using $N-V=s(t^- - t^+)$:

$$t^+[V] = -\frac{\eta}{1+\eta}\frac{N-V}{s}, \qquad t^-[V] = \frac{1}{1+\eta}\frac{N-V}{s}; \tag{19}$$

Since the difference between schedule delay costs at t_s and t^+ is $\beta(t^+-t_s)$, while prices must be the same, the toll must equal this difference:

$$\rho[V] = \beta\frac{\eta}{1+\eta}\frac{V}{s}. \tag{20}$$

Because t_s and t_e are the same as without tolling, the price still equals $\delta\cdot N/s$, but travel cost is ρ lower during the tolled period. Total cost equals total price (i.e. price times N) minus the toll revenue of $\rho(N-V)=\rho\cdot s(t^- - t^+)$:

$$TC[V] = TP - TR[V] = \frac{\beta}{s}\frac{\eta}{1+\eta}\left(N^2 - V\left(N-V\right)\right). \tag{21}$$

Minimising (21) gives an optimal V that is equal to half the total number of users: $V = N/2$. Inserting this V into (21) and some algebra shows that total cost is a quarter lower than in the NT case, and thus the relative efficiency is 1/2.

A.2. ADL coarse tolling under homogeneous preferences

Now, there is a mass departure at t^-. After the last mass user arrives, the peak ends and there is no more travel. The peak is shifted to later: otherwise, the expected price in the mass would be below that in the other periods.[10] There are V no-toll payers, and M of these use the mass. The expected price for a mass user equals the schedule delay cost at $t-$ plus the travel time and schedule delay costs due to the expected time, $M/(2\cdot s)$, it takes to pass the bottleneck. The M is determined that the untolled early price should equal the expected price in the mass.

[10] If $\alpha>\gamma$ there is normal queuing after the mass, and the ADL and Laih models have the same total costs and optimal V (Lindsey et al., 2012)

The timings of the tolled period still follow (19), but the formulas for t_s and t_e are different. By equating the prices of the three periods and using $N/s=t_e-t_s$, we get:

$$\rho[V]=\frac{\mu+\eta}{2+\eta+\mu}\beta\frac{V}{s}, \tag{22}$$

$$\begin{aligned} t_s[V]&=-\frac{\eta}{1+\eta}\frac{N}{s}+\frac{V}{s}\frac{1}{1+\eta}\frac{\eta+\mu}{2+\eta+\mu},\\ t_e[V]&=\frac{1}{1+\eta}\frac{N}{s}+\frac{V}{s}\frac{1}{1+\eta}\frac{\eta+\mu}{2+\eta+\mu}. \end{aligned} \tag{23}$$

This makes total cost:

$$\begin{aligned} TC[V]&=TP[V]-TR[V]=-\beta\cdot t_S[V]\cdot N+\rho[V](N-V)\\ &=\frac{\beta}{s}\frac{\eta}{1+\eta}\left(N^2-V\left(N-V\right)-V^2\cdot\frac{1}{\eta}\frac{\eta-\mu}{2+\eta+\mu}\right). \end{aligned} \tag{24}$$

Minimising total costs gives:

$$V=N\cdot\frac{\eta\left(2+\eta+\mu\right)}{\left(1+\eta\right)\left(\eta+\mu\right)}>\frac{N}{2}. \tag{25}$$

The gain of coarse tolling is now higher than in the Laih model, and the relative efficiency equals:

$$\omega_{ADL}=\frac{1}{2}\left(1+\frac{\eta-\mu}{(1+\eta)(\eta+\mu)}\right)>\frac{1}{2}\cdot \tag{26}$$

A.3. Braking model of coarse tolling under homogeneous preferences

The braking starts at t^b, and the first braker waits a time Δt until the toll is lifted at t^-. The peak now lasts $N/s+\Delta t$. As prices must be the same during all used arrival times, Δt must be such that the extra travel costs of the first braker equals the toll paid by the last toll-payer: $\Delta t=\rho/(\alpha+\gamma)$. The prices at t_s and t^+ must be equal, and thus ρ follows the same formula as earlier. The schedule delay costs at t_s and t_e again have to be equal, and those at t^+ and t^b must also be equal; this gives:

$$\begin{aligned} t_S[V]&=-\eta\cdot t_e[V],\\ t_e[V]&=\frac{1}{1+\eta}\left(\frac{N}{s}+\Delta t[V]\right)=\frac{1}{1+\eta}\frac{1}{s}\left(N-V\frac{\eta}{\mu+\eta(2+\eta+\mu)}\right), \end{aligned} \tag{27}$$

Combining al this gives a total cost of:

$$\begin{aligned} TC[V]&=TP[V]-TR[V]=-\beta\cdot t_S[V]\cdot N-\rho[V](N-V)\\ &=\frac{\beta}{s}\frac{\eta}{1+\eta}\left(N^2-V\left(N-V\right)+V^2\cdot\frac{\eta}{\mu+\eta(\eta+\mu)}\right); \end{aligned} \tag{28}$$

Minimising total cost results in:

$$V=\frac{\mu+\eta(\eta+\mu)}{\eta+\mu+\eta(\eta+\mu)}\frac{N}{2}<\frac{N}{2}. \tag{29}$$

Together (28) and (29) imply that the relative efficiency is:

$$\omega_{Braking}=\frac{1}{2}\left(1-\frac{\eta}{(1+\eta)(\eta+\mu)}\right)<\frac{1}{2}. \tag{30}$$

Appendix B: Coarse tolling under proportional heterogeneity

B.1: All three models of coarse tolling

Proof of proposition 1. The tolled period allows users to have a lower travel time and/or schedule delay, but for this, they have to pay the toll. Naturally, users with higher values are more willing to pay for this, and thus self-select to the tolled period.

Self-selection lowers costs, thus with the limited self-selection of coarse tolling there is a lower gain than with the fully-separated self-ordering of the FB toll. □

Proof of proposition 2. The reason for proposition 2 is the same as for the ordering without tolling. In equilibrium, the travel time function has a slope of $\alpha_i/\beta_i\equiv\mu$ before t^* and $-\alpha_i/\gamma_i\equiv\eta\cdot\mu$ thereafter. These ratios are by assumption the same for all, and thus, within a period, the price for a type is constant and types travel jointly. □

B.2. ADL model

For β^* users, the price in the untolled early period, the expected price in the mass, and the tolled price need to be equal. This implies $-\beta^*(-t^s+t^+)=\gamma\cdot t^-+(\mu+\eta)\beta^*(t^e-t^-)/2=\rho.$ The arrival period of the mass lasts t^e-t^-=*M*/*s*, and the untolled early period t^+-t^s =(*V*−*M*)/*s*. Using this, we get:

$$\rho[V]=\beta^*[V]\frac{V}{s}\frac{\eta+\mu}{1+\eta+\mu}, \tag{31}$$

$$M[V]=\frac{V}{1+\eta+\mu}. \tag{32}$$

The sum of the price over all users or total price (*TP*) follows:

$$\begin{aligned}TP[V]&=\int_0^V P^U[x]dx+\int_V^N P^T[x]dx=-t^s[V]\cdot\int_0^V\beta_i[x]dx+\rho[V]\cdot(N-V)-t^+[V]\cdot\int_V^N\beta_i[x]dx\\&=-t^s[V]\cdot\beta^L[V]\cdot V+\rho[V]\cdot(N-V)-t^+[V]\cdot\beta^H[V]\cdot\left(N-V\right).\end{aligned}$$

Using all this and a toll revenue of *TR*=*ρ*(*N*−*V*), we get the total cost eq. (13) in text.

The first order condition for minimising total cost is:[11]

$$\frac{\partial TC}{\partial V}=-\frac{E[B]\cdot\eta}{1+\eta}\frac{N}{s}+(\beta^{*}[V]\cdot\frac{1}{s}+\cdot\beta^{L}[V])\cdot V\cdot\frac{\eta+\mu}{2+\eta+\mu}=0\,. \tag{33}$$

For a general distribution of β, there is no closed-form solution, as $\beta^{L}[V]$ and $\beta^{*}[V]$ have no closed-form solution. Still, for a uniform distribution, with a mean of E[β] and range of d:

$$\mathrm{V}_{\mathrm{Uiniform}}=N\left(\frac{1}{3}-\frac{2\cdot\mathrm{E}[\beta]}{3\cdot d}+\sqrt{\frac{1}{9}+\frac{2\cdot E[\beta]}{9d}\left(\frac{2\cdot E[\beta]}{d}+\frac{\eta+6\cdot\eta/(\mu+\eta)-2}{1+\eta}\right)}\right); \tag{34}$$

which is a simplification of eq. (68) in Xiao et al. (2011).

B.3. Laih model

The f.o.c. for total cost minimisation is:

$$\frac{\partial TC}{\partial V}=\frac{\eta}{(1+\eta)s}\left(\beta^{*}[V]-\int_{V}^{N}\beta_{i}[x]dx\right)=\frac{\eta}{(1+\eta)s}\left(\beta^{*}[V]\cdot V-\beta^{H}[V](N-V)\right)=0; \tag{35}$$

which is a simpler than in the ADL model, but still does not have a closed-form solution. Yet, with a uniform distribution, it does, and again V is larger than with homogeneity:

$$V_{\mathrm{uniform}}=\frac{2\cdot d-4\cdot E[\beta]+\sqrt{d^{2}+2\cdot d\cdot E[\beta]+4\cdot E[\beta]\cdot E[\beta]}}{3\cdot d}\frac{N}{2}>V_{\mathrm{homogeneity}}=\frac{N}{2}. \tag{36}$$

Inserting this V into (34) and some algebra results in the relative efficiency (12) in text.

B.4. Braking model

The Δt is determined by that for the β^{*} type the prices when arriving as the last tolled user and as the first untolled-late users must be the same, and thus $\Delta t(\alpha^{*}+\gamma^{*})\equiv\Delta t\cdot(\mu+\eta)\cdot\beta^{*}$ should equal ρ. The timing t^{s} and t^{e} as well as t^{+} and t^{-} follow the same equations as homogeneity, although the equation for Δt differs. Finally, the toll can be determined from the condition $\rho=\beta^{*\cdot}(t^{+}-t^{s})$:

$$\rho[V]=\beta^{*}[V]\frac{V}{s}\frac{\eta+\mu}{\mu/\eta+\eta+\mu}.$$

The condition for minimising total costs in (14) is:

[11] Here $\partial\left(\beta^{H}[V]\cdot(N-V)\right)/\partial V=\partial\left(\int_{V}^{N}\beta_{i}[x]dx\right)/\partial V=-\beta^{*}[V]$, and $\partial\left(\beta^{L}[V]\cdot V\right)/\partial V=\partial\left(\int_{0}^{V}\beta_{i}[x]dx\right)/\partial V=\beta^{*}[V]$.

$$\begin{aligned}\frac{\partial TC}{\partial V} &= \frac{\eta}{1+\eta}\left(\left(\frac{\partial \Delta t[V]}{\partial V}\beta^{L}[V]\cdot V+\Delta t[V]\cdot\beta^{*}[V]\right)+\frac{1}{s}\left(\beta^{*}[V]-\beta^{H}[V]\left(N-V\right)\right)\right)=0\\ &=\frac{\eta}{1+\eta}\frac{1}{s}\left(\frac{V}{\mu/\eta+\eta+\mu}\left(\beta^{L}[V]\cdot V+\beta^{*}[V]\right)+\left(\beta^{*}[V]-\beta^{H}[V]\left(N-V\right)\right)\right)=0.\end{aligned} \tag{37}$$

Again, there is no closed-form solution for a general distribution. But for a uniform distribution:

$$V=\frac{N}{3}\left(1-2\cdot\frac{E[\beta]}{d}+\sqrt{1+2\frac{E[\beta]}{d}\left(2\left(\frac{E[\beta]}{d}\right)+\left(1-\frac{3\cdot\eta}{(1+\eta)(\eta+\mu)}\right)\right)}\right). \tag{38}$$

Appendix C: Coarse tolling and *α* heterogeneity

C.1. Generals of coarse tolling under α heterogeneity

The price for period without a mass departure follows the same pattern as in the NT equilibrium. For the tolled period, the price is:

$$P_i^T=\rho+\frac{\delta}{s}\left(\int_{\underline{\alpha}}^{\alpha_i} n_j^T\ d\alpha_j+\alpha_i\int_{\alpha_i}^{\bar{\alpha}}(n_j^T/\alpha_j)\ d\alpha_j\right), \tag{39}$$

where n_j^T is the density of users in tolled period *T* with α_j. The sum of the densities of type *j* users travelling in the three periods equals the total density for *j*: $n_j^T+n_j^{UE}+n_j^{UL}=n_j\equiv f_j[\alpha_j]\cdot N$. If a types uses the tolled period, a fraction $\eta/(1+\eta)$ uses arrives early and tolled, the remainder tolled and late. The prices in the untolled periods *without a mass departure* are:

$$P_i^{UE}=-\beta\cdot t^{+}+\frac{\beta}{s}\left(\int_{\underline{\alpha}}^{\alpha_i} n_j^{UE}d\alpha_j+\alpha_i\int_{\alpha_i}^{\bar{\alpha}}(n_j^{UE}/\alpha_j)\ d\alpha_j\right), \tag{40}$$

$$P_i^{UL}=\gamma\cdot t^{-}+\frac{\gamma}{s}\left(\int_{\underline{\alpha}}^{\alpha_i} n_j^{UL}\ d\alpha_j+\alpha_i\int_{\alpha_i}^{\bar{\alpha}}(n_j^{UL}/\alpha_j)\ d\alpha_j\right). \tag{41}$$

With a mass departure, the price follows:

$$P_i^{mass}=\gamma\cdot t^{-}+(\alpha_i+\gamma)\frac{M}{2\cdot s}=\gamma\cdot t^{-}+\beta(\mu_i+\eta)\frac{M}{2\cdot s}. \tag{42}$$

Proof of proposition 3: The mass-departure price is linear in α_i, while in the other periods it is strictly concave. Hence, if in contradiction of prop. 3, a group of types (e.g. all types with a α_i between 4 and 8) used the mass and another period, the prices in those periods could not be the same for all of them. This proves that this contradiction is not in equilibrium. □

Proof of proposition 4: Given (39)-(41), if the types *j* with $\alpha_i<\alpha_j$ that use these multiple periods were not spilt in fixed proportions over them, then *i*'s prices could not be the same in all these periods, as *i*'s price in period *k* depends on the weighed mean for period *k* of $1/\alpha_j$ over all types with α_i<α_j : $\int_{\alpha_i}^{\bar{\alpha}} (n_j^k / \alpha_j)\, d\alpha_j$. This effect is multiplied by δ in the tolled period, β in the untolled early period and γ in the untolled late period, and thus users need not be shared equally over periods. It is allowable that some types with $\alpha_k>\alpha_i$ do not use these periods at all, as then they do not directly affect the periods' prices. □

Proof of proposition 5: Again, we prove this by contradiction. Suppose that high-α users only drove tolled and faced the price in (39), and low-α users only drove in the untolled early period and had a price following (41). (However, similar arguments hold for any violation of prop. 5). Then, there would be an indifferent type with α_1^* whose prices in these periods would be equal. In the untolled period, α_1^* users would only face a schedule delay cost, as they would have the highest α_i of all untolled types; in the tolled period, they would only face the toll and travel time costs and would have the lowest price of all tolled users. If a user with $\alpha_i>\alpha_1^*$ would then move to the untolled early period, she would face the same zero travel time as a α_1^* user and thus the same price. Hence, this lowers her price, and proves that the violating set-up is not an equilibrium. □

The timings of the peak follow the same formulas from (19) as with homogeneity and proportional heterogeneity. If, of the V no-toll payers, V_L arrive late after t^*, these timings follow

$$\begin{aligned} t_s &= -\frac{\eta}{\eta+1}\frac{N-V}{s} - \frac{V-V_L}{s}, \\ t_e &= \frac{\eta}{\eta+1}\frac{N-V}{s} + \frac{V_L}{s}; \end{aligned} \tag{43}$$

where V_L is unknown up-front in the ADL model, but can be calculated in the other models. As with homogeneity, the toll is determined by equalising prices at t_s and t^+:

$$\rho = \beta\frac{V-V_L}{s}. \tag{44}$$

C.2. Laih model under α heterogeneity

Here, V_L (the number of users who arrive untolled and late) equals $V/(1+\eta)$ (as otherwise the prices of the early and late untolled periods would differ), and the coarse toll is the same as with homogeneity. The prices are found by filling in (39)-(41).

C.3. ADL model under α heterogeneity

Proof of proposition 7: If in contradiction of proposition 7 the lowest-α users where in the mass, their price from (42) would be:

$$p_{Mass} = = (\alpha_i+\gamma)\frac{M}{2\cdot s} + \gamma\left(\frac{N-V}{s}\frac{1}{1+\eta}\right). \tag{45}$$

Moving to the untolled early period and arrive at t^+, gives an out-of-equilibrium price of:

$$P_{out\text{-}of\text{-}equilibrium} = \delta\frac{N-V}{s} + \frac{\alpha_i\cdot\gamma}{1+\eta}\frac{(N-V)}{s\left(\bar{\alpha}-\alpha_2^*\right)}Ln\left[\frac{\bar{\alpha}}{\alpha_2^*}\right] + \alpha_i\cdot\beta\frac{V-M-\eta\frac{N-V}{1+\eta}}{s\left(\alpha_2^*-\alpha_1^*\right)}Ln\left[\frac{\alpha_2^*}{\alpha_1^*}\right]. \tag{46}$$

Here, α_1^* indicates the type that in the candidate equilibrium is indifferent between using the mass and the early-untolled period. The α_2^* is the type that is first to use both the tolled and the untolled early period, and these indifferent values follow $\underline{\alpha} < \alpha_1^* \le \alpha_2^* < \bar{\alpha}$. Eq. (46) is found by replacing in the limits of the integrals of (40) the α_i's with α_1^* (as this gives the schedule delays and travel times moving users would face) and simplifying using the uniform distribution.

It is attractive for a type *i* mass user to move out of equilibrium if:

$$2\cdot\alpha\cdot\gamma\cdot Ln\left[\frac{\bar{\alpha}}{\alpha_2^*}\right] + (1+2\eta)\Big(\Big(\underline{\alpha}-\alpha_1^*(\alpha+\gamma\Big) + 2\alpha\cdot\beta Ln\left[\frac{\alpha_2^*}{\alpha_1^*}\right]\Big) < 0\,; \tag{47}$$

Using that, in the contradicting equilibrium, for the indifferent α_1^* users the prices in the untolled early and mass periods should be the same, we get:

$$Ln\left[\alpha_2^*\right] = \frac{1}{2\alpha_1^*(\beta+\gamma)}\left(\left(2\cdot\eta\cdot\alpha_1^* + \alpha_1^* + \gamma + 2\gamma\eta\right)\left(\alpha_1^* - \underline{\alpha}\right) + 2\alpha_i\gamma\left(Ln\left[\alpha_1^*\right] - Ln\left[\underline{\alpha}\right]\right) + 4\cdot\gamma\cdot\alpha_1^*\cdot Ln\left[\alpha_1^*\right]\right). \tag{48}$$

Inserting this condition into (47) gives that violating the candidate equilibrium is attractive for a mass user if: $\alpha_i < \alpha_1^*$. Hence, for all types who strictly use the mass it is attractive to move out of the mass, and this completes the proof. □

The equilibrium price in the mass (for users with $\alpha_2^* \ge \alpha_i \ge \alpha_3^*$) is:

$$P_i^M = \gamma\cdot t^- + (\alpha_i+\gamma)\frac{M}{2\cdot s} = \delta\frac{N-V}{s} + (\alpha_i+\gamma)\frac{M}{2\cdot s}. \tag{49}$$

We attain the other prices by filling in (40) for the early-untolled period:

$$\begin{aligned}
P_i^1 &= \delta\frac{N-V}{s} + \frac{\beta}{s}\left(\left(\int_{\underline{\alpha}}^{\alpha_i} n_j d\alpha_j + \alpha_i\int_{\alpha_i}^{\alpha_1^*}\frac{n_j}{\alpha_j}\,d\alpha_j\right) + \frac{\eta}{1+2\eta}\left(\alpha_i\int_{\alpha_1^*}^{\alpha_2^*}\frac{n_j}{\alpha_j}\,d\alpha_j\right) + \frac{\eta}{1+2\eta}\left(\alpha_i\int_{\alpha_3^*}^{\bar{\alpha}}\frac{n_j}{\alpha_j}\,d\alpha_j\right)\right), && \alpha_i \le \alpha_1^*;\\
P_i^2 &= \delta\frac{N-V}{s} + \frac{\beta}{s}\left(\left(\int_{\underline{\alpha}}^{\alpha_1^*} n_j\,d\alpha_j\right) + \frac{\eta}{1+2\eta}\left(\int_{\alpha_1^*}^{\alpha_i} n_j d\alpha_j + \alpha_i\int_{\alpha_i}^{\alpha_2^*}\frac{n_j}{\alpha_j}\,d\alpha_j\right) + \frac{\eta}{1+2\eta}\left(\alpha_i\int_{\alpha_3^*}^{\bar{\alpha}}\frac{n_j}{\alpha_j}\,d\alpha_j\right)\right), && \alpha_1^* \le \alpha_i \le \alpha_2^*;\\
P_i^3 &= \delta\frac{N-V}{s} + \frac{\beta}{s}\left(\left(\int_{\underline{\alpha}}^{\alpha_1^*} n_j\,d\alpha_j\right) + \frac{\eta}{1+2\eta}\left(\int_{\alpha_1^*}^{\alpha_2^*} n_j d\alpha_j\right) + \frac{\eta}{1+2\eta}\left(\int_{\alpha_3^*}^{\alpha_i} n_j d\alpha_j + \alpha_i\int_{\alpha_i}^{\bar{\alpha}}\frac{n_j}{\alpha_j}\,d\alpha_j\right)\right), && \alpha_3^* \le \alpha_i.
\end{aligned} \tag{50}$$

Just as in the Laih model, having a group that only uses the untolled period only raises total cost. Hence, in the numerical optimisation, the optimal size of Group 1 is zero. Nevertheless, Group 3 does have a positive size in the numerical example and the types in this group use the tolled and untolled early period and have higher values of time than the mass users.

Define N^H as the number of users with $\alpha_i \geq \alpha_3^*$.We then have 3 unknowns: V, M, and N^H. We can calculate the indifferent values using the conditions $F[\alpha_3^*]\cdot N = N - N^H$, $(F[\alpha_3^*] - F[\alpha_2^*])\cdot N = M$ and $F[\alpha_1^*]\cdot N = V - M - (N - V)\cdot \eta/(1+\eta)$, but only if we define the distribution form. However, even then, there is not closed-form solution for the optimal V, M, and N^H.

C.4. Braking model under α heterogeneity

Proof of proposition 9. That the types with the highest values use all periods follows from proposition 5. That the lowest-values types only travel tolled follows from proposition 6 and that in optimum the untolled early period is shorter than the tolled early period.

The type that is first to brake at t^b, also travels during the early-untolled period; we will call its arrival time in the early-untolled period t^w. Travel time during any period $k = \{T, UE, UL\}$ for type i is (with n_j^k being the density of type j in period k)

$$\begin{aligned} &\frac{\beta}{s}\int_{\alpha_i}^{\bar{\alpha}} \frac{n_j^k}{\alpha_j}\, d\alpha_j && \text{if } t \leq t^*, \\ &\frac{\eta\cdot\beta}{s}\int_{\alpha_i}^{\bar{\alpha}} \frac{n_j^k}{\alpha_j}\, d\alpha_j && \text{if } t > t^*. \end{aligned} \tag{51}$$

Following proposition 4, for types that travel untolled early and late, it must be the case that $n_j^{UE} = \eta \cdot n_j^{UL}$. Combining this with (51) implies that that the travel times at t^w and t^- are the same. Hence, for the prices at t^w and t^- to be equal, their schedule delays must be the same, and thus $t^w = -\eta \cdot t^-$. The prices at t^+ and t^b must also be identical, and this implies $t^+ = -\eta \cdot t^b$. Using all this, we get $t^w = t^+ - \Delta t/\eta$. Consequently, t^w is before the end of the untolled period: $t^w < t^+ < 0$. Accordingly, there are intermediate types that use the *early*-untolled period, but not the *late*-untolled period. □

We find the prices by filling in the general equation (39) for the tolled period, as all types use this period. Types that only travel tolled are indicated by superscript 1 and have a price:

$$P_i^1 = \rho + \frac{\beta}{s}\left(\left(\int_{\underline{\alpha}}^{\alpha_i} n_j \, d\alpha_j \right) + \alpha_i \left(\int_{\alpha_i}^{\alpha_1^*} \frac{n_j}{\alpha_j} d\alpha_j + \frac{\eta}{1+2\eta}\int_{\alpha_1^*}^{\alpha_2^*} \frac{n_j}{\alpha_j}\, d\alpha_j + \frac{\eta}{2+2\eta}\int_{\alpha_2^*}^{\bar{\alpha}} \frac{n_j}{\alpha_j}\, d\alpha_j \right)\right), \qquad \alpha_i < \alpha_1^*. \tag{52}$$

For the intermediate group, the price is:

$$P_i^2 = \rho + \frac{\beta}{s}\left(\left(\int_{\underline{\alpha}}^{\alpha_1^*} n_j \; d\alpha_j + \frac{\eta}{1+2\eta} \int_{\alpha_1^*}^{\alpha_i} n_j \; d\alpha_j \right) + \alpha_i \left(\frac{\eta}{1+2\eta} \int_{\alpha_i}^{\alpha_2^*} \frac{n_j}{\alpha_j} d\alpha_j + \frac{\eta}{2+2\eta} \int_{\alpha_2^*}^{\bar{\alpha}} \frac{n_j}{\alpha_j} \; d\alpha_j \right) \right), \qquad \alpha_1^* \le \alpha_i < \alpha_2^*; \quad (53)$$

and for the highest values it is:

$$P_i^3 = \rho + \frac{\beta}{s}\left(\left(\int_{\underline{\alpha}}^{\alpha_1^*} n_j d\alpha_j + \frac{\eta}{1+2\eta} \int_{\alpha_1^*}^{\alpha_2^*} n_j d\alpha_j + \frac{\eta}{2+2\eta} \int_{\alpha_2^*}^{\alpha_i} n_j d\alpha_j \right) + \alpha_i \left(\frac{\eta}{2+2\eta} \int_{\alpha_i}^{\bar{\alpha}} \frac{n_j}{\alpha_j} d\alpha_j \right) \right), \qquad \alpha_2^* \le \alpha_i. \quad (54)$$

Appendix D: Coarse tolling and *γ* heterogeneity

D.1. All 3 models of coarse tolling

With *γ* heterogeneity, it easiest to minimize total cost to the level of the coarse toll, ρ, and thus now *V* is implied by *ρ*. Proposition 10 implies that, just as with homogeneity, prices at t_s and t^+ should be equal. As travel times are zero at these arrival moments, this means:

$$t_s = -\left(1 - F\left[\eta_1^*\right]\right) N \big/ s , \quad (55)$$

$$t^+ = t_s + \rho / \beta = -\left(1 - F\left[\eta_1^*\right]\right) N \big/ s + \rho / \beta . \quad (56)$$

The solution to η_2^* follows the same condition in all 3 models:

$$1 - F[\eta_1^*] - \int_{\eta_2^*}^{\eta_1^*} \left(\eta_j \cdot f[\eta_j] \right) d\eta_j = \frac{s \cdot \rho}{\beta \cdot N} \cdot \quad (57)$$

The type η_2^* users are the last to arrive during the tolled period at t^-. Formula (57) is found by equating the price for type η_2^* following (16) with the sum of the toll and schedule delay cost at t^- (i.e. with $\rho + \eta_2^* \cdot \left(F\left[\eta_1^*\right] - F\left[\eta_2^*\right] \right) N \big/ s$).

D.2. The Laih model under γ heterogeneity

In the Laih model, η_1^* can be found by equating, for the lowest-*γ* users, the price in (17) with their schedule delay costs for arrival on $t_e = F\left[\eta_1^*\right] \cdot N / s$:

$$1 - F[\eta_1^*] = \int_{\underline{\eta}}^{\eta_1^*} f[\eta_j] \left(\eta_j \right) d\eta_j . \quad (58)$$

This condition is the same as in the NT and FB equilibria and, accordingly, the η_1^* is the same and the timings of the peak also remain the same as before. This in turn implies that coarse tolling leaves prices unchanged in the Laih model. Using constraints (57) and (58) to determine the indifferent ratios, one can then optimise the system by minimising total cost to *ρ*.

D.3. The ADL model under γ heterogeneity

Different from in the Laih model, the η_1^* is now derived using that type η_2^* should be indifferent between the tolled period and using the mass:

$$\frac{1}{2}\left(\eta_2^* + \mu\right)F\left[\eta_2^*\right] = \frac{\rho \cdot s}{\beta \cdot N}; \tag{59}$$

D.4. The Braking model under γ heterogeneity

To find the indifferent ratios we again use that for type $\underline{\eta}$ the price should equal the schedule delay cost when arriving at $t_e = F\left[\eta_1^*\right] \cdot N / s$:

$$\underline{\eta} \cdot \Delta t \cdot \frac{s}{N} = \left(1 - F[\eta_1^*]\right) - \int_{\underline{\eta}}^{\eta_1^*} \left(f[\eta_j] \cdot \eta_j\right) d\eta_j. \tag{60}$$

This is the same condition as in the Laih model but for the addition of the cost from braking.

References

Arnott, R., de Palma, A., Lindsey, R., 1988. Schedule delay and departure time decisions with heterogeneous commuters. Transportation Research Record 1197, 56–67.

Arnott, R., de Palma, A., Lindsey, R., 1990. Economics of a bottleneck. Journal of urban Economics 27(1), 111–130.

Arnott, R., de Palma, A., Lindsey, R., 1993. A structural model of peak-period congestion: a traffic bottleneck with elastic demand. American Economic Review 83(1), 161–79.

Arnott, R., de Palma, A., Lindsey, R., 1994. The welfare effects of congestion tolls with heterogeneous commuters. Journal of Transport Economics and Policy 28(2), 139–161.

Arnott, R, Kraus, M., 1995. Financing capacity in the bottleneck model. *Journal of Urban Economics* 38(3), 272–290.

Börjesson, M., Kristoffersson, I., 2012. Estimating welfare effects of congestion charges in real world settings. CTS Working Paper 2012:13.

de Palma, A., Lindsey, R, 2002. Congestion pricing in the morning and evening peaks: A comparison using the Bottleneck Model. In: Proceedings of the 39th Annual Conference of the Canadian Transportation Research Forum: 2002 Transportation Visioning - 2002 and Beyond, Vancouver, Canada, 9–12 May 2004, 179–193.

Fosgerau, M., 2011. How a fast lane may replace a congestion toll. Transportation Research Part B 45(6), 845–851.

Fosgerau, M., Small, K.A., 2013. Hypercongestion in downtown metropolis. Journal of Urban Economics 76, 122–134.

Hall, J.D., 2013. Pareto improvements from lexus lanes: the case for pricing a portion of the lanes on congested highways.. In: proceedings of the Kuhmo NECTAR Conference on Transportation Economics: Annual Conference of the International Transportation Economics Association 2013.

Koster, P.R., Koster, H. 2013. Commuters' Preferences for Fast and Reliable Travel. Tinbergen Institute Discussion Paper 13-075

Laih, C.H., 1994. Queuing at a bottleneck with single and multi-step tolls. Transportation Research Part A, 28(3), 197–208.

Laih, C.H., 2004. Effects of the optimal step toll scheme on equilibrium commuter behavior. Applied Economics, 36(1), 59–81.

Lindsey, R., 2004a. Existence, uniqueness, and trip cost function properties of user equilibrium in the bottleneck model with multiple user classes. Transportation Science 38(3), 293–314.

Lindsey, R., 2004b. The welfare-distributional impacts of congestion pricing on a road network. In: Proceedings of the 39th Annual Conference of the Canadian Transportation Research Forum: 2004 Transportation Revolutions, Calgary, Canada, 9-12 May 2004, 149–163.

Lindsey, C.R., van den Berg, V.A.C., Verhoef, E.T., 2012. Step tolling with bottleneck queuing congestion. Journal of Urban Economics, 72(1), 46–59.

Newell, G.F., 1987. The morning commute for nonidentical travellers. Transportation Science 21(2), 74–88.

Small, K.A., Verhoef, E.T., 2007. The Economics of Urban Transportation. London: Routledge.

Small, K.A., Winston, C., Yan, J., 2005. Uncovering the distribution of motorists' preferences for travel time and reliability. Econometrica 73(4), 1367–1382.

van den Berg, V.A.C., 2012. Step-tolling with price-sensitive demand: Why more steps in the toll make the consumer better off. Transportation Research Part A, 46(10), 1608–1622

van den Berg, V.A.C., Verhoef, E.T., 2011a. Winning or Losing from Dynamic Bottleneck Congestion Pricing? The Distributional Effects of Road Pricing with Heterogeneity in Values of Time and Schedule Delay. Journal of Public Economics, 95(7–8), 983–992.

van den Berg, V.A.C., Verhoef, E.T., 2011b. Congestion tolling in the bottleneck model with heterogeneous values of time. Transportation Research Part B 45(1), 60–70.

Verhoef, E.T., Small, K.A., 2004. Product differentiation on roads: constrained congestion pricing with heterogeneous users. Journal of Transport Economics Policy 38(1), 127–156.

Vickrey, W.S., 1969. Congestion theory and transport investment. American Economic Review (Papers and Proceedings) 59(2), 251–260.

Vickrey, W.S., 1973. Pricing, metering, and efficiently using urban transportation facilities. Highway Research Record, 476, 36–48.

Xiao, F., Shen, W., Zhang, H.M., 2012. The morning commute under flat toll and tactical waiting. Transportation Research Part B 46(10), 1346–1359.

Xiao, F., Qian, Z., Zhang, H.M., 2011. The morning commute problem with coarse toll and nonidentical commuters. Networks and Spatial Economics 11(2), 343–369.

Zhang, L., Levinson, D. M., Zhu, S., 2008. Agent-based model of price competition, capacity choice, and product differentiation on congested networks. Journal of Transport Economics and Policy 42(3), 435–461.