Netw Spat Econ (2011) 11:343–369
DOI 10.1007/s11067-010-9141-8

The Morning Commute Problem with Coarse Toll and Nonidentical Commuters

Feng (Evan) Xiao · Zhen (Sean) Qian · H. Michael Zhang

Published online: 12 August 2010

Abstract This paper studies the morning commute problem under a flat peak-period toll (coarse toll) within the context of heterogeneous commuters. All the possible cumulative departure curves resulting from different choices of toll level and charging period are examined. The optimal toll schemes are then derived from minimizing the total travel cost of all commuters, excluding toll cost. We prove that at the optimum there will be no queue or capacity waste at the bottleneck at both the starting and ending points of the charging period for the type of Value-Of-Time (VOT) distribution considered in the paper. Moreover, the optimal coarse toll scheme is pareto-improving. Different from the homogeneous case, which can be regarded as a special case of the heterogeneous case, price discrimination occurs when commuters have different VOT. The optimal solution depends on the units in which the system cost is measured and we find that commuters in the middle pack of the VOT distribution are worse off by higher toll charges if the system cost is measured in money instead of time. A numerical example is provided at the end for demonstration.

Keywords Morning commute · Nonidentical commuters · Heterogeneity · Value-of-time · Coarse toll

1 Introduction

The morning commute problem has been widely studied since Vickrey published his classical paper in the late 1960's (Vickrey 1969). Equipped with a simple deterministic queuing model now known as the *bottleneck* or *point queue* model (e.g., Nie and Zhang 2005), Vickrey successfully explained the morning rush hour congestion by defining an equilibrium based on commuters' departure time

F. (E.) Xiao · Z. (S.) Qian · H. M. Zhang (✉)
Department of Civil & Environmental Engineering, University of California, Davis One Shields Avenue, Davis, CA 95616, USA
e-mail: hmzhang@ucdavis.edu

decisions: commuters arriving at work too early or too late will experience more schedule delay penalty, yet those arriving closer to the work starting time have to spend more time waiting in the queue. In the next several decades, Vickrey's bottleneck model was broadly extended by assuming different work starting times (e.g., Daganzo 1985), elastic demand (e.g., Arnott et al. 1993), heterogeneous commuters, and so on. Readers can refer to Ramadurai et al.'s paper (2010) for a comprehensive literature review related to this subject.

The previous studies, however, assume a continuous strategy space of commuters' departure time choices, which in real life is discrete in nature. This discreteness may come from the discrete strategy space (Levinson 2005) or the discrete change of the travel cost, e.g. a multi-step coarse toll during the rush hour (Arnott et al. 1990b; Laih 1994). Although in theory an optimal continuously changing toll scheme can totally eliminate the queuing delay at the bottleneck (Vickrey 1969) and time-dependent fine toll has been widely utilized to optimize the dynamic networks in theoretical studies (Wie and Tobin 1998; Lin et al. 2010), it can hardly be implemented in reality, not only because of the difficulty in collecting all the information required to derive the optimal toll rate, but the confusion it may cause to the commuters with continuously changing toll rate. In practice the coarse toll is more widely adopted.

The peak period coarse toll problem on a single bottleneck, where a flat toll is charged during part of the morning commuting period, has been studied in the literature in the context of homogeneous commuters. For the optimal toll charge and charging period, Arnott et al. (1990b) pointed out that the queue at the bottleneck at the starting and ending points of the charging period should be eliminated. Based on this important property, the optimal toll level, optimal starting and ending times of the tolling period, as well as the total system cost can all be easily calculated. The optimal toll level was found to be independent of the Value-Of-Time (VOT) attached to queuing delay. However, they did not provide details of how the queuing profile changes with respect to toll level and the choices of the starting and ending times of the tolling period, neither did they consider commuter heterogeneity. According to their numerical example, the one-step coarse toll provides slightly less than half of the efficiency of the fine (first-best) toll. Later they extended these results to a one-to-one, bi-bottleneck parallel network (Arnott et al. 1990a). Bernstein and El sanhouri (1994) corrected an error in that paper by announcing that if the demands are inter-dependent, the optimal toll level should depend on all the three VOT parameters attached to queuing, schedule-early and schedule-late delays; Laih (1994) revisited the single-bottleneck coarse toll problem by providing an easier way to calculate the optimal coarse toll without discussing the explicit evolution of the queue. Unfortunately, the methodology proposed in that paper is merely an approximation of the precise solution. His analysis based on the assumption that the flat toll will not alter the trip price of each commuter, which is generally not true. In fact, from our analysis we can see that an arbitrarily high coarse toll may induce a period that no one departs from home or even a period that no one passes the bottleneck. As a result, the coarse toll does alter the trip cost of each commuter. Especially, when the coarse toll scheme is optimized, everyone benefits, even if they have different values of time. Therefore, Laih's conclusion that for a one-step toll scheme, "the optimal toll level is half of the maximum optimal time-varying toll and can at most eliminate half of the total queuing time" no longer holds. Recently, Knockaert et al. (2009) studied the single step coarse

toll on a single bottleneck with inelastic demand. Different from our study, which assumes an anonymous coarse toll scheme for heterogeneous commuters, they investigated the efficiency improvement by differentiating the level and timing of the coarse toll across two groups of travelers, who still have the same VOT.

The commuter heterogeneity has been addressed differently in the literature due to various assumptions made. One way to classify these studies is based on if their models are discrete (there are finite classes of commuters) or continuous (the number of the classes is infinite). Most previous studies assume finite multi-class users: Van der Zijpp and Koolstra (2002) provided a generic algorithm that solves the departure time choice equilibrium given heterogeneous departure time preferences, arbitrary origin-bound and destination-bound rescheduling cost functions, and arbitrary queuing cost functions. Ramadurai et al. (2010) developed a linear complementarity formulation for solving the single bottleneck problem in discrete time and user classes. Lindsey (2004) investigated the existence and uniqueness of departure-time user equilibrium in the bottleneck model with multi-user classes. Li et al. (2008) studied a model combining the route, departure time and parking location choices. In that model, they introduced multiple groups of travelers, with each group having its own unit costs of queuing delay, arriving early and late penalties. Unlike those discrete models, Newell (1987) obtained certain analytical results from assuming a continuously distributed VOT. The queuing pattern was derived for a certain class of cost models and it was shown that a stable commuting pattern exists and is dictated by a certain fraction of travelers.

Existing models also differ by their definitions of heterogeneity: Newell (1987) defined "nonidentical" as different ratios of queuing time cost and schedule delay cost for each person. Since tolling was not considered, the departure time choice is only dependent on the distribution of this ratio and has nothing to do with the absolute VOT. Arnott et al. (1988) analyzed the departure time decisions of morning commuters who differ in three different ways: travel time and schedule delay costs, relative costs of schedule-early and late delay and work starting time. Huang (2000) dealt with pricing and modal split in a competitive mass transit/highway system with two groups of commuters that differ in their disutility from travel time, schedule-early delay and transit crowding. Instead of considering a coarse toll during the peak period on a bottleneck with fixed demand, most of the studies focus on a uniform toll throughout the whole rush hour, which makes the problem much easier to solve. Since a uniform toll covering the whole time period will not influence the departure time choices of the commuters when the number of commuters is fixed, people either assume an elastic demand (Braid 1989) or a parallel link or travel mode competing for the demand with the bottleneck (Tabuchi 1993; Braid 1996, Danielis and Marcucci 2002). The VI formulation proposed by Ramadurai et al. (2010) is capable of tackling a general assumption of heterogeneity among commuters, who have different values attached to not only queuing delay but schedule-early and schedule-late delays. However, they didn't examine the optimization problem with pricing.

In this paper we analytically solve the optimal one-step coarse toll scheme for a single bottleneck with nonidentical commuters. The study can be considered as an extension to Arnott et al.'s work (1990b) by involving heterogeneity of commuters.[1] To reasonably simplify the problem and obtain insightful analytical results, we

[1] It can be shown that Arnott et al.'s model is a special case of our heterogeneous model.

introduce just one random variable to describe the heterogeneity, which is dependent on the income level of the commuters. Different from the methodology used in Arnott et al. (1990b), we formulate a nonlinear optimization problem to solve the resulting equilibrium under a general setting of VOT distribution. We show how the departure profile evolves with respect to different choices of toll level and tolling period, and the changes in each individual commuter's cost after the coarse toll was applied, which have not been explicitly provided in the previous studies. By introducing the heterogeneity, we are also able to examine how the optimal departure pattern changes when the toll operator makes tradeoffs between cost and time and how the coarse toll benefits the commuters with different income levels.

2 No-toll equilibrium in the morning commute

In this section we first give a description for the no-toll equilibrium (NTE) at a single bottleneck during the morning rush hour. Although these results are well known, we include them here for completeness because it introduces the necessary concepts and problem setting for our coarse toll problem. In all subsequent discussions, we assume all the morning commuters have the same work starting time, t^*. The cumulative number of arrivals at the bottleneck by time t is $A(t)$ and the waiting time in queue for any commuter who passes the bottleneck at time t is $w(t)$. The total number of people commuting during the morning peak is N and the passing rate (bottleneck capacity) at the bottleneck is s. Without loss of generality, we assume that there is no travel time cost other than the queuing time cost at the bottleneck. Thus a commuter arrives at the bottleneck as soon as s/he departs from home and arrives at work immediately after leaving the bottleneck. First, the arrival rate at the bottleneck exceeds the passing rate and a queue builds up from time t_q; After the arrival rate goes down below the bottleneck capacity at point B, the queue dissipates linearly till it disappears at time $t_{q'}$. From the definition, equilibrium is obtained when no individual has an incentive to change his/her departure time. The cumulative departure curve (which in the rest of this paper we briefly call "the profile"), is defined as the cumulative departures from home as a function of time. The profile at NTE is drawn in Fig. 1.

We assume α is the unit monetary value attached to queuing delay time, β is the unit monetary value of schedule-early delay and γ is the unit monetary value of schedule-late delay. In accordance with empirical evidences and for the existence and uniqueness of the equilibrium, we assume the relation $\gamma>\alpha>\beta$ holds. Here to locate the profile accurately, we provide each endpoint of the departure curve a coordinate. The queue starting time t_q is set to be the origin, i.e. $t_q=0$. As described in Vickrey's paper (1969), at NTE the slope of $\overline{OB}$ is equal to $\alpha s/(\alpha-\beta)$ and the slope of $\overline{BC}$ is equal to $\alpha s/(\alpha+\gamma)$. Thus we can analytically solve the coordinates of point B

$$x_1 = \frac{\gamma(\alpha-\beta)}{\alpha(\gamma+\beta)}\frac{N}{s} \tag{1}$$

$$y_1 = \frac{N\gamma}{\gamma+\beta} \tag{2}$$

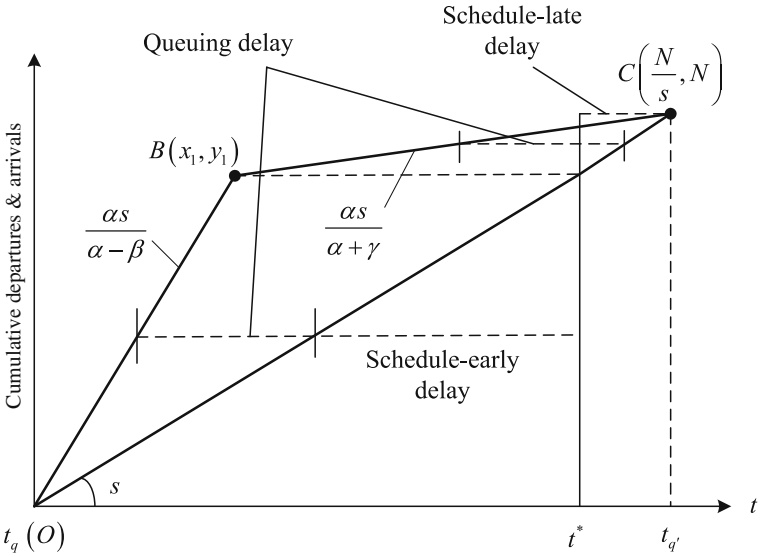

Fig. 1 Profile of NTE

The work starting time t^* is equal to the x-axis coordinate of the point on line $\overline{OC}$ (the queue discharge curve) which has y_1 as the y-axis coordinate. Thus it's not hard to obtain that

$$t^* = \frac{y_1}{s} = \frac{\gamma}{\gamma+\beta}\frac{N}{s} \tag{3}$$

The total travel cost of an individual who arrives at time t comprises two parts, the queuing delay cost and the schedule delay cost. A general expression of an individual's travel cost when s/he passes the bottleneck at time t is

$$C(t) = \alpha w(t) + \max\left(\beta\left(t^* - t\right), \gamma\left(t - t^*\right)\right) \tag{4}$$

If commuters are identical, at NTE every commuter will incur the same travel cost, which is equal to the schedule-early delay t^* of the first individual who experiences no queuing delay. Thus from Eq. (3), each commuter's travel cost at NTE can be calculated as

$$C = \beta t^* = \frac{\beta\gamma}{\gamma+\beta}\frac{N}{s} \tag{5}$$

And the total system cost at NTE is

$$\overline{TC} = \frac{\beta\gamma}{\gamma+\beta}\frac{N^2}{s} \tag{6}$$

In the real world, commuters differ in their VOT. The income level of an individual largely determines how much the individual's VOT, α, could be, yet the relative value α/β depends mainly on how flexible one's work schedule is. The relationship between income level and the flexibility of the job may vary: sometimes highly paid white-collar workers have more flexible work hours compared with blue-collar workers; yet sometimes due to the job's specific nature, high income

commuters may still have rigid work schedules, while low income commuters could have flexible work schedules. Nevertheless, it's still reasonable to believe that people who have a higher valuation of schedule delay will generally have a higher valuation of the time spent waiting at the bottleneck. In the absence of adequate empirical evidence, it's difficult to establish a certain relationship between α and α/β. To reasonably simplify the problem, we only focus on the heterogeneity in the valuation of time causing by the income level, but assume that everyone has the same work flexibility and the ratio between the unit penalties for the schedule-early delay and schedule-late delay is the same for all the commuters, i.e. α/β=constant, β/γ= constant, and α follows a distribution

$$F(\omega) = \Pr\{\alpha \leq \omega\} \tag{7}$$

Then the cost of the *vth* person becomes

$$C(v,t) = \alpha(v)\left(w(t) + \max\left(\eta_1\left(t^* - t\right), \eta_2\left(t - t^*\right)\right)\right) \tag{8}$$

where $\beta=\alpha\eta_1$, $\gamma=\alpha\eta_2$. From our assumption, it holds that $0<\eta_1<1<\eta_2$. For convenience, we arrange the commuters in increasing order of α: $\alpha(v)$ gives the *vth* person's value of queuing delay and $\alpha(v)$ is monotone and increasing in *v*. It's worth noting that without a toll, based on our assumption that α/β and β/γ are constant for all the commuters, the order in which they depart in the equilibrium is indeterminate.

An optimal time-dependent toll can be found to totally eliminate the queuing delay in the system. Under this optimal toll, the departure is evenly distributed throughout the time interval $\left(t_q, t_{q'}\right)$, at the rate of the bottleneck capacity, *s*. Arnott et al. (1994) observed that at system optimum, the order of the commuters is dependent on the absolute value of β and γ. Under our assumption of heterogeneity, β and γ are both proportional to α. Thus in our case, commuters with higher value of α will depart closer to the work starting time t^* under the optimal time-dependent toll.

3 Departure profiles under coarse toll

In reality it is impractical to implement a continuously changing optimal toll because of the difficulty in collecting all the information needed for deriving the toll, and the confusion it could cause the public with its too frequent toll rate changes. On the other hand, an approximate form to such a toll, which divides the peak periods into several tolling intervals and a flat toll is charged in each tolling interval, can be implemented with ease (Two examples are the multi-step tolling scheme on the State Route 91 express lane in Orange County in California, U.S.A and the Stockholm congestion charge in Sweden). In this paper, we deal with a special case of such coarse tolls, one with only a single toll rate and tolling period. The solution of this coarse toll problem provides insights to more refined coarse tolls with multiple toll rates and tolling periods.

A coarse toll is defined to be a flat charge ρ to the commuters passing the bottleneck within a time interval $[t^+, t^-]$(Arnott et al. 1990b). Since the coarse toll is defined as a rush hour tolling scheme, it's reasonable to assume that the toll is applied at the time

after the queue forms and before work starts, $t^+ \in \left[t_q, t^*\right]$ and lifted after work starts, $t^- \in \left[t^*, t_{q'}\right]$. Thus every selection of the three parameters (ρ, t^+, t^-) represents a tolling pattern, which also determines a unique departure profile.

After a coarse toll is imposed, only those with high VOT will travel within $[t^+, t^-]$. If we assume that the *Vth* commuter is the commuter with the lowest VOT among those traveling within the tolling period, then this commuter will have the same VOT $\alpha(v)$ with the highest VOT among those people who travels outside $[t^+, t^-]$, i.e. there's no difference to the *Vth* commuter whether he/she chooses to travel inside or outside. The proof is straightforward: given a toll rate ρ, we assume that at the equilibrium the difference of delay costs between traveling inside and outside, which is measured by generalized queuing time (since we assume β,γ are both proportional to α, we can translate the schedule delay into equivalent queuing delay), is Δt. Then the *Vth* commuter, where V is given by $\Delta t = \rho/\alpha(V)$, will experience the same travel cost no matter traveling inside or outside $[t^+, t^-]$. For any commuter who has a VOT $\alpha(v) < \alpha(V)$, s/he will stay outside because the cost of traveling outside is lower than traveling inside, reversely.

Undoubtedly, compared with the no-toll case queue lengths at the bottleneck within time interval $[t^+, t^-]$ will be reduced after a flat toll is charged. Some of the commuters who travel inside $[t^+, t^-]$ previously with low VOT will be forced to travel outside of it, either before t^+ or after t^-, yet the profile within $[t^+, t^-]$ is still similar with the original profile in NTE, because a uniform toll has no effect on the departure time choices. For those commuters who travel outside $[t^+, t^-]$, they also follow the same departure rate as in NTE. However, it's not clear what the profile will be like around the starting and ending points of the tolling period.

3.1 The profile without capacity waste

When the toll is relatively low, after some initial departures there could be a while no one departs before t^+, but commuters pass the bottleneck all the time (the capacity of the bottleneck is fully utilized), as shown in Fig. 2.

In this figure, let B denote the last departure leaving the bottleneck before tolling, C the first departure after B, D the tolling start time, E the departure who arrives at work on time, F the last tolled departure and G the last departure. From the definition of equilibrium, the last person who passes the bottleneck before toll starts should have the same cost as the first person who pays the toll. Thus the commuter departing at point B experiences a longer queuing time of $\overline{BD}$ than the commuter departing at point C who experiences a queuing time $\overline{CD}$ and an additional toll ρ. $s/(1-\eta_2)$ and $s/(1+\eta_2)$ represent respectively the departure rates for the commuters who experience queuing delay and schedule delay. We define $m = \overline{BC} - \overline{BD}$ to indicate if the capacity of the bottleneck is fully used during the whole commuting period. Here m stands for the difference between the departure gap before the tolling period and the queuing delay of the last commuter who passes the bottleneck before toll. If $m<0$, which is the case here, the capacity is not wasted; if $m>0$, there is capacity waste, which will be discussed later. It's not hard to obtain the relationship that

$$\overline{BC} = y_1/s + m - x_1. \tag{9}$$

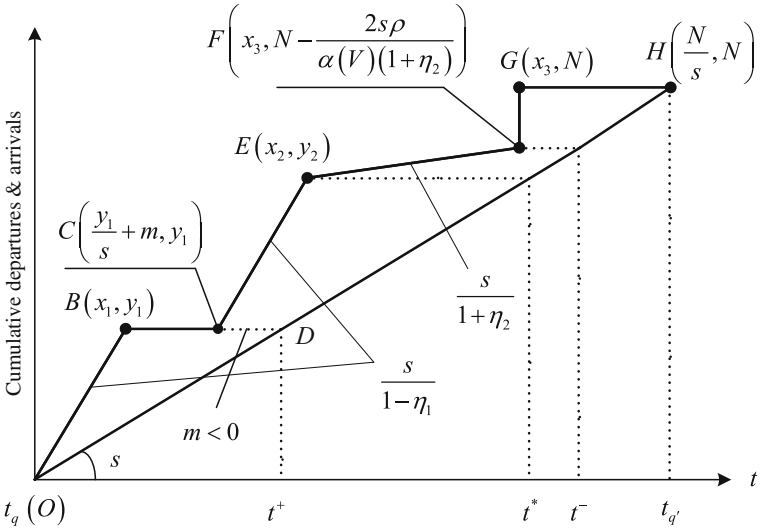

Fig. 2 Profile 1 without capacity waste

And because the two commuters departing at B and C experience the same schedule-early delay, the toll should be equal to the difference between the queuing delay cost, i.e.

$$\rho = \alpha(V)\overline{BC} \tag{10}$$

Similarly, at the time when the toll is lifted, the last person who pays the toll should have the same travel cost as the first person who passes the bottleneck after the toll is lifted. Under the equilibrium, the rest of commuters will depart together immediately after the toll is lifted. Otherwise, the commuter who passes the bottleneck later will experience both more queuing delay and schedule delay, which does not satisfy the equilibrium condition. Since the position in the queue for these commuters is random, everyone has the same expectation of queuing and schedule-late delays. The expected total travel cost is ρ units higher than the total cost of the commuters traveling inside the tolling period. From 0, we observe that the expected travel cost for passing the bottleneck after t^- should be equal to that of the middle commuter who arrives simultaneously at the bottleneck after the toll is lifted. It can be easily calculated that the size of the number of commuters passing the bottleneck after the tolling period is

$$\overline{FG} = 2s\rho/(\alpha(V)(1+\eta_2)) \tag{11}$$

3.2 Profiles with capacity waste

When the toll is set too high or charging time interval is over-stretched, there could be a while that no one travels through the bottleneck, which we refer to as "capacity waste". As we will show in the following, the capacity waste could happen not only at the starting point but also the ending point of the toll.

3.2.1 Capacity waste at the starting point of the toll

This profile represents the situation when there is a while that no one travels through the bottleneck immediately after the toll is imposed, which could be caused by either a high toll level or a t^+ far from t^* or both (See Fig. 3).

Now at t^+, the last person who passes the bottleneck without paying the toll experiences a longer schedule-early delay than the first person who pays the toll. In this case, Eq. (9) still holds, except that $m \geq 0$. There will be an early time interval in the tolling period that no one departs from home until a commuter's schedule-early delay decreases sufficiently to compensate for the toll charge. Thus in this profile, the capacity waste exists during the tolling period $[t^+, t^-]$. The toll should be equal to the sum of the differences between the queuing and schedule-early delay costs, i.e.

$$\rho = \alpha(V)\left(\overline{BD} + \eta_1 \overline{CD}\right) \tag{12}$$

3.2.2 Capacity waste at the ending point of the toll

Similar with the case of capacity waste at the starting point of the toll, when toll is too high or t^- is far away from t^*, there could also be capacity waste at the ending point of the toll (See Fig. 4).

3.2.3 Other profiles

Obviously, there could be another possible profile which combines the two profiles with capacity wastes in Figs. 3 and 4, e.g. capacity waste exists at both the starting and the ending points of the toll (as shown in Fig. 5).

And for the extreme situation that when the toll is sufficiently high, there could even be a while around t^* that no one will pass the bottleneck (as shown in Fig. 6).

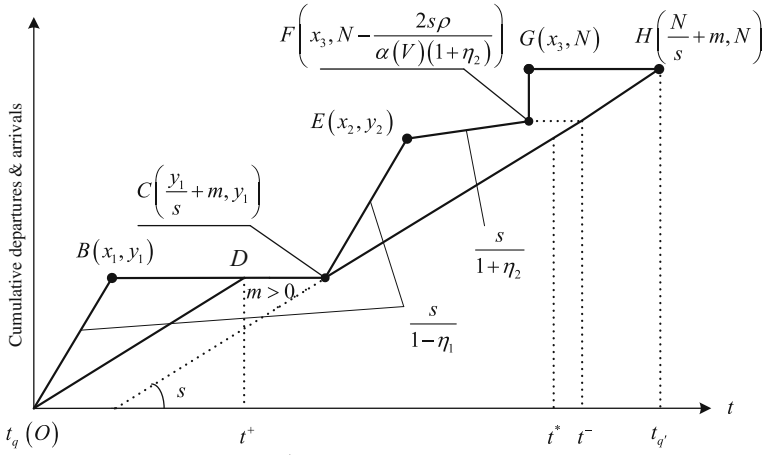

Fig. 3 Profile 2 with capacity waste at t^+

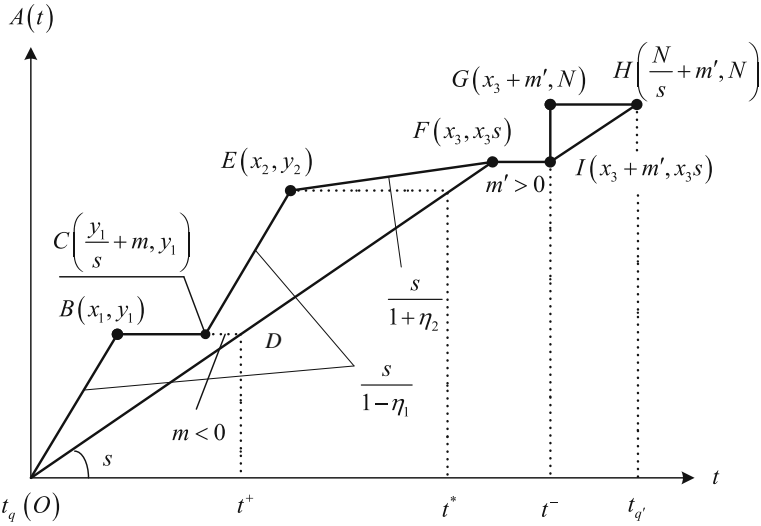

Fig. 4 Profile 3 with capacity waste at t^-

4 The optimal coarse toll scheme

As we have discussed in Section 3, there could be totally five profiles under different settings of coarse toll. In this section, we want to find the optimal profile through solving a series of optimization problems. For the convenience of discussion, we first define several quantities used later in this section

$$A_1 = \int_0^V \alpha(v)dv \tag{13}$$

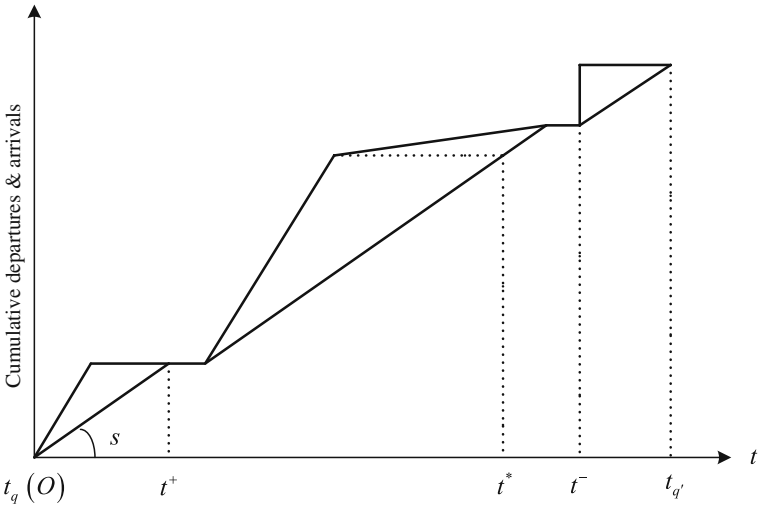

Fig. 5 Profile 4 with capacity waste at both t^+ and t^-

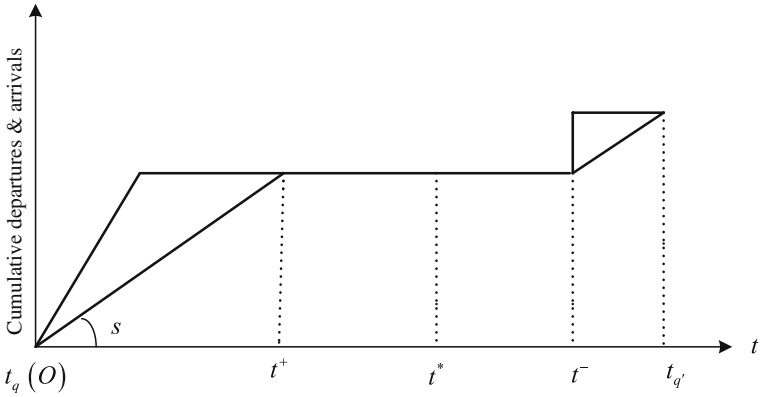

Fig. 6 Extreme profile 5 when no one passes the bottleneck between t^+ and t^-

$$A_2 = \int_V^N \alpha(v)dv \tag{14}$$

$$K = A_1 + A_2 = \int_0^N \alpha(v)dv \tag{15}$$

where V is the number of the commuters who travel outside the charging time interval $[t^+, t^-]$, which, in Figs. 2, 3, 4, equals to $y_1 + \overline{FG}$.

4.1 The optimal coarse toll scheme for profile 1

When the toll level and tolling period are such that profile 1 results (See Fig. 2), one can find the optimal toll scheme by minimizing the system cost, TC (which takes into consideration both delay cost and toll revenue), subject to the constraints that are defined from this profile. With A_1,A_2 and K introduced earlier, the resulting nonlinear constrained optimization problem is:

$$\min_{V,x_2,y_2} TC = \frac{\eta_1 y_2}{s} A_1 + \left(\frac{y_2}{s} - x_2\right) A_2 = \frac{\eta_1 A_1 + A_2}{s} y_2 - A_2 x_2 \tag{16}$$

s.t.

$$\frac{y_1}{x_1} = \frac{s}{1-\eta_1} \tag{17}$$

$$\frac{y_2 - y_1}{x_2 - \frac{y_1}{s} - m} = \frac{s}{1-\eta_1} \tag{18}$$

$$\frac{N - \frac{2s\rho}{\alpha(V)(1+\eta_2)} - y_2}{x_3 - x_2} = \frac{s}{1+\eta_2} \tag{19}$$

$$\frac{y_1}{s} + m - x_1 = \frac{\rho}{\alpha(V)} \tag{20}$$

$$m \leq 0 \tag{21}$$

$$\frac{N - \frac{2s\rho}{\alpha(V)(1+\eta_2)}}{x_3} \geq s \tag{22}$$

$$V = \frac{2s\rho}{\alpha(V)(1+\eta_2)} + y_1 \tag{23}$$

The objective function (16) calculates the total system cost excluding toll revenue, which is equal to the total delay cost (including queuing delay and schedule delay). The first term in the objective function calculates the total delay cost of the commuters traveling outside the charging time interval by looking at the first commuter passing the bottleneck with only schedule delay; while the second term calculates the total delay cost of the commuter traveling inside the interval by looking at the commuter who arrives at work on time, experiencing only queuing delay. Constraints (17), (18) and (19) guarantee the departure rates satisfying the equilibrium. Constraint (20) is from Eqs. (9) and (10). Constraints (22) ensures that the queue at time t^- is greater than or equal to 0. In other words, constraints (17)–(22) ensures that the toll scheme produces a profile consistent with the one shown in Fig. 2. Constraint (23) states that the total number of commuters whose VOT are greater than $\alpha(V)$ should be equal to the number of commuters traveling within the time interval $[t^+, t^-]$. From some straightforward algebraic manipulations, the above optimization problem can be simplified (See Appendix A for details)

$$\begin{aligned} \min_{V, x_3, \rho} TC = {} & \frac{(1+\eta_2)\eta_1 K}{\eta_1+\eta_2}\left(\frac{\rho}{\alpha(V)(1-\eta_1)} + \left(\frac{N}{s} - \frac{2\rho}{\alpha(V)(1+\eta_2)}\right) - \frac{1}{1+\eta_2}x_3\right) \\ & - \left(\frac{K}{1-\eta_1} - A_1\right)\frac{\rho}{\alpha(V)} \end{aligned} \tag{24}$$

s.t.

$$x_3 \leq \frac{N}{s} - \frac{2\rho}{\alpha(V)(1+\eta_2)} \tag{25}$$

$$m = \frac{\rho}{\alpha(V)}\frac{1+2\eta_1+\eta_2}{1+\eta_2} - \frac{\eta_1 V}{s} \tag{26}$$

$$m \leq 0 \tag{27}$$

Obviously, the simplified problem still has three variables $(x_3, V, \rho)$to be solved. From the first-order optimality conditions one can show that constraint (25) should be binding, which indicates that the queue at time t^- is equal to 0 at the optimum.

Therefore the problem can be further simplified by removing constraints (25) and (26)

$$\min_{V,m} TC = \frac{\eta_1\eta_2 KN}{(\eta_1+\eta_2)s} + \left(\frac{1+\eta_2}{1+2\eta_1+\eta_2}A_1 - \frac{\eta_2 K}{\eta_1+\eta_2}\right)m + \eta_1\left(\frac{1+\eta_2}{1+2\eta_1+\eta_2}A_1 - \frac{\eta_2 K}{\eta_1+\eta_2}\right)\frac{V}{s} \tag{28}$$

s.t.

$$m \le 0 \tag{29}$$

We call this Problem I.

4.2 The optimal coarse toll scheme for profile 2

Under profile 2 (See Fig. 3) we cannot simply tell if the system cost is reduced by moving t^+ to the right: on one hand, the capacity waste and the total delay cost of each individual are reduced; yet on the other hand, the area of the triangle $\overline{OBD}$ is increased, which indicates less toll revenue and more deadweight loss of queuing delay. However, one can find the optimal toll scheme for this profile by solving a nonlinear optimization problem

$$\min_{V,x_2,y_2} TC = \eta_1\left(\frac{y_2}{s}+m\right)A_1 + \left(\frac{y_2}{s}+m-x_2\right)A_2 \tag{30}$$

s.t.

$$\frac{y_1}{x_1} = \frac{s}{1-\eta_1} \tag{31}$$

$$\frac{y_2-y_1}{x_2-\frac{y_1}{s}-m} = \frac{s}{1-\eta_1} \tag{32}$$

$$\frac{N-\frac{2s\rho}{\alpha(V)(1+\eta_2)}-y_2}{x_3-x_2} = \frac{s}{1+\eta_2} \tag{33}$$

$$\frac{y_1}{s}+\eta_1 m - x_1 = \frac{\rho}{\alpha(V)} \tag{34}$$

$$m \ge 0 \tag{35}$$

$$\frac{N-\frac{2s\rho}{\alpha(V)(1+\eta_2)}}{x_3-m} \ge s \tag{36}$$

$$V = \frac{2s\rho}{\alpha(V)(1+\eta_2)} + y_1 \tag{37}$$

The total system cost is calculated by summing up the total delay costs of the commuters passing the bottleneck within and outside the tolling interval. Constraint (37) ensures that the total number of commuters whose VOT are greater than $\alpha(V)$ should be equal to the number of commuters traveling within the time interval $[t^+, t^-]$. The problem can be further simplified (See Appendix B for details)

$$\min_{V,x_2,y_2} TC = \frac{(1+\eta_2)\eta_1 K}{\eta_1+\eta_2}\left(\frac{\rho}{\alpha(V)(1-\eta_1)} + \left(\frac{N}{s} - \frac{2\rho}{\alpha(V)(1+\eta_2)}\right) - \frac{1}{1+\eta_2}x_3 + m\right) - \left(\frac{K}{1-\eta_1} - A_1\right)\frac{\rho}{\alpha(V)} \tag{38}$$

s.t.

$$x_3 \le \frac{N - \frac{2s\rho}{\alpha(V)(1+\eta_2)}}{s} + m \tag{39}$$

$$m = \frac{\rho}{\eta_1\alpha(V)}\frac{1+2\eta_1+\eta_2}{1+\eta_2} - \frac{V}{s} \tag{40}$$

$$0 \le m \le \frac{\rho}{\eta_1\alpha(V)} \tag{41}$$

Again, from first-order optimality conditions one finds that constraint (39) is binding, indicating that there is no queue at time t^- at the optimum. By replacement of variables, we obtain the following simplified problem, which we refer as Problem II.

$$\min_{V,m} TC = \frac{\eta_1\eta_2 K}{\eta_1+\eta_2}\frac{N}{s} + \left(\frac{\eta_1(1+\eta_2)}{1+2\eta_1+\eta_2}A_1 + \frac{\eta_1(1+\eta_2)K}{\eta_1+\eta_2}\right)m + \eta_1\left(\frac{1+\eta_2}{1+2\eta_1+\eta_2}A_1 - \frac{\eta_2 K}{\eta_1+\eta_2}\right)\frac{V}{s} \tag{42}$$

s.t.

$$0 \le m \le \frac{\rho}{\eta_1\alpha(V)} \tag{43}$$

4.3 The optimal coarse toll scheme for profile 3

Similarly, under profile 3 (See Fig. 4) we still cannot tell if the system cost is reduced by simply moving t^- to the left: On the one hand, the capacity waste and the total delay cost of each individual are reduced; while on the other hand, the number of commuters traveling after the toll is increased, which indicates more deadweight loss of queuing delay. If we assume the capacity waste before t^- is m', we have

$$\frac{2\eta_2}{1+\eta_2}m' = x_3 - \left(\frac{N}{s} - \frac{2s\rho}{\alpha(V)(1+\eta_2)}\right) \tag{44}$$

Thus when capacity waste exists at the ending point of the toll, it has to be satisfied that

$$x_3 \geq \frac{N}{s} - \frac{2\rho}{\alpha(V)(1+\eta_2)} \tag{45}$$

The optimal toll scheme for this profile can be solved by the following nonlinear optimization problem:

$$\min_{V,x_2,y_2} TC = \frac{\eta_1 y_2}{s} A_1 + \left(\frac{y_2}{s} - x_2\right) A_2 = \frac{\eta_1 A_1 + A_2}{s} y_2 - A_2 x_2 \tag{46}$$

s.t.

$$\frac{y_1}{x_1} = \frac{s}{1-\eta_1} \tag{47}$$

$$\frac{y_2 - y_1}{x_2 - \frac{y_1}{s} - m} = \frac{s}{1-\eta_1} \tag{48}$$

$$\frac{x_3 s - y_2}{x_3 - x_2} = \frac{s}{1+\eta_2} \tag{49}$$

$$\frac{y_1}{s} + m - x_1 = \frac{\rho}{\alpha(V)} \tag{50}$$

$$m \leq 0 \tag{51}$$

$$x_3 \geq \frac{N}{s} - \frac{2\rho}{\alpha(V)(1+\eta_2)} \tag{52}$$

$$V = \frac{2s\rho}{\alpha(V)(1+\eta_2)} + y_1 \tag{53}$$

By simplifying this problem, we have

$$\min_{V,x_3,\rho} TC = \left(A_1 - \frac{\eta_2 K}{\eta_1 + \eta_2}\right) \frac{\rho}{\alpha(V)} + \frac{\eta_1 \eta_2 K}{\eta_1 + \eta_2} x_3 \tag{54}$$

s.t.

$$x_3 \geq \frac{N}{s} - \frac{2\rho}{\alpha(V)(1+\eta_2)} \tag{55}$$

$$m = \frac{\rho}{\alpha(V)} \frac{1 + 2\eta_1 + \eta_2}{1+\eta_2} - \frac{\eta_1 V}{s} \tag{56}$$

$$m \leq 0 \tag{57}$$

Again, from first-order optimality conditions one finds that constraint (55) is binding, which indicates that at optimum, the capacity waste does not exist. By replacing the variables, we find this problem reduces to Problem I.

4.4 The unified optimal coarse toll problem

In the previous discussion we have already shown that the queue at the ending point of the toll will be eliminated by optimizing the system cost. Here we further combine Problems I and II together to form a unified problem. The benefit is that constraints (29) and (43) related to m will be removed. The optimal coarse toll scheme for a bottleneck with nonidentical commuters can thus be solved by the following nonlinear optimization problem

$$\min_{V} TC = \frac{\eta_1\eta_2 K}{\eta_1+\eta_2}\frac{N}{s} + \lambda|m| + \eta_1\left(\frac{1+\eta_2}{1+2\eta_1+\eta_2}A_1 - \frac{\eta_2 K}{\eta_1+\eta_2}\right)\frac{V}{s} \tag{58}$$

s.t.

$$\lambda = \begin{cases} \dfrac{\eta_2 K}{\eta_1+\eta_2} - \dfrac{1+\eta_2}{1+2\eta_1+\eta_2}A_1, & m \le 0 \\ \dfrac{\eta_1(1+\eta_2)}{1+2\eta_1+\eta_2}A_1 + \dfrac{\eta_1(1+\eta_2)K}{\eta_1+\eta_2}, & 0 < m \le \dfrac{\rho}{\eta_1\alpha(V)} \end{cases} \tag{59}$$

Because $0\eta_1 1\eta_2$, the term

$$\frac{(1+2\eta_1+\eta_2)\eta_2}{(\eta_1+\eta_2)(1+\eta_2)} = \frac{\eta_2+2\eta_1\eta_2+\eta_2^2}{\eta_1+\eta_2+\eta_1\eta_2+\eta_2^2} = 1 + \frac{\eta_1(\eta_2-1)}{\eta_1+\eta_2+\eta_1\eta_2+\eta_2^2} > 1 \tag{60}$$

And because $A_1 \le K$, from (60) we have

$$\frac{\eta_2 K}{\eta_1+\eta_2} - \frac{1+\eta_2}{1+2\eta_1+\eta_2}A_1 = \frac{1+\eta_2}{1+2\eta_1+\eta_2}\left(\frac{(1+2\eta_1+\eta_2)\eta_2 K}{(\eta_1+\eta_2)(1+\eta_2)} - A_1\right) > 0 \tag{61}$$

Since the coefficient of m is always greater than zero, to reduce the system cost m has to be 0. The two profiles then reduce to the same profile without capacity waste or queue at the starting time of the toll and the minimized total system cost is the same.

Following a similar deduction procedure, it's not hard to conclude that profile 4 in Fig. 5 will reduce to the same profile when optimized and profile 5 in Fig. 6 is not optimal. To save space, the proofs are not presented here.

From all the above discussions, we are thus able to conclude that for heterogeneous commuters, where heterogeneity is defined by Eq. (8), there will be no queue or capacity waste at times t_s^+ and t_s^-, when the total system cost is minimized with optimal coarse toll level ρ_s and optimal tolling window $\left(t_s^+, t_s^-\right)$.

The optimal coarse toll level and the charging time interval can be calculated when the system cost is measured in money

$$\rho_m^s = \frac{(1+\eta_2)\eta_1}{1+2\eta_1+\eta_2}\frac{V}{s}\alpha(V) \tag{62}$$

$$t_s^+ = \frac{1+\eta_2}{1+2\eta_1+\eta_2}\frac{V}{s} \tag{63}$$

$$t_s^- = \frac{N}{s} - \frac{2\eta_1}{1+2\eta_1+\eta_2}\frac{V}{s} \tag{64}$$

The minimum total system cost is

$$TC_{\min} = \frac{\eta_1\eta_2 KN}{(\eta_1+\eta_2)s} - V^2\alpha(V)\frac{\eta_1(1+\eta_2)}{(1+2\eta_1+\eta_2)s} \tag{65}$$

Substituting (62) into (28) we have

$$TC(V) = \frac{\eta_1\eta_2 KN}{(\eta_1+\eta_2)s} + \eta_1\left(\frac{1+\eta_2}{1+2\eta_1+\eta_2}A_1 - \frac{\eta_2 K}{\eta_1+\eta_2}\right)\frac{V}{s} \tag{66}$$

It can be shown that $\frac{\partial TC(V)}{\partial V}|_{V=0} < 0$, $\frac{\partial TC(V)}{\partial V}|_{V=N} > 0$ and $\frac{\partial^2 TC(V)}{\partial V^2} > 0$, which ensures optimal V falls in [0,N]. By taking $\frac{\partial TC(V)}{\partial V} = 0$, we obtain the first-order optimality condition with only V as variable:

$$\int_0^V \alpha(v)dv + V\alpha(V) = \frac{\eta_2(1+2\eta_1+\eta_2)}{(\eta_1+\eta_2)(1+\eta_2)}\int_0^N \alpha(v)dv \tag{67}$$

Equation (67) implicitly determines the amount of commuters V who are traveling outside the tolling period after toll is imposed, as long as the VOT distribution of the commuter population is known. Generally we cannot get the explicit expression of V except for simple forms of VOT distribution. For instance, if VOT follows a uniform distribution, i.e. $\alpha(v) = b + av$, $a>0$, $b>0$ we can obtain that

$$V = \frac{\sqrt{4b^2 + 6a\frac{\eta_2(1+2\eta_1+\eta_2)}{(\eta_1+\eta_2)(1+\eta_2)}\left(bN + \frac{1}{2}aN^2\right)} - 2b}{3a} \tag{68}$$

5 Discussions and examples

5.1 A special case: identical commuters

A special case is that the commuters are identical. If every commuter has the same α, the number of commuters who pass through the bottleneck outside $[t^+, t^-]$ can be easily solved from Eq. (67)

$$V = \frac{\gamma(\gamma+\alpha+2\beta)}{2(\alpha+\gamma)(\beta+\gamma)}N > \frac{N}{2} \tag{69}$$

Equation (69) states that more commuters will choose to pass the bottleneck outside $\left[t_s^+, t_s^-\right]$ to avoid the toll. In other words, to minimize the system cost, the

tolling time window cannot be longer than half of the entire morning commute period. Solving for the optimal toll gives

$$\rho^s = \frac{\beta\gamma}{2(\beta+\gamma)}\left(\frac{N}{s}\right) \tag{70}$$

which is consistent with the optimal coarse toll level obtained by Arnott et al. (1990b).

Furthermore, we use a numerical example to show the change in system cost with respect to the combination choices of ρ and m when all the commuters have the same VOT. For comparison, we adopt the same VOT as the example provided in Arnott et al's paper (1990b): α=6.4, β=3.9 and γ=15.21. To compare the total costs under different pairs of (ρ,m), we assume the demand N=100 and the passing rate s=50. The results are shown in Fig. 7.

We notice that no matter how much the coarse toll is, the system cost is always minimized at m=0. The inspiration from this result is that once a toll level is decided, the tolling time interval has to be carefully chosen: when the time interval is too long, the capacity of the bottleneck cannot be fully utilized, while if the time interval is too short, an additional amount of deadweight loss of queuing delay is induced. The optimum is obtained at a toll level of \$3.1, which is consistent with the result in Arnott et al. (1990b). The total travel cost saved by the optimal coarse toll is 27.08%.

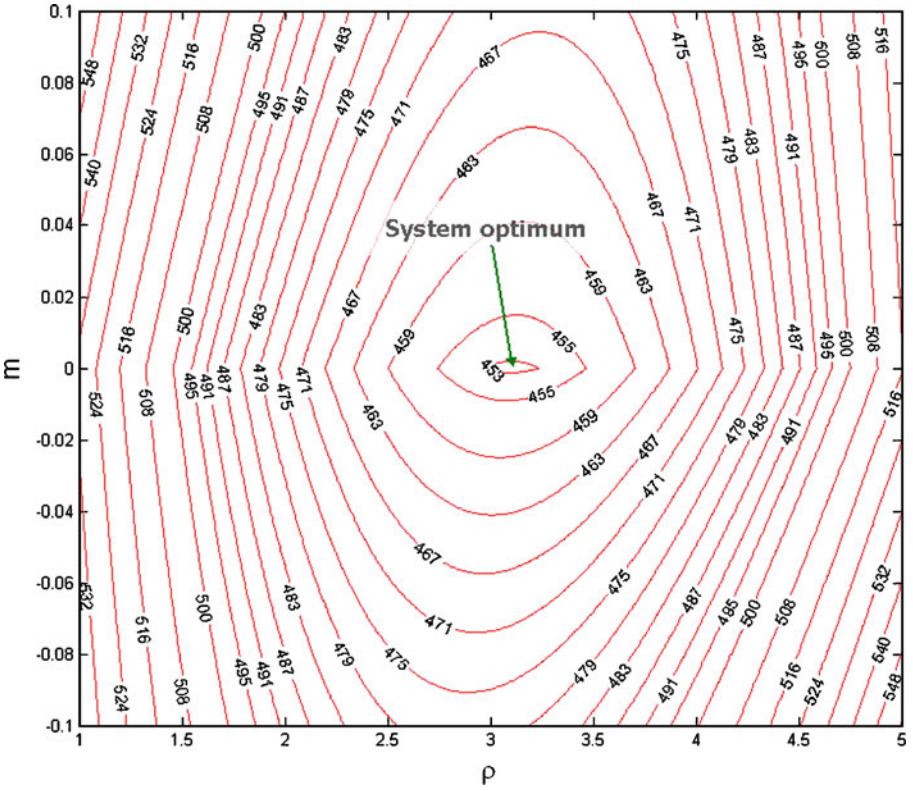

Fig. 7 System cost with respect to ρ and m

5.2 Minimizing total system cost in time unit

It is well known that in a static transportation network, the optimal flow pattern derived from minimizing the total system cost measured in time could be different from that derived by minimizing the total system cost measured in money (Yang and Huang 2002). We note that the same phenomenon can be found here. Under our assumption of heterogeneity, if we minimize the total system cost in time unit using the coarse toll, the optimal solution will be identical with what we get from the homogeneous case. We provide the optimal toll for minimizing the total generalized travel time without proof here

$$\rho_t^s = \frac{\eta_1\eta_2 N}{2s(\eta_1+\eta_2)}\alpha\left(\frac{\eta_2(\eta_2+1+2\eta_1)}{2(1+\eta_2)(\eta_1+\eta_2)}N\right) \tag{71}$$

We can see ρ_t^s is generally different from the optimal toll level ρ_m^s corresponding to minimal total cost measured in monetary unit. Another observation can be made by comparing the results from homogeneous and heterogeneous cases. To be comparable, we assume the parameter values attached to queuing and schedule delays in the homogeneous case are the expectations of the VOT distributions in the heterogeneous case, represented by $\overline{\alpha}, \overline{\beta}, \overline{\gamma}$. Then from Eq. (70) the optimal toll for homogeneous case becomes

$$\rho^s = \frac{\eta_1\eta_2}{2(\eta_1+\eta_2)}\left(\frac{N}{s}\right)\overline{\alpha} \tag{72}$$

Noting that $\int_0^V \alpha(v)dv \le V\alpha(V)$ ("=" obtained when $\alpha(V)$ is a constant), from Eq. (67) we have

$$V\alpha(V) \ge \frac{\eta_2(1+2\eta_1+\eta_2)}{2(\eta_1+\eta_2)(1+\eta_2)}\int_0^N \alpha(v)dv \tag{73}$$

Substituing Eqs. (62) and (72) into (73), we have

$$\frac{\rho_m^s}{\rho^s} \ge \frac{\int_0^N \alpha(v)dv}{N\overline{\alpha}} = 1 \tag{74}$$

Inequality (74) implies that for any kind of VOT distribution $\alpha(V)$, simply assuming homogeneity over the commuters will lead to an under-estimation of the optimal toll level. From Eqs. (62)–(64) we also can obtain that the under-estimation of the toll level will correspondingly lead to a longer charging time interval.

We know that the total system cost under NTE is

$$\overline{TC} = \frac{\eta_1\eta_2 KN}{(\eta_1+\eta_2)s} \tag{75}$$

By comparing Eqs. (65) and (75), we find the coarse toll scheme reduces the total system cost. But since everyone has a different VOT, naturally we will ask the question: does the coarse toll scheme reduce everyone's travel cost in the monetary unit? We will answer this question in Section 5.3.

5.3 The pareto-improving property of the optimal coarse toll

The total system cost excluding toll revenue can be reduced by the coarse toll, but the toll scheme may still increase the travel costs of some individuals. In this section we examine the performance the coarse toll from the perspective of individual commuters.

For the homogeneous case, since all the commuters are identical, to see if the coarse toll is pareto-improving, we only need to examine whether the total delay cost under the optimal coarse toll, TC_{delay}, is reduced. It can be easily proven that

$$TC_{delay} = \left(\frac{\beta\gamma}{\beta+\gamma}\frac{N^2}{s}\right)\left(1-\frac{(\gamma-\alpha)\beta}{2(\beta+\gamma)(\gamma+\alpha)}\right)\overline{TC} \tag{76}$$

Thus for the homogeneous case, the optimal coarse toll is pareto-improving.

For heterogeneous case, before tolling the *vth* commuter's travel cost is

$$C(v) = \frac{\eta_1\eta_2}{\eta_1+\eta_2}\frac{N}{s}\alpha(v) \tag{77}$$

After the optimal coarse toll is imposed, the commuters traveling outside $[t^+, t^-]$ experience the same reduction of generalized travel time. Thus we only need to observe the cost of the first commuter who experienced only schedule early delay y_2/s. By solving constraints (17), (18), (19), (20) and (22) with m=0 (See Appendix C for details), we have the generalized travel time for the commuters traveling outside $[t^+, t^-]$ be

$$T_{out} = \eta_1\frac{y_2}{s} = \frac{\eta_1\eta_2}{\eta_1+\eta_2}\frac{N}{s} - \frac{\eta_1^2(\eta_2-1)}{(\eta_1+\eta_2)(1+2\eta_1+\eta_2)}\frac{V}{s} \tag{78}$$

Thus the change of travel cost for the *vth* commuter who travels outside the tolling window is

$$\Delta C_{out} = -\frac{\eta_1^2(\eta_2-1)}{(\eta_1+\eta_2)(1+2\eta_1+\eta_2)}\frac{V\alpha(v)}{s} < 0 \tag{79}$$

Recall that at equilibrium, the *Vth* commuter will incur the same travel cost whether he chooses to travel outside or inside the tolling window $[t^+, t^-]$, and since a uniform toll will not influence the departure time choices for the commuters traveling inside the tolling window, we obtain the generalized travel time for the commuters traveling inside the tolling window by examining the *Vth* commuter

$$T_{in} = \frac{\eta_1\eta_2}{\eta_1+\eta_2}\frac{N}{s} - \frac{\eta_1^2(\eta_2-1)}{(\eta_1+\eta_2)(1+2\eta_1+\eta_2)}\frac{V}{s} - \frac{\rho}{\alpha(V)} \tag{80}$$

Thus for the *vth* commuter who travels inside the tolling window, the change of travel cost can be calculated as

$$\begin{aligned}\Delta C_{in} &= \left(\frac{\eta_1\eta_2}{\eta_1+\eta_2}\frac{N}{s} - \frac{\eta_1^2(\eta_2-1)}{(\eta_1+\eta_2)(1+2\eta_1+\eta_2)}\frac{V}{s} - \frac{\rho}{\alpha(V)}\right)\alpha(v) + \rho \\ &= -\frac{\eta_1^2(\eta_2-1)}{(\eta_1+\eta_2)(1+2\eta_1+\eta_2)}\frac{V\alpha(v)}{s} - \rho\left(\frac{\alpha(v)}{\alpha(V)} - 1\right) < 0\end{aligned} \tag{81}$$

To sum up Eqs. (79) and (81), the changes of travel cost for nonidentical commuters across the population are given as a piece-wise linear function:

$$\Delta C = \begin{cases} -\dfrac{\eta_1^2(\eta_2 - 1)}{(\eta_1 + \eta_2)(1 + 2\eta_1 + \eta_2)} \dfrac{V}{s}\alpha(v), & v \le V \\ -\dfrac{\eta_1\eta_2}{\eta_1 + \eta_2} \dfrac{V}{s}\alpha(v) + \rho, & V < v \le N \end{cases} \tag{82}$$

We can see from Eqs. (79) and (81) that ΔC is always negative for the whole population, which implies that the optimal coarse toll is still pareto-improving in the heterogeneous case. We also observe form Eq. (82) that the coefficient of $\alpha(v)$ in the second function is greater than first one, which indicates that for the commuters traveling inside the tolling period $[t^+, t^-]$, their benefits from the coarse toll increase more quickly than those traveling outside as their VOTs become higher. The latter aspect is shown in Fig. 8 for the special case that VOT follows a uniform distribution.

5.4 Efficiency gain from the coarse toll: a comparison

We propose a numerical example to compare the optimal solutions under different problem settings. We assume the demand N=100 and the capacity of the bottleneck s=50. The value attached to queuing delay $\alpha(v)$=0.128v, and η_1=0.609, η_2=2.377. Thus α is uniformly distributed within [0,12.8] and we can calculate the means $\overline{\alpha} = 6.4$, $\overline{\beta} = 3.9$ and $\overline{\gamma} = 15.21$, which are equal to the values used in the example for the homogeneous case. We list the optimal solutions when total system cost is measured in time and monetary units in Table 1. For comparison, the results for the homogeneous case are also listed.

From Table 1, if we measure the system cost in money, we find the heterogeneous case has a shorter tolling period and higher toll charge which is consistent with the conclusion derived from inequality (74). Moreover, for this special example of uniform VOT distribution, the total percentage saving of the heterogeneous case is around 40%, greater than that calculated under the homogeneous case, which implies that the benefit of coarse toll will be underestimated if heterogeneity is not considered.

Figure 9 shows the difference between the two optimal profiles. Without loss of generality, the work starting time is set to be the origin point. We observe that for

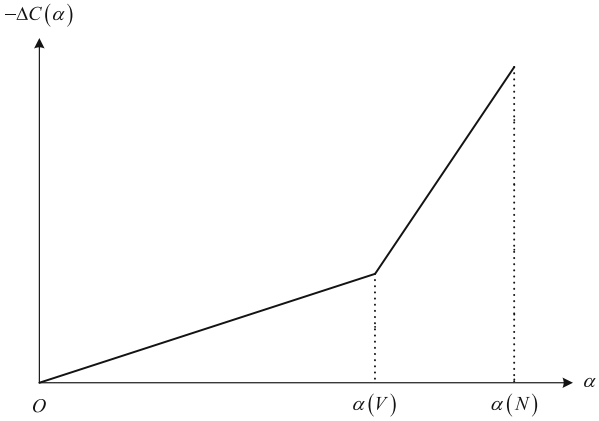

Fig. 8 Utility (=−cost) change with VOT across the population

Table 1 Optimal solutions

	Homogeneous	Heterogeneous (in time)	Heterogeneous (in money)
Total saving	27.08%	27.08%	40.06%
Commuters travel inside	45.84%	45.84%	39.91%
Length of toll interval	1.40	1.40	1.34
Optimal toll level	3.10	3.36	4.14
Commute starting time	−1.526	−1.526	−1.518
Toll revenue	142.3	154.1	165.2

both cases the departure time of the first person will move to the right compared with NTE, which implies that both coarse-toll schemes are cost-efficient. Less commuters are charged in the peak period and the total percentage cost saving is greater for the heterogeneous case.

When minimizing the total cost in time unit, we get the same profile with the homogeneous case (See columns 1 and 2 in Table 1). In this special example, the distribution of the benefits of the coarse toll differentiates across the population when the system performance is measured by money instead of time. More benefits are given to the two tails of the population (the poor and the rich), while the middle class receives less benefits because the higher toll forces them to change their commute from inside to outside of the tolling period and the saving from not paying the toll cannot offset the increase in their delay costs.

6 Conclusions

This study investigates how a single-step coarse toll affects the departure profiles in the morning commute and the properties of the optimal coarse toll when commuters have diverse values of time. All possible departure patterns under various levels of toll rate and tolling durations are thoroughly examined, from which the optimal coarse toll scheme is obtained through minimizing the total system cost. We prove that under the

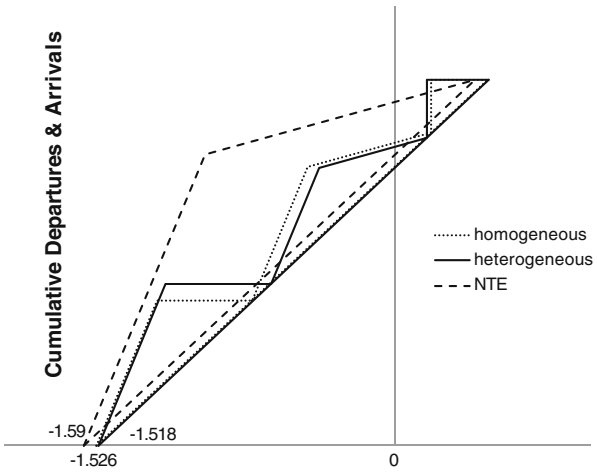

Fig. 9 Profiles for numerical examples

optimal coarse toll, there will be no waste of capacity at both the starting and ending times of the tolling period. Although this optimal coarse toll scheme cannot completely eliminate congestion, it has the dual advantages of simplicity and congestion relief: the flat toll is easy to implement and a suitably chosen toll level and tolling window can make every commuter better off than before such a toll is levied.

There could also be a range of optimal coarse toll schemes, if the system cost function is weighed differently between money and time. We cannot tell if the optimal toll charged is higher or lower when changing the weights, because it also depends on the form of VOT distribution. But we know that a higher toll charge will narrow the tolling window and for those who are forced to transfer from inside to outside of the tolling window, they will have an increase in their generalized cost.

We also show that the coarse toll problem under identical commuters considered by Arnott et al's work (1990b) is a special case of our heterogeneous model. By considering heterogeneity, we found that the toll level and toll revenue all increase, while the number of commuters being tolled is smaller. The tolling period, and the total system cost are all being reduced if the cost/benefit is measured in monetary terms. The consideration of heterogeneity also delays the starting time of the morning commute when money is used to measure costs and benefits.

The problem can be extended in several ways: 1) The assumption of proportionality in characterizing heterogeneity can be relaxed. We can assume that the α, β and γ parameters have separate distributions and are independent with each other. This, however, will complicate the problem and makes it harder, if not impossible, to obtain analytical results, and even to define what one means by optimum; 2) Multi-step coarse toll may be considered as a more general case of the one-step coarse toll. In fact when each step becomes infinitely small, the multi-step toll approaches to a fine toll. Clearly, more efficiency gain can be achieved at the price of identifying many more possible profiles, thanks to the increase of degrees of freedom in the optimization; and 3) Instead of a single bottleneck, we may consider a corridor with multiple bottlenecks and multiple OD pairs, rather than a one-to-one network with a single bottleneck. The problem becomes to determine not only how much and when to charge the toll, but also where to charge the toll.

Acknowledgements We wish to thank Prof. Richard Arnott for his valuable comments to an earlier version of this paper, and the Sustainable Transportation Center at University of California Davis for its financial support through a faculty research grant to the senior author, H. M. Zhang. The authors assume full responsibility for the contents of the paper.

Appendices

Appendix A. Simplifying the problem corresponding to profile 1

From Eqs. (17) and (20) we have

$$y_1 = \frac{s}{\eta_1}\left(\frac{\rho}{\alpha(V)} - m\right) \tag{83}$$

From Eqs. (18) and (83) we have

$$y_2 = \frac{s}{1-\eta_1}\left(x_2 - \frac{\rho}{\alpha(V)}\right) \tag{84}$$

Substituting (84) into the objective function (16) we have

$$\min_{V,x_2,\rho} TC = \frac{\eta_1 K}{1-\eta_1}x_2 - \left(\frac{K}{1-\eta_1} - A_1\right)\frac{\rho}{\alpha(V)} \tag{85}$$

From Eqs. (19) and (84) we have

$$x_2 = \frac{1}{\frac{s}{1-\eta_1} - \frac{s}{1+\eta_2}}\left(\frac{s\rho}{\alpha(V)(1-\eta_1)} + \left(N - \frac{2s\rho}{\alpha(V)(1+\eta_2)}\right) - \frac{s}{1+\eta_2}x_3\right) \tag{86}$$

From Eqs. (23) and (83) we also obtain that

$$m = \frac{\rho}{\alpha(V)}\frac{1+2\eta_1+\eta_2}{1+\eta_2} - \frac{\eta_1 V}{s} \tag{87}$$

Combining (85), (86) and (87) we can simplify the problem as follows

$$\begin{aligned}\min_{V,x_3,\rho} TC = {} & \frac{(1+\eta_2)\eta_1 K}{\eta_1+\eta_2}\left(\frac{\rho}{\alpha(V)(1-\eta_1)} + \left(\frac{N}{s} - \frac{2\rho}{\alpha(V)(1+\eta_2)}\right) - \frac{1}{1+\eta_2}x_3\right) \\ & - \left(\frac{K}{1-\eta_1} - A_1\right)\frac{\rho}{\alpha(V)}\end{aligned} \tag{88}$$

s.t.

$$x_3 \le \frac{N}{s} - \frac{2\rho}{\alpha(V)(1+\eta_2)} \tag{89}$$

$$m = \frac{\rho}{\alpha(V)}\frac{1+2\eta_1+\eta_2}{1+\eta_2} - \frac{\eta_1 V}{s} \tag{90}$$

$$m \le 0 \tag{91}$$

Appendix B. Simplify the problem corresponding to profile 2

From Eqs. (31) and (34) we have

$$y_1 = \frac{s\rho}{\eta_1\alpha(V)} - ms \tag{92}$$

Since y_1 has to be nonnegative, we have

$$m \le \frac{\rho}{\eta_1\alpha(V)} \tag{93}$$

From Eqs. (32) and (92) we have

$$y_2 = \frac{s}{1-\eta_1}\left(x_2 - \frac{\rho}{\alpha(V)}\right) - ms \tag{94}$$

Substituting (94) into (30) we have

$$\min_{V,x_2,y_2} TC = \frac{\eta_1 K}{1-\eta_1}x_2 - \left(\frac{K}{1-\eta_1} - A_1\right)\frac{\rho}{\alpha(V)} \tag{95}$$

We find that the objective function follows exactly the same form of (85).
From Eqs. (33) and (94) we have

$$x_2 = \frac{1}{\frac{s}{1-\eta_1} - \frac{s}{1+\eta_2}}\left(\frac{s\rho}{\alpha(V)(1-\eta_1)} + \left(N - \frac{2s\rho}{\alpha(V)(1+\eta_2)}\right) - \frac{s}{1+\eta_2}x_3 + ms\right) \tag{96}$$

From Eqs. (92) and (37) we also obtain that

$$m = \frac{\rho}{\eta_1\alpha(V)}\frac{1+2\eta_1+\eta_2}{1+\eta_2} - \frac{V}{s} \tag{97}$$

Combining (95)–(97), we can simplify the problem as follows

$$\begin{aligned}\min_{V,x_2,y_2} TC = {} & \frac{(1+\eta_2)\eta_1 K}{\eta_1+\eta_2}\left(\frac{\rho}{\alpha(V)(1-\eta_1)} + \left(\frac{N}{s} - \frac{2\rho}{\alpha(V)(1+\eta_2)}\right) - \frac{1}{1+\eta_2}x_3 + m\right) \\ & - \left(\frac{K}{1-\eta_1} - A_1\right)\frac{\rho}{\alpha(V)}\end{aligned} \tag{98}$$

s.t.

$$x_3 \le \frac{N - \frac{2s\rho}{\alpha(V)(1+\eta_2)}}{s} + m \tag{99}$$

$$m = \frac{\rho}{\eta_1\alpha(V)}\frac{1+2\eta_1+\eta_2}{1+\eta_2} - \frac{V}{s} \tag{100}$$

$$0 \le m \le \frac{\rho}{\eta_1\alpha(V)} \tag{101}$$

Appendix C. Calculation of travel time reduction

To solve y_2 we only need to solve the following group of equations:

$$\begin{cases} \frac{y_1}{x_1} = \frac{s}{1-\eta_1} \\ \frac{y_1}{s} - x_1 = \frac{\rho}{\alpha(V)} \\ \frac{N - \frac{2s\rho}{\alpha(V)(1+\eta_2)}}{x_3} = s \\ \frac{y_2 - y_1}{x_2 - \frac{y_1}{s}} = \frac{s}{1-\eta_1} \\ \frac{N - \frac{2s\rho}{\alpha(V)(1+\eta_2)} - y_2}{x_3 - x_2} = \frac{s}{1+\eta_2} \end{cases} \tag{102}$$

There are five unknown independent variables and five independent equations. By the first two equations we have

$$y_1 = \frac{s\rho}{\eta_1 \alpha(V)} \tag{103}$$

Substituting the third equation and (103) into the last two equations we have

$$y_2 = \frac{\eta_2 N}{\eta_1 + \eta_2} - \frac{s\rho(\eta_2 - 1)}{\alpha(V)(1+\eta_2)(\eta_1+\eta_2)} \tag{104}$$

We have already known the toll level expressed by Eq. (62), thus we have

$$y_2 = \frac{\eta_2 N}{\eta_1 + \eta_2} - \frac{(\eta_2 - 1)\eta_1 V}{(\eta_1+\eta_2)(1+2\eta_1+\eta_2)} \tag{105}$$

References

Arnott R, Palma AD, Lindsey R (1988) Schedule delay and departure time decisions with heterogeneous commuters. Transp Res Rec 1197:56–67

Arnott R, Palma AD, Lindsey R (1990a) Departure time and route choice for the morning commute. Transp Res B Methodol 24:209–228

Arnott R, Palma AD, Lindsey R (1990b) Economics of a bottleneck. J Urban Econ 27:111–130

Arnott R, Palma AD, Lindsey R (1993) A structural model of peak-period congestion: a traffic bottleneck with elastic demand. Am Econ Rev 83:161–179

Arnott R, Palma AD, Lindsey R (1994) A welfare congestion tolls with heterogeneous commuters. J Transp Econ Policy 28:139–161

Bernstein D, El sanhouri I (1994) A note on departure time and route choice for the morning commute. Transp Res B 28:391–394

Braid RM (1989) Uniform versus peak-load pricing of a bottleneck with elastic demand. J Urban Econ 26:320–327

Braid RM (1996) Peak-load pricing of a transportation route with an unpriced substitute. J Urban Econ 40:179–197

Daganzo CF (1985) Uniqueness of a time-dependent equilibrium distribution of arrivals at a single bottleneck. Transp Sci 19:29–37

Danielis R, Marcucci E (2002) Bottleneck road congestion pricing with a competing railroad service. Transp Res E 38:379–388

Huang HJ (2000) Fares and tolls in a competitive system with transit and highway: the case with two groups of commuters. Transp Res E 36:267–284

Knockaert J, Verhoef E, Rouwendal J (2009) Bottleneck congestion: differentiating the coarse charge. Vrije Universiteit Amsterdam, Christos Evangelinos, Bernhard Wieland, Technische Universität Dresden, Working paper

Laih CH (1994) Queuing at a bottleneck with single-Step and multistep tolls. Transp Res A 28:197–208

Levinson D (2005) Micro-foundations of congestion and pricing: a game theory perspective. Transp Res A 39:691–704

Li ZC, Lam HK, Wong SC, Huang HJ, Zhu DL (2008) Reliability evaluation for stochastic and time-dependent networks with multiple parking facilities. Netw Spat Econ 8:355–381

Lin DY, Unnikrishnan A, Waller ST (2010) A dual variable approximation based heuristic for dynamic congestion pricing. Netw Spat Econ. doi:10.1007/s11067-009-9124-9

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

Newell GF (1987) The morning commute for nonidentical travelers. Transp Sci 21:74–88

Nie XJ, Zhang HM (2005) A comparative study of some macroscopic link models used in dynamic traffic assignment. Netw Spat Econ 5(1):89–115

Ramadurai G, Ukkusuri S, Zhao J, Pang JS (2010) Linear complementary formulation for the multi-user class single bottleneck problem. Transp Res B 44:193–214

Tabuchi T (1993) Bottleneck congestion and modal split. J Urban Econ 34:414–431

Van der Zijpp N, Koolstra K (2002) Multiclass continuous-time equilibrium model for departure time choice on single-bottleneck network. Transp Netw Model 2002(1783):134–141

Vickrey WS (1969) Congestion theory and transport investment. Am Econ Rev 59:251–260

Wie BW, Tobin RL (1998) Dynamic congestion pricing models for general traffic networks. Transp Res B 32:313–327

Yang H, Huang HJ (2002) The multi-class, multi-criteria traffic network equilibrium and systems optimum problem. Transp Res B 38:1–15