Efficiency of a highway use reservation system for morning commute

Wei Liu, Hai Yang, Yafeng Yin

Abstract
This paper examines the design and efficiency of a highway use reservation system where commuters need reservations to access a highway facility at specific times. We show that, by accommodating reservation requests to the level that the highway capacity allows, traffic congestion can be relieved. Generally, a more differentiated design of the reservation system yields a higher reduction of travel cost and thus achieves a higher efficiency. The efficiency bound of the system is established. We also show that braking or tactical waiting behaviors of drivers would cause a loss of efficiency, which thus need be proactively accommodated. Given that user heterogeneity cause further loss of efficiency, we explore how two specific types of user heterogeneity affect the system efficiency. Auction-based reservation is then proposed to mitigate the efficiency loss.

1. Introduction
Reservation or booking is not a new concept, which indeed has been widely used in the transportation industry such as railway, airline, liner shipping and parking (Yang et al., 2013; Liu et al., 2014a,b), and some other industries such as hotel and entertainment. This reservation or booking concept can be applied to congestible infrastructures such as highway. One can imagine that, under a reservation system for highway use, travelers need to make a reservation (in advance) in order to pass a highway with a recurring bottleneck. Furthermore, a highway use reservation request may need travelers to specify a time before or after which they will pass the bottleneck. For given time intervals, by accommodating reservation requests to the level that the highway capacity (of the bottleneck) allows, traffic congestion is likely to be relieved.

Based on a stated-preference survey, Akahane and Kuwahara (1996) quantitatively evaluated a trip reservation system that adjusts departure times of travelers on motorways in a similar way as train seat reservations. Wong (1997) then conceptually described the booking system for highway use to control traffic and improve road performance. Koolstra (1999) studied the potential benefits of slot reservation for highway by analyzing the difference between the user equilibrium and the system optimal departure times. de Feijter et al. (2004) proposed trip booking as a method for improving travel time reliability and increasing the effective usage of road capacity. A highway space inventory control system was presented in Teodorovic et al. (2005) and Edara and Teodorovic ́ (2008). More recently, Zhao et al. (2010) introduced a downtown space
reservation system to facilitate the mitigation of traffic congestion in a cordon-based downtown area. Under such a system, people who want to drive into the cordon-based downtown area are required to make reservations in advance. While all these papers have discussed the advantages and issues related to trip reservation systems, this paper conducts a more comprehensive economic analysis of the highway use reservation system. Under such a system, travelers using a highway bottleneck are required to make reservations in advance, free access is no longer available. For a given time interval, the number of vehicles that the highway can accommodate is limited (depends on the highway capacity). Therefore, the system manager can reject reservation requests to ensure the reservations not beyond highway capacity thus queue can be avoided and traffic congestion can be relieved. It is then of our interest to study how the reservation system, which allocates highway space to potential users at different time intervals and accepts or rejects user reservation requests based on the highway capacity and the number of reservations made, can help to decrease commuters’ total travel cost and improve system performance. Efficiency of such a system will also be examined. Furthermore, some factors affecting performance of the system will be discussed, including the limited differentiation of reservations, braking or tactical waiting behaviors of commuters, and user heterogeneity in values of travel time and schedule penalties.

The remainder is organized as follows. Section 2 reviews the basic model formulation of the morning commute. In Section 3, the reservation system for highway use in the morning commute is introduced, the resulting traffic equilibrium and the system efficiency in reducing travel cost are then discussed. Section 4 discusses the efficiency loss due to braking or tactical waiting behaviors of commuters. In Section 5, the efficiency loss due to different types of user heterogeneity is discussed, and auction-based reservation is then introduced to mitigate the loss. Lastly, Section 6 concludes the paper.

2. Basic model

In this paper, we design and analyze the reservation system in a highly simplified setting of morning commute. The bottleneck model (Vickrey, 1969) is adopted to capture the congestion dynamics in the morning commute. Smith (1984) and Daganzo (1985) established the existence and uniqueness of the time-dependent equilibrium distribution of arrivals at a single bottleneck respectively. It has been extended to consider demand elasticity (Arnott et al., 1993; Yang and Huang, 1997), stochastic capacity and demand (Arnott et al., 1999; Lindsey, 2009; Xiao et al., 2013a, 2014), heterogeneous commuters (Newell, 1987; Arnott et al., 1994; Lindsey, 2004; Doan et al., 2011; Liu and Nie, 2011; van den Berg and Verhoef, 2011; van den Berg, 2014), integration of both morning and evening peak hour commutes (Zhang et al., 2005, 2008; Gonzales and Daganzo, 2013; Daganzo, 2013). Thanks to its tractability, the bottleneck model has been applied to generate policy insights for various congestion mitigation strategies e.g., congestion pricing (Arnott et al., 1990; Laih, 1994; Xiao et al., 2012; Yang and Meng, 1998), tradable mobility credits (Nie and Yin, 2013; Tian et al., 2013; Xiao et al., 2013b), variable speed limits (Liu et al., 2015), and hybrid Pareto-improving strategies (Daganzo and Garcia, 2000; Xiao and Zhang, 2013).

Consider a continuum of \( N \) commuters traveling from home to workplace via a highway every morning. Commuters have a common preferred arrival time \( \tau^* \) at the destination. Travel cost of a commuter, including travel time cost and schedule delay cost, by departing at time \( t \) is given by

\[
c(t) = \alpha \cdot T(t) + \beta \cdot \max\{0, \tau^* - t - T(t)\} + \gamma \cdot \max\{0, t + T(t) - \tau^*\},
\]  

where \( T(t) \) is the travel time at departure time \( t \); \( \alpha \) is the value of unit travel time, and \( \beta \) and \( \gamma \) are the schedule penalties for a unit time of early and late arrival respectively. Without loss of generality, we assume the travel time, \( T(t) \), only contains the queuing delay at the highway bottleneck whose service capacity is \( s \). Thus, \( T(t) = q(t)/s \), where \( q(t) \) is the queue the commuter departing at time \( t \) would face at the bottleneck. Consistent with empirical observations, we assume that \( \gamma > \alpha > \beta > 0 \). We also denote \( \delta = \beta\gamma/(\beta + \gamma) \).

Fig. 1. Commuting equilibrium without policy intervention.
Under the above standard setting from the literature, the morning commuting equilibrium can be depicted by Fig. 1, where the blue solid line and red dashed line represent the cumulative departure from home and arrival at the destination respectively. Note that in Fig. 1, we have

$$r_1 = \frac{\alpha}{\alpha - \beta} s; \quad r_2 = \frac{\alpha}{\alpha + \gamma} s,$$

which are the departure rates from home for commuters arriving at destination earlier or later than $t^*$ respectively. At equilibrium, total travel cost of all commuters is

$$TC_0 = \delta \frac{N^2}{s},$$

which increases with the total number of commuters.

3. The highway use reservation system

In this section, we will first introduce the reservation system, and then discuss the commuting equilibrium under such a system and examine the efficiency of the system in reducing total travel cost.

3.1. Reservation system

Since the capacity of the highway bottleneck is limited, the reservation system should control the number of reservations allowed for a given time interval such that traffic congestion and queueing is controlled. Under a reservation system for highway use, a commuter needs to make a reservation (in advance) in order to pass the highway bottleneck as shown in Fig. 2. Furthermore, when a commuter makes a reservation, he or she needs to specify a time before or after which he or she will pass the bottleneck (the system can provide a series of alternatives in advance, and commuters only need to pick up one of them as presented later). Automated vehicle identification sensors located at the highway bottleneck will check whether a passing vehicle has made a reservation (as depicted in Fig. 2). If not, the vehicle will suffer a significant amount of penalty fee. Practically, a certain portion of the capacity (lanes not for reservation) should be reserved for those who do not have a reservation. Those who have unexpected travel needs may use these lanes. However, there might be significant queuing delays at the unreserved lane(s). Also, travelers without reservations might choose public transportation, which is similar to that for rationed travelers under road space rationing policy (Liu et al., 2014c). Further study might consider how the reservation system can work in a multi-modal transportation network, particularly when the public transit operator is responsive to the highway capacity and service as considered in Zhang et al. (2014).

Generally, one can imagine that, during the duration with relatively high potential demand, commuters should make reservations early enough to secure reservations (probably before they leave home). In this paper, we consider the commuting equilibrium of the morning peak, and it is assumed that all commuters will make reservations before they pass the highway bottleneck to avoid the penalty.

Under a reservation system, the system manager can set a series of time points, i.e., $\{t_1, t_2, \ldots, t_k, t', t_{k+1}, \ldots, t_{n-1}, t_n\}$ represented in a chronologically increasing order, which divides the duration of $[t_1, t_n]$ into $n$ intervals, with the $i$th denoted by $\Delta t_i$. We then have

$$\Delta t_i = \begin{cases} t_{i+1} - t_i & 1 \leq i \leq k - 1 \\ t' - t_i & i = k \\ t_i - t' & i = k + 1 \\ t_{i-1} - t_i & k + 1 \leq i \leq n \end{cases}$$

For all commuters, they can make reservations, and choose to pass the highway bottleneck before time $t$ where $t = t_1, t_2, \ldots, t_k, t'$, or choose to pass the highway bottleneck after time $t$ where $t = t', t_{k+1}, \ldots, t_n$. Generally, commuters would prefer a reservation with more flexibility, i.e., for early arrival, a larger $t_i$; and for late arrival, a smaller $t_i$. Thus, if the reservation service is operated in the way of first-come-first-served, travelers will start making reservations with $t_i$ closer to $t'$, and arrive at their destination closer to $t'$. To avoid queuing, reservation requests accepted should be within the highway capacity. This means, the travel requests accepted to pass through the highway bottleneck before $t = t'$ should be within $M_k = s \cdot \Delta t_k$, and those before $t = t_k$ should be $M_{k-1} = s \cdot \Delta t_{k-1}$ and so on; while the travel requests accepted to pass through the highway bottleneck after $t = t'$ should be within $M_{k+1} = s \cdot \Delta t_{k+1}$, and those after $t = t_{k+1}$ should be $M_{k+2} = s \cdot \Delta t_{k+2}$ and so on.

Taking a reservation request for passing the highway bottleneck before $t = t_i$ where $i \leq k$ as an example, Fig. 3 depicts a simple first-come-first-served allocation procedure that can be used by the highway infrastructure manager to accommodate reservation requests.

In practice, there could be a considerable amount of year-long commuters. For these commuters, it might be more efficient and convenient to let them have fixed reservations on a weekly or monthly basis. Thus, they do not have to make online reservations every day. However, as they may have specific needs in a particular day, the system might allow them to cancel
their usual reservations in advance and re-book a new one more suitable for them. Also, they may use lanes which are not for reservation. Besides, there are quite many practical issues need to be addressed before the implementation of the reservation system, which, however, is beyond the scope of this paper. For later analysis, without loss of generality, we consider the whole morning peak is divided into $n$ time intervals equally, i.e., $\Delta t_i$ is constant, where $i = 1, 2, \ldots, n$. Note that there are $k$ early arrival intervals and $n/k$ late arrival intervals. For a given $n$, $k$ should be appropriately chosen to improve efficiency, which will be discussed in next section.

### 3.2. Efficiency of the reservation system

Now let’s examine the design of reservation system (scheme) and the corresponding commuting equilibrium in detail. For commuters with a reservation that allows them to pass the highway bottleneck after $t' \text{ or } t_i$ ($i \geq k + 1$), we assume all of them will depart just at $t'$ or $t_i$ and join the queue, and their arrival order will be decided randomly. This mass departure assumption is firstly used in Arnott et al. (1990), and followed by many (e.g., Xiao et al., 2011) for the step toll problem. Note that, in the step toll problem, mass departure arises since travelers want to reduce schedule delays (so they join a mass, and everyone has the same probability to be the first one to arrive) while they can avoid a higher toll (so they depart after the toll drops). For travelers in the mass departure, travel cost is an expectation, which equals the average cost of all travelers in the mass, from a long term perspective. In the equilibrium under reservation system, mass departure arises as travelers also want to reduce schedule delay. However, in this case, commuters are forced to travel after certain time points as they do not have reservations to travel at an earlier time. There are alternative assumptions for mass departure, e.g., travelers can wait on a separate lane without impeding others who want to pass the bottleneck (Laih, 1994, 2004), which might be considered in future research; and braking or tactical waiting in Lindsey et al. (2012) and Xiao et al. (2012), which will be discussed in Section 4.

Under the above setting, the departure time equilibrium can be depicted by Fig. 4. Note that the reservation system might be used only for the congested peak hour (the travel duration shown in Fig. 4) while the highway is for the whole day. However, if it is also used for uncongested non-peak hours, commuters are always able to make a reservation as the highway is not congested and there is still available capacity. As long as we assume $\gamma > 2$, which is consistent with empirical studies, travelers in a specific mass have no incentive to depart at the time when the last commuter in this mass arrives even if he or she can enjoy zero queue. This is because, the expectation cost of traveling in the mass is lower than that by departing at the
time when last commuter in the mass arrives. For the same reason, the commuters in the last mass (nth group) will not choose to travel after the end of peak as shown in Fig. 4. For more details, one can refer to Lindsey et al. (2012).

Total travel cost under the commuting equilibrium depicted in Fig. 4 can be expressed as follows:

$$TC_n = \left( \alpha \cdot \frac{1}{2} \cdot \frac{N}{ns} - \frac{N}{nr} \right) + \beta \cdot \frac{kN}{2} \cdot \frac{N}{n} + \left( \frac{\alpha + \gamma}{2} \cdot \frac{n - kN}{2s} + \gamma \cdot \frac{n - kN}{2s} \right) \cdot \frac{n - k}{n}.$$ (5)

In Eq. (5), the first term is travel cost of early arrival commuters, while the second term is the travel cost of late arrival commuters. Note that $k/n$ should be optimally determined thus Eq. (5) is minimized for a given $n$. Based on traditional analysis of the bottleneck model on road pricing, it is expected that $k/n \rightarrow \gamma/(\beta + \gamma)$ for relatively large $n$. Taking the first-order partial derivative of Eq. (5) with respect to $k/n$, and let it be zero, we then obtain the optimal $(k/n)^* = \gamma/(\beta + \gamma) + (\alpha - \beta)/(2n(\beta + \gamma))$. It is straightforward to see that $(k/n)^* > \gamma/(\beta + \gamma)$. This is because, the mass departure of late arrival commuters is less efficient as the queuing delay is larger (one may refer to Fig. 4, for the same amount of commuters, the queuing triangle for late arrival is larger than that for early arrival). Thus we should have more early arrivals. However, as we set $n \rightarrow \infty$, we have $k/n \rightarrow \gamma/(\beta + \gamma)$, and then $TC_n \rightarrow 0.5\delta n^2/s$. Compared to the original departure/arrival equilibrium without policy intervention (which is shown in Fig. 1), the reservation system would yield 50% reduction in total travel cost. The reservation system achieves the same efficiency as the well-known first-best time-varying pricing scheme. Note that, in this case with $n \rightarrow \infty$, the queue at the highway bottleneck is eliminated completely and the queuing delay cost is thus reduced to zero. The schedule delay cost is equal to half of total travel cost in the original departure/arrival equilibrium, which is $TC_o = \delta N^2/s$.

As it is practically infeasible to set a $n \rightarrow \infty$, we now derive the efficiency loss due to a finite number of $n$. The loss is formally defined as follows:

$$\theta_n = \frac{\min \{TC_n\} - TC_\infty}{TC_o - TC_\infty},$$ (6)

where $\min \{TC_n\}$ is the minimum total travel cost achievable by a $n$-step reservation system, which is determined by Eq. (5) by setting $k/n = \gamma/(\beta + \gamma) + (\alpha - \beta)/(2n(\beta + \gamma))$. Therefore, $\min \{TC_n\} - TC_\infty$ is the loss of travel cost reduction of a $n$-step reservation system, and $TC_o - TC_\infty$ is the travel cost reduction under the reservation system with an infinite number of steps. We then have the following proposition.

**Proposition 3.1.** The efficiency loss defined in Eq. (6) satisfies the following:

$$\theta_n < \frac{\alpha + \gamma}{\beta \gamma} \frac{1}{n}.$$ (7)

**Proof.** For given $n$, if we set $k = \frac{\gamma}{\beta + \gamma} n$, then

$$\frac{TC_n - TC_\infty}{TC_o - TC_\infty} = \frac{\alpha + \gamma}{\beta \gamma} \frac{1}{n}.$$ (8)

Since $\min \{TC_n\} < TC_n$, we then have Eq. (7) holds. □

Proposition 3.1 implies that the efficiency loss due to a limited number of steps of the reservation system can be bounded. When one can set the number of steps $n \rightarrow \infty$, we have $k/n = \gamma/(\beta + \gamma) + (\alpha - \beta)/(2n(\beta + \gamma))$, therefore $\min \{TC_n\} \rightarrow TC_n$, where $TC_n$ is determined from Eq. (5) by setting $k/n = \gamma/(\beta + \gamma)$. This implies that as $n$ becomes larger,
the upper bound in Eq. (7) approaches the exact efficiency loss defined in Eq. (6). Further considering \( \alpha < \gamma \), Proposition 3.1 indicates \( \theta_e < 2/n \). Accordingly, for a reservation system of 10 steps, i.e., \( n = 10 \), the efficiency loss would be less than 20%. If we consider \( \gamma/\alpha \approx 2 \), then the efficiency loss would be within 15%.

For given number of step, \( n \), the optimal \( k/n = \gamma/(\beta + \gamma) + (\alpha - \beta)/(2\gamma(\beta + \gamma)) \) relies on commuters’ value of time and schedule preference. However, the performance of the reservation system with a finite number of steps (for the homogeneous case) is robust to the variation of the value of time of commuters, i.e., \( (\alpha + \gamma)/\gamma \) will always be less than 2 and \( \theta_e < 2/n \).

4. Braking or tactical waiting behaviors of travelers

In the previous section, we assume that late arrival commuters would depart immediately once they are allowed to pass the highway bottleneck, which yields a mass departure. The mass departure assumption ignores that commuters have an incentive to depart earlier than allowed and stand by to wait for their turns at the highway bottleneck to enjoy a lower travel cost (by arriving earlier, travelers can enjoy a lower schedule delay). Such behavior is called as braking or tactical waiting in the literature of the bottleneck models with step tolls, e.g., Lindsey et al. (2012) and Xiao et al. (2012). Note that in this study we do not consider travelers can wait on a separate lane without impeding others who want to pass the bottleneck at the current time interval (Laih, 1994, 2004). For both the step toll and reservation systems, braking or tactical waiting arises as incentive to depart earlier than allowed and stand by to wait for their turns at the highway bottleneck, which yields a mass departure. The mass departure assumption ignores that commuters have an incentive to depart earlier than allowed and stand by to wait for their turns at the highway bottleneck to enjoy a lower travel cost (by arriving earlier, travelers can enjoy a lower schedule delay). Such behavior is called as braking or tactical waiting in the literature of the bottleneck models with step tolls, e.g., Lindsey et al. (2012) and Xiao et al. (2012). Note that in this study we do not consider travelers can wait on a separate lane without impeding others who want to pass the bottleneck at the current time interval (Laih, 1994, 2004). For both the step toll and reservation systems, braking or tactical waiting arises as incentive to depart earlier than allowed and stand by to wait for their turns at the highway bottleneck, which yields a mass departure. The mass departure assumption ignores that commuters have an incentive to depart earlier than allowed and stand by to wait for their turns at the highway bottleneck to enjoy a lower travel cost (by arriving earlier, travelers can enjoy a lower schedule delay). Such behavior is called as braking or tactical waiting in the literature of the bottleneck models with step tolls, e.g., Lindsey et al. (2012) and Xiao et al. (2012). Note that in this study we do not consider travelers can wait on a separate lane without impeding others who want to pass the bottleneck at the current time interval (Laih, 1994, 2004). For both the step toll and reservation systems, braking or tactical waiting arises as incentive to depart earlier than allowed and stand by to wait for their turns at the highway bottleneck, which yields a mass departure. The mass departure assumption ignores that commuters have an incentive to depart earlier than allowed and stand by to wait for their turns at the highway bottleneck to enjoy a lower travel cost (by arriving earlier, travelers can enjoy a lower schedule delay). Such behavior is called as braking or tactical waiting in the literature of the bottleneck models with step tolls, e.g., Lindsey et al. (2012) and Xiao et al. (2012). Note that in this study we do not consider travelers can wait on a separate lane without impeding others who want to pass the bottleneck at the current time interval (Laih, 1994, 2004).

Without loss of generality, we assume the commuters are divided into \( n \) groups equally. Furthermore, by foreseeing the braking or tactical waiting behaviors of commuters, the design of the reservation system (for the late arrival traffic) would be different from that discussed in Section 3. Indeed, \( \{t_1, t_2, \ldots, t_n, t_{n+1}, \ldots, t_{n-1}, t_n\} \) can then be determined as follows:

\[
t_i = \begin{cases} 
  t^* - (k - i + 1) \cdot \frac{N}{n} & i \leq k \\
  t^* - \left( \frac{N}{n} - \frac{N}{mr} \right) + (i - k) \cdot \frac{N}{mr} - \frac{N}{n} & i > k + 1
\end{cases}
\]  

(9)

The \( i \)-th group (i.e., early arrival) is required to pass the bottleneck before \( t_{i+1} \) when \( i < k \) or \( t^* \) when \( i = k \); while the \( i \)-th group (i.e., late arrival) is required to pass the bottleneck after \( t_i \). As can be seen, for early arrival, we have \( t_j - t_{i+1} = t^* - t_i = N/\bar{n} \) where \( 2 \leq j \leq k \). However, for late arrival, \( \Delta t = N/nr > N/\bar{n} \) where \( k + 1 \leq j \leq n \) and the rates \( r_1 = z\alpha/\alpha - \beta \) and \( r_2 = z\alpha/\alpha + \gamma \), indicating the number of reservation requests accepted is less than the capacity, and there is capacity waste due to the braking or tactical waiting. Note that, in Eq. (9), the last group of early arrival, i.e., \( k \)-th group, is allowed to pass the bottleneck before \( t^* \), and the first group of late arrival, i.e., \( (k + 1) \)-th group, is not allowed to pass the bottleneck before \( t_{k+1} \), but not \( t^* \) (which is different from the settings in Section 3). Under the current setting, the commuting equilibrium can be depicted by Fig. 5.

Total travel cost under the commuting equilibrium depicted in Fig. 5 is

\[
TC'_{\text{T}} = \left( \alpha \cdot \frac{1}{2} \left( \frac{N}{n} - \frac{N}{mr} \right) + \beta \cdot \frac{k}{2} \frac{N}{n} \right) \cdot k \frac{N}{n} + \left( \alpha \cdot \frac{1}{2} \left( \frac{N}{mr} - \frac{N}{n} \right) + \gamma \right) \cdot \left( \frac{N}{n} \frac{N}{mr} - \frac{N}{n} \right) \cdot (n - k) \frac{N}{n}
\]

(10)

In Eq. (10), the first term is the travel cost of early arrival commuters, while the second term is the travel cost of late arrival commuters. For given \( n \), to minimize Eq. (10), we should set \( k/n = (\alpha + \gamma)(\gamma - \beta)/n/(\gamma \beta + (\alpha + \gamma) \gamma) + 1/2n \). If we consider that the system manager can set \( n \rightarrow \infty \), we have \( k/n \rightarrow (\alpha + \gamma)(\gamma - \beta)/n/(\gamma \beta + (\alpha + \gamma) \gamma) > \gamma/(\beta + \gamma) \), which implies that when anticipating braking or tactical waiting of travelers, we should allow more early arrival traffic than the case in Section 3 to reduce
the inefficiency incurred by the behavior. To see the efficiency loss due to braking or tactical waiting, we examine
\[ \Delta TC_1 = TC_n^* - TC_n, \]
where \( TC_n \) and \( TC_n^* \) are defined by Eqs. (5) and (10) respectively, which is
\[ \Delta TC_1 = \frac{1}{2} \left( \frac{n-k}{n} \right) N^2 \cdot \left[ (\gamma - \alpha) \cdot \frac{\gamma}{2} \cdot ((n-k) \cdot \gamma - 2\beta) \right]. \] (11)

Note that the first terms in Eq. (5) and in Eq. (10) are identical, which are the travel cost of all early arrival commuters, of which the number is \( \frac{1}{2} N \). Therefore, Eq. (11) is also the travel cost difference of late arrival commuters, of which the number is \( \frac{n-k}{n} \cdot N \).

Note that for ease of comparison, we look at the cost difference \( \Delta TC_1 \) for the same values of \( k \) and \( n \). Suppose \( k = \lambda n \) where \( 0 < \lambda \leq 1 \), for a given \( n \), to minimize \( TC_1 \) given by Eq. (5), we should set \( \lambda = \gamma/(\beta + \gamma) + (\alpha - \beta)/(2n(\beta + \gamma)) \); to minimize \( TC^*_n \), given by Eq. (10), we should set \( \lambda = (\alpha + \gamma)(\gamma - \beta)/n)/(\alpha(\beta + (\alpha - \gamma)\gamma) + 1/2n \). However, no matter how we set \( \lambda \), \( \Delta TC_1 > 0 \), suggesting that efficiency loss due to commuters’ braking or tactical waiting is inevitable. Moreover, when \( n \to \infty \), \( \Delta TC_1 \to 0.5(1-\lambda)^2\gamma^2N^2/3\alpha \), which indicates that even under the ideal system with an infinite number of steps, commuters’ tactical waiting would yield an efficiency loss. In this case with \( n \to \infty \), when \( \lambda \to 1 \), i.e., \( k \to n \), or \( \gamma \to \infty \) (late arrival is note allowed), \( \Delta TC_1 \to 0 \). This is because the efficiency loss comes from late arrival commuters’ braking or tactical waiting behavior. When late arrival is not allowed, there would be no tactical waiting, as well as the efficiency loss.

Let \( \alpha = 9.91 \text{ (EUR$)} \), \( \beta = 4.66 \text{ (EUR$)} \) and \( \gamma = 14.48 \text{ (EUR$)} \) (from Tseng et al., 2005). Fig. 6 depicts how the efficiency loss defined in Eq. (11) vary with both \( k \) and \( n \). Note that, in Eq. (11), \( N^2/s \) is constant for all combinations of \( k \) and \( n \), which is thus omitted in the example. Fig. 6 shows the contours of \( \Delta TC_1 \) in the two-dimension domain \((n, k)\). As can be seen, the contours are almost straight lines, implying \( \Delta TC_1 \) is approximately the same for given \( \lambda = k/n \), particularly when \( n \) becomes larger. Besides, as \( \lambda = k/n \) increases, the efficiency loss term \( \Delta TC_1 \) decreases. This is because, increasing \( \lambda = k/n \) indicates less late arrival travelers.

5. User heterogeneity

In this section, we will discuss how two specific types of user heterogeneity would affect performance of the reservation system. For ease of analysis, we focus on the ideal case that the system manager can set \( n \to \infty \). When the value of time (VOT) varies among commuters but the schedule delay penalty parameters \( \beta \) and \( \gamma \) are identical, then user heterogeneity would not influence the performance of the reservation system. Such user heterogeneity only affects the departure order of commuters, e.g., in the original equilibrium when there is no policy intervention, those with lower values of time will choose to pass the bottleneck near the desired work start time. However, in the ideal case when queue is eliminated, whoever reserves to pass the bottleneck close to the work start time makes no difference to the performance of the reservation system. However, as reported in empirical studies (e.g., Tseng and Verhoef, 2008), not only commuters’ VOT, but also schedule delay penalty can vary. In this case, under the socially preferable commuting equilibrium (total travel cost is minimized), commuters with higher \( \beta \) and \( \gamma \) should be enabled to obtain reservations with more flexibility, i.e., their departure and arrival should be closer to \( t^* \). Unfortunately, without the help of other instruments, the first-come-first-served reservation system would fail to do so and thus lose efficiency.
5.1. Proportional heterogeneity

Assume commuters’ VOT, \( \alpha \), continuously increases from \( \underline{\alpha} \) to \( \overline{\alpha} \) and follows the following cumulative distribution function: \( F(x) = Pr(\alpha \leq x) \). The corresponding density function is denoted by \( f(x) \), which is positive for \( \alpha \in [\underline{\alpha}, \overline{\alpha}] \) and is zero otherwise. Let \( F(x) = 1 - F(x) \). For a specific commuter, his or her early and late arrival penalty is proportional to VOT, i.e., \( \beta = \rho_1 \alpha \) and \( \gamma = \rho_2 \alpha \), where \( \rho_1 < 1 < \rho_2 \) and \( \rho_1 \) and \( \rho_2 \) are identical for all commuters. This is similar to the heterogeneity considered in Vickrey (1973), Xiao et al. (2011) and others. Note that the two ratios measure the willingness of a commuter to reduce schedule delays by increasing travel time, and will determine the arrival order of commuters.

When no policy is introduced, the departure/arrival equilibrium will be identical to that in the homogeneous case, which is depicted in Fig. 1. For a commuter with VOT of \( \alpha \), travel cost would be

\[
C_{o}(\alpha) = \rho \alpha \frac{N}{S},
\]

where \( \rho \) is given by

\[
\rho = \frac{\rho_1 \rho_2}{\rho_1 + \rho_2}.
\]

Total travel cost at this no-policy equilibrium is

\[
TC_{o} = \int_{\underline{\alpha}}^{\alpha} \rho \alpha \frac{N}{S} f(\alpha) N d\alpha.
\]

Under the system optimum (SO), the queue should be eliminated and the schedule delay cost should be minimized. Therefore, commuters with a larger \( \alpha \) should arrive closer to the desired arrival time, \( t' \). In this case, for a commuter with VOT of \( \alpha \), travel cost is

\[
C_{w}(\alpha) = \rho \alpha \frac{N}{S} F(\alpha).
\]

Total travel cost at SO is then

\[
TC_{so} = \int_{\underline{\alpha}}^{\alpha} \rho \alpha \frac{N}{S} F(\alpha) f(\alpha) N d\alpha.
\]

Unfortunately, the reservation system proposed above cannot differentiate commuters with respect to their VOTs. We now consider two cases: (i) the average case, i.e., for each specific reservation, all commuters have the same probability of obtaining it; (ii) the worst case, i.e., commuters with higher VOT obtain the reservation with less flexibility and arrives farther from \( t' \). For case (i), from a long-term perspective, the expected total travel cost is given by

\[
TC_{av} = \int_{\underline{\alpha}}^{\alpha} \rho \alpha \frac{N}{S} F(\alpha) f(\alpha) N d\alpha.
\]

In case (ii), the travel cost of a specific commuter with VOT of \( \alpha \) is:

\[
C_{w}(\alpha) = \rho \alpha \frac{N}{S} F(\alpha).
\]

And consequently the total travel cost is

\[
TC_{w} = \int_{\underline{\alpha}}^{\alpha} \rho \alpha \frac{N}{S} F(\alpha) f(\alpha) N d\alpha.
\]

Now we are ready to present the following proposition.

**Proposition 5.1.** Total travel costs satisfy the following:

\[
TC_{so} < TC_{av} < TC_{w} < TC_{o}.
\]

**Proof.** See Appendix A. □

The above proposition indicates that the proposed reservation system can reduce total travel cost, although it cannot differentiate commuters with different VOTs. Its worst-case performance still yields a cost reduction as compared to that when no policy is introduced. Besides, we note that, in the average case, i.e., case (i), total travel cost is a half of that under the no policy equilibrium, i.e., \( TC_{av} = 0.5 TC_{o} \). This indicates the reservation system would yield a 50% saving in an average sense. Note that, the cost reduction of SO is larger than 50%, which is larger than that in the homogeneous case. This is because when travelers are heterogeneous, SO reduces schedule delays as well as queueing.
Proposition 5.2. \( \frac{TC_W}{TC_{so}} < \frac{2}{3} \)

Proof. See Appendix B. □

Although the first-come-first-served reservation system would lose efficiency due to user heterogeneity, the inefficiency is upper bounded. If \( \bar{x}/\bar{z} < 1 \), the efficiency loss approaches zero, and \( TC_w \to TC_{so} \). If \( \bar{x}/\bar{z} = 2 \), \( TC_w \leq 2TC_{so} \) holds. Since \( TC_{so} + TC_{no} = TC_o \), from Proposition 5.2, one can verify that \( TC_w \leq (\bar{x}/(\bar{x} + \bar{z}))TC_o \). This indicates even in the worst case, the reservation system can at least reduce a certain percentage of cost, i.e., \( \frac{\bar{x}}{\bar{x} + \bar{z}} \). If \( \bar{x}/\bar{z} = \bar{x} = 1 \), this percentage approaches 50%. If \( \bar{x}/\bar{z} = 2 \), the percentage is 33%.

To further look at the efficiency loss due to user heterogeneity, we consider \( x \) follows a uniform distribution, i.e., the density function \( f(x) = \frac{1}{\bar{x} - \bar{z}} \) for \( x \in [\bar{z}, \bar{x}] \), and \( f(x) = 0 \) otherwise. Then we have the following proposition regarding the efficiency of the reservation system.

Proposition 5.3. When \( x \sim U[\bar{z}, \bar{x}] \), we have

\[
\frac{1}{3} \leq \frac{TC_{so}}{TC_o} < \frac{1}{2} \frac{TC_{av}}{TC_o} = \frac{1}{2} \frac{TC_w}{TC_o} < \frac{2}{3} \frac{TC_{so}}{TC_{io}} < \frac{1}{2} \frac{TC_w}{TC_{io}} < 2. \tag{21}
\]

Proof. Note that \( TC_{av} = 0.5TC_o \) always holds. When \( x \sim U[\bar{z}, \bar{x}] \), from Eqs. (14), (16) and (19), it can be shown that

\[
\frac{TC_{so}}{TC_o} = \frac{1}{3} \left( 1 + \frac{\bar{z}}{\bar{x} + \bar{z}} \right); \quad \frac{TC_w}{TC_o} = \frac{1}{3} \left( 1 + \frac{\bar{x}}{\bar{x} + \bar{z}} \right); \quad \frac{TC_w}{TC_{so}} = \frac{2\bar{x} + \bar{z}}{\bar{x} + 2\bar{z}}. \tag{22}
\]

With Eq. (22), one can verify that

\[
\frac{\bar{x}}{\bar{z}} \Rightarrow \frac{TC_{so}}{TC_o}; \quad \frac{TC_w}{TC_o}; \quad \frac{TC_w}{TC_{so}}. \tag{23}
\]

Since \( 1 < \frac{\bar{x}}{\bar{z}} < \infty \), with Eq. (23), one can verify that Eq. (21) holds. □

Proposition 5.3 provides a meaningful measure of efficiency loss when VOT is uniformly distributed. In this case, the efficiency loss would solely depend on the ratio of \( \bar{x}/\bar{z} \), as shown in Eq. (22). Consistent with Proposition 5.1, when \( \bar{x}/\bar{z} \to 1 \), since commuters are almost homogeneous, the efficiency loss approaches zero, i.e., \( TC_w/TC_{so} \to 1 \). However, when \( \bar{x}/\bar{z} \to \infty \), unlike Proposition 5.2, the efficiency loss can still be bounded, i.e., \( TC_w/TC_{so} \to 2 \), and even in the worst case, travel cost can be reduced by 33%, i.e., \( TC_w/TC_o \to 2/3 \). Fig. 7 shows how \( TC_{so}/TC_o \), \( TC_{av}/TC_o \), and \( TC_w/TC_o \) vary with the ratio of \( \bar{x}/\bar{z} \). In the figure, one can easily verify the first two results in Eq. (23). Also, the first three conditions in Eq. (21) can be verified. As the ratio of \( \bar{x}/\bar{z} \) increases, the worst case becomes less preferable while SO achieves more travel cost reduction.

![Fig. 7. Ratios of TCso/TCo, TCav/TCo and TCw/TCo under different \bar{x}/\bar{z}](image-url)
5.2. Non-proportional heterogeneity

We now turn to consider another type of user heterogeneity where commuters may have different $p_1$ and $p_2$. To focus on the heterogeneity in $p_1$ and $p_2$, we consider an identical VOT for all commuters. It is further assumed that $\eta = p_2/p_1$ is identical for all commuters. However, $p_1$ continuously increases from $p_1^l$ to $p_1^u$ with a cumulative distribution function of $F_1(x) = \Pr(p_1 \leq x)$. We denote $F_1(x) = 1 - F_1(x)$. Note that, even $p_1$ and $p_2$ can vary, it is assumed $p_1 < 1 < p_2$, then we see that, $p_1^2 < 1$ and $p_2^2 > \eta p_1^2 > 1$. Arranging commuters in a decreasing order of $p_1$, and thus $\rho_1(m)$ specifies the mth commuter’s $p_1$, which can be given as follows:

$$\rho_1(m) = F_1^{-1}\left(1 - \frac{m}{N}\right).$$  \hfill (24)

This type of user heterogeneity is similar to that considered in van den Berg and Verhoef (2011) and Tian et al. (2013).

Under this setting, at the no-policy equilibrium, commuters’ arrivals are ordered with respect to their $\rho_1$. More specifically, commuters with lower $\rho_1$ will arrive farther away from the desired arrival time $t^*$. This is because a lower $\rho_1$ means the commuter is less willing to reduce schedule delay by increasing travel time, or is more willing to reduce travel time by increasing schedule delay. The commuter with $\rho_1$ would experience a travel cost as follows:

$$C_w(\rho_1) = \int_{\rho_1^l}^{\rho_1^u} \frac{\eta}{1 + \eta} \cdot wz \cdot \frac{N}{S} f_1(w) \cdot dw + \frac{\eta}{1 + \eta} \cdot \rho_1 \cdot \frac{N}{S} \cdot F_1(\rho_1).$$ \hfill (25)

Total travel cost at the no-policy equilibrium is

$$TC_o = \int_{\rho_1^l}^{\rho_1^u} C_w(\rho_1) f_1(w) N dw.$$ \hfill (26)

Similarly, under the system optimum, queue should be eliminated and schedule delay cost should be minimized. Therefore, commuters with larger $\rho_1$ should arrive closer to the desired arrival time $t^*$. The travel cost of a commuter with $\rho_1$ is

$$C_w(\rho_1) = \frac{\eta}{1 + \eta} \cdot \rho_1 \cdot \frac{N}{S} \cdot F_1(\rho_1).$$ \hfill (27)

Total travel cost at this system optimum is

$$TC_{so} = \int_{\rho_1}^{\rho_1^u} C_w(\rho_1) f_1(w) N dw.$$ \hfill (28)

Similar to the above section, we consider two cases, i.e., the average and worse cases. In the average case, the expected total travel cost is given by the following equation:

$$TC_{av} = \int_{\rho_1^l}^{\rho_1^u} \frac{\eta}{1 + \eta} \cdot wz \cdot \frac{1}{2} \cdot \frac{N}{S} f_1(w) N dw.$$ \hfill (29)

In the worst case, the travel cost of for a specific commuter with $\rho_1$ is

$$C_w(\rho_1) = \frac{\eta}{1 + \eta} \cdot \rho_1 \cdot \frac{N}{S} \cdot F_1(\rho_1).$$ \hfill (30)

Further, the worst-case total travel cost is

$$TC_w = \int_{\rho_1}^{\rho_1^u} C_w(\rho_1) f_1(w) N dw.$$ \hfill (31)

**Proposition 5.4.** $TC_{so} < TC_{av} < TC_w$.

**Proof.** See Appendix C. \hfill \Box

It is straightforward that $TC_w < TC_o$. However, it is worth mentioning that $TC_w$ is not necessarily less than $TC_{av}$. In the worst case, commuters with a larger $\rho_1$ arrive farther away from $t^*$, and experience relatively large schedule delays, which hurts the system performance substantially.

**Proposition 5.5.** $\frac{TC_{av}}{TC_w} \leq \frac{\rho_1^u}{\rho_1^l}$.
To further examine the efficiency loss due to this type of user heterogeneity, we assume two special distributions of \( \rho_1 \): a uniform distribution thus \( F_1(x) = (x - \rho_1^\text{i}) / (\rho_1^\text{u} - \rho_1^\text{i}) \) for \( \rho_1^\text{i} \leq \rho_1 \leq \rho_1^\text{u} \), and a Binomial distribution such that \( F_1(\rho_1^\text{i}) = \omega \) and \( F_1(\rho_1^\text{u}) = 1 - \omega \).

**Proposition 5.6.** When \( \rho_1 \sim U[\rho_1^\text{i}, \rho_1^\text{u}] \), we have

\[
\frac{TC'_w}{TC_o} = \frac{1}{2} \cdot \frac{TC'_w}{TC_o} < \frac{3}{4} \cdot \frac{TC'_w}{TC_o} < \frac{\omega}{2} \cdot \frac{TC'_w}{TC_o} < 2.
\]

**Proof.** When \( \rho_1 \sim U[\rho_1^\text{i}, \rho_1^\text{u}] \), from Eqs. (26), (28), (29) and (31), it can be shown that

\[
\frac{TC'_w}{TC_o} = \frac{1}{2} \cdot \frac{TC'_w}{TC_o} = \frac{3}{4} \left( \frac{\rho_1^u + \rho_1^i}{\rho_1^u + 2\rho_1^i} \right); \quad \frac{TC'_w}{TC_o} = \frac{2\rho_1^u + \rho_1^i}{\rho_1^u + \rho_1^i}.
\]

With Eq. (22), one can verify that

\[
\rho_1^u = \rho_1^i \Rightarrow \frac{TC'_w}{TC_o} = \frac{2\rho_1^u + \rho_1^i}{\rho_1^u + \rho_1^i} = 1.
\]

Since \( 1 < \rho_1^u/\rho_1^i < \infty \), with Eq. (33), one can verify Eq. (32) holds. □

From **Proposition 5.6**, we see that when commuters’ schedule delay penalty is uniformly distributed, the reservation system can still reduce total travel cost even if \( \rho_1^u/\rho_1^i \to \infty \). Furthermore, under this type of user heterogeneity, the relative efficiency of the reservation system decreases as compared to that under the proportional heterogeneity, i.e., \( TC'_w/TC_o > TC'_o/TCo, TC'_w/TC'_o > TC'_o/TCo \) and \( TC'_w/TC'_o > TC'_o/TCo \). This can be verified by comparing Eq. (32) with Eq. (21). This is because in the no policy equilibrium under the non-proportional heterogeneity, commuters’ arrival is ordered in a preferable way so that the schedule delay cost is minimized already. Thus the potential for the reservation system to reduce travel cost is smaller. Another point to highlight is that under both types of heterogeneity, the worst-case total travel cost can be approximately twice of the travel cost associated with SO.

Given \( \rho_1^i \) and \( \rho_1^u \), under the binomial distribution, the variance of \( \rho_1 \) is relatively large. We have:

\[
\begin{align*}
TC'_o &= \frac{1}{2} \cdot \frac{TC'_o}{TC_o} \cdot \frac{N^2}{s} \left( \frac{\omega}{2} \rho_1^i + (1 - \omega)^2 \rho_1^u \right) \\
TC'_w &= \frac{1}{2} \cdot \frac{TC'_w}{TC_o} \cdot \frac{N^2}{s} \left( \omega \rho_1^i + (1 - \omega) \rho_1^u \right) \\
TC'_w &= \frac{1}{2} \cdot \frac{TC'_w}{TC_o} \cdot \frac{N^2}{s} \left( 2\omega - \omega^2 \right) \rho_1^i + (1 - \omega)^2 \rho_1^u \\
TC'_o &= \frac{1}{2} \cdot \frac{TC'_o}{TC_o} \cdot \frac{N^2}{s} \left( 2\omega - \omega^2 \right) \rho_1^i + (1 - \omega^2) \rho_1^u.
\end{align*}
\]

From Eq. (35), one can see that it is possible that \( TC'_w > TC'_o \) or \( TC'_w > TC'_o \). More specifically, it can be shown that \( TC'_w > TC'_o \) if and only if \( \omega < 2/3 \) and \( \rho_1^i/\rho_1^u > (1 + 2w - 3\omega^2)/(2\omega - 3\omega^2) \). When \( \omega \to 0 \), \( (1 + 2w - 3\omega^2)/(2\omega - 3\omega^2) \to \infty \), which indicates that \( \rho_1^i/\rho_1^u > (1 + 2w - 3\omega^2)/(2\omega - 3\omega^2) \) is unlikely to hold. In addition, \( TC'_w > TC'_o \) if and only if \( \omega < 1/2 \) and \( \rho_1^i/\rho_1^u > (1 + w - 2\omega^2)/(\omega - 2\omega^2) \). Similarly, when \( \omega \to 0 \), \( (1 + w - 2\omega^2)/(\omega - 2\omega^2) \to \infty \), which indicates that \( \rho_1^i/\rho_1^u > (1 + w - 2\omega^2)/(\omega - 2\omega^2) \) is also unlikely to hold. The above analysis then implies that if either \( \rho_1 \) is concentrated at the upper bound \( \rho_1^u \) or lower bound \( \rho_1^i \), i.e., \( \omega \to 1 \) or \( \omega \to 0 \), even the worst-case performance of the reservation system yields improvement over the do-nothing case, i.e., \( TC'_w > TC'_o \). A more comprehensive example is given in Fig. 8, which compares the total costs under SO, average case, worst case and the do-nothing case (labeled as OUE, i.e., original user equil.) for different ratios of \( \rho_1^u/\rho_1^i \). Note that, total costs are normalized in the way that when all travelers have identical \( \rho_1 \) equal to the lower bound, total travel cost is one.

As shown in Fig. 8, when \( \rho_1^u/\rho_1^i = 1 \), the SO, average-case and worst-case cost curves coincide with each other, and all of them are below the curve representing the no-policy case. As the ratio increases, the three curves diverge. When this ratio is not too large, e.g., \( \rho_1^u/\rho_1^i = 2 \), even the worst case is still socially preferable than the case of no policy intervention. However, when this ratio becomes larger, the worst case or even the average case can be worse, e.g., \( \rho_1^u/\rho_1^i = 6 \) and \( \rho_1^u/\rho_1^i = 15 \). The last case with \( \rho_1^u/\rho_1^i = 10.000 \) is used to approximate the case where \( \rho_1^u/\rho_1^i \to \infty \). It is evident that the SO is always preferable than OUE in all cases, even \( \rho_1^u/\rho_1^i \) becomes very large.
5.3. Auction-based reservation

As discussed above, the inability of the first-in-first-served reservation system to differentiate commuters with respect to their values of time or schedule preferences would lead the system to lose efficiency. In this section, we now incorporate auctioning into the system to help reveal commuters’ preferences and realize the optimal allocation of reservations among commuters. In our discussion below, we will use the proportional heterogeneity as an example to demonstrate how auctioning can help. For ease of analysis, we focus on the ideal system with an infinite number of reservation steps, i.e., \( n \to \infty \).

We consider that all highway use reservations are allocated or sold to commuters through an auction. It is possible that reservations of the same features for a period (e.g., a month, a quarter) can be bundled for auctioning, thus commuters making recurring trips do not need to participate in auction every day. Here, for simplicity and without loss of generality, we only consider the auction for reservations in one single day. The auction is conducted in the following way: the starting price is set to be zero, and ends when no participant is willing to bid further; at that point the highest bidder pays their bid. This is known as the English auction or open ascending price auction. If the auction is conducted in an online electronic system, it is possible for a commuter to set a maximum price for the auction, and the system will make bids for him or her accordingly. It is assumed that a commuter would bid based on his or her own value of time and schedule preference.

For convenience, we arrange commuters in a decreasing order of \( \alpha \) and thus \( \alpha(m) \) specifies the \( m \)th commuter’s \( \alpha \), which can be given as follows:

\[
\alpha(m) = F^{-1}\left(1 - \frac{m}{N}\right).
\]  
(36)

Furthermore, for ease of presentation, we denote the reservation requires commuters to pass the highway bottleneck before or after a certain time \( t \) by \( t \)th reservation. By using the \( t \)th reservation to travel, since all queues are eliminated, the \( m \)th commuter will experience a travel cost equal to

\[
c_p(m) = \alpha(m) \cdot \max\{\rho_1(t' - t), \rho_2(t - t')\}.
\]  
(37)

Note that in Eq. (37), \( t \) should be within \([t_s, t_e]\), where \( t_s = t' - \rho_2 N/(\rho_1 + \rho_2)s \) and \( t_e = t' + \rho_1 N/(\rho_1 + \rho_2)s \), which represent the starting and ending times of the peak respectively. Suppose that for the morning trip, a commuter can always choose to travel outside of the peak, e.g., he or she can depart earlier than the peak begins and no reservation is required or reservation is always available at no cost. Then, by traveling outside the peak, i.e., passing the highway bottleneck just before \( t_s \) or after \( t_e \), the \( m \)th commuter would experience a travel cost equal to

\[
c_{\text{op}}(m) = \rho \alpha(m) \frac{N}{S}.
\]  
(38)

Then, the commuters’ willingness to pay for a reservation during the peak would be the savings from making a reservation during the peak compared to travel off-peak, i.e.,
\[ p(m, t) = c_{pm}(m) - c_p(m). \]

From Eqs. (37)–(39), it can be verified that

\[ \frac{\partial p(m, t)}{\partial m} = \frac{\partial z(m)}{\partial m} \left[ \frac{N}{s} - \max \left\{ \rho \left( t' - t \right), \rho_2 \left( t - t' \right) \right\} \right]. \]

where \( t \in [t_s, t_e] \) and

\[ \frac{\partial p(m, t)}{\partial t} = \begin{cases} \left( \frac{z(m)\rho_1}{\rho_2} \right) & t_s \leq t \leq t' \\ -z(m)\rho_2 & t' < t \leq t_e \end{cases}. \]

Since \( z(m) \) decreases in \( m \) and \( t \in [t_s, t_e] \), then we see Eq. (40) is negative, indicating commuters with higher VOT are willing to pay more for the same reservation. Also note from Eq. (41), the \( n \)th reservation with a \( t \) closer to \( t' \) is always preferable for all commuters. Based on the above analysis, we then can conclude that generally the commuters with higher VOT are willing to and finally will get the reservation closer to \( t' \). More specifically, the \( m \)th commuter will either obtain the reservation with \( t = t' - \rho_2m/(\rho_1 + \rho_2)s \) or \( t = t' + \rho_1m/(\rho_1 + \rho_2)s \). With Eq. (39), the price he or she will bid can be determined as follows:

\[ p(m) = \rho z(m) \frac{N - m}{s}. \]

In Eq. (42), the \( z(m) \) is determined by Eq. (36), which indicates that the resulting equilibrium price tightly relates to the distribution of value of time among the population. The travel cost the \( m \)th commuter would experience will be the summation of the travel time and schedule delay costs, i.e., \( \rho z(m) \frac{N}{s} \), and the price he pay for the reservation, i.e., \( p(m) \), which is \( \rho z(m) \frac{N}{s} \).

The above discussion shows the potential for auction to help the reservation system. Further study might explore for an implementable mechanism of the auction-based reservation system, similar to the auction mechanism for tradable network permit discussed in Wada and Akamatsu (2011), or the hybrid mechanism of auction and capacity control discussed in Wada and Akamatsu (2013). The auction-based reservation system is a mix of the price-based regulation and quantity-based regulation. For general comparisons between price-based regulation and quantity-based regulation in economics or transportation, one may refer to Weitzman (1974), Laffont (1977) and Shirmohammadi et al. (2013). It is worth mentioning that, compared with the tradable network permits system discussed in Akamatsu et al. (2006, 2007), the proposed reservation system is more of a quantity-based regulation as the reservations are not tradable. If we allow the reservation to be tradable, the reservation system would be similar to the tradable network permits system. However, our study here provides a more comprehensive analysis of efficiency of the reservation system.

### 6. Concluding remarks

This paper examines the reservation system for highway use in the morning commute. The reservation system allocates the highway space to potential users during different time intervals and accepts or rejects users’ reservation requests based on the highway capacity and the number of reservations made. Since highway use reservation requests beyond highway capacity are rejected or abandoned, queuing at the highway bottleneck can be reduced and traffic congestion is then relieved. In the ideal case where we can set system with an infinite number of reservation steps, the reservation system approaches the same efficiency as the well-known first-best time-varying toll. Efficiency loss due to a limited number of steps of a practical system is bounded.

We also show that braking or tactical waiting behavior of travelers would lead the reservation system to lose efficiency, which then should be proactively accommodated. User heterogeneity is another factor that affects the efficiency of the reservation system. Two types of user heterogeneity, i.e., proportional and non-proportional, and the associated efficiency losses are discussed. For proportional heterogeneity, even the worst case under the reservation system is still preferred rather than the no policy equilibrium. However, this is not true for the case with non-proportional heterogeneity. Note that our efficiency analysis of the cases with user heterogeneity assumes the system manager can set \( n \rightarrow \infty \), which would be impractical in reality. If we consider \( n \) is finite, additional efficiency loss may arise as the system cannot always guarantee that travelers with higher value of time can enjoy less queuing time. This might be a direction for further research.

In this study, auction-based reservation is proposed to reduce the inefficiency due to user heterogeneity. Further research may consider to develop practically implementable auction mechanisms for the reservations, such as those for vehicle licenses implemented in Chinese cities, e.g., Shanghai, Guangzhou, and Shenzhen. In this case, a considerable amount of reservations might be auctioned on a weekly or monthly basis for those regular travelers, while some capacity (as well as reservations) should be used to accommodate travelers on a daily basis. Besides, various practical issues related to implementation should be discussed to ensure that the reservation system is efficient and user friendly.

Also, as travel demand is intrinsically uncertain, reservation requests of travelers can provide additional information for the system manager to better manage a highway bottleneck. Our future study will examine potential benefits from accommodating travel demand uncertainty.

In addition to the factors discussed in the current paper, there are many others which may lead the system to lose efficiency, such as no show of travelers with reservations, time and/or resources that travelers and system planner have to
dedicate to making a reservation. Since no show is a waste of highway capacity, strategies such as over-booking in the airline industry should be developed to mitigate the inefficiency.

Acknowledgements

The authors would like to thank anonymous referees for their useful comments on an earlier version of the paper. The authors are also grateful to Associate Editor, Prof. Henry Liu at University of Michigan, Ann Arbor, for handling the paper. The work described in this paper was supported by grants from Hong Kong’s Research Grants Council under project HKUST620712. Yafeng Yin would like to thank the support from the U.S. National Science Foundation (CNS-1239364; CMMI-1256106; CMMI-1362631).

Appendix A. Proof of Proposition 5.1

Under the user heterogeneity setting in Section 5.1, the commuters can be rearranged in the order of decreasing \( a_i \), the \( m \)th commuter’s VOT is then

\[
 a(m) = F^{-1}\left(1 - \frac{m}{N}\right),
\]

where \( m \in [0, N] \), \( a(0) = \bar{a} \) and \( a(N) = \bar{a} \). The total travel cost at the no-policy equilibrium given in Eq. (14) can then be rewritten as

\[
 T_{Co} = \int_0^N \rho a(m) \frac{N}{s} dm.
\]

At the system optimum, the \( m \)th commuter would arrive at destination either at \( t^* - \rho m/\rho_1 s \) or \( t^* + \rho m/\rho_2 s \). Total travel cost given in Eq. (16) can be rewritten as

\[
 T_{so} = \int_0^N \rho a(m) m \frac{N}{s} dm.
\]

At the worst case, the \( m \)th commuter would arrive at the destination either at \( t^* - \rho(N-m)/\rho_1 s \) or \( t^* + \rho(N-m)/\rho_2 s \). The total travel cost given in Eq. (19) can be rewritten as

\[
 T_{w} = \int_0^N \rho a(m) \frac{N - m}{s} dm.
\]

In addition, the total travel cost in the average case given in Eq. (17) can be rewritten as

\[
 T_{av} = \int_0^N \rho a(m) \frac{1}{2} N \frac{N}{s} dm.
\]

From Eqs. (44)–(47), it can be easily shown that \( T_{so} < T_{Co} \), \( T_{w} < T_{Co} \) and \( T_{av} < T_{Co} \), since \( m \leq N, N - m \leq N \) and these two inequalities strictly hold for some \( m \), and \( 0.5N < N \). Since \( T_{av} = 0.5T_{Co} \), and \( T_{w} + T_{so} = T_{Co} \), to prove the proposition, it suffices to show that \( T_{so} < T_{w} \). From Eqs. (45) and (46), we see that

\[
 T_{so} = \int_0^N \rho a(m) m \frac{N}{s} + \rho a(N-m) \frac{N-m}{s} dm
\]

\[
 T_{w} = \int_0^N \rho a(m) \frac{N - m}{s} + \rho a(N-m) \frac{m}{s} dm.
\]

For \( m \leq 0.5N \), one can verify that

\[
 \rho a(m) \frac{m}{s} + \rho a(N-m) \frac{N-m}{s} \leq \rho a(m) \frac{N-m}{s} + \rho a(N-m) \frac{m}{s}.
\]

Note that the inequality in Eq. (49) strictly holds for some \( m \). With Eq. (48), we then conclude that \( T_{so} < T_{w} \). This completes the proof. \( \square \)

Appendix B. Proof of Proposition 5.2

From Eqs. (45) and (46), we know that

\[
 T_{so} \geq \rho \bar{a} \int_0^N \frac{m}{s} dm = \rho \bar{a} \frac{1}{2} N^2.
\]

\[
 T_{w} \leq \rho \bar{a} \int_0^N \frac{N-m}{s} dm = \rho \bar{a} \frac{1}{2} N^2.
\]
From Eqs. (50) and (51), we conclude that

\[
\frac{TC_w}{TC_{so}} \leq \frac{2}{N}.
\]  
(52)

This completes the proof. □

Appendix C. Proof of Proposition 5.4

In Section 5.2, commuters are rearranged in the order of decreasing \( \rho_1 \), the \( m \)th commuter will either arrive at their destination at \( t' = \eta m / (1 + \eta) s \) or \( t' + m / (1 + \eta) s \). The total travel cost at the no-policy equilibrium given in Eq. (26) can then be rewritten as

\[
TC_{so} = \frac{\eta}{1 + \eta} \rho_1 \int_0^N \left( \frac{m}{s} + \int_m^N \frac{1}{s} dx \right) dm.
\]  
(53)

At system optimum, the \( m \)th commuter would arrive at the destination either at \( t' = \eta m / (1 + \eta) s \) or \( t' + m / (1 + \eta) s \). Total travel cost given in Eq. (28) can be rewritten as

\[
TC_{so} = \frac{\eta}{1 + \eta} \rho_1 \int_0^N \frac{m}{s} dm.
\]  
(54)

At the worst case, the \( m \)th commuter would arrive at the destination either at \( t' = \eta(N - m) / (1 + \eta) s \) or \( t' + (N - m) / (1 + \eta) s \). Total travel cost given in Eq. (31) can be rewritten as

\[
TC_w = \frac{\eta}{1 + \eta} \rho_1 \int_0^N \frac{N - m}{s} dm.
\]  
(55)

In addition, the total travel cost in the average case given in Eq. (29) can be rewritten as

\[
TC_{av} = \frac{\eta}{1 + \eta} \rho_1 \int_0^N \frac{N}{2s} dm.
\]  
(56)

With Eqs. (54)–(56), the remaining proof is similar as that of Proposition 5.1. □

Appendix D. Proof of Proposition 5.5

From Eqs. (50) and (51), we know that

\[
TC_{so} \geq \frac{\eta}{1 + \eta} \rho_1 \int_0^N \frac{m}{s} dm = \frac{\eta}{1 + \eta} \rho_1 \frac{1}{2} N^2.
\]  
(57)

\[
TC_w \leq \frac{\eta}{1 + \eta} \rho_1 \int_0^N \frac{N - m}{s} dm = \frac{\eta}{1 + \eta} \rho_1 \frac{1}{2} N^2.
\]  
(58)

From Eqs. (57) and (58), we conclude that

\[
\frac{TC_w}{TC_{so}} \leq \frac{\rho_1}{\rho_1}.
\]  
(59)

This completes the proof. □

References


