TI 2024-025/VIII
Tinbergen Institute Discussion Paper

Preference heterogeneity in a dynamic flow congestion model

Xiaojuan Yu[1]
Vincent A.C. van den Berg[2]
Erik T. Verhoef[3]

[1] Vrije Universiteit Amsterdam and Zhongnan University

[2] Vrije Universiteit Amsterdam and Tinbergen Institute

[3] Vrije Universiteit Amsterdam and Tinbergen Institute

Preference heterogeneity in a dynamic flow congestion model

Xiaojuan Yu[a,b*], Vincent A.C. van den Berg[b,c], Erik T. Verhoef[b,c]

[a] School of Business Administration, Zhongnan University of Economics and Law, Wuhan 430073, China
[b] Department of Spatial Economics, VU Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
[c] Tinbergen Institute, Gustav Mahlerplein 117, 1082 MS Amsterdam, The Netherlands

Abstract

We study how preference heterogeneity affects travel behavior and congestion pricing in a dynamic flow congestion model. We formulate and solve a multi-point optimal control problem using a Hamiltonian-based method to derive the social optimum. The properties of the travel equilibrium are explored analytically, particularly for travelers' arrival rates, arrival intervals, congestion externalities, and tolls. In the absence of tolling, the arrival order is determined by the ratio of the value of time (VOT) to the value of schedule delay, as in the bottleneck model. However, unlike the bottleneck model, the same holds for the social optimum when only the VOT differs across users, as travel delays will not be fully eliminated. In social optimum, the arrival rate, travel delay, and toll jump discontinuously at the boundary time between user types, but these discontinuities do not undermine the stability of the socially optimal equilibrium. Assessment of the distributional effects indicates that users with a lower VOT always lose from tolling, whereas users with a higher VOT may gain or lose from tolling. The latter depends on the type and degree of heterogeneity, the elasticity of travel delay with respect to arrival rate, and the number of users for both types. Compared to the bottleneck model, tolling is less beneficial for society and hurts users more. Our findings reveal the significance of the type of congestion and preference heterogeneity when assessing the implementation of congestion tolling.

Keywords: Dynamic flow congestion; Bottleneck model; Preference heterogeneity; Congestion pricing; Distributional effects; Optimal control

* Corresponding author.

Email addresses: x.yu@vu.nl, v.a.c.vanden.berg@vu.nl, e.t.verhoef@vu.nl.

1. Introduction

Traffic congestion is a large problem in many cities worldwide, particularly during peak hours. Different dynamic congestion models have been proposed to study the dynamics of congestion. While it is natural that there are various types of models, with different strengths and weaknesses, there is a danger in some models gaining dominance. In particular, the Vickrey(1969) bottleneck model has dominated the literature. Its specifics make it a natural favorite (see Small and Verhoef, 2007; Small, 2015; and Li et al., 2020 for extensive reviews), but it is important to remain alert on the question how general the insights derived from it are?

Recently, there has been a growing interest in flow-based dynamic congestion models more generally, including also Bathtub and MFD models (e.g., Arnott, 2013; Yildirimoglu et al., 2015; Liu and Geroliminis, 2016; Arnott and Buli, 2018; Verhoef and Silva, 2018; Bao et al., 2019; Long and Szeto, 2019; Verhoef, 2020; Arnott and Kilani, 2022; Beojone and Geroliminis, 2023). This literature complements an already sizable literature employing bottleneck congestion models, with insights on how the nature of dynamic congestion - i.e. flow versus bottleneck congestion-affects policy recommendations and welfare consequences of congestion management strategies.

Preference heterogeneity can significantly alter the overall and distributional impacts of congestion policies. Earlier studies using the bottleneck model found differences in behavior across heterogeneous users, particularly in their departure-time choices (e.g., Arnott et al., 1988, 1994; Lindsey, 2004; Wu and Huang, 2015; Liu et al., 2015) and in response to congestion tolls (Cohen, 1987; Arnott et al., 1994; Van den Berg and Verhoef, 2011a, b; Van den Berg, 2014; Wu and Huang, 2014; Chen et al., 2015; Sun et al., 2020; Guo et al., 2023; Van den Berg, 2024). However, existing studies on preference heterogeneity primarily use static or bottleneck congestion. Little is known about the role of preference heterogeneity in dynamic flow congestion, particularly for the distributional and welfare effects of congestion pricing.

In an influential paper, Chu (1995) showed how dynamic flow-congestion models produce insights that may deviate from, and hence enrich, those of the bottleneck model. We extend his approach to include various forms of preference heterogeneity, and analyze the overall and distributional effects of congestion pricing. We believe that our approach contributes valuable insights for real world congestion policies, which likely involve both bottleneck and flow congestion.[1]

Against this background, our study examines the effects of preference heterogeneity in a

[1] Mun (1999, 2002) extended Chu's work into a model with both flow and bottleneck congestion.

dynamic flow-congestion model. We focus on three questions. (i) How will preference heterogeneity affect dynamic equilibrium behavior and congestion? (ii) What are the socially optimal tolls under heterogeneous preferences? (iii) What are travelers' responses to such a toll, and how are users affected by tolls?

To study these questions, we use the model of Chu (1995, 1999).[2] The model's essential feature is that the travel delay depends on the instantaneous arrival flow at the arrival moment, without congestion interaction between users with different arrival times.[3] Earlier studies have shown that, in the bottleneck model, a triangular, time-varying toll can fully eliminate queues and decentralize the first-best optimum. Under homogeneity, the generalized price (i.e., travel cost plus toll) is the same as without tolling (Arnott et al., 1988, 1994; Van den Berg and Verhoef, 2011a, b). Conversely, in (dynamic) flow congestion, travel delays persist, and under homogeneity, optimal tolling raises the generalized price (Chu, 1995). We include preference heterogeneity into the dynamic flow congestion model, and will show how the simultaneous change in travel delay and toll complicates the derivation, thereby altering the properties of socially optimal equilibrium.

We consider two types of heterogeneity: 'ratio' and 'proportional' heterogeneity. Ratio heterogeneity means there is heterogeneity in the ratio of the value of time (VOT) to the value of schedule delay. The VOT is the ratio of the marginal utility of travel time to the marginal utility of income, and similarly for the value of schedule delay. Ratio heterogeneity measures how people differ in how they trade off travel time and schedule delay; or, in other words, how they differ in how flexible they are in terms of when to arrive. This heterogeneity could, for example, stem from variations in job type, trip purpose, or family status, since these differences alter how flexible people are (Van den Berg and Verhoef, 2011a; Hall, 2018; Van den Berg, 2024). Conversely, proportional heterogeneity varies all values in a fixed proportion. It could stem from heterogeneity in the marginal utility of income caused by income differences[4] (Van den Berg and Verhoef, 2011a). As we will see, preference heterogeneity leads to significant differences in the properties of the travel equilibrium compared to those of the bottleneck model. These are

[2] Tractability is a major challenge for more elaborate dynamic models. Even the Agnew (1977) and bathtub models, which assume spatial homogeneity, typically do not have closed-form solutions. As with continuous time continuous space models, such as the car-following model or the hydrodynamic Lighthill-Whitham-Richards (LWR) model, these models require numerical methods to be solved.

[3] This differs from the bottleneck model, whereby departures during queuing also affect the travel times of users departing later. The Chu model thus describes a dynamic equilibrium under flow congestion, which is particularly relevant when there are no strict and predictable queues, and speeds, densities, and flows are below their free flow values throughout the facility or network.

[4] The value of time is the (absolute of) ratio of the marginal utility of time to marginal utility of income, and similar for the values of schedule delay. So, changing the marginal utility income, changes all values for the same percentage. It is normal that the marginal utility of income falls with the income of people., Hence, pure proportional heterogeneity can stem from income, if income does not affect the marginal utilities of time and schedule delay directly.

important for policy making in practice.

In a recent study, Long and Szeto (2019) considered general heterogeneity in values of time and schedule delay in a dynamic flow congestion model, while also incorporating an environmental externality. Their un-tolled setting is like our case of ratio heterogeneity, but they only analyze the arrival order and do not explore the analytical properties of the equilibrium and congestion effects. Furthermore, they analyze a second-best toll, using partially analytical methods and partially numerical methods. Conversely, we use pure analytical methods to analyze and solve for the first-best toll and the travel equilibrium under separate ratio and proportional heterogeneity, using dynamic optimization with Kuhn-Tucker Hamiltonians. We then illustrate the effects in a numerical model.

Our paper makes three main contributions. *First*, we present a flow-based congestion model with heterogeneous preferences and derive closed-form solutions for the equilibrium. The properties of the equilibrium are explored analytically, particularly for travelers' arrival rates, arrival intervals, and congestion externalities. In the absence of tolling, although bottleneck and flow congestion approaches lead to the same arrival orders of different types of users, the travel patterns and congestion effects are substantially different.

Second, we derive the social optimum, which minimizes total travel cost. To do so, we formulate an optimal control problem and solve it using Hamiltonians, adding Khun-Tucker conditions to determine when a certain type travels. We find that user types self-separate over arrival time, even when scheduling preferences are the same and only the VOT differs. Specifically, under ratio heterogeneity—for which only the VOT differs and scheduling preferences are the same—travelers with a lower VOT travel in the center of the peak period, whereas, under proportional heterogeneity, travelers with a higher VOT travel in the center. Discontinuities in arrival rate, travel delay, and toll occur at the boundary time between types but do not undermine the stability of the toll-supported social optimum. This is markedly different from the bottleneck model.

Third, we conduct an analytical investigation of the distributional effects of tolling on travelers and compare them for different types of heterogeneity. As the travel delay cannot be eliminated in this model, the efficiency gains from tolls are smaller than in the bottleneck model. Travelers with a lower VOT always lose from tolling, and users with a higher VOT may gain or lose from tolling, depending on the type of heterogeneity and the parameters. Hence, compared to the bottleneck model, tolling is less beneficial for society and hurts users more: the reduction of aggregate travel delay here requires acceptance of higher schedule delays. Tolling is less attractive for users and has different distributional effects.

The remainder of this paper is organized as follows. The next section presents our flow-based congestion model under heterogeneity. Section 3 studies the travel equilibrium and congestion externality without tolling under two-type ratio heterogeneity, and Section 4 examines the social optimum in this setting. Section 5 studies the model under proportional heterogeneity. Section 6 compares our results with those from the bottleneck model. Section 7 turns to the numerical model and presents extensive sensitivity analyses. Section 8 briefly discusses other forms of heterogeneity. Section 9 concludes the paper.

2. Basic model

This section presents a general formulation of the flow-based congestion model for many types of users (or 'types' for brevity) and any form of discrete heterogeneity in the values of time and schedule delay, with a homogeneous preferred arrival time. Suppose that every morning a fixed number of travelers travel from home to a workplace along a single road with a fixed capacity per hour, *K*. All travelers wish to arrive at their workplace at an identical preferred arrival time, t^*. Those who arrive early or late encounter a schedule delay cost. Each traveler chooses their arrival time based on a trade-off between the travel delay cost, schedule delay cost, and possibly a toll, in order to minimize their travel price (i.e., travel cost plus toll).

2.1 A general formulation of travel time and costs

Let $T(t)$ represent the travel time at arrival time *t*. Following Chu (1995), we assume that a traveler's speed on that road is constant over time during the trip and depends only on the arrival rate at the road's exit when the trip is completed. This avoids complications from other congestion technologies, as the model ignores congestion interactions between individuals traveling at different moments, regardless of how close these moments are.[5] Therefore, the travel speed at arrival time *t* is set by the arrival flow through a power function (or BPR function). The travel time function, $T(t)$, is

$$T(t) = T(f_1(t), f_2(t), ..., f_n(t); K) = T_f + \left(\frac{\sum_i f_i(t)}{K} \right)^{\chi}, \tag{1}$$

where $f_i(t)$ is the arrival flow of type *i* at arrival time *t*, and χ governs the curvature of the

[5] As discussed in footnote 2, tractability is a major challenge for more elaborate dynamic flow congestion models.

travel delay relation $T(t)$ and represents the elasticity of travel delay with respect to arrival flow. For analytical convenience and without much loss of generality, the free-flow travel time, T_f, is, for now, normalized to zero. The numerical analysis will consider a positive value.

Let $c_i(t)$ represent the travel cost of type *i* travelers arriving at *t*. It consists of the travel time cost and schedule delay cost of arriving early or late. Under the conventional assumptions of a linear schedule delay cost function and a constant value of travel delays, $c_i(t)$ equals:

$$c_i(t) = \alpha_i \cdot T(f_1(t), f_2(t), ..., f_n(t); K) + \begin{cases} \beta_i \cdot (t^* - t) & \text{if } \mathrm{t} \le \mathrm{t}^* \\ \gamma_i \cdot (t - t^*) & \text{if } \mathrm{t}>\mathrm{t}^* \end{cases}, \tag{2}$$

where α_i is the VOT for type *i*. β_i is the value of schedule delay early for type *i*: it gives the shadow price of an arrival one hour earlier than is most preferred; γ_i is the corresponding value for late arrivals. Under the plausible assumption that early arrivers prefer ending the trip over continuing it, $\alpha_i > \beta_i$ should hold. Following convention, denote δ_i as a composite scheduling preference parameter, with $\delta_i = \beta_i \gamma_i / (\beta_i + \gamma_i)$.

In the absence of tolling, travelers choose their arrival times to minimize their own travel cost. Let t_{si} denote the arrival time of the first type-*i* traveler, and t_{ei} the arrival time of the last type-*i* traveler. The total travel cost for type i is the integral of the product of the arrival rate and travel cost over arrival time t (between t_{si} and t_{ei}). The total travel cost, TC, equals the sum of the travel costs of the different types:

$$TC = \sum_i \int_{t_{s_i}}^{t_{ei}} f_i(t) \cdot c_i(t) dt. \tag{3}$$

2.2 Social optimum

The no-toll equilibrium is not efficient due to uninternalized congestion externalities. Time-varying tolling can be applied to internalize the congestion externality. Let $\tau(t)$ denote the time-varying toll charged for arrival at *t*. Let $p_i(t)$ denote the travel price of type-*i* travelers arriving at *t*, encompassing the toll, travel time cost,[6] and schedule delay cost:

[6] For the conventional bottleneck model, the optimal time-varying toll eliminates queuing and thus travel time delays. However, as previously stated, in the dynamic flow congestion model, the optimal time-varying toll will typically not fully eliminate travel delays.

$$p_i(t)=\tau(t)+\alpha_i\cdot T(f_1(t),f_2(t),...,f_n(t);K)+\begin{cases}\beta_i\cdot(t^*-t) & \text{if t} \le \text{t}^* \\ \gamma_i\cdot(t-t^*) & \text{if t>t}^*\end{cases}. \tag{4}$$

We assume that the demand per type of user is fixed. The social regulator then chooses $f_i(t)$, t_{si} and t_{ei} to minimize the total travel cost under the constraint that all users should arrive in their arrival intervals. This leads to the following total travel cost minimization problem:

$$\begin{aligned}&\min_{\substack{f_1(t),f_2(t),...,f_n(t),\\ t_{s1},t_{s2},...,t_{sn},t_{e1},t_{e2}...,t_{en}}} \quad TC \\ &s.t. \quad \int_{t_{si}}^{t_{ei}} f_i(t)dt=N_i,\ i=1,2,...,n\end{aligned}, \tag{5}$$

where the total travel cost, *TC*, is defined in (3).

2.3 Definition of dynamic equilibrium

In the dynamic equilibrium, no traveler can reduce her generalized price by unilaterally changing her arrival time. This implies that all travelers within a type incur the same price for their chosen arrival times and face equal or higher travel prices at any other time. Moreover, all N_i users of any type *i* should arrive and thus $\int_{t_{si}}^{t_{ei}} f_i(t)dt=N_i$.

In the absence of tolling, solving $dc_i(t)/dt=0$ yields:

$$\frac{dT}{dt}=\begin{cases}\beta_i/\alpha_i, & \text{for } i \text{ type users with positive early arrivals at moment } t \\ -\gamma_i/\alpha_i, & \text{for } i \text{ type users with positive late arrivals at moment } t\end{cases}. \tag{6}$$

The travel delay should increase at a rate of β_i/α_i within type *i*'s early-arrival interval and decrease at a rate of γ_i/α_i within type *i*'s late-arrival interval. This condition mimics that for bottleneck congestion (e.g., Arnott et al., 1988).

For the first-best social optimum, the toll pattern also matters and solving $dp_i(t)/dt=0$ yields:

$$\frac{d\tau(t)}{dt}+\alpha_i\cdot\frac{dT(t)}{dt}=\begin{cases}\beta_i, & \text{for } i \text{ type users with positive early arrivals at moment } t \\ \gamma_i, & \text{for } i \text{ type users with positive late arrivals at moment } t\end{cases}, \tag{7}$$

implying that to secure a dynamic equilibrium, for type *i* the sum of the travel delay cost and toll needs to increase at rate β_i for early arrivals and decrease at rate γ_i for late arrivals. Condition (7) highlights the main difference between Chu's model and the bottleneck model under heterogeneity. As travel delays will not be eliminated, the simultaneous changes in travel delay and toll in (7) make travel patterns ambiguous, thereby elevating the complexity of solving for the social optimal equilibrium.

To enhance the transparency of the results, we now focus on two types: H and L. Without loss of generality, we consider the H type to have a higher VOT than L type (i.e., $\alpha_H > \alpha_L$). We thus obtain 'ratio heterogeneity' with identical scheduling preferences, and obtain 'proportional heterogeneity' when β_i and γ_i vary in a fixed proportion over the two types. Other forms of heterogeneity are briefly discussed in Section 8.

3. Two-type ratio heterogeneity in dynamic flow congestion: no tolling

Our first form of preference heterogeneity, ratio heterogeneity, captures differences in how travelers value travel time versus schedule delays or how flexible they are regarding arrival times.[7] This heterogeneity could, for instance, result from differences in type of job, family status, or the purpose of the trip.

We introduce ratio heterogeneity by letting the VOT, α_i vary while the other values remain fixed. As noted, this section considers two types of users: H and L. The H type has a higher VOT, and the L type a lower VOT. The H type cares relatively more about travel time losses than about when to arrive (i.e., the schedule delay). These drivers are thus relatively more flexible in when to arrive, and consequently have a lower queuing tolerance. This section explores travel behavior without tolling, focusing on travel equilibrium in terms of arrival order, arrival rate, and travel cost. We also examine congestion externalities caused by the different user types.

3.1 Arrival order

Following Arnott et al. (1988) and Van den Berg and Verhoef (2011b), we construct iso-cost curves representing combinations of travel delay (along the vertical axis) and schedule delay (along the horizontal axis), resulting in a constant travel cost over time. By condition (6), travel delay increases at a rate of β/α_i for early arrivals, and decreases at a rate of γ/α_i for late arrivals. As a result of a lower VOT, the iso-cost line for the L type is steeper than that for the H type. Fig. 1 illustrates the associated equilibrium iso-cost curves. The solid lines represent the

[7] Following Van den Berg and Verhoef (2011a), we denote such heterogeneity as ratio heterogeneity since the VOT varies relative to the values of schedule delay, and the queuing intolerance as given by the ratios α/β and α/γ varies over time.

equilibrium travel delay, while the dashed lines depict the out-of-equilibrium continuation of the iso-cost function. Although users do not arrive at these times, the dashed lines indicate the necessary travel delay for them to incur the same travel cost. The equilibrium arrival order is summarized in Proposition 1.

Proposition 1. Under ratio heterogeneity for which only the VOT differs, in the un-tolled equilibrium, users with a lower VOT travel in the center of the peak period, and users with a higher VOT in the shoulder of the peak period.
Proof. See Appendix A. □

The arrival order in Proposition 1 is consistent with that in the bottleneck model. The intuition behind Proposition 1 is that users with a higher VOT are more willing to accept larger schedule delays in exchange for shorter travel times, or, equivalently, have a lower queuing tolerance. The two types arrive separately in time since the evolution of travel times by arrival time that keeps one type arriving at that moment in equilibrium pushes the other type toward its own window.

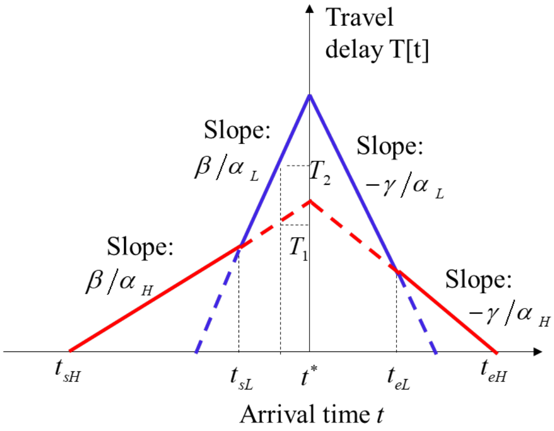


Fig. 1. Travel delay under ratio heterogeneity without tolling.
Note: L type users travel in the center and H type users travel in the shoulder of the peak period.

3.2 Travel equilibrium

As a result of the separated travel patterns, the travel cost can be further formulated as:

$$c_i(t)=\alpha_i\cdot\left(\frac{f_i(t)}{K}\right)^{\chi}+\begin{cases}\beta\cdot(t^*-t), & \text{for } i \text{ type users with positive early arrivals at } t\\ \gamma\cdot(t-t^*), & \text{for } i \text{ type users with positive late arrivals at } t\end{cases}. \tag{8}$$

In dynamic equilibrium, the H type arrives within the interval $[t_{sH},t_{sL}]\cup(t_{eL},t_{eH}]$, and the L type arrives within the interval $(t_{sL}, t_{eL}]$, i.e., $\int_{t_{sH}}^{t_{sL}} f_H(t)dt + \int_{t_{eL}}^{t_{eH}} f_H(t)dt = N_H$ and $\int_{t_{sL}}^{t_{eL}} f_L(t)dt = N_L$ hold. Solving the associated equilibrium conditions, we can derive the resulting travel equilibrium. Detailed derivations can be found in Appendix B.

Despite flow and bottleneck congestion having the same arrival order, their travel equilibriums differ significantly. The equilibrium travel costs for the two user types are:

$$\begin{cases} c_H = \alpha_L \left((\frac{1}{\chi}+1) \cdot \left(\frac{\delta N_H}{\alpha_H K} + \frac{\delta N_L}{\alpha_L K} \right) \right)^{\frac{\chi}{\chi+1}} + \left(\frac{\delta N_H}{\alpha_H K} \cdot \frac{\chi+1}{\chi} \right)^{\frac{\chi}{\chi+1}} \cdot (\alpha_H - \alpha_L), \\ c_L = \alpha_L \left((\frac{1}{\chi}+1) \cdot \left(\frac{\delta N_H}{\alpha_H K} + \frac{\delta N_L}{\alpha_L K} \right) \right)^{\frac{\chi}{\chi+1}}. \end{cases} \tag{9}$$

Taking the difference between c_H and c_L, the travel cost difference between the different types of users is:

$$\Delta c = c_H - c_L = \left(\frac{\delta N_H}{K} \cdot \frac{\chi+1}{\chi} \right)^{\frac{\chi}{\chi+1}} (\alpha_H)^{\frac{1}{\chi+1}} \cdot (1 - \frac{\alpha_L}{\alpha_H}) \geq 0. \tag{10}$$

The cost difference, Δc, rises with the degree of ratio heterogeneity, α_H / α_L.[8] It should be noted that under bottleneck congestion, Δc increases linearly with α_H / α_L, whereas under flow congestion, the power of the speed-flow function also matters, as shown by the term $(\alpha_H)^{\frac{1}{\chi+1}}$.

3.3 Congestion externality with ratio heterogeneity

Under bottleneck congestion, Van den Berg and Verhoef (2011b) found that L type causes higher congestion effects, which mimics the analytical expression for homogeneous users, whereas H type causes lower congestion effects. In the present flow congestion model, congestion effects are:

$$\begin{aligned} &\frac{\partial c_H}{\partial N_L} = \frac{\partial c_L}{\partial N_L} = \left(\frac{\chi+1}{\chi} \cdot \left(\frac{\delta N_H}{\alpha_H K} + \frac{\delta N_L}{\alpha_L K} \right) \right)^{-\frac{1}{\chi+1}} \cdot \frac{\delta}{K}, \quad \frac{\partial c_L}{\partial N_H} = \frac{\alpha_L}{\alpha_H} \cdot \frac{\partial c_L}{\partial N_L}, \\ &\frac{\partial c_H}{\partial N_H} = \frac{\partial c_L}{\partial N_H} + \left(\frac{\delta N_H}{\alpha_H K} \cdot \frac{\chi+1}{\chi} \right)^{-\frac{1}{\chi+1}} \cdot \frac{\delta(\alpha_H - \alpha_L)}{\alpha_H K} > \frac{\partial c_H}{\partial N_L} \end{aligned}. \tag{11}$$

This implies that L type imposes an equal congestion effect on both types of users. Perhaps

[8] Denote $\alpha_H / \alpha_L = m$. Taking the derivative of (10) with respect to m yields $\partial(\Delta c)/\partial m > 0$. Specifically, when $\alpha_H = \alpha_L$, all users have the same arrival patterns and travel cost.

surprisingly, but matching the general notion that preference heterogeneity tends to dampen congestion as travelers are less inclined to impede each other's travel windows, H-type users impose a lower congestion effect on L types than they do on themselves.

The marginal external cost, MEC_i, of type i equals the sum of the congestion effects it imposes on all travelers, which can be calculated as $MEC_i = \partial c_H / \partial N_i \cdot N_H + \partial c_L / \partial N_i \cdot N_L$. Combining (11) and further taking the difference between MEC_H and MEC_L, we find the following property in the relationship between the MEC of the different types.

Proposition 2. Under dynamic flow congestion with ratio heterogeneity, when $\frac{\alpha_H}{\alpha_L} \le \frac{(1+N_L/N_H)^{\chi+1}-1}{N_L/N_H}$ is satisfied, $MEC_L \ge MEC_H$ holds; otherwise, when $\frac{\alpha_H}{\alpha_L} > \frac{(1+N_L/N_H)^{\chi+1}-1}{N_L/N_H}$ is satisfied, $MEC_L < MEC_H$.

Proof. See Appendix C. □

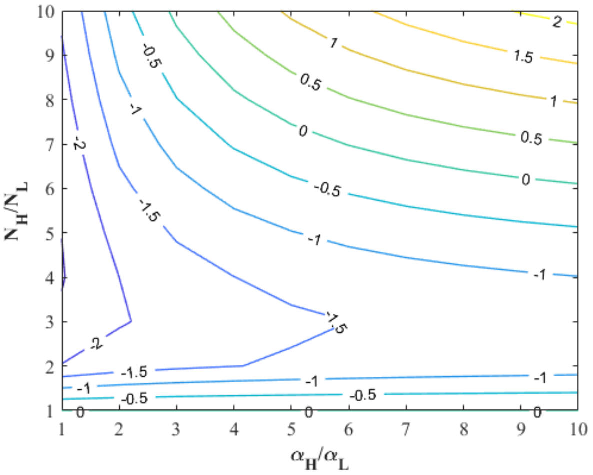
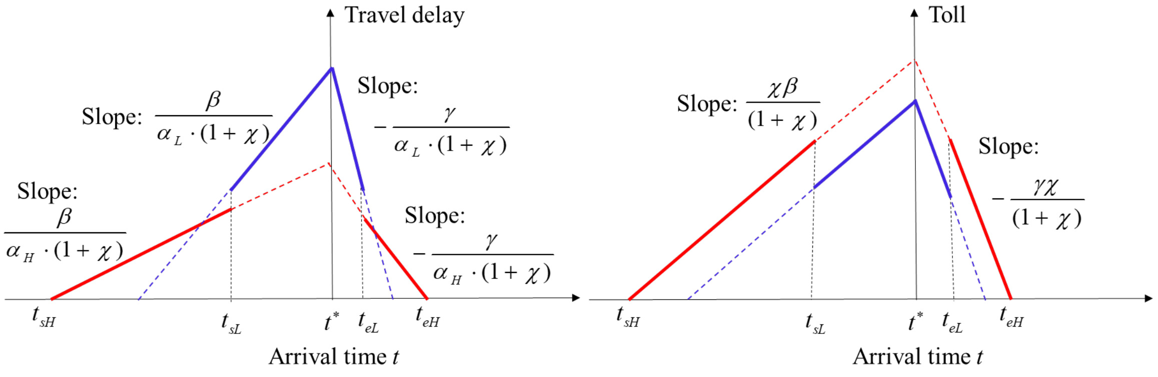


Fig. 2. Contour plot of the difference in marginal external cost of the types, MEC$_H$ – MEC$_L$, when $\chi = 4$. Note: A positive number implies that the H type causes a larger externality.

Proposition 2 indicates that the L type may impose higher or lower MECs than H type, depending on the degree of ratio heterogeneity, α_H/α_L, the ratio of N_L to N_H, and power χ. This finding differs from the bottleneck model, in which the MEC caused by L type is always higher than the MEC of H type. Hence, ignoring flow congestion will overestimate the MEC imposed by L-type users. Fig. 2 illustrates the result of Proposition 2 when $\chi = 4$. When the degree of ratio heterogeneity and N_H/N_L are large, H-type users tend to have a higher MEC; otherwise, L-type users tend to have a higher MEC.

4. Two-type ratio heterogeneity in dynamic flow congestion: social optimum

Fig. 3 shows the resulting travel delays and tolls in the optimum with two-type ratio heterogeneity. We will derive it below. Seeing this figure first can help understand the mathematics that follow. We also point to four surprising features: 1) the two types travel separately over time, with those with low values traveling in the center; 2) the travel delay and toll are discontinuous at the times separating the types; 3) the social optimum cannot remove all travel delays, unlike in the bottleneck model; and 4) the slopes of the travel delay and toll are related to those in the bottleneck model, but also depend on the power χ.



(a) Travel delay (b) Optimal toll

Fig. 3. Equilibrium travel delay and toll under the social optimum with ratio heterogeneity. Note: L type users travel in the center, and H type users travel in the shoulder of the peak period.

4.1 Arrival order

We will find that under ratio heterogeneity, it is socially optimal to have users traveling separately over time. Moreover, an anonymous toll—in which at any moment *t* does not differ across types—also naturally results in the dynamic user equilibrium, in which no traveler can reduce their travel price by unilaterally altering the arrival time.

Solving $dp_i(t)/dt = 0$ yields $d\tau(t)/dt + \alpha_i \cdot dT(t)/dt = \beta$ for early arrivals, and similar (with γ) for late arrivals. Obviously, this condition cannot be true for H and L simultaneously.

This is the main difference between the bottleneck and Chu models under ratio heterogeneity with identical scheduling preferences.[9]

As travelers travel separately, two potential arrival orders may occur: one, with either H in the center and L in the shoulders or, two, the other way around. However, because of the higher VOT of H type, scheduling H type in the shoulder of the peak period will reduce the total cost. Therefore, there is a single pattern of departure of the types in the optimum. The equilibrium arrival order is summarized in the following proposition.[10]

Proposition 3. Under dynamic flow congestion with ratio heterogeneity, time-varying tolling leads different types of users to travel separately. Specifically, L-type users travel in the center of the peak period, and H-type users travel in the shoulder of the peak period. This is the same self-separation as in the un-tolled equilibrium.
Proof. See Appendix D. □

The findings for Proposition 3 differ from previous bottleneck studies, such as Arnott et al. (1994), in which all users travel jointly in the social optimum under this type of ratio heterogeneity. This difference arises because, with flow congestion, the optimum cannot eliminate travel delay.

4.2 Formulation of the social optimum problem

Given the arrival order of different types of users, we still need to pinpoint the full optimum in terms of arrival flows and windows. The social optimum problem can be formulated as a corresponding optimal control problem. Travelers of H type arrive within $[t_{sH}, t_{sL}] \cup (t_{eL}, t_{eH}]$ in the shoulders of the peak period when travel times are relatively short. Travelers of L type arrive within $(t_{sL}, t_{eL}]$ in the center of the peak when travel times are long but schedule delays low. Let $A_H(t) = \int_{t_{sH}}^{t} f_H(t)dt$ and $A_L(t) = \int_{t_{sL}}^{t} f_L(t)dt$ denote their cumulative arrivals. The regulator's optimization problem is:

[9] One could ensure joint travel using a type-specific toll. So that, say, an H-type user pays a toll of 2.00 when arriving at 8:44, and an L-type user pays 1.50 when arriving at the same time. This is very difficult to implement in practice. And, as we will see, it is not needed: it is optimal for the types to travel separately.

[10] It could also be possible that types travel separately by alternating. Say, first H-type users, then L-type users, then H again, then L, and then H, and so on. Such an outcome, however, always lowers the total costs of moving to one of the suggested potential equilibria.

$$\min_{\substack{f_H(t),f_L(t),t_{sH},\\ t_{sL},t_{eH},t_{eL}}} TC = \int_{t_{sH}}^{t^*} f_H(t)\cdot c_H(t)dt + \int_{t^*}^{t_{eH}} f_H(t)\cdot c_H(t)dt + \int_{t_{sL}}^{t_{eL}} f_L(t)\cdot c_L(t)dt \ . \tag{12}$$

subject to the equations of motion:

$$\begin{cases} \dfrac{dA_H(t)}{dt} = f_H(t) \\ \dfrac{dA_L(t)}{dt} = f_L(t) \end{cases} . \tag{13}$$

and the following constraints:

$$f_H(t) \geq 0 \tag{14}$$

$$f_L(t) \geq 0 \tag{15}$$

$$A_H(t_{sH}) = 0, A_H(t_{eH}) = N_H, A_L(t_{sL}) = 0, A_L(t_{eL}) = N_L \quad (N_H,\ N_L \text{ given}) \tag{16}$$

$$t_{sH}, t_{eH}, t_{sL}, t_{eL} \text{ chosen freely.} \tag{17}$$

Eq. (13) governs the arrivals of type *i*. Conditions (14) and (15) stipulate that the arrival rate cannot be negative. Condition (16) specifies initial and terminal values for cumulative arrivals, which means all users arrive within their travel period. Lastly, (17) establishes that the timings of the arrival windows are free and need to be determined.

4.3 Maximizing the Hamiltonian

We turn the *TC* minimization into equivalent free-end-time Hamiltonian maximization.[11] The introduction of heterogeneity adds multiple switching points to that with homogeneity (e.g., Yang and Huang, 1997; Mun, 1999, 2002). When solving the optimal solution, special transversality conditions are required at the switching points.

For ease of exposition, we set it up with the two types in separate periods, which we then show to be optimal. We define the following Hamiltonian for the two types:

$$\begin{aligned} H_H(t) &= -c_H(t)\cdot f_H(t) + \lambda_H(t)\cdot\frac{dA_H(t)}{dt}, \\ H_L(t) &= -c_L(t)\cdot f_L(t) + \lambda_L(t)\cdot\frac{dA_L(t)}{dt}. \end{aligned} \tag{18}$$

[11] We also solved (12)-(17) directly with a Lagrangian, without turning it into a Hamiltonian (see footnote 12). This gives the same outcome. However, for readers equipped with basic knowledge of dynamic optimization methods, the Hamiltonian approach is more intuitive and streamlined to follow.

Here, the cumulative number of arrivals, $A_i(t)$, is a state variable, the arrival rate, $f_i(t)$, a control variable, and $\lambda_i(t)$ a costate variable that measures the shadow cost of the control. We will see that $\lambda_i(t)$ is the marginal social cost of *i*, and that it is constant over time. Then, after adding constraints that the arrival rate is non-negative, $f_i(t) \geq 0$, we get the following Kuhn-Tucker (KT) Hamiltonian:

$$L(t) = H_H(t) + H_L(t) + \theta_H(t) f_H(t) + \theta_L(t) f_L(t), \tag{19}$$

where $\theta_i[t]$ is the shadow price of the KT constraint.

Maximizing (19) implies that the costate variable evolves according to the following equation of motion:

$$\dot{\lambda}_H = -\frac{\partial L}{\partial A_H} = 0 \qquad \text{[equation of motion for } \lambda_H\text{]}, \tag{20}$$

$$\dot{\lambda}_L = -\frac{\partial L}{\partial A_L} = 0 \qquad \text{[equation of motion for } \lambda_L\text{]}. \tag{21}$$

With $t_{sH}, t_{eH}, t_{sL}, t_{eL}$ chosen freely, and the arrival rate being discontinuous at t_{sL} and t_{eL}, the transversality conditions at these times are:

$$H_H(t_{sH}) = 0,\ H_H(t_{eH}) = 0 \quad \text{[transversality condition for } t_{sH}, t_{eH}\text{]}, \tag{22}$$

$$H_H(t_{sL}) = H_L(t_{sL}) \qquad \text{[transversality condition for } t_{sL}\text{]}, \tag{23}$$

$$H_H(t_{eL}) = H_L(t_{eL}) \qquad \text{[transversality condition for } t_{eL}\text{]}. \tag{24}$$

Condition (22) implies that the MEC is zero at the start and end of the peak period. It also means that switching the marginal user to an even earlier (later) moment does not lower costs. Finally, it dictates that the arrival rate is zero when the first and last travelers arrive. Equations (23) and (24) ensure the optimality of t_{sL} and t_{eL}. Otherwise, these times would not be optimal, and the regulator could reduce total travel costs by adjusting the schedule. We note that when optimizing directly with Lagrangians without turning (12) into Hamiltonians, the first-order conditions of t_{sL} and t_{eL} imply the same results as the last two transversality conditions.

It remains for us to determine the optimal path of $f_i(t)$. The optimal arrival rates follow:

$$\frac{\partial L}{\partial f_H(t)} = -c_H(t) - \frac{\partial c_H(t)}{\partial f_H(t)} \cdot f_H(t) + \lambda_H(t) + \theta_H(t) = 0, \tag{25}$$

$$\frac{\partial L}{\partial f_L(t)} = -c_L(t) - \frac{\partial c_L(t)}{\partial f_L(t)} \cdot f_L(t) + \lambda_L(t) + \theta_L(t) = 0, \tag{26}$$

$$\frac{\partial L}{\partial \theta_H(t)} = f_H(t) \geq 0,\ \theta_H(t) \geq 0, f_H(t) \cdot \theta_H(t) = 0, \tag{27}$$

$$\frac{\partial L}{\partial \theta_L(t)} = f_L(t) \geq 0,\ \theta_L(t) \geq 0, f_L(t) \cdot \theta_L(t) = 0. \tag{28}$$

The KT condition shows that if $f_i(t)$ is positive and finite during an open time interval containing *t*, then $\theta_i(t) = 0$. It also implies that the types must travel separately: it is impossible for these conditions to hold if f_L and f_H are positive at the same time. Conditions (20)-(28) form a multi-point optimal control problem.[12]

Lemma 3. (i) At the start and end of the peak, arrival rates are zero: $f_H(t_{sH}) = f_H(t_{eH}) = 0$.

(ii) At the separation times of the types, the arrival rate is discontinuous, jumping upward at t_{sL} and downward at t_{eL}. It follows: $\dfrac{f_H(t_{sL})}{f_L(t_{sL})} = \dfrac{f_H(t_{eL})}{f_L(t_{eL})} = \left(\dfrac{\alpha_L}{\alpha_H}\right)^{\frac{1}{\chi+1}} < 1$.

(iii) At t_{sL} and t_{eL}, $f_H(t) \cdot MEC_H(t) = f_L(t) \cdot MEC_L(t)$ holds, indicating that the 'collective' instantaneous MEC, weighted by the arrival rate, is equal for both types of users.

Proof. See Appendix E. □

Lemma 3 provides insights for travel delays. Specifically, at the beginning and end of the

[12] We could also transfer problem (12)-(17) into a two-step Lagrangian optimization. Given the unknown continuity of the travel delay, let $\left(f_L(t_{sL})/f_H(t_{sL})\right)^{\chi} = r$. From $-\beta t_{sL} = \gamma t_{eL}$ and $p_L(t_{sL}) = p_L(t_{eL})$, we can obtain $\left[f_L(t_{sL})/f_H(t_{sL})\right]^{\chi} = \left[f_L(t_{eL})/f_H(t_{eL})\right]^{\chi} = r$. Specifically, *r*=1 means the travel delay curve is continuous; otherwise, it is discontinuous at the interchange points. Then the *TC* minimization problem can be studied as a two-step optimization. In the first step, for any given *r*, travelers decide their arrival times by minimizing the travel cost. In the second step, the regulator decides *r* to minimize the total travel cost. This approach yields the same outcome as our Hamiltonian-based method.

peak period, the travel delay is zero. At the switching moments between types, there is a sudden increase in travel delay at t_{sL} and a corresponding decrease at t_{eL} to attain the social optimum.

Optimality requires the necessary condition that the sum of one's private cost and the MEC should not be lower outside one's own arriving window than within it. The intuition is that in the social optimum, a marginal shift of the interchange moment cannot lead to a decrease in the total social cost. Lemma 3(ii) can confirm its satisfaction. Moreover, Lemma 3(iii) indicates that in the social optimum, the marginal social costs of H type exceed those of L type, which is consistent with (33).

4.4 Optimal toll and equilibrium travel equilibrium

We now turn to the resulting toll and equilibrium in the social optimum. It will be shown that the toll rule is consistent with the toll expression in Chu (1995) of homogeneous users, and boils down to an intuitive dynamic generalization of the well-known Pigouvian MEC toll rule.

Proposition 4. In dynamic equilibrium, the optimal toll equals the MEC but exhibits non-continuity at the switching point. It can be expressed as:

$$\tau(t)=\begin{cases}\dfrac{\partial c_L(t)}{\partial f_L(t)}\cdot f_L(t)=\alpha_L\chi\cdot\left(\dfrac{f_L(t)}{K}\right)^{\chi}, & \mathrm{t}\in(\mathrm{t}_{sL},\mathrm{t}_{eL}]\\ \dfrac{\partial c_H(t)}{\partial f_H(t)}\cdot f_H(t)=\alpha_H\chi\left(\dfrac{f_H(t)}{K}\right)^{\chi}, & \mathrm{t}\in[\mathrm{t}_{sH},\mathrm{t}_{sL}]\cup(\mathrm{t}_{eL},\mathrm{t}_{eH}]\end{cases}. \tag{29}$$

Specifically, at t_{sL} the toll jumps downward and at t_{eL} the toll jumps upward.

Proof. See Appendix F. □

Remark. In equilibrium, (i) the toll increases at a rate of $\beta\chi/(1+\chi)$ for early arrivals and decreases at a rate of $\gamma\chi/(1+\chi)$ for late arrivals; (ii) the travel delay increases at a rate of $\beta/(\alpha_i(1+\chi))$ for early arrivals and decreases at a rate of $\gamma/(\alpha_i(1+\chi))$ for late arrivals. These rates of changes are consistent with patterns found in Chu (1995). As discussed in Proposition 3, neither type would feel tempted to enter the other type's window, given the toll in those windows would be anonymous, implying that the discontinuities in travel delay and toll do not undermine the stability of the social optimal equilibrium.

So far, we have established the toll pattern based on arrival order. The remaining question is the travel equilibrium, which includes the timing of arrivals for different types, the arrival rate, and the resulting travel price. The derivation is given in Appendix G.[13] Specifically, applying the toll in (29) yields the following arrival rate patterns:

$$f_H(t)=\begin{cases} K\cdot\left(\dfrac{\beta\cdot(t-t_{sH})}{\alpha_H\cdot(1+\chi)}\right)^{\frac{1}{\chi}}, \text{ if } \mathrm{t}\in[\mathrm{t}_{sH},\mathrm{t}_{sL}] \\ K\cdot\left(\dfrac{\gamma_H\cdot(t_{eH}-t)}{\alpha_H\cdot(1+\chi)}\right)^{\frac{1}{\chi}}, \text{ if } \mathrm{t}\in(\mathrm{t}_{eL},\mathrm{t}_{eH}] \end{cases}, \tag{30}$$

$$f_L(t)=\begin{cases} K\cdot\left(\dfrac{\beta(t-t_{sL})}{\alpha_L\cdot(1+\chi)}+\left(\dfrac{\delta N_H}{\alpha_L K\chi}\right)^{\frac{\chi}{\chi+1}}\right)^{\frac{1}{\chi}}, \text{ if } \mathrm{t}\in(\mathrm{t}_{sL},\mathrm{t}^*] \\ K\cdot\left(\dfrac{\gamma(t_{eL}-t)}{\alpha_L\cdot(1+\chi)}+\left(\dfrac{\delta N_H}{\alpha_L K\chi}\right)^{\frac{\chi}{\chi+1}}\right)^{\frac{1}{\chi}}, \text{ if } \mathrm{t}\in(\mathrm{t}^*,\mathrm{t}_{eL}] \end{cases}. \tag{31}$$

Equations (30) and (31) show that the arrival rates display nonlinearity. We can get the timing of the peak by combining the above with the transversality conditions and that everyone should travel. In contrast to the bottleneck model, tolling shifts the earliest departure to earlier and the latest departure later. These timings correspond to the following prices:

$$\begin{cases} p_H=\alpha_H\left(1-\left(\dfrac{\alpha_L}{\alpha_H}\right)^{\frac{1}{\chi+1}}\right)(1+\chi)\left(\dfrac{\delta N_H}{\alpha_H K\chi}\right)^{\frac{\chi}{\chi+1}}+\alpha_L\cdot(1+\chi)\left(\dfrac{\delta N_H}{\alpha_L K\chi}+\dfrac{\delta N_L}{\alpha_L\chi K}\right)^{\frac{\chi}{\chi+1}} \\ p_L=\alpha_L\cdot(1+\chi)\cdot\left(\dfrac{\delta N_H}{\alpha_L K\chi}+\dfrac{\delta N_L}{\alpha_L\chi K}\right)^{\frac{\chi}{\chi+1}} \end{cases}. \tag{32}$$

Taking the derivative of (32) with respect to the number of different types of users yields $\left(\partial p_L/\partial N_H\right)/\left(\partial p_L/\partial N_L\right)=1$ and $\left(\partial p_H/\partial N_H\right)/\left(\partial p_H/\partial N_L\right)>1$. All users exert the same price effects on L-type users, since L type arrives in the center of the peak period and the toll varies at the same slope. However, an extra L-type user has a smaller impact on type H's travel price than an extra type-H user. The difference in marginal price effects is caused by the jump in the toll at

[13] Mangasarian (1966) asserts that necessary conditions are also sufficient when the Hamiltonian is concave in state and control variables, given the satisfaction of an appropriate constraint qualification. It can be verified that the derived optimum minimizes the total cost globally. The proof is available upon request.

the switching moments between types.

Remark. Taking the difference between p_H and p_L yields:

$$\Delta p = p_H - p_L = \alpha_H \left(1-\left(\frac{\alpha_L}{\alpha_H}\right)^{\frac{1}{\chi+1}}\right)(1+\chi)\left(\frac{\delta N_H}{\alpha_H K\chi}\right)^{\frac{\chi}{\chi+1}} > 0, \tag{33}$$

implying that H-type users experience a higher travel price than L-type users, due to the peak spreading. Specifically, when $\alpha_H = \alpha_L$, the price in (32) converges to the price found in Chu (1995). Note that in the bottleneck model different user types incur an equal travel price.

Proposition 5. (Distributional effects) Under dynamic flow congestion with ratio heterogeneity, both types of users lose from tolling.
Proof. See Appendix H. □

Compared to the bottleneck model, where H type users are not affected, and L type lose from tolling (Van den Berg and Verhoef, 2011a), Proposition 5 shows that in the presence of dynamic flow congestion proposed by Chu (1995), all users lose from tolling, due to the fact that the travel delay cannot be fully eliminated by tolling, and the peak period is widened. We note that as the BPR power χ approaches infinity, the congestion function becomes increasingly similar to that of the bottleneck model, and we indeed find H type users will remain unaffected by tolling. Consequently, ignoring flow congestion would overlook the loss of H type users.

5. Two-type proportional heterogeneity in dynamic flow congestion

The second type of heterogeneity we explore is proportional heterogeneity, as introduced by Vickrey (1973). In this context, all three values of time and schedule delay vary proportionally: $\alpha_i = \mu \cdot \beta_i$ and $\gamma_i = \eta \cdot \beta_i$, where μ and η are homogeneous ratios.[14] This type of heterogeneity may reflect income differences, whereby a higher income would reduce the marginal utility of income and thus increase all values proportionally. In our analysis, we focus on a two-type case in which H type has higher values, but both H and L types share the same ratio of VOT to values of schedule delay. The following subsections investigate the no-tolling and

[14] Consequently, $\alpha_H/\alpha_L = \beta_H/\beta_L = \gamma_H/\gamma_L$ holds under proportional heterogeneity.

social optimum, respectively.

5.1 Two-type proportional heterogeneity in dynamic flow congestion: no tolling

We start with no-toll equilibrium. Since the ratios β_i/α_i and γ_i/α_i are identical across types, according to (6), travel delay follows the same triangular pattern. As illustrated in Fig. 4, the iso-cost curves for the two types overlap, indicating that they travel jointly, not separately, in time.

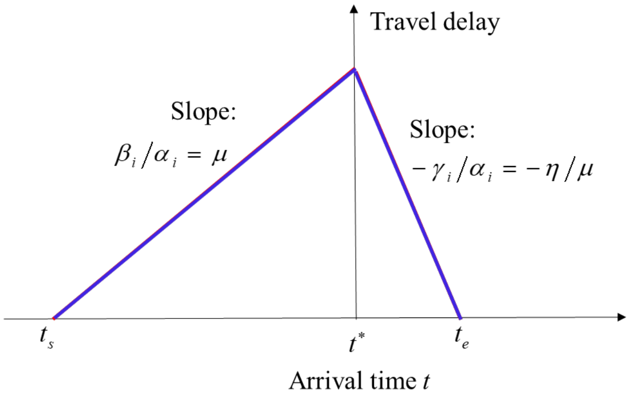


Fig. 4. Travel delay under proportional heterogeneity without tolling.
Note: All users travel jointly.

As a result of the joint travel of the two different types, the travel cost can be expressed as:

$$c_i(t) = \alpha_i \cdot \left(\frac{f_H(t)+f_L(t)}{K}\right)^{\chi} + \begin{cases} \beta_i \cdot (t^* - t) & \text{i=H, L; } t \le t^* \\ \gamma_i \cdot (t - t^*) & \text{i=H, L; } t>t^* \end{cases}. \tag{34}$$

At equilibrium, travelers of the same type incur equal travel costs regardless of their arrival times. With the first and last travelers encountering no travel delay, we can determine the start and end times of travel,[15] yielding the equilibrium travel cost as:

$$c_H = \alpha_H \left(\frac{\delta_H(N_H+N_L)}{\alpha_H K} \cdot \frac{1+\chi}{\chi}\right)^{\frac{\chi}{\chi+1}}, \; c_L = \alpha_L \left(\frac{\delta_H(N_H+N_L)}{\alpha_H K} \cdot \frac{1+\chi}{\chi}\right)^{\frac{\chi}{\chi+1}}. \tag{35}$$

These costs indeed, again, reduce to those reported by Chu when we make the two groups equal. Notably, H type incurs a higher travel cost than L type, due to the larger VOT. Specifically, with a constant mean value for all users' VOT, this difference increases with the degree of proportional

[15] By solving the equilibrium conditions, the start and end times of the peak period can be derived as:

$$t_{sH} = t_{sL} = -\frac{\alpha_H}{\beta_H}\left(\frac{\delta_H(N_H+N_L)}{\alpha_H K} \cdot \frac{1+\chi}{\chi}\right)^{\frac{\chi}{\chi+1}}, \quad t_{eH} = t_{eL} = \frac{\alpha_H}{\gamma_H}\left(\frac{\delta_H(N_H+N_L)}{\alpha_H K} \cdot \frac{1+\chi}{\chi}\right)^{\frac{\chi}{\chi+1}}.$$

heterogeneity, α_H/α_L.

The resulting total travel cost is:

$$TC = \left(\alpha_H N_H + \alpha_L N_L\right) \cdot \left(\frac{\delta_H (N_H + N_L)}{\alpha_H K} \cdot \frac{1+\chi}{\chi} \right)^{\frac{\chi}{\chi+1}}. \tag{36}$$

With a fixed mean VOT, the degree of proportional heterogeneity has no impact on total travel cost. All users travel jointly, experiencing no gains or losses from heterogeneity.[16]

By taking the derivate of the travel cost with respect to the number of travelers, we establish the relationship of the congestion externality in the absence of time-varying tolling:

$$\frac{\partial c_L}{\partial N_H} = \frac{\partial c_L}{\partial N_L} = \frac{\alpha_L}{\alpha_H} \frac{\partial c_H}{\partial N_H} = \frac{\alpha_L}{\alpha_H} \frac{\partial c_H}{\partial N_L} \text{ with } \frac{\partial c_H}{\partial N_H} = \frac{\delta_H}{K} \left(\frac{\delta_H (N_H + N_L)}{\alpha_H K} \cdot \frac{1+\chi}{\chi} \right)^{-\frac{1}{\chi+1}}. \tag{37}$$

implying that under proportional heterogeneity, H type causes a smaller congestion externality for L type than for themselves, with a ratio of α_L/α_H. The single reason for the difference is that L-type users value additional time losses at a lower α_L. In contrast, L type causes less congestion externality for themselves than for H type, with the same ratio of α_L/α_H. This results in an equal MEC of both types:

$$MEC_H = MEC_L = \frac{\delta_H}{K} \cdot \left(N_H + \frac{\alpha_L}{\alpha_H} N_L \right) \cdot \left(\frac{\delta_H (N_H + N_L)}{\alpha_H K} \cdot \frac{1+\chi}{\chi} \right)^{-\frac{1}{\chi+1}}. \tag{38}$$

Eq. (38) implies that under proportional heterogeneity, as all users travel jointly, they have the same MECs.

Proposition 6. Under proportional heterogeneity in the absence of tolling, $MEC_H = MEC_L$ always holds, with the MEC depending on the degree of proportional heterogeneity, the number of travelers of each type, and the elasticity of travel delay with respect to travel flow.

5.2 Two-type proportional heterogeneity in dynamic flow congestion: social optimum

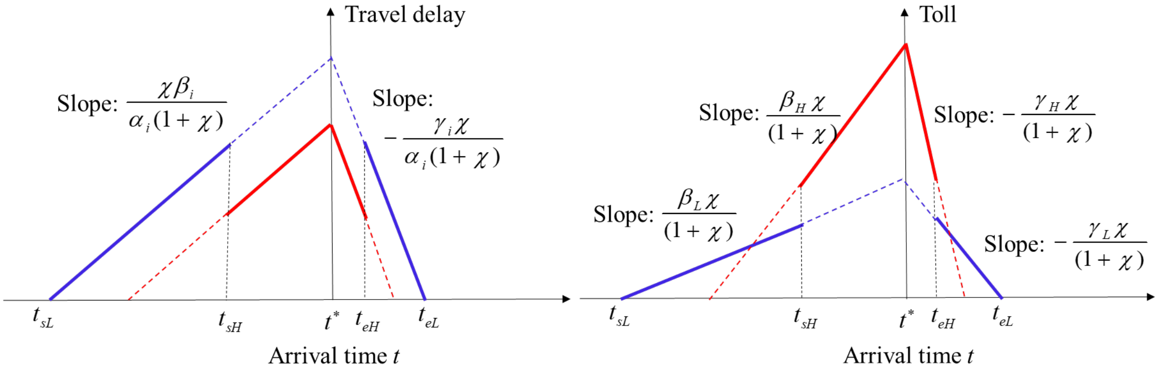
Similar to the case of ratio heterogeneity, travel delay will not be eliminated by an optimal time-varying toll under dynamic flow congestion. In the dynamic equilibrium, users of the same

[16] Substituting (35) into the travel cost function of (34), we can derive the joint arrival rate of different types Nonetheless, due to the joint travel, the arrival rate for a certain type cannot be distinguished without further assumptions.

type should have identical prices. Solving $dp_i(t)/dt = 0$ yields that for early arrivals $d\tau(t)/dt + \alpha_i \cdot dT(t)/dt = \beta_i$ holds, and for late arrivals $d\tau(t)/dt + \alpha_i \cdot dT(t)/dt = -\gamma_i$ holds. Different types of users cannot travel jointly if the toll varies over time.

Solving the corresponding optimal control problem of total cost minimization, we find that the toll rule under ratio heterogeneity still applies. The time-varying toll equals the MEC caused by different types of users. Again, due to the non-continuity of the arrival rate, different types of users impose a different MEC at the switching time. Solving the associated equilibrium conditions, we can derive the travel equilibrium in the social optimum. Detailed derivations are shown in Appendix I. Here we will concentrate on presenting the main insights.

In the social optimum, exchanging travel delays for tolls involves greater amounts per user closer to t^*, which is particularly attractive for high-VOT travelers. Fig. 5 illustrates the equilibrium in which the travel delay slope remains consistent for both user types. However, the toll for H type changes more rapidly than for L type, due to the larger value of schedule delay. Notably, in contrast to ratio heterogeneity, the travel delay curve now features a downward jump, whereas the toll curve exhibits an upward jump at the switching point for early arrivals, and vice versa for late arrivals. The properties of the arrival rate, travel delay, and toll are summarized in Proposition 7.



(a) Travel delay (b) Optimal toll

Fig. 5. Travel time and time-varying tolling under proportional heterogeneity.

Note: H type users travel in the center and L type users travel in the shoulder of the peak period.

Proposition 7. In the social optimum with proportional heterogeneity: (i) travelers with a higher VOT travel in the center of the peak period, and travelers with a lower VOT travel in the shoulder;

(ii) at the switching moments between types, there is a sudden decrease in travel delay at t_{sH} and a corresponding increase at t_{eH} to attain the social optimum; and (iii) there is a sudden increase in toll at t_{sH} and a corresponding decrease at t_{eH}.

Proof. See Appendix I. □

Specifically, the equilibrium travel price is:

$$\begin{cases} p_L = \alpha_L \cdot (1+\chi)\left[\left(\dfrac{\delta_H N_H}{\alpha_H \chi K} + \dfrac{\delta_L N_L}{\alpha_H \chi K}\right)^{\frac{\chi}{1+\chi}} + (\dfrac{\delta_L N_L}{\alpha_L \chi K})^{\frac{\chi}{1+\chi}} - (\dfrac{\delta_L N_L}{\alpha_H \chi K})^{\frac{\chi}{1+\chi}}\right] \\ p_H = \alpha_H \cdot (1+\chi) \cdot \left(\dfrac{\delta_H N_H}{\alpha_H \chi K} + \dfrac{\delta_L N_L}{\alpha_H \chi K}\right)^{\frac{\chi}{1+\chi}} \end{cases}. \tag{39}$$

As the H type travel in the center of the peak period, incurring lower travel delays with a higher toll, they experience a higher travel price than L type.[17] Taking the derivative of (39) with respect to the number of each type of users yields the ratio of marginal price effects being $(\frac{\partial p_H}{\partial N_H}) \Big/ (\frac{\partial p_H}{\partial N_L}) = \delta_H / \delta_L > 1$ and $(\frac{\partial p_L}{\partial N_H}) \Big/ (\frac{\partial p_L}{\partial N_L}) > 1$.[18] Consequently, an additional H-type user adds more to both types' travel price than an additional L-type user. This happens because H type users have a larger slope in the optimal toll schedule.

Comparing (39) and (35), we find that under proportional heterogeneity, users may gain or lose from tolling. The specific results are summarized in the following proposition.

Proposition 8. (**Distributional effects**) Under proportional heterogeneity, H-type users may gain or lose from tolling, depending on the degree of proportional heterogeneity, the distribution of user types, and the elasticity of travel delay concerning arrival flow. Specifically, when $(1+\chi)^{\frac{1}{\chi+1}} > \left(\dfrac{\delta_H N_H + \delta_H N_L}{\delta_H N_H + \delta_L N_L}\right)^{\frac{\chi}{\chi+1}}$ is satisfied, H-type users lose from tolling; otherwise, they gain. Conversely, L-type users always lose from tolling.

Proof. See Appendix J.

[17] Taking the difference between p_L and p_H in (39) yields:

$$p_L - p_H = (\alpha_L - \alpha_H) \cdot (1+\chi)\left(\frac{\delta_H N_H}{\alpha_H \chi K} + \frac{\delta_L N_L}{\alpha_H \chi K}\right)^{\frac{\chi}{1+\chi}} + \alpha_L \cdot (1+\chi)\left((\frac{\delta_L N_L}{\alpha_L \chi K})^{\frac{\chi}{1+\chi}} - (\frac{\delta_L N_L}{\alpha_H \chi K})^{\frac{\chi}{1+\chi}} \right) < 0.$$

[18] $\partial p_L \big/ \partial N_H = \partial p_L \big/ \partial N_L + \left(\dfrac{\delta_L N_L}{\alpha_L \chi K}\right)^{\frac{\chi}{1+\chi}-1} \dfrac{\delta_L}{K} - \left(\dfrac{\delta_L N_L}{\alpha_H \chi K}\right)^{\frac{\chi}{1+\chi}-1} \dfrac{\delta_L}{K} \dfrac{\alpha_L}{\alpha_H} > \partial p_L \big/ \partial N_L$.

Proposition 8 indicates that under proportional heterogeneity, the distributional effect on H type is ambiguous. In fact, the outcome depends on the balance between the benefits from the self-separation and the loss from the toll: if the former effect dominates, H type users benefit from tolling; otherwise, they lose. Fig. J1 shows that $p_H - c_H$ increases with N_H / N_L and decreases with α_H / α_L. Specifically, $p_H > c_H$ is more likely to happen when N_H / N_L is relatively large, and $p_H < c_H$ is more likely to happen when N_H / N_L is relatively small. As for the L type, they always lose from tolling due to the peak widening. Again, it should be noted that, when the BPR power goes to infinity, H type benefit from tolling and L type are unaffected, which is consistent with the distributional effects found with the bottleneck model (e.g., Vickrey, 1973; Van den berg and Verhoef, 2011a). Therefore, ignoring the dynamic flow congestion will underestimate the adverse effects of pricing on travelers.

6. Comparison with the bottleneck model

It is insightful at this point to highlight some differences between the present flow-based model and the bottleneck model.

(i) Unlike the fixed arrival rate at capacity that applies to the bottleneck model, flow congestion has a varying arrival rate, which is continuous without tolling but has discrete discontinuities in the social optimum. Here, the intuition is that the equality of marginal social cost that characterizes the social optimum typically implies a discontinuity in travel delay and hence in flow at the moment that the groups switch. Specifically, in contrast to the bottleneck model, in which tolling eliminates travel delays, flow congestion presents discontinuities in both the travel delay and toll.

(ii) For ratio heterogeneity with identical scheduling preferences, the arrival order in the social optimum differs between the two models. In the bottleneck model, types travel jointly. Under flow congestion, different types travel separately: users with a lower VOT travel in the center, whereas users with a higher VOT travel in the shoulder of the peak period. The intuition is that the lower VOT makes the remaining time losses in the central peak less harmful for the lower-VOT travelers.

(iii) In the un-tolled equilibrium, the congestion effects differ from those in the bottleneck model. In particular, under ratio heterogeneity, the bottleneck model shows that L-type users always impose a lower congestion externality than H-type users (e.g., Van den Berg, 2011b). However, in the flow congestion setting, L-type users may impose a higher or lower congestion externality than H-type, which depends on the degree of ratio heterogeneity and the number of each type of users.

(iv) The distributional effects of tolling also differ significantly. Prior bottleneck studies indicated that under ratio heterogeneity, L-type users lose from tolling and H-type users are unaffected (e.g., Arnott et al., 1994; de Palma and Lindsey, 2002); under proportional heterogeneity, tolling benefits H-type users, whereas L-type users are unaffected (e.g., Vickrey, 1973). In contrast, we find when dynamic flow congestion is considered, under ratio heterogeneity, both types of users lose from tolling, whereas under proportional heterogeneity, L type always loses, and H type may benefit or lose from tolling, depending on the elasticity of travel delay with respect to arrival rate, the number of each type of users, and the preference parameters.

7. Numerical model

7.1 Base-case calibration

Here, we try to keep the result comparable with previous works (Van den Berg and Verhoef, 2011a, b). We consider 9000 users with a free-flow travel time of 30 min. The capacity, K, is such that the un-tolled travel costs are the same as in a bottleneck model with a capacity of 3600. Following Chu (1995), the elasticity of travel delay with respect to arrival rate is $\chi = 4.08$. The number of H- and L-type users is 4500 each.

The average VOT is €10.00/h, which is close to the official Dutch average (Kouwenhoven et al., 2014; Knoope, 2023). The values of schedule delay follow from their ratios to the VOT in Small (1982) and Arnott et al. (1993). Therefore, under ratio heterogeneity, we use α_L = €6.5/h, α_H = €13.5/h, β = €6.09/h, and γ = €23.76/h. The calibration targets discussed result in a choice of $K = 4141$.

Under proportional heterogeneity, to avoid too small a value of schedule delay early, we use α_L = €9.48/h, α_H = €10.52/h, β_L = €4.73/h, β_H = €5.26/h, γ_L = €18.95/h, and γ_H = €21.05/h. Again, for ease of comparison, we keep the un-tolled travel cost for the L-type the same as under bottleneck congestion, which now results in $K = 4391$.[19]

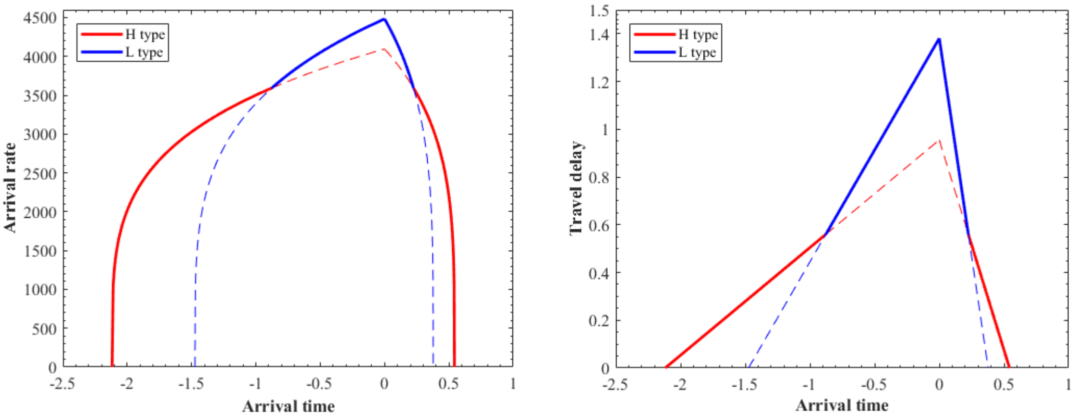
[19] With ratio heterogeneity, we have $K = \frac{\chi+1}{\chi} \cdot \left(\frac{\delta N_H}{\alpha_H} + \frac{\delta N_L}{\alpha_L} \right) \cdot \left(\delta_L \cdot (N_L + \frac{\alpha_L}{\alpha_H} \cdot N_H) \Big/ (\alpha_L s) \right)^{-\frac{\chi+1}{\chi}} = 4141$. With proportional heterogeneity, we have $K = \frac{\delta_H N}{\alpha_H} \cdot \frac{\chi+1}{\chi} \left(\frac{s}{N} \frac{\beta_L \alpha_H}{\delta_L \beta_H} \right)^{\frac{\chi+1}{\chi}} = 4391$.

7.2 Base-case numerical model under ratio heterogeneity

Fig. 6 depicts the arrival rate and travel delay in the no-toll equilibrium. L type travels in the center of the peak period, and H type in the shoulder. The arrival rate displays continuous nonlinearity. Given the power of the BPR function, this is needed to achieve linearity of travel delays, as shown in Fig. 6(b). The lower VOT for L-type users results in faster changes in travel delay than for H-type users.

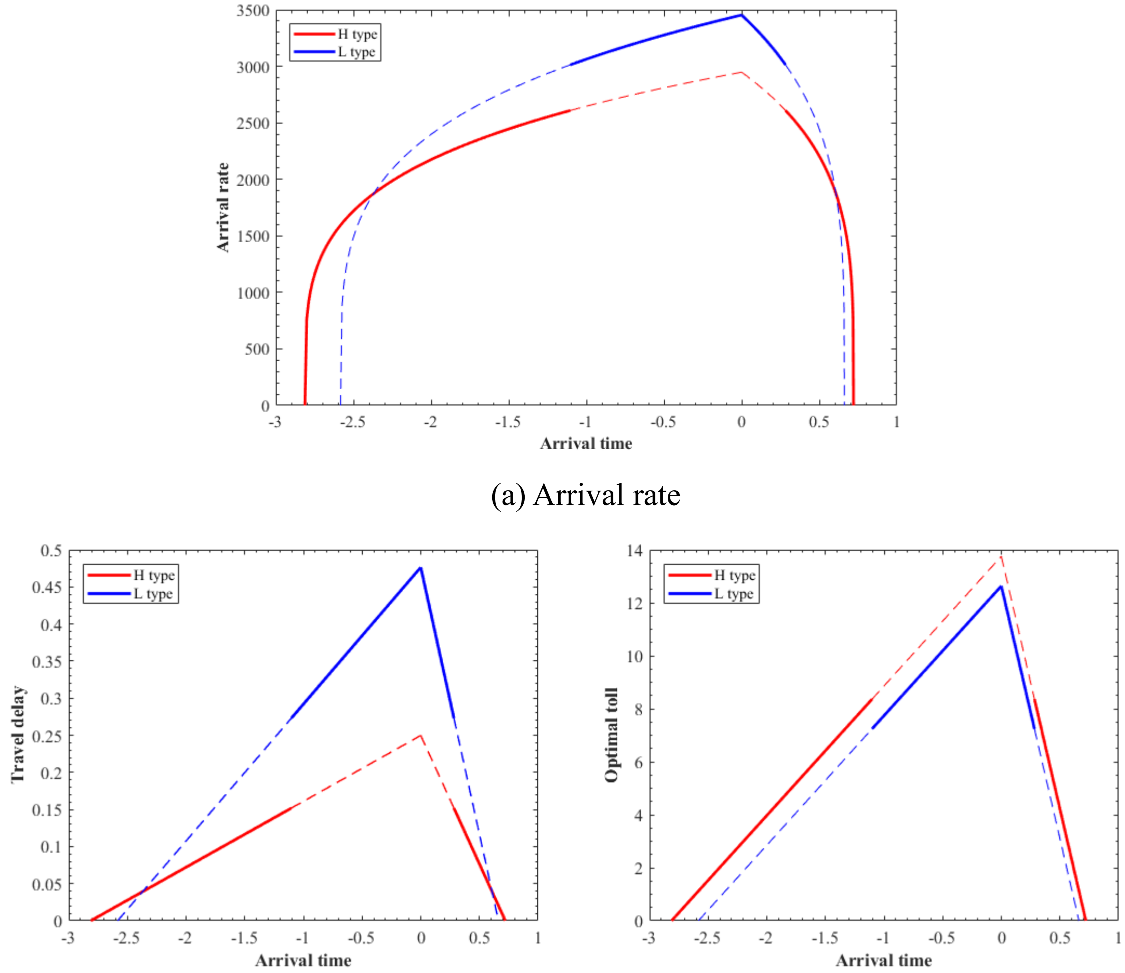
In the social optimum, the arrival rate, travel delay, and toll all show discontinuities, as presented in Fig. 7. Fig. 7(a) illustrates an upward jump in arrival rate at time -1.01 and a downward jump at time 0.28. Fig. 7(b) shows that the travel delay for L-type users changes more quickly than that of H-type users, with an upward jump at time -1.10 and a downward jump at time 0.28. Fig. 7(c) reveals a consistent slope of the toll for both user types, jumping downward at time -1.10 and upward at time 0.28.

The detailed outcomes for the ratio heterogeneity case are given in Table 1. For presentation purpose, 'H' represents H type users and 'L' represents L type users. Some interesting observations emerge from these results. (i) With bottleneck congestion, optimal tolling keeps the peak duration unchanged. Conversely, with flow congestion, it induces peak widening. (ii) With bottleneck congestion, H-type drivers are unaffected and L-type drivers lose from optimal tolling. In contrast, with flow congestion, all users lose. (iii) With bottleneck congestion, optimal tolling eliminates 100% of travel delay costs, reducing H-type schedule delay costs by 33.3% and increasing L-type schedule delay costs by 100%. With our flow congestion, optimal tolling reduces travel delay costs by 72.8% for the H-type and 61.7% for the L-type; while it increases schedule delay costs by 30.5% for the H-type and 27.6% for the L-type, both due to peak widening.



(a) Arrival rate (b) Travel delay

Fig. 6. Un-tolled equilibrium patterns with ratio heterogeneity.



(a) Arrival rate

(b) Travel delay

(c) Optimal toll

Fig. 7. Equilibrium patterns under social optimum with ratio heterogeneity.

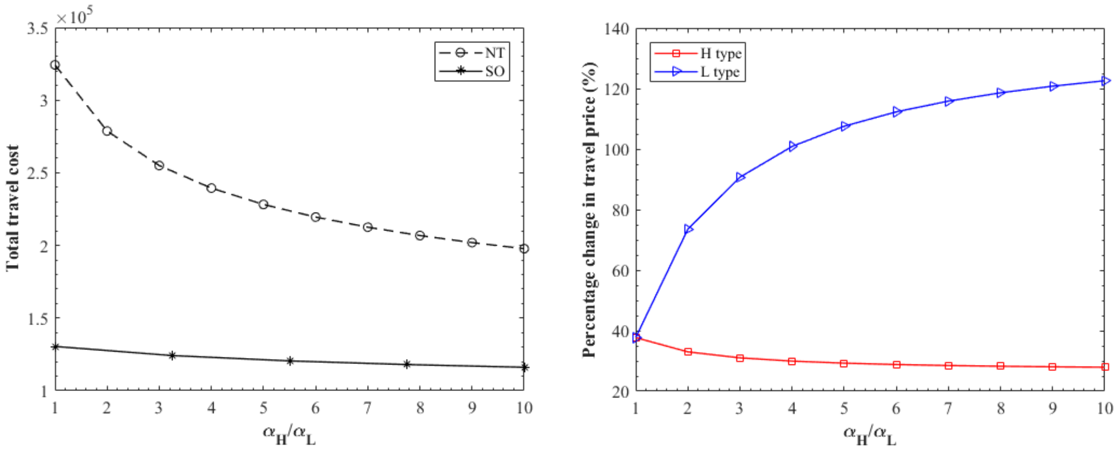
Table 1. Outcomes of ratio heterogeneity under base-case calibration

	Bottleneck model		Flow congestion model	
	No toll	Social optimum	No toll	Social optimum
Arrival interval H	[-1.99,-0.10], [0.26,0.51]	[-1.99,0.51]	[-2.12,0.88], [0.22,0.54]	[-2.81,-1.10], [0.28,0.72]
Arrival interval L	[-0.10,0.26]	[-1.99,0.51]	[-0.88,0.22]	[-1.10,0.72]
Private travel cost H	18.87	12.81	19.65	19.23
Private travel cost L	12.23	9.31	12.23	8.98
Travel price H	18.87	18.87	19.65	23.88
Travel price L	12.23	15.37	12.23	18.98
Travel delay cost H	3.03	0	4.19	1.14
Travel delay cost L	5.95	0	6.40	2.45
Schedule delay cost H	9.09	6.06	8.70	11.35
Schedule delay cost L	3.03	6.06	2.57	3.28
Total travel cost	139,950	99,540	143,432	126,998

7.3 Varying the degree of ratio heterogeneity

We vary the degree of ratio heterogeneity, α_H/α_L, by maintaining the mean VOT at 10, and keeping the values of schedule delay unchanged.[20] Fig. 8(a) shows the impact on total travel cost. Increasing ratio heterogeneity reduces total travel cost, both with and without tolling, reflecting the benefit of travelers' self-ordering. With higher ratio heterogeneity, the difference between no tolling and tolling narrows due to the faster reduction in travel delay in the absence of tolling compared to the social optimum.[21]

Fig. 8(b) depicts the distributional effects of tolling by measuring the percentage change in generalized prices. Positive values indicate increased prices caused by tolling. Both user types lose from tolling, L type facing the greater impact. As ratio heterogeneity increases, the relative price rise from tolling increases for L type and decreases for H type. A more detailed decomposition of the equilibrium travel price is provided in Appendix K.



(a) Total travel cost (b) Distributional effects

Fig. 8. Effects of ratio heterogeneity.

Note: Distributional effects are measured by $(p_i^{SO} - c_i^{NT})/c_i^{NT} \cdot 100$.

7.4 Varying the elasticity of travel delay with respect to arrival rate

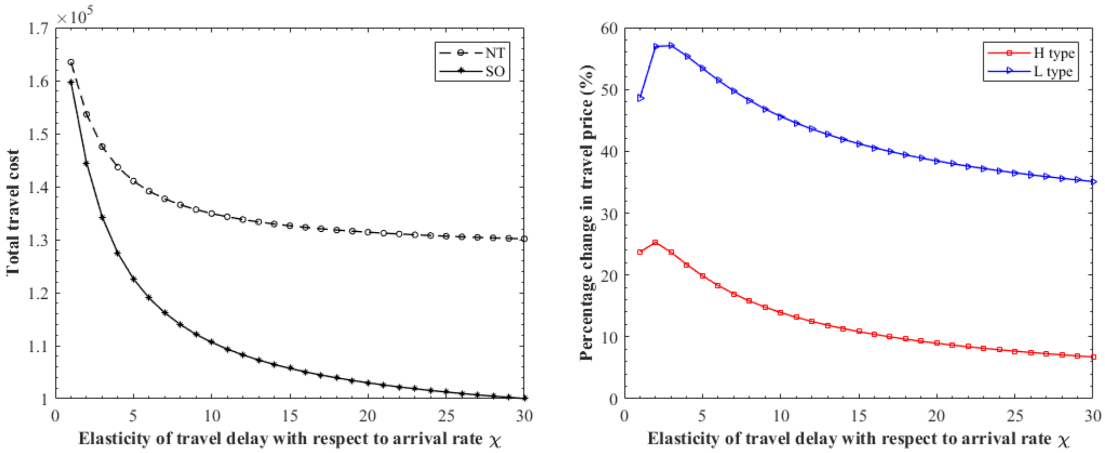
Fig. 9 depicts the effects of varying χ, which gives the elasticity of travel delay with respect to arrival rate. Fig. 9(a) shows that as χ increases, total travel cost decreases, especially with tolling. Indeed, the shift in travel interval reduces the travel delay cost, causing the schedule delay cost to rise initially and then fall. Overall, the decrease in travel delay cost dominates, leading to

[20] This average value of time is close to the official Dutch average (Kouwenhoven et al., 2014).

[21] In the absence of tolling, the travel delay varies with a rate of β/α_i (or γ/α_i), whereas in the social optimum the travel delay varies with a rate of $\beta/(\alpha_i\,(1+\chi))$ for early arrivals and $\gamma/(\alpha_i\,(1+\chi))$ for late arrivals.

a reduction in total travel cost with an increase in χ. In the social optimum, optimal tolling results in a greater decrease in total travel cost compared to no tolling, and the gap widens with χ, reflected in the increasing toll schedule. This matches the intuitive notion that with a higher χ, and thus a more strongly curved travel delay function, fewer drivers would have to move away from the busiest moments to obtain significant gains in travel time reductions.

As for the distributional effects, Fig. 9(b) shows that as χ increases, the percentage price increase from tolling first rises and then falls. This implies that for a relatively small χ, the travel price without tolling decreases more rapidly due to the dominating effect of peak widening caused by tolling. However, for χ exceeding 3, the travel price with tolling decreases more quickly, as the flatter travel delay induced by the increasing χ dominates. Specifically, when varying χ from 0 to 1000, we can observe that as χ approaches infinity, the distributional effects converge to the results in the bottleneck model: H type users are unaffected and L type users lose from tolling.



(a) Total travel cost

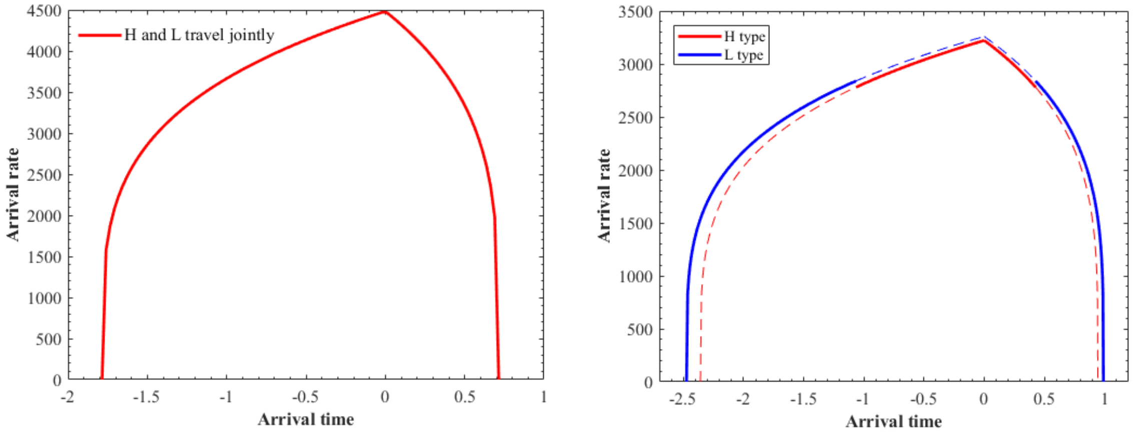
(b) Distributional effects

Fig. 9. Effects of elasticity of travel delay with respect to arrival rate χ.

7.5 Numerical analysis under proportional heterogeneity

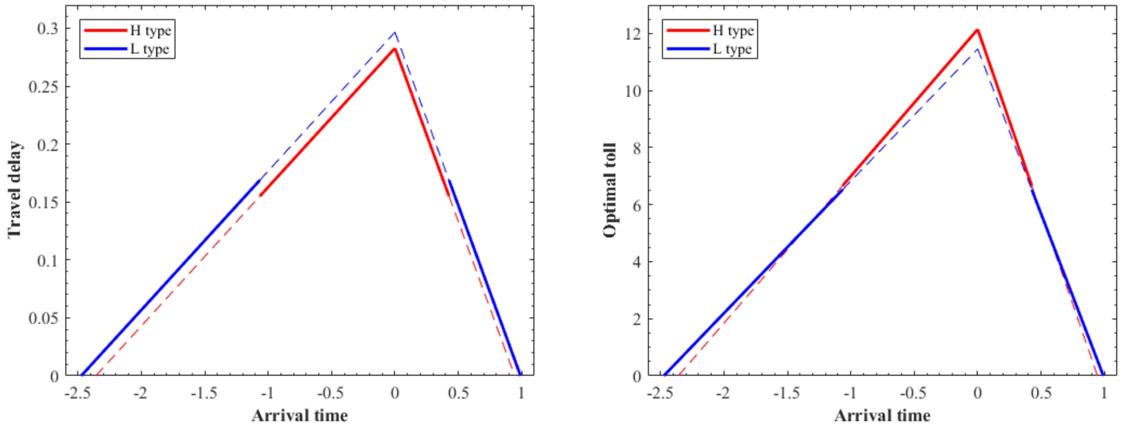
We now turn to proportional heterogeneity and first look at the results in the base case. Fig. 10(a) depicts the joint arrival rate without tolling, displaying continuous nonlinear patterns starting from a rate of 0 at time -1.79, peaking at 4470 at time 0, and reaching 0 again at 0.71. Fig. 10(b) depicts the arrival rate in the social optimum. Now the rate is discontinuous and nonlinear. Notably, there is a downward jump at time -1.07 and an upward jump at time 0.43.

Fig. 11 depicts the travel delay and toll in the social optimum. The identical linear travel delay slope for both types (Fig. 11(a)) is due to the equal ratio of the VOT to the value of schedule delay. Our flow congestion set-up causes an upward jump in delays at time -0.17 and a downward jump at time 0.43. In the toll schedule (Fig. 11(b)), the toll for H-type users changes more rapidly than for L-type users due to the higher value of schedule delay. Although difficult to see, the toll jumps upward at switching time -1.07 and drops at time 0.43.



(a) Without tolling (b) Social optimum

Fig. 10. Arrival rate under proportional heterogeneity.



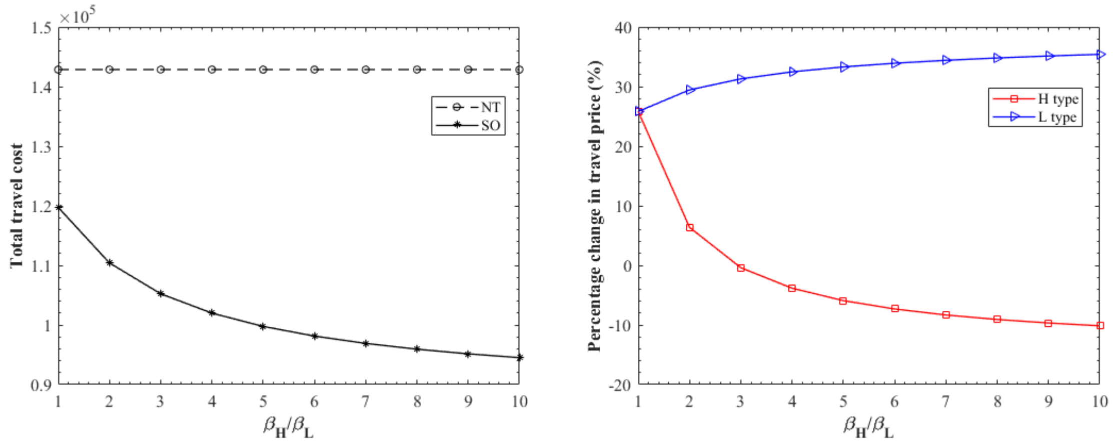
(a) Travel delay (b) Optimal toll

Fig. 11. Travel delay and toll under social optimum with proportional heterogeneity.

Fig. 12 varies the degree of proportional heterogeneity by keeping the average VOT at 10, and the ratios β_i/α_i and γ_i/α_i identical across types.[22] Fig. 12(a) shows that in the social optimum, total cost decreases with the degree of proportional heterogeneity, reflecting the benefits of diverging preferences for reducing congestion. Without tolling, proportional heterogeneity has no impact on total cost, as all users travel jointly. In Fig. 12(b), L-type users always lose from tolling, with more intensified losses as proportional heterogeneity increases.

[22] The degree of proportional heterogeneity is measured by β_H/β_L, which indeed equals α_H/α_L.

Conversely, for H-type users, the loss from tolling decreases with proportional heterogeneity. Notably, when the proportional heterogeneity exceeds a threshold of around 3, tolling even becomes beneficial for H-type users.

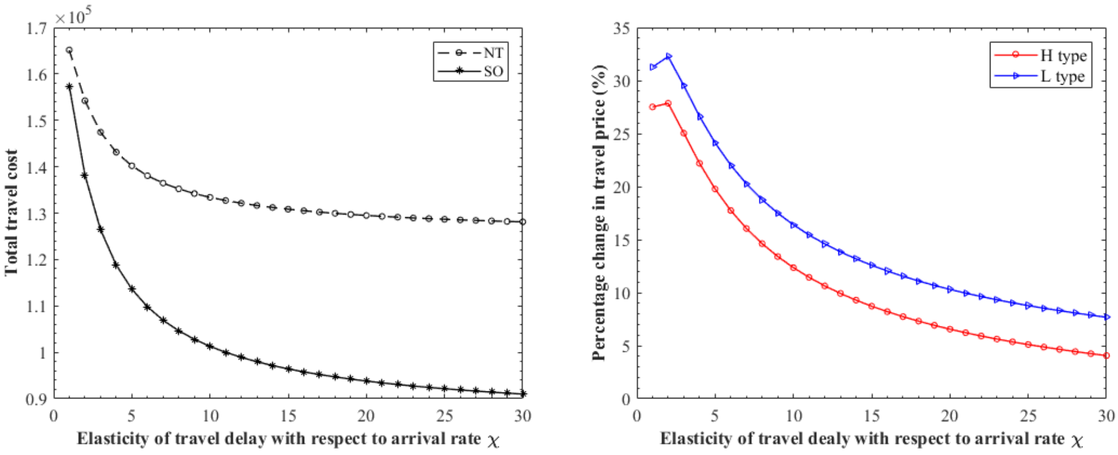


(a) Total travel cost (b) Distributional effects

Fig. 12. Effects of proportional heterogeneity.

Note: Price effects are measured by $(p_i^{SO} - c_i^{NT})/c_i^{NT} \cdot 100$.

Fig. 13 varies the elasticity of travel delay with respect to arrival rate χ. With increased elasticity, total travel cost decreases, especially with tolling. With regard to distributional effects, the percentage change in travel price first increases, and then decreases. Indeed, although travel price decreases with elasticity, the reduction is more significant without tolling when χ is small. As χ increases, the decrease with tolling exceeds that without tolling. Again, when varying χ from 0 to 1000, we can observe that as χ approaches infinity, the distributional effects converge to the results in the bottleneck model: H type users benefit from tolling and L type users are unaffected.



(a) Total travel cost (b) Distributional effect

Fig. 13. Effects of BPR power χ under proportional heterogeneity.

8. Discussion of other forms of heterogeneity

After examining ratio and proportional heterogeneity separately, deriving equilibrium when both types coexist is relatively straightforward. In the un-tolled equilibrium, the arrival order aligns with that in bottleneck congestion: travelers with a higher β_i/α_i and γ_i/α_i arrive in the center of the peak period (e.g., Arnott et al., 1988, 1994). In the social optimum, unlike under bottleneck congestion, dynamic flow congestion could result in two possible optima: i) travelers with a higher β_i/α_i (or γ_i/α_i for late arrivals) travel in the center of the peak period (as with ratio heterogeneity); and (ii) travelers with a higher β_i (or γ_i for late arrivals) travel in the center (as with proportional heterogeneity). The global optimum depends on the specific parameters.

More general heterogeneity in multiple dimensions may be a combination of the results under the various types of separate heterogeneities mentioned above, if t^* is homogeneous. But there may exist several possible equilibria. If t_i^* is heterogeneous, the equilibrium will become much more complicated—as was shown for the Chu model by Verhoef (2020) and the bottleneck model by Hall (2021, 2023). It seems interesting to investigate this in a future study.

When adding more and more types of users for the same form of heterogeneity, it is to be expected that discontinuities in travel delays and tolls will decrease, and that they disappear for continuous forms of heterogeneity. Even though we report them as features of the first-best optimum, we may not believe that, in reality, these discontinuities would be significant. However, the effects of heterogeneity on the overall and distributional effects of tolling will probably remain qualitatively similar.

9. Conclusion

This paper investigated how preference heterogeneity affects travel behavior and congestion pricing in a dynamic flow-congestion framework. We extended the Chu (1995) model, in which speed is assumed constant during the trip and a function alone of the flow at the road's exit. The properties of travel equilibrium, congestion externality, and congestion pricing were explored analytically and compared under different types of heterogeneity. The welfare and distributional effects of tolling were examined explicitly.

Arrival rates, travel delays, and toll patterns differ significantly from those in the bottleneck model. In the absence of tolling, the arrival rate varies over time, unlike the constant arrival rate in a bottleneck. In the social optimum of our discrete groups model, preference heterogeneity introduces discontinuities in the arrival rate, travel delay, and toll. The slope of the travel delay and the toll depends not only on preference parameters, but also on the elasticity of travel delay with respect to the arrival rate.

In most instances, heterogeneous travelers have the same arrival order as in the bottleneck model, except for the social optimum under ratio heterogeneity. We showed that users with a lower VOT travel in the center and users with a higher VOT travel in the shoulder of the peak period. This contrasts sharply with the bottleneck model, in which travelers arrive in a pooled equilibrium.

Under ratio heterogeneity, users with a higher VOT may impose a greater or lower congestion externality than users with a lower VOT, depending on the degree of ratio heterogeneity and the numbers of each type of user. At the boundary time between types, the collective instantaneous MEC, weighted by the arrival rate, is equal for both types of users. This contrasts with the bottleneck literature, which suggests that users with a lower VOT impose a lower congestion externality (e.g., Van den Berg and Verhoef, 2011b). Under proportional heterogeneity, different types of users have the same MEC, as in the bottleneck model. However, the MECs differ between the models; although, in the limit, as the elasticity of travel delay with respect to arrival rate approaches infinity, the MECs become the same.

Finally, in our dynamic flow congestion setting, the efficiency gains from tolling are smaller than those in the bottleneck model. Tolling always harms users with a lower VOT. The impact of tolling on users with a higher VOT is ambiguous, depending on the type of heterogeneity and the parameters, but does not result in a worse outcome than for users with a lower VOT. Compared to the bottleneck model, tolling is less attractive for users and has different distributional effects. Our results thus confirm that it is important to consider not only preference heterogeneity, but also the congestion type when analyzing user travel behavior and assessing the implementation of congestion pricing.

Although this paper provides some new insights into dynamic congestion with heterogeneous users, some further interesting extensions should be made in future studies. An obvious follow-up question is how step tolling would perform (see, e.g., Lindsey et al., 2012; Van den Berg, 2014). Additionally, it would be interesting to consider multiple dimensions of heterogeneity or have different congestion models. In particular, Verhoef (2020) and Hall (2021, 2023) have shown that adding heterogeneity into the preferred arrival time can lead to an enormous change in the effects of tolling. Another pertinent question would how results change if there were continuous distributions of preferences or more discrete groups. Finally, the congestion model adopted in this paper ignores the direct congestion interactions between travelers arriving at different moments. An important avenue for future research would be to consider alternative congestion technologies in which interactions between travelers arriving at different instants are still present. So, for example, the bathtub/MFD model or Mun (1999, 2002)'s model.

References

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

Arnott, R., Buli, J. (2018). Solving for equilibrium in the basic bathtub model. Transportation Research Part B: Methodological, 109, 150-175.

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

Arnott, R., de Palma, A., Lindsey, R. (1993). A structural model of peak-period congestion: A traffic bottleneck with elastic demand. The American Economic Review, 161-179.

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

Arnott, R., Kilani, M. (2022). Social optimum in the basic bathtub model. Transportation Science, 56(6), 1505-1529.

Agnew, C.E. (1977). The theory of congestion tolls. Journal of Regional Science, 17(3), 381-393.

Bao, Y., Verhoef, E. T., Koster, P. (2019). Regulating dynamic congestion externalities with tradable credit schemes: Does a unique equilibrium exist?. Transportation Research Part B: Methodological, 127, 225-236.

Beojone, C. V., Geroliminis, N. (2023). A dynamic multi-region MFD model for ride-sourcing with ridesplitting. Transportation Research Part B: Methodological, 177, 102821.

Chen, H., Nie, Y. M., Yin, Y. (2015). Optimal multi-step toll design under general user heterogeneity. Transportation Research Part B: Methodological, 81, 775-793.

Chu, X. (1995). Endogenous trip scheduling: The Henderson approach reformulated and compared with the Vickrey approach. Journal of Urban Economics, 37(3), 324-343.

Chu, X. (1999). Alternative congestion technologies. Regional Science and Urban Economics, 29(6), 697-722.

Cohen, Y. (1987). Commuter welfare under peak-period congestion tolls: Who gains and who loses? International Journal of Transport Economics, 14(3), 239-266.

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

Guo, R. Y., Yang, H., Huang, H. J. (2023). The day-to-day departure time choice of heterogeneous commuters under an anonymous toll charge for system optimum. Transportation Science, 57(3), 661-684.

Hall, J.D. (2018). Pareto improvements from Lexus Lanes: The effects of pricing a portion of the lanes on congested highways. Journal of Public Economics, 158, 113-125.

Hall, J.D. (2021). Can tolling help everyone? Estimating the aggregate and distributional consequences of congestion pricing. Journal of the European Economic Association, 19(1), 441-474.

Hall, J.D. (2023). Inframarginal travelers and transportation policy. SSRN working paper, 3424097.

Knoope, M. 2023. Nieuwe waarderingskengetallen voor reistijd, betrouwbaarheid en comfort (December 2023). Kennisinstituut voor Mobiliteit. Accesed from https://www.kimnet.nl/binaries/kimnet/documenten/publicaties/2023/12/04/nieuwe-waarderingskengetallen-voor-reistijd-betrouwbaarheid-en-comfort/Nieuwe+waarderingskengetallen+voor+reistijd%2C+betrouwbaarheid+en+comfort.pdf on 18 February 2024.

Kouwenhoven, M., de Jong, G.C., Koster, P., van den Berg, V.A.C., Verhoef, E.T., Bates, J., Warffemius, P.M. (2014). New values of time and reliability in passenger transport in The Netherlands. Research in Transportation Economics, 47, 37-49.

Li, Z.C., Huang, H.J., Yang, H. (2020). Fifty years of the bottleneck model: A bibliometric review and future research directions. Transportation Research Part B: Methodological, 139, 311-342.

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

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

Liu, Y., Nie, Y. M., Hall, J. (2015). A semi-analytical approach for solving the bottleneck model with general user heterogeneity. Transportation research part B: Methodological, 71, 56-70.

Long, J., Szeto, W. Y. (2019). Congestion and environmental toll schemes for the morning commute with heterogeneous users and parallel routes. Transportation Research Part B: Methodological, 129, 305-333.

Mangasarian, O.L. (1966). Sufficient conditions for the optimal control of nonlinear systems. SIAM Journal on Control, 4(1), 139-152.

Mun, S. I. (1999). Peak-load pricing of a bottleneck with traffic jam. Journal of Urban Economics, 46(3), 323-349.

Mun, S. I. (2002). Bottleneck congestion with traffic jam: a reformulation and correction of earlier result. Working paper, Graduate School of Economics, Kyoto University, Kyoto, Japan. Accessed from https://www.econ.kyoto-u.ac.jp/~mun/papers/Bottleneck0920.pdf on 25 March 2024.

Peer, S., Verhoef, E. T. (2013). Equilibrium at a bottleneck when long-run and short-run scheduling preferences diverge. Transportation Research Part B: Methodological, 57, 12-27.

Silva, H. E., Lindsey, R., de Palma, A., van den Berg, V. A.C. (2017). On the existence and uniqueness of equilibrium in the bottleneck model with atomic users. Transportation Science, 51(3), 863-881.Small, K. A. (1982). The scheduling of consumer activities: work trips. The

American Economic Review, 72(3), 467-479.

Small, K. A. (2015). The bottleneck model: An assessment and interpretation. Economics of Transportation, 4(1-2), 110-117.

Small, K., Verhoef, E. T. (2007). The economics of urban transportation. Routledge.

Sun, J., Wu, J., Xiao, F., Tian, Y., Xu, X. (2020). Managing bottleneck congestion with incentives. Transportation research part B: Methodological, 134, 143-166.

van den Berg, V.A.C. (2014). Coarse tolling with heterogeneous preferences. Transportation Research Part B: Methodological, 64, 1-23.

van den Berg, V.A.C. (2024). Self-financing roads under coarse tolling and preference heterogeneity. Transportation Research Part B: Methodological, 182, 102909.

van den Berg, V.A.C., Verhoef, E.T. (2011a). Winning or losing from dynamic bottleneck congestion pricing?: The distributional effects of road pricing with heterogeneity in values of time and schedule delay. Journal of Public Economics, 95(7-8), 983-992.

van den Berg, V.A.C., Verhoef, E.T. (2011b). Congestion tolling in the bottleneck model with heterogeneous values of time. Transportation Research Part B: Methodological, 45(1), 60-78.

Verhoef, E.T. (2020). Optimal congestion pricing with diverging long-run and short-run scheduling preferences. Transportation Research Part B: Methodological, 134, 191-209.

Verhoef, E. T., Silva, H. E. (2017). Dynamic equilibrium at a congestible facility under market power. Transportation Research Part B: Methodological, 105, 174-192.

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

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

Wu, W.X., Huang, H.J. (2014). Finding anonymous tolls to realize target flow pattern in networks with continuously distributed value of time. Transportation Research Part B: Methodological, 65, 31-46.

Wu, W.X., Huang, H.J. (2015). An ordinary differential equation formulation of the bottleneck model with user heterogeneity. Transportation Research Part B: Methodological, 81, 34-58.

Yang, H., Huang, H.J. (1997). Analysis of the time-varying pricing of a bottleneck with elastic demand using optimal control theory. Transportation Research Part B, 31 (6), 425-440.

Yildirimoglu, M., Ramezani, M., Geroliminis, N. (2015). Equilibrium analysis and route guidance in large-scale networks with MFD dynamics. Transportation Research Procedia, 9, 185-204.

Acknowledgments

The work described in this paper was supported by a grant from the National Natural Science Foundation of China (72201278). Any remaining errors are ours.

Appendix

Appendix A. Proof of Proposition 1

In Fig. 1, if an H-type user chooses to arrive at time t during $(t_{sL}, t_{eL}]$, the desired travel delay to keep the original equilibrium travel cost, T_1, is obtained from the crossover point of the dashed line and the vertical line at time t. However, the actual travel delay T_2 is determined by the travel delay of L type, implying that H type arriving at $(t_{sL}, t_{eL}]$ will experience a higher travel delay and, consequently, a higher travel cost. Therefore, they will prefer not to arrive during $(t_{sL}, t_{eL}]$ but in one of the two shoulder periods. Similar logic explains why the two shoulder periods are less attractive for L-type users. □

Appendix B. Derivation of the no-toll equilibrium under ratio heterogeneity

We first look at H-type users. According to $c_H(t) = c_H(t_{sH}) = c_H(t_{eH})$, the arrival rate of H type can be expressed by t_{sH} and t_{eH}:

$$f_H(t) = \begin{cases} K \cdot \left(\dfrac{\beta(t - t_{sH})}{\alpha_H} \right)^{\frac{1}{\chi}}, & \text{for t} \in [t_{sH}, t_{sL}] \\ K \cdot \left(\dfrac{\gamma(t_{eH} - t)}{\alpha_H} \right)^{\frac{1}{\chi}}, & \text{for t} \in (t_{eL}, t_{eH}] \end{cases} . \tag{B1}$$

Substituting (B1) into $\int_{t_{sH}}^{t_{sL}} f_H(t)dt + \int_{t_{eL}}^{t_{eH}} f_H(t)dt = N_H$ yields:

$$\frac{1}{\beta}\left(\frac{\beta(t_{sL} - t_{sH})}{\alpha_H} \right)^{\frac{1}{\chi}+1} + \frac{1}{\gamma}\left(\frac{\gamma(t_{eH} - t_{eL})}{\alpha_H} \right)^{\frac{1}{\chi}+1} = \frac{N_H}{\alpha_H K} \cdot \left(\frac{1}{\chi} + 1 \right). \tag{B2}$$

Combining $-\beta t_{sH} = \gamma t_{eH}$ and $-\beta t_{sL} = \gamma t_{eL}$ yields $(t_{sL} - t_{sH}) = \frac{\gamma}{\beta} \cdot (t_{eH} - t_{eL})$. Substituting it into (B2) yields:

$$t_{eH} - t_{eL} = \frac{\alpha_H}{\gamma} \cdot \left(\frac{\delta N_H}{\alpha_H K} \cdot \frac{\chi + 1}{\chi} \right)^{\frac{\chi}{\chi+1}}, \quad t_{sL} - t_{sH} = \frac{\alpha_H}{\beta} \cdot \left(\frac{\delta N_H}{\alpha_H K} \cdot \frac{\chi + 1}{\chi} \right)^{\frac{\chi}{\chi+1}}. \tag{B3}$$

Now we look at L-type users. Note that at t_{sL} and t_{eL}, all users have the same travel delay. Substituting (B1) into the travel delay function and combining the user equilibrium conditions, we can obtain the arrival rates for L-type users, expressed as:

$$f_L(t)=\begin{cases} K\cdot\left(\dfrac{\beta}{\alpha_L}(t-t_{sL})+\dfrac{\beta\cdot(t_{sL}-t_{sH})}{\alpha_H}\right)^{\frac{1}{\chi}}, \ \text{for t} \in (\text{t}_{sL},\text{t}^*] \\ K\cdot\left(\dfrac{\gamma}{\alpha_L}(t_{eL}-t)+\dfrac{\gamma(t_{eH}-t_{eL})}{\alpha_H}\right)^{\frac{1}{\chi}}, \ \ \text{for t} \in (\text{t}^*,\text{t}_{eL}] \end{cases}. \tag{B4}$$

Substituting (B3) into (B4) and inserting $\int_{t_{sL}}^{t_{eL}} f_L(t)dt = N_L$ yields:

$$\left(\frac{\gamma(t_{eH}-t_{eL})}{\alpha_H}+\frac{\gamma}{\alpha_L}t_{eL}\right)^{\frac{1}{\chi}+1}-\left(\frac{\gamma(t_{eH}-t_{eL})}{\alpha_H}\right)^{\frac{1}{\chi}+1}=(\frac{1}{\chi}+1)\frac{\delta N_L}{\alpha_L K}. \tag{B5}$$

Substituting (B3) into (B5), we can derive t_{eL} as:

$$t_{eL}=\frac{\alpha_L}{\gamma}\left((\frac{1}{\chi}+1)\cdot(\frac{\delta N_H}{\alpha_H K}+\frac{\delta N_L}{\alpha_L K})\right)^{\frac{\chi}{\chi+1}}-\frac{\alpha_L}{\gamma}\Psi_H, \tag{B6}$$

with $\Psi_H=\left[\dfrac{\delta N_H}{\alpha_H K}\cdot\left(\dfrac{1}{\chi}+1\right)\right]^{\frac{\chi}{\chi+1}}$. Combining (B6), (B3) and $-\beta t_{sL}=\gamma t_{eL}$ and further solving leads to the following arrival times of the first and last traveler of a type:

$$\begin{cases} t_{sH}=-\dfrac{\alpha_L}{\beta}\left((\dfrac{1}{\chi}+1)\cdot\left(\dfrac{\delta N_H}{\alpha_H K}+\dfrac{\delta N_L}{\alpha_L K}\right)\right)^{\frac{\chi}{\chi+1}}-\dfrac{\Psi_H}{\beta}\cdot(\alpha_H-\alpha_L),\ t_{sL}=-\dfrac{\alpha_L}{\beta}\left((\dfrac{1}{\chi}+1)\cdot\left(\dfrac{\delta N_H}{\alpha_H K}+\dfrac{\delta N_L}{\alpha_L K}\right)\right)^{\frac{\chi}{\chi+1}}+\dfrac{\alpha_L}{\beta}\Psi_H, \\ t_{eL}=\dfrac{\alpha_L}{\gamma}\left((\dfrac{1}{\chi}+1)\cdot\left(\dfrac{\delta N_H}{\alpha_H K}+\dfrac{\delta N_L}{\alpha_L K}\right)\right)^{\frac{\chi}{\chi+1}}-\dfrac{\alpha_L}{\gamma}\Psi_H,\ t_{eH}=\dfrac{\alpha_L}{\gamma}\left((\dfrac{1}{\chi}+1)\cdot\left(\dfrac{\delta N_H}{\alpha_H K}+\dfrac{\delta N_L}{\alpha_L K}\right)\right)^{\frac{\chi}{\chi+1}}+\dfrac{\Psi_H}{\gamma}\cdot(\alpha_H-\alpha_L). \end{cases} \tag{B7}$$

The equilibrium travel cost for H-type users can be derived by $c_H=-\beta t_{sH}$ and $c_L=-\beta t_{sL}+\alpha_L\cdot T(t_{sL})$, as presented in (9).

Appendix C. Proof of Proposition 2

Substituting (11) into $MEC_i=\dfrac{\partial c_H}{\partial N_i}N_H+\dfrac{\partial c_L}{\partial N_i}N_L$ yields the marginal external costs as:

$$\begin{cases} MEC_H=\left(\dfrac{\chi+1}{\chi}\cdot\left(\dfrac{\delta N_H}{\alpha_H K}+\dfrac{\delta N_L}{\alpha_L K}\right)\right)^{-\frac{1}{\chi+1}}\cdot\dfrac{\delta\alpha_L(N_H+N_L)}{\alpha_H K}+\left(\dfrac{\delta N_H}{\alpha_H K}\cdot\dfrac{\chi+1}{\chi}\right)^{-\frac{1}{\chi+1}}\cdot\dfrac{\delta(\alpha_H-\alpha_L)N_H}{\alpha_H K}, \\ MEC_L=\left(\dfrac{\chi+1}{\chi}\cdot\left(\dfrac{\delta N_H}{\alpha_H K}+\dfrac{\delta N_L}{\alpha_L K}\right)\right)^{-\frac{1}{\chi+1}}\cdot\dfrac{\delta(N_H+N_L)}{K}. \end{cases} \tag{C1}$$

Taking the difference between MEC_H and MEC_L yields:

$$MEC_L - MEC_H = \left(\frac{\chi+1}{\chi} \cdot \left(\frac{\delta N_H}{\alpha_H K} + \frac{\delta N_L}{\alpha_L K} \right) \right)^{-\frac{1}{\chi+1}} \cdot (1-\frac{\alpha_L}{\alpha_H}) \cdot \frac{\delta(N_H+N_L)}{K} - \left(\frac{\delta N_H}{\alpha_H K} \cdot \frac{\chi+1}{\chi} \right)^{-\frac{1}{\chi+1}} \cdot (1-\frac{\alpha_L}{\alpha_H}) \frac{\delta N_H}{K} < 0 , \tag{C2}$$

which can be further simplified as $\left(\frac{N_H}{\alpha_H} + \frac{N_L}{\alpha_L} \right)^{-\frac{1}{\chi+1}} \cdot (N_H + N_L) < \left(\frac{N_H}{\alpha_H} \right)^{-\frac{1}{\chi+1}} N_H$. This also implies $\frac{\alpha_H}{\alpha_L} < \frac{\left(1 + N_L / N_H\right)^{(\chi+1)} - 1}{N_L / N_H}$.

Appendix D. Proof of Proposition 3

Proposition 3 is proven by contradiction. We assume H-type users travel in the center of the peak period, with H type arriving within $(t_{sH}, t_{eH}]$ and L type arriving within $[t_{sL}, t_{sH}] \cup (t_{eH}, t_{eL}]$. The marginal effect of H type on travel delay is $\frac{dT(t)}{df_H(t)} = \frac{\chi}{K} \cdot \left(\frac{f_H(t)}{K} \right)^{\chi-1}$. Given the higher VOT for H type users, moving a marginal H-type user to L type's interval reduces the total travel delay cost. Consider arrival time t_{sH} and an arbitrary small value ε; if we move a marginal H user from $t_{sH} + \varepsilon$ to $t_{sH} - \varepsilon$, the decrease in the travel delay cost for H type users exceeds the increase for L type users. Consequently, total travel costs decrease with this rescheduling, indicating that H type arriving in the center of the peak period is suboptimal. □

Appendix E. Proof of Lemma 3

The equation of motion of λ_i in (20) and (21) shows that this shadow cost should be constant over time: $\lambda_H(t) = \overline{\lambda_H}$, $\lambda_L(t) = \overline{\lambda_L}$. Inserting into (22) and combining (27)-(28), we obtain:

$$\begin{aligned} H_H(t_{sH}) &= -c_H(t_{sH}) \cdot f_H(t_{sH}) + \lambda_H(t_{sH}) \cdot f_H(t_{sH}) = \frac{\partial c_H[t]}{\partial f_H[t]}[t_{sH}] \cdot (f_H[t_{sH}])^2 = 0 \\ H_H(t_{eH}) &= -c_H(t_{eH}) \cdot f_H(t_{eH}) + \lambda_H(t_{eH}) \cdot f_H(t_{eH}) = \frac{\partial c_H[t]}{\partial f_H[t]}[t_{eH}] \cdot (f_H[t_{eH}])^2 = 0 \end{aligned} . \tag{E1}$$

According to the travel cost expression, the only way for (E1) to hold is $f_H(t_{sH}) = f_H(t_{eH}) = 0$.

This completes the proof of part (i) of the lemma.

The t_{sL} and t_{eL} separate different types, and $f_i(t_{sL}) \neq 0$ and $f_i(t_{eL}) \neq 0$ both hold. At these points $\theta_H(t_{sL}) = \theta_L(t_{sL}) = \theta_H(t_{eL}) = \theta_L(t_{eL}) = 0$ holds. Combining (25) and (26) yields:

$$\begin{aligned} &\overline{\lambda_H} = c_H[t_{sH}] = c_H[t_{eH}],\ \mathrm{t} \in [\mathrm{t}_{sH},\mathrm{t}_{sL}] \cup (\mathrm{t}_{eL},\mathrm{t}_{eH}], \\ &\overline{\lambda_L} = c_L[t_{sL}] + \frac{\partial c_L[t]}{\partial f_L[t]}[t_{sL}] \cdot f_L[t_{sL}],\ \mathrm{t} \in (\mathrm{t}_{sL},\mathrm{t}_{eL}]. \end{aligned} \cdot \tag{E2}$$

By combining (18), (23), and (24), we can further obtain the relationship between $f_H(t_{sL})$ and $f_L(t_{sL})$, and between $f_H(t_{eL})$ and $f_L(t_{eL})$:

$$\frac{f_H(t_{sL})}{f_L(t_{sL})} = \frac{\overline{\lambda_L} - c_L(t_{sL})}{\overline{\lambda_H} - c_H(t_{sL})} = \left(\frac{\alpha_L}{\alpha_H}\right)^{\frac{1}{\chi+1}},\ \frac{f_H(t_{eL})}{f_L(t_{eL})} = \frac{\overline{\lambda_L} - c_L(t_{eL})}{\overline{\lambda_H} - c_H(t_{eL})} = \left(\frac{\alpha_L}{\alpha_H}\right)^{\frac{1}{\chi+1}}, \tag{E3}$$

This relationship is presented in Lemma 3(ii). Lemma 3(iii) follows as a result of conditions (23) and (24).

Appendix F. Proof of Proposition 4

Proof. The toll starts at zero at t_{sH} and becomes zero again at t_{eH}. The travel price for H-type users equals $\overline{\lambda_H} = c_H[t_{sH}] = c_H[t_{eH}]$. Hence, when $t \in [t_{sH}, t_{sL}] \cup (t_{eL}, t_{eH}]$, the optimal toll is:

$$\tau(t) = \overline{\lambda_H} - c_H(t) = c_H[t] + \frac{\partial c_H[t]}{\partial f_H[t]} \cdot f_H[t] - c_H(t) = \frac{\partial c_H[t]}{\partial f_H[t]} \cdot f_H[t] = \alpha_H \chi \left(\frac{f_H(t)}{K}\right)^{\chi}. \tag{F1}$$

Similarly, the travel price for L-type users is $p_L(t) = c_L(t) + \tau(t) = c_L(t_{sL}) + \tau(t_{sL})$. Combining (E2), we can obtain:

$$\tau(t) = \frac{\partial c_L[t]}{\partial f_L[t]} \cdot f_L[t] = \alpha_L \chi \cdot \left(\frac{f_L(t)}{K}\right)^{\chi},\ \mathrm{t} \in (\mathrm{t}_{sL},\mathrm{t}_{eL}], \tag{F2}$$

as presented in Eq. (29).

According to Lemma 3, the properties of the toll at the separation times of the types, t_{sL} and t_{eL}, can be further derived. □

Appendix G. Derivation of the equilibrium in social optimum under ratio heterogeneity

To solve the travel equilibrium, we first consider H-type users, as they travel in the shoulder of the peak period. In equilibrium, the following constraints should be satisfied: (i) At t_{sH} and t_{eH}, the schedule delay cost is the same, i.e., $-\beta_H t_{sH} = \gamma_H t_{eH}$; (ii) H-type users arrive during $[t_{sH}, t_{sL}] \cup (t_{eL}, t_{eH}]$, i.e., $\int_{t_{sH}}^{t_{sL}} f_H(t)dt + \int_{t_{eL}}^{t_{eH}} f_H(t)dt = N_H$; (iii) For H-type users, the travel price at any time t is the same as that at t_{sH} for early arrivals and at t_{eH} for late arrivals.

Solving the above equilibrium conditions yields the relationship between the travel interval and the arrival rate for H-type users, which is presented in (30), with

$$t_{sL} - t_{sH} = \frac{\alpha_H}{\beta} \cdot (1+\chi)^{\frac{1}{\chi+1}} \Psi_H, \; t_{eH} - t_{eL} = \Psi_H \cdot \frac{\alpha_H}{\gamma} \cdot (1+\chi)^{\frac{1}{\chi+1}}. \tag{G1}$$

To further solve the specific travel interval, we now consider L-type users. The travel price for L-type users is:

$$p_L(t) = \begin{cases} \alpha_L \cdot (1+\chi) \cdot \left[\dfrac{f_L(t)}{K}\right]^{\chi} - \beta t, \ \text{if t} \in (\text{t}_{sL}, \text{t}^*] \\ \alpha_L \cdot (1+\chi) \cdot \left[\dfrac{f_L(t)}{K}\right]^{\chi} + \gamma t, \ \text{if t} \in (\text{t}^*, \text{t}_{eL}] \end{cases}. \tag{G2}$$

Similar to H type users, for L-type users, in equilibrium, the following equilibrium conditions should be satisfied: $-\beta t_{sL} = \gamma t_{eL}$, $\int_{t_{sL}}^{t_{eL}} f_L(t)dt = N_L$, $p_L(t_{sL}) = p_L(t_{eL}) = p_L(t)$, and the transversality condition $\dfrac{f_L(t_{sL})}{f_H(t_{sL})} = \dfrac{f_L(t_{eL})}{f_H(t_{eL})} = \left(\dfrac{\alpha_H}{\alpha_L}\right)^{\frac{1}{\chi+1}}$. The arrival rate of L-type users can be expressed as a function of t_{sL} and t_{eL}, as presented in (31). Substituting (G1) and $-\beta t_{sL} = \gamma t_{eL}$ into (31), and solving $\int_{t_{sL}}^{t_{eL}} f_L(t)dt = N_L$, we can obtain:

$$\begin{cases} t_{sH} = \dfrac{\alpha_L}{\beta}(1+\chi)\left[\left(\dfrac{\delta N_H}{\alpha_L K \chi}\right)^{\frac{\chi}{\chi+1}} - \left(\dfrac{\delta N_H}{\alpha_L K \chi} + \dfrac{\delta N_L}{\alpha_L \chi K}\right)^{\frac{\chi}{\chi+1}}\right] - \dfrac{\alpha_H}{\beta}(1+\chi)\left(\dfrac{\delta N_H}{\alpha_H K \chi}\right)^{\frac{\chi}{\chi+1}} \\ t_{sL} = \dfrac{\alpha_L}{\beta}(1+\chi)\left[\left(\dfrac{\delta N_H}{\alpha_L K \chi}\right)^{\frac{\chi}{\chi+1}} - \left(\dfrac{\delta N_H}{\alpha_L K \chi} + \dfrac{\delta N_L}{\alpha_L \chi K}\right)^{\frac{\chi}{\chi+1}}\right] \\ t_{eL} = \dfrac{\alpha_L \cdot (1+\chi)}{\gamma}\left[\left(\dfrac{\delta N_H}{\alpha_L K \chi} + \dfrac{\delta N_L}{\alpha_L \chi K}\right)^{\frac{\chi}{\chi+1}} - \left(\dfrac{\delta N_H}{\alpha_L K \chi}\right)^{\frac{\chi}{\chi+1}}\right] \\ t_{eH} = \dfrac{\alpha_L \cdot (1+\chi)}{\gamma}\left[\left(\dfrac{\delta N_H}{\alpha_L K \chi} + \dfrac{\delta N_L}{\alpha_L \chi K}\right)^{\frac{\chi}{\chi+1}} - \left(\dfrac{\delta N_H}{\alpha_L K \chi}\right)^{\frac{\chi}{\chi+1}}\right] + \dfrac{\alpha_H}{\gamma}(1+\chi)\left(\dfrac{\delta N_H}{\alpha_H K \chi}\right)^{\frac{\chi}{\chi+1}} \end{cases}. \tag{G3}$$

From $p_H(t) = p_H(t_{sH}) = -\beta t_{sH}$ and $p_L(t) = p_L(t^*)$, we can derive the equilibrium travel price, as presented in (32). □

Appendix H. Proof of Proposition 5

Based on (32) and (9), for the L type, taking the difference between p_L and c_L yields:

$$p_L - c_L = \alpha_L \cdot (1+\chi) \cdot \left(\frac{\delta N_H}{\alpha_L K \chi} + \frac{\delta N_L}{\alpha_L \chi K} \right)^{\frac{\chi}{\chi+1}} - \alpha_L \left((\frac{1}{\chi}+1) \cdot \left(\frac{\delta N_H}{\alpha_H K} + \frac{\delta N_L}{\alpha_L K} \right) \right)^{\frac{\chi}{\chi+1}} . \tag{H1}$$

For presentation purposes, let $\alpha_H / \alpha_L = u$ and $N_H / N_L = v$. Eq. (H1) can be rewritten as:

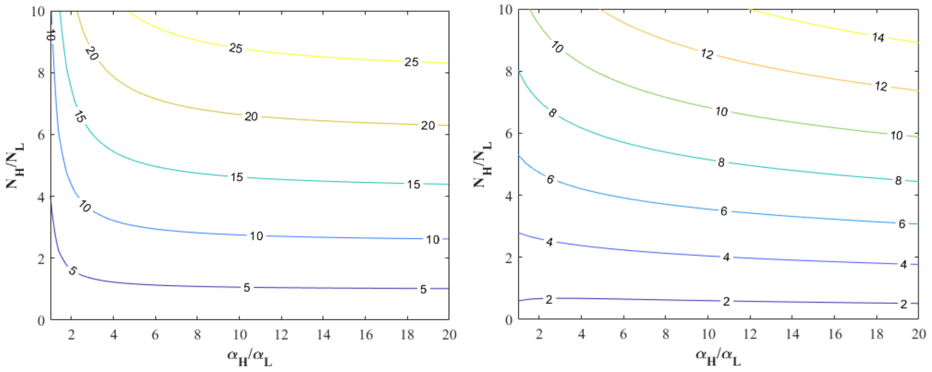
$$F(u) = p_L - c_L = \alpha_L \cdot (1+\chi) \cdot \left((v+1) \frac{\delta N_L}{\alpha_L K \chi} \right)^{\frac{\chi}{\chi+1}} - \alpha_L \left((1+\chi) \cdot \left(\frac{v}{u} + 1 \right) \frac{\delta N_L}{\alpha_L K \chi} \right)^{\frac{\chi}{\chi+1}} . \tag{H2}$$

Taking the derivative of $F(u)$ with respect to u yields:

$$\begin{aligned} \frac{\partial F(u)}{\partial u} &= \alpha_L \frac{\chi}{\chi+1} \left((1+\chi) \cdot \left(\frac{v}{u} + 1 \right) \frac{\delta N_L}{\alpha_L K \chi} \right)^{\frac{\chi}{\chi+1}-1} (1+\chi) \frac{\delta N_L}{\alpha_L K \chi} \frac{v}{u^2} \\ &= \left((1+\chi) \cdot \left(\frac{v}{u} + 1 \right) \frac{\delta N_L}{\alpha_L K \chi} \right)^{\frac{\chi}{\chi+1}-1} \frac{\delta N_L}{K} \frac{v}{u^2} > 0 \end{aligned} . \tag{H3}$$

Specifically, when $u = 1$, we can easily obtain $F(1) > 0$. Hence, $p_L - c_L > 0$ always holds, and, as the degree of ratio heterogeneity increases, the difference in the travel price between the different types becomes larger.

Using the same logic, we can prove $p_H - c_H > 0$ holds. Fig. H1 further gives a numerical illustration under χ=4.



(a) L type users

(b) H type users

Fig. H1. Contour plot of the sign-determined part of $p_i - c_i$ under χ=4.

Appendix I. Equilibrium under social optimum with proportional heterogeneity

(i) Arrival order

Similar to ratio heterogeneity, the arrival order in Proposition 7 can be proven by contradiction. The properties of the travel delay and toll at the switching time are derived from the transversality condition in Eq. (I2). Details are available upon request.

(ii) Solving the travel equilibrium

Formulating the corresponding optimal control problem and solving the Hamiltonian yield the same toll rule as ratio heterogeneity. The travel price function thus becomes:

$$p_i(t) = \alpha_i \cdot (1+\chi) \cdot \left(\frac{f_i(t)}{K} \right)^{\chi} + \begin{cases} \beta_i \cdot (t^* - t) & \text{for early arrivals of } i \text{ type} \\ \gamma_i \cdot (t - t^*) & \text{for late arrivals of } i \text{ type} \end{cases} . \tag{I1}$$

Then we start to solve the equilibrium. At t_{sH} and t_{eH}, the transversality condition shows:

$$\left(\frac{f_H(t_{sH})}{f_L(t_{sH})} \right)^{\chi+1} = \left(\frac{f_H(t_{eH})}{f_L(t_{eH})} \right)^{\chi+1} = \frac{\alpha_L}{\alpha_H} . \tag{I2}$$

We first look at L-type users. According to $p_L(t) = -\beta t_{sL} = \gamma t_{eL}$ and $\int_{t_{sL}}^{t_{sH}} f_L(t)dt + \int_{t_{eH}}^{t_{eL}} f_L(t)dt = N_L$, the arrival rate for L-type users can be expressed as:

$$f_L(t) = \begin{cases} K \cdot \left(\frac{\beta_L (t - t_s)}{\alpha_L \cdot (1+\chi)} \right)^{\frac{1}{\chi}}, \text{ for } t \in [t_{sL}, t_{sH}] \\ K \cdot \left(\frac{\gamma_L (t_e - t)}{\alpha_L \cdot (1+\chi)} \right)^{\frac{1}{\chi}}, \text{ for } t \in (t_{eH}, t_{eL}] \end{cases} . \tag{I3}$$

Combining (I2), H type's arrival rates at t_{sH} and t_{eH} are, respectively:

$$f_H(t_{sH}) = K \cdot \left(\frac{\alpha_L}{\alpha_H} \right)^{\frac{1}{\chi+1}} \cdot \left(\frac{\beta_L (t_{sH} - t_s)}{\alpha_L \cdot (1+\chi)} \right)^{\frac{1}{\chi}}, f_H(t_{eH}) = K \cdot \left(\frac{\alpha_L}{\alpha_H} \right)^{\frac{1}{\chi+1}} \cdot \left(\frac{\gamma_L (t_e - t_{eH})}{\alpha_L \cdot (1+\chi)} \right)^{\frac{1}{\chi}} . \tag{I4}$$

Substituting (I3) into $\int_{t_{sL}}^{t_{sH}} f_L(t)dt + \int_{t_{eH}}^{t_{eL}} f_L(t)dt = N_L$ yields:

$$\int_{t_{sL}}^{t_{sH}} \left(\frac{\beta_L (t - t_s)}{\alpha_L \cdot (1+\chi)} \right)^{\frac{1}{\chi}} dt + \int_{t_{eH}}^{t_{eL}} \left(\frac{\gamma_L (t_e - t)}{\alpha_L \cdot (1+\chi)} \right)^{\frac{1}{\chi}} dt = \frac{N_L}{K} . \tag{I5}$$

Combining (I5) with $-\beta_L t_{sL} = \gamma_L t_{eL}$, $-\beta_H t_{sH} = \gamma_H t_{eH}$ and $\gamma_H / \beta_H = \gamma_L / \beta_L$ gives the travel lengths for early and late arrivals of L-type users:

$$t_{sH}-t_{sL}=\frac{\alpha_L\cdot(1+\chi)}{\beta_L}\left(\frac{\delta_L N_L}{\alpha_L\chi K}\right)^{\frac{\chi}{1+\chi}},\ \ t_{eL}-t_{eH}=\frac{\alpha_L\cdot(1+\chi)}{\gamma_L}\left(\frac{\delta_L N_L}{\alpha_L\chi K}\right)^{\frac{\chi}{1+\chi}}. \tag{I6}$$

Next, we look at H-type users. According to (I1), the travel price at t_{sH} and t_{eH} is,

$$p_H(t_{sH})=\alpha_H\cdot(1+\chi)\cdot\left(\frac{f_H(t_{sH})}{K}\right)^{\chi}-\beta_H t_{sH}\ \ ,\ \ \text{and}\ \ p_H(t_{eH})=\alpha_H\cdot(1+\chi)\cdot\left(\frac{f_H(t_{eH})}{K}\right)^{\chi}+\gamma_H t_{eH}\ \ ,$$

where $f_H(t_{sH})$ and $f_H(t_{eH})$ are given by (I4). Combining $p_H(t)=p_H(t_{sH})$ and $p_H(t)=p_H(t_{eH})$, the arrival rate for H-type users can be further expressed as:

$$f_H(t)=\begin{cases}K\cdot\left[\dfrac{\beta_H(t-t_{sH})}{\alpha_H\cdot(1+\chi)}+\left(\dfrac{\alpha_L}{\alpha_H}\right)^{\frac{\chi}{\chi+1}}\cdot\left(\dfrac{\delta_L N_L}{\alpha_L\chi K}\right)^{\frac{\chi}{1+\chi}}\right]^{\frac{1}{\chi}}, \text{ for } t\in(t_{sH},t^*] \\ K\cdot\left[\dfrac{\gamma_H(t_{eH}-t)}{\alpha_H\cdot(1+\chi)}+\left(\dfrac{\alpha_L}{\alpha_H}\right)^{\frac{\chi}{\chi+1}}\cdot\left(\dfrac{\delta_L N_L}{\alpha_L\chi K}\right)^{\frac{\chi}{1+\chi}}\right]^{\frac{1}{\chi}}, \text{ for } t\in(t^*,t_{eH}]\end{cases}. \tag{I7}$$

Substituting (I7) into $\int_{t_{sH}}^{t^*}f_H(t)dt+\int_{t^*}^{t_{eH}}f_H(t)dt=N_H$ yields:

$$\int_{t_{sH}}^{t^*}\left[\frac{\beta_H(t-t_{sH})}{\alpha_H\cdot(1+\chi)}+\left(\frac{\alpha_L}{\alpha_H}\right)^{\frac{\chi}{\chi+1}}\cdot\frac{\beta_L(t_{sH}-t_s)}{\alpha_L\cdot(1+\chi)}\right]^{\frac{1}{\chi}}dt+\int_{t^*}^{t_{eH}}\left[\frac{\gamma_H(t_{eH}-t)}{\alpha_H\cdot(1+\chi)}+\left(\frac{\alpha_L}{\alpha_H}\right)^{\frac{\chi}{\chi+1}}\cdot\frac{\gamma_L(t_e-t_{eH})}{\alpha_L\cdot(1+\chi)}\right]^{\frac{1}{\chi}}dt=\frac{N_H}{K}. \tag{I8}$$

Using (I6) to further simplify (I8), we can obtain:

$$\frac{-\beta_H t_{sH}}{\alpha_H\cdot(1+\chi)}=\left(\frac{\delta_H N_H}{\alpha_H\chi K}+\frac{\delta_L N_L}{\alpha_H\chi K}\right)^{\frac{\chi}{1+\chi}}-\left(\frac{\delta_L N_L}{\alpha_H\chi K}\right)^{\frac{\chi}{1+\chi}}. \tag{I9}$$

As a result, $$t_{sH}=\frac{\alpha_H}{\beta_H}\cdot(1+\chi)\left(\frac{\delta_L N_L}{\alpha_H\chi K}\right)^{\frac{\chi}{1+\chi}}-\frac{\alpha_H}{\beta_H}\cdot(1+\chi)\left(\frac{\delta_H N_H}{\alpha_H\chi K}+\frac{\delta_L N_L}{\alpha_H\chi K}\right)^{\frac{\chi}{1+\chi}}. \tag{I10}$$

Combining (I10) and (I6), we can further derive the expression of t_{sL}, t_{eL} and t_{eH}. Substituting them into (I3) and (I7) yields the arrival rates. The equilibrium travel price for H type is obtained from $p_H(t)=p_H(t^*)$, and for L type is derived from $p_L=-\beta_L t_{sL}$, as presented in (39). This completes the derivation of the equilibrium. □

Appendix J. Proof of Proposition 8

According to (39) and (35), for H-type users, solving $p_H > c_H$ yields:

$$\alpha_H \cdot (1+\chi) \cdot \left(\frac{\delta_H N_H}{\alpha_H \chi K} + \frac{\delta_L N_L}{\alpha_H \chi K} \right)^{\frac{\chi}{1+\chi}} > \alpha_H \left(\frac{\delta_H (N_H + N_L)}{\alpha_H K} \cdot \frac{1+\chi}{\chi} \right)^{\frac{\chi}{\chi+1}}, \tag{J1}$$

which can be simplified as: $(1+\chi)^{\frac{1}{\chi+1}} > \left(\frac{\delta_H N_H + \delta_H N_L}{\delta_H N_H + \delta_L N_L} \right)^{\frac{\chi}{\chi+1}}$. To gain more insights, we use simulation.[23] Fig. J1 plots the contour of $p_H - c_H$ under $\chi = 4$. We can see $p_H > c_H$ is more likely to happen when N_H / N_L is relatively large.

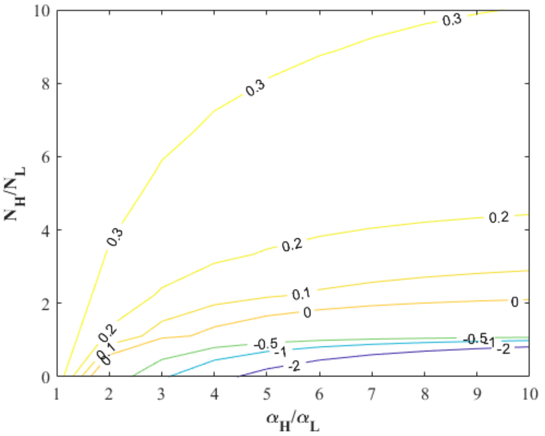


Fig. J1. Contour of $p_H - c_H$ under $\chi = 4$.

For L-type users, solving $p_L > c_L$ and simplifying further gives:

$$\left(\frac{\delta_H N_H + \delta_L N_L}{\alpha_H} \right)^{\frac{\chi}{1+\chi}} + \left(\frac{\delta_L N_L}{\alpha_L} \right)^{\frac{\chi}{1+\chi}} - \left(\frac{\delta_L N_L}{\alpha_H} \right)^{\frac{\chi}{1+\chi}} > \left(\frac{\delta_H (N_H + N_L)}{\alpha_H} \right)^{\frac{\chi}{\chi+1}} (1+\chi)^{-\frac{1}{\chi+1}}, \tag{J2}$$

which also means

$$(1+\chi)^{\frac{1}{\chi+1}} > \frac{\left(\frac{\delta_H (N_H + N_L)}{\alpha_H} \right)^{\frac{\chi}{\chi+1}}}{\left(\frac{\delta_H N_H + \delta_L N_L}{\alpha_H} \right)^{\frac{\chi}{1+\chi}} + \left(\frac{\delta_L N_L}{\alpha_L} \right)^{\frac{\chi}{1+\chi}} - \left(\frac{\delta_L N_L}{\alpha_H} \right)^{\frac{\chi}{1+\chi}}}. \tag{J3}$$

[23] Let $u = \alpha_H / \alpha_L$ and $v = N_H / N_L$, inequality (J2) can be rewritten as $(1+\chi)^{\frac{1}{\chi+1}} > \left(\frac{u(v+1)}{uv+1} \right)^{\frac{\chi}{\chi+1}}$.

To prove (J3) always holds, let $\alpha_H/\alpha_L = u$ and $N_H/N_L = v$; (J3) can be rewritten as $(1+\chi)^{\frac{1}{\chi+1}} > \dfrac{\left(u(v+1)\right)^{\frac{\chi}{\chi+1}}}{\left(uv+1\right)^{\frac{\chi}{1+\chi}} + u^{\frac{\chi}{\chi+1}} - 1}$. Let $F(v) = (1+\chi)^{\frac{1}{\chi+1}}\left[\left(uv+1\right)^{\frac{\chi}{1+\chi}} + u^{\frac{\chi}{\chi+1}} - 1\right] - \left(u(v+1)\right)^{\frac{\chi}{\chi+1}}$.

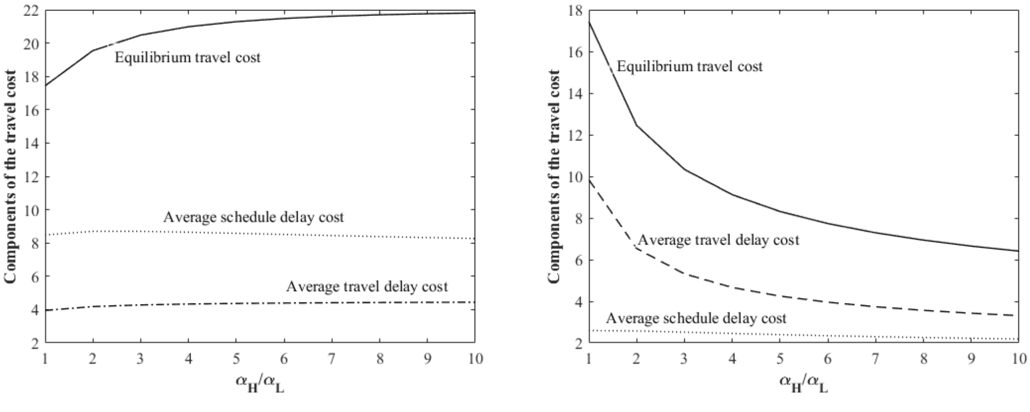
Taking the derivative of $F(v)$ with respect to v yields:

$$\frac{\partial F}{\partial v} = \frac{\chi u}{1+\chi}(1+\chi)^{\frac{1}{\chi+1}}\left(uv+1\right)^{-\frac{1}{1+\chi}} - \frac{\chi u}{1+\chi}\left(u(v+1)\right)^{-\frac{1}{\chi+1}}. \tag{J4}$$

From $\left(\dfrac{uv+1}{1+\chi}\right) < \left(u(v+1)\right)$, we can further obtain: $\left(\dfrac{uv+1}{1+\chi}\right)^{-\frac{1}{1+\chi}} > \left(u(v+1)\right)^{-\frac{1}{\chi+1}}$. Hence, $(1+\chi)^{\frac{1}{\chi+1}}\left(uv+1\right)^{-\frac{1}{1+\chi}} > \left(u(v+1)\right)^{-\frac{1}{\chi+1}}$ holds, which means $\partial F/\partial v > 0$. Considering $v \ge 0$ and $F(0) = (1+\chi)^{\frac{1}{\chi+1}}u^{\frac{\chi}{\chi+1}} - u^{\frac{\chi}{\chi+1}} > 0$, we thus show that $F(v) > 0$ always holds. Therefore, $p_L > c_L$.□

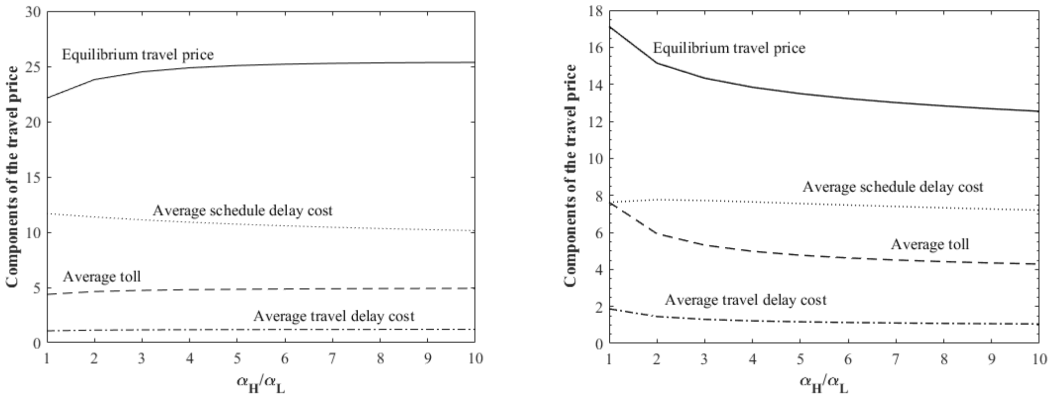
Appendix K. Decomposition of the equilibrium travel price under ratio heterogeneity

Fig. K1 presents a detailed decomposition of equilibrium travel price for different types of users, with varying degrees of ratio heterogeneity. It can be seen that increased ratio heterogeneity raises the travel price for H type users, as shown in Fig. K2(a) and Fig. K2(c), and lowers the travel price for L type users, as shown in Fig. K2(b) and Fig. K2(d). Overall, L type users are more impacted by tolling.



(a) H type users (without tolling)

(b) L type users (without tolling)



(c) H type users (social optimum) (d) L type users (social optimum)

Fig. K1. Decomposition of the travel price under ratio heterogeneity.