Transportation Research Part B 139 (2020) 523–538

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

Emergent hypercongestion in Vickrey bottleneck networks☆

Dario Frascaria[a,*], Neil Olver[b,c], Erik Verhoef[d,e]

[a] Department of Econometrics & Operations Research, Vrije Universiteit Amsterdam Netherlands
[b] Department of Mathematics, London School of Economics and Political Science United Kingdom
[c] Centrum Wiskunde & Informatica, Amsterdam United Kingdom
[d] Department of Spatial Economics, Vrije Universiteit Amsterdam Netherlands
[e] Tinbergen Institute Netherlands

ARTICLE INFO

Article history:
Received 21 December 2019
Revised 23 July 2020
Accepted 29 July 2020

Keywords:
Hypercongestion
Vickrey bottlenecks
Spaceless vertical queues
Arbitrary networks
Homogeneous users
Optimal (first-best) pricing

ABSTRACT

Hypercongestion—the phenomenon that higher traffic densities can reduce throughput—is well understood at the link level, but has also been observed in a macroscopic form at the level of traffic networks; for instance, in morning rush-hour traffic into a downtown core. In this paper, we show that macroscopic hypercongestion can occur as a purely emergent effect of dynamic equilibrium behavior on a network, even if the underlying link dynamics (we consider Vickrey bottlenecks with spaceless vertical queues) do not exhibit hypercongestion.

1. Introduction

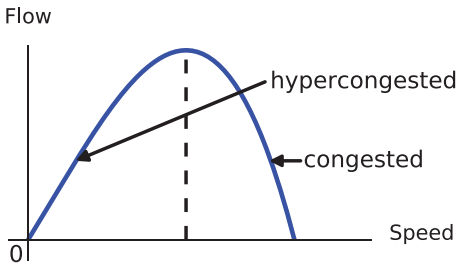
Empirical studies have shown that traffic throughput can be lower during peak hours of high demand than during less congested hours. This phenomenon is called *hypercongestion*. Hypercongestion has been studied for single links, as well as for networks as a whole. For single links, it is understood through the *fundamental diagrams of traffic flow* (Fig. 1 and (c)) that relate traffic density, traffic speed and traffic flow, and these mechanisms are of particular importance in modelling highway traffic flow. In highway traffic, it is a well-understood and intuitive fact that speed is negatively related to density. Since flow is the product of speed and density, maximum flow is achieved at some intermediate value of density. Increasing density beyond that point (or equivalently, decreasing speed) yields to decreased throughput; this is hypercongestion.

For networks as a whole, representing, for example, the downtown core of a city, "macroscopic" versions of the fundamental traffic diagram have been empirically observed (e.g., (Geroliminis and Daganzo, 2008; Daganzo et al., 2011)). Fig. 1(d) reproduces Fig. 9 from (Daganzo et al., 2011): here the vertical axis represents the number of vehicles passing, per lane and per unit of time, the points of measurement and the horizontal axis represents the density of the network, which is proportional to the number of vehicles present in the network.

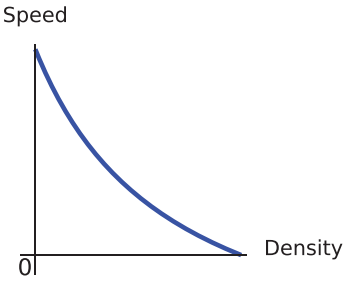
The effect has been discussed in the context of *bathtub models* (Arnott, 2013; Fosgerau, 2015; Arnott et al., 2016), and indeed is a key motivation for these models. Here, speed and density are *uniform* across the space, even though agents have different routes with different lengths. With this strong spatial homogeneity assumption, the network structure of

☆ Partially supported by NWO TOP grant 614.001.510 and NWO Vidi grant 016.Vidi.189.087.
* Corresponding author.
E-mail address: d.frascaria@vu.nl (D. Frascaria).

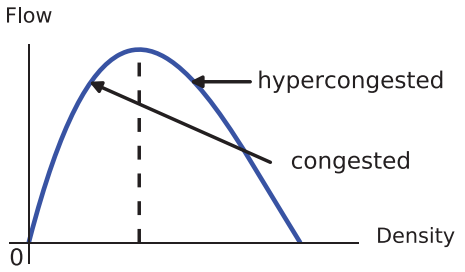
https://doi.org/10.1016/j.trb.2020.07.010



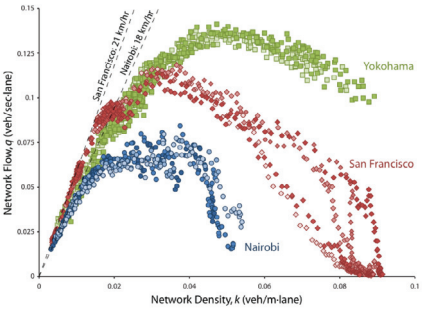
(a) Flow-speed diagram.



(b) Speed-density diagram.



(c) Flow-density (or fundamental) diagram.



(d) Macroscopic fundamental diagram from [Daganzo et al., 2011, Fig.9].

Fig. 1. Fundamental diagrams of traffic flow.

the city and its road structure is abstracted away and network interactions[1] are completely absent. A negative relationship (e.g., Greenshields' linear relation) is then prescribed between the (spatially averaged) density and speed. This leads to hypercongestion when density exceeds some critical value, for the same reason as for a single link.

While these models do an excellent job of matching empirically observed behavior such as the macroscopic fundamental diagram of Fig. 1(d), they do not provide an explanation for the source of this negative relationship, and hence the source of hypercongestion. In particular, they typically lack explicit modelling of entry into and exit out of the bathtub, while even for single-link models it has been shown that hypercongestion can only build up if exit capacity is restricted (i.e., there is a downstream bottleneck), while entry capacity should not be below the downstream exit capacity (e.g., (Verhoef, 2001; 2003)). To the extent that the bathtub is modeled as a simple congestible facility, one might expect the same type of necessary conditions for hypercongestion to occur in these models, which would justify an explicit consideration of entry and exit mechanisms. In any case, it seems worthwhile to search for a deeper explanation of the precise causes of hypercongestion in urban traffic, as this may produce new insights on optimal transport (pricing) policies.

The most obvious explanation for this relationship between averaged density and speed would be that the relationship holds at the level of individual links. In other words: if hypercongestion occurs at the level of individual links, we would expect it to occur in the macroscopic level as well. Hypercongestion at the link level can be considered as expected on highways entering the city, given that downstream capacity is limited. But *within* the city, is less evident that one should expect substantial per-link hypercongestion. This is especially relevant given that the Vickrey bottleneck model (Vickrey, 1969) is generally considered to be a good "workhorse" dynamic economic model for the most heavily congested links in the context of urban traffic. However, in its purest form, this model does not exhibit the flow drops that are characteristic for hypercongestion at the link level.

An alternative source for hypercongestion is through spillback (coined "triggerneck congestion" by Vickrey (1969)). If traffic congestion in one part of the network can block upstream intersections, it is relatively easy to construct examples where hypercongestion occurs.[2] In highly congested situations, the potential for "gridlock" caused by spillback effects is a very plausible mechanism for hypercongestion in cities. But what about lower levels of congestion where the impact of spillback is nonexistent or small? What about the role of dynamic route choices and network interactions before triggerneck

[1] Temporary and spatially separate events that affect each other.

[2] For example, consider a highway with two consecutive off-ramps, with different destinations. If the later off-ramp becomes sufficiently congested, the resulting queue can block traffic exiting on the first off-ramp, resulting in a lower overall network capacity.

congestion sets in and, therewith, the role of network design? Can that in itself be a source of hypercongestion at the level of the network? In light of these considerations, we ask the following natural question.

Question. Can hypercongestion occur in settings where (1) no hypercongestion occurs at the link level, in that the flow-speed relationship does not exhibit a backward-bending curve; and (2) there are no spillback effects?

Our assumptions in asserting whether dynamic network behavior can produce hypercongestion through route interaction are therefore conservative: if, in our model, hypercongestion is observed at the network level, it can be fully attributed to the network interaction of multiple users, all aiming to minimize their travel costs.

Thus, in this work, we consider pure bottleneck congestion dynamics, with spaceless vertical queues as introduced by Vickrey (1969) where, at the link level, an increase in density never corresponds to a decrease of flow. We combine this with endogenous departure time choices; users desire to arrive at work at a particular time, and incur *scheduling cost* for arriving either early or late. This is a very standard model, also introduced by Vickrey (1969), and elaborated in many works since (see Small (2015) for a survey). Most of the attention in the transportation economics literature has been on single link bottleneck models. Here, it is well-known that hypercongestion effects do *not* occur (Arnott et al., 1990). Extending from a single link to multiple perfectly parallel links (i.e., a collection of completely separate routes from the origin to the destination) does not introduce any new behavior: hypercongestion still does not occur. In this work, we will consider arbitrary network topologies, which have received much less attention in the literature of hypercongestion.

We distinguish two aspects of hypercongestion that generally coincide for regular single-link models, but that become useful to distinguish in our network setting.

Speed-flow hypercongestion. We consider this form of hypercongestion to occur when a backward-bending curve manifests in a macroscopic equivalent of Fig. 1(a) or (c), using appropriate macroscopic equivalents of speed, density and flow (throughput). We will make this precise, and detail how we measure these quantities, in Section 2.

Generally, we would expect that the network becomes more congested as the desired arrival time approaches, after which it decreases. So this form of hypercongestion may show itself as a period of time before the desired arrival where throughput starts to *decrease*, or alternatively, a period after the desired arrival time where the throughput *increases*.

Throughput hypercongestion. This form of hypercongestion is exhibited when the imposition of optimal (first-best) pricing to decentralize the social ("system") optimal departure pattern leads to an increase in the (time-averaged) arrival flow at the destination, compared to the original no-toll equilibrium. Put differently: tolls chosen to minimize the average cost (journey time cost plus scheduling cost) lead not just to a reduction of these *average costs* (a triviality), but also to a reduction in the *generalized price*, in which the toll is also included. Thus, as phrased by Arnott (2013) in his discussion of the bathtub model, "[in] very congested cities optimal tolling would still benefit commuters even if the toll revenue were completely squandered!".

A reduction in the generalized price in a dynamic equilibrium with homogeneous users necessarily implies that the duration of the peak—the time between the first and last departure—decreases. Note that bottleneck models for heterogeneous users have already shown that some users can gain from the imposition of optimal tolls before revenues are recycled. But these gains do not result from an increase in rate of arrivals at the destination, and hence a shortening of peak duration, but rather from the replacement of travel delays as a dynamic equilibrium-restraint mechanism by tolls. This tends to benefit users with a high value of time, for whom paying with time is relative less attractive than paying with money (van den Berg and Verhoef, 2011). In the present model, the reduction in travel price stems from an increase in the physical network usage and we obtain it under the, for this perspective conservative, assumption of homogeneous preferences.

Agraph*Our results We answer the main question with a resounding *yes*. We show that both forms of hypercongestion can occur, even in very small networks, and with all traffic having the same origin and destination. Our result is perhaps surprising, and seems to differ from what was generally believed. For example, Arnott et al. (2016, page 1) write on the model of bottleneck congestion:

> "While it has proved very adaptable and has generated a host of useful insights, as a model of downtown traffic congestion it is flawed since it rules out hypercongestion, assuming instead that under congested conditions aggregate traffic flow is constant."

Our primary goal is to show simply that hypercongestion *can* occur in a model with bottleneck congestion, as soon as the network structure is explicitly taken into consideration; and, that is true even if spillback congestion is ruled out. We do not examine realistic city-sized instances or compare with empirical data. But we do show that the effect is robust, and does not require precise numerical tuning of the instance. Further, the effect can be fairly significant; for example, we can exhibit a decrease of the average generalized price using optimal tolling of approximately 20%. We examine some aspects of the sensitivity of the hypercongestion effect to various parameters in Section 4.

We also emphasize upfront that unlike for a single bottleneck, the two forms of hypercongestion are no longer equivalent. In particular, it is possible for speed-flow hypercongestion to occur while throughput hypercongestion does not, and vice versa.

The interesting feature of our result is that we obtain increased throughput and reduced generalized prices for finite tolls that do not fully close down certain links, like one might expect from Braess-type networks, and that satisfy necessary conditions for first-best optimal (dynamic) congestion pricing. This has important policy implications regarding the acceptability of optimal tolls.

In parallel-link networks with pure bottlenecks, the generalized price under optimal tolls is the same as for the untolled equilibrium. Our main result shows that it can strictly decrease in general. The reader may wonder whether optimal tolls can ever strictly increase the generalized price in this model. In Section 5, we show that this is indeed possible. Generalized prices increasing due to tolls are what commentators typically have in mind when opposing road pricing. It is noteworthy that in our model this intuitive notion may be confirmed, but may also be rejected.

Policy implications Our findings are relevant for various pertinent policy questions in urban transport.

First, we provide a framework in which we can separate hypercongestion as arising from network interactions from hypercongestion as it may result from local, link-specific travel time functions. Insights into causes and consequences of hypercongestion are of great interest for efficient policy makings for cities around the world. Traffic congestion is without doubt one of the main challenges facing contemporary cities and hypercongestion is a particularly severe and wasteful phenomenon, representing traffic conditions for which the same flow can be reached at a higher-speed, lower-density configuration. Therefore, improving the understanding of hypercongestion helps better formulating policies to combat the most severe types of congestion.

Second, we show how optimal pricing may not only eliminate queuing but in addition decrease generalized prices before toll revenues are redistributed. This is an important question for the political and social acceptability of pricing. Among multiple reasons, the mere fact that congestion pricing normally brings societal benefits at the price of raising the generalized price for road users is often seen as a dominant cause for the very limited societal acceptability of road pricing (e.g., Small and Verhoef (2007)). Therefore, finding instances of road pricing that addresses the most severe type of congestion while reducing the generalized price of travel before recycling tax revenues is a particularly attractive feature.

Third, we provide insight into how lessons that can be learned from spatially homogeneous bathtub or MFD models translate into link-specific, spatially differentiated congestion pricing, as it is likely to be implemented in real applications: namely, whenever road pricing policies will not be defined to be homogeneous over space (as will be true for the archetype bathtub model), but will instead be differentiated over key links and bottlenecks in the whole network.

2. Model and preliminaries

Modelling networks of Vickrey bottlenecks, with topologies more complicated than parallel link networks, requires more care than extensions to the purely parallel link setting, with or without differences in free-flow travel times. The perspective we take here, follows some relatively recent treatments in the operations research literature (e.g., Cominetti et al. (2015); Koch and Skutella (2011)), which has its roots in the classical study of so-called *flows over time*. We avoid discussion of some (unimportant, for our current purposes) technicalities. The reader may wish to consult (Cominetti et al., 2015) for a more in-depth treatment, albeit for exogenously chosen departure times. For our setting there are variational inequalities formulations and algorithms to compute the dynamic equilibria (see (Friesz and Han, 2019) for a survey), but these formulations are less convenient for our purposes.

Network structure and link dynamics Within the economics literature, the Vickrey bottleneck model has been principally studied in the setting of single or purely parallel links. Here, there are multiple choices of separate roads between a common origin and a common destination. We will be particularly concerned with more complicated network structures than purely parallel links. This will be described by a given directed graph G, with node set V and arc set E. We will still restrict ourselves to the setting of a common origin, which we will denote by $s \in V$, and a common destination $t \in V$.

Each link $e \in E$ has a *capacity* ν_e and a *free-flow travel time* τ_e associated with it. We can describe the flow on a link e by functions $f_e^+, f_e^- : \mathbb{R} \to \mathbb{R}_+$; $f_e^+(\xi)$ denotes the *inflow rate* into arc ξ at time ξ, and similarly $f_e^-(\xi)$ the *outflow rate*. (So for example, the total amount of flow that has entered arc e by time ξ is $\int_{-\infty}^{\xi} f_e^+(t)dt$.) We require that these functions be measurable and defined almost everywhere (for the equilibrium flows we will shortly consider, they will in fact be piecewise constant). Traffic is treated as divisible, and vehicles therefore as atomistic and a continuous flow, and so these functions can take on arbitrary nonnegative values, without any integrality restrictions.

Flow through a link can never exceed the capacity, meaning that if the inflow rate exceeds the capacity, a (FIFO) *queue* forms at the entrance to the link. We use $z_e(\xi)$ to denote the total mass of the queue on link e at time ξ. A particle that enters the link at time ξ will experience a queueing delay of $z_e(\xi)/\nu_e$, since the queue will evacuate at the maximum possible rate, i.e. the capacity ν_e.[3] The particle will thus depart the link at time

$$T_e(\xi) := \xi + z_e(\xi)/\nu_e + \tau_e. \tag{1}$$

The evolution of the queue is described via the link dynamics

$$\frac{d}{d\xi} z_e(\xi) = \begin{cases} f_e^+(\xi) - \nu_e & \text{if } z_e(\xi) > 0 \\ \max\{f_e^+(\xi) - \nu_e, 0\} & \text{if } z_e(\xi) = 0 \end{cases}. \tag{2}$$

[3] Once inside the network, users want to arrive at destination as soon as possible, a feature that is self-evident for arrivals after the most desired moment, and consistent for early arrivals with the conventional assumption that the value of schedule early delays (β) is smaller than the value of travel time (α).

The outflow rate at time $\xi + \tau_e$ is also clear: if there is no queue, it will be equal to the inflow rate, and if there is a queue, then it will be equal to the capacity.

$$f_e^-(\xi + \tau_e) = \begin{cases} f_e^+(\xi) & \text{if } z_e(\xi) = 0 \\ \nu_e & \text{if } z_e(\xi) > 0 \end{cases} . \tag{3}$$

Note that we implicitly assume that the link e feeding into the bottleneck has a zero length. Where this is not realistic, we could introduce an uncongested link directly upstream of e to maintain the assumption of a zero travel time of e. We will only consider *vertical* queues: there are no constraints on the length of a queue, and it does not interfere with traffic not using the link in question (there is no *spillback*). So one can equivalently think of the queue as being formed at the exit of the (spaceless) link.

Once flow leaves a particular arc $e = vw$, it must immediately depart on arcs leaving w; no waiting at nodes is allowed.[4] This is captured by the following *flow conservation* constraints: at any moment ξ, and any v aside from s and t, the total flow arriving at v at time ξ equals the total flow leaving. In other words, define the *net flow*

$$\nabla_v(\xi) := \sum_{e \in \delta^-(v)} f_e^-(\xi) - \sum_{e \in \delta^+(v)} f_e^+(\xi),$$

where $\delta^+(v)$ (respectively $\delta^-(v)$ denotes the set of arcs leaving (respectively entering) node v. Then $\nabla_v(\xi) = 0$ for all $v \in V \backslash \{s, t\}$ and all $\xi \in \mathbb{R}$. The value $-\nabla_s(\xi)$ describes the departure rate at time ξ, and need not be zero.

Given a collection $(f_e^+)_{e \in E}$, corresponding outflow functions f_e^- and queue functions z_e can be deduced from (2) and (3). If $(f_e^+)_{e \in E}$ and $(f_e^-)_{e \in E}$ together satisfy the flow conservation constraints, we say that $(f_e^+)_{e \in E}$ is a *flow over time*. The *total value* of this flow is defined to be $-\int_{-\infty}^{\infty} \nabla_s(\xi) d\xi$; this is the total mass of particles which depart from s.

User costs and choices Each individual user is considered to control a negligible fraction of the total flow, so that there are an infinite number of infinitesimal, atomistic and purely price-taking users. We think of each user as being able to freely choose both their departure time, and their route through the network. The joint choices of all these users should then induce a corresponding flow over time. We will restrict our attention to a setting with completely homogeneous users. Firstly, all users will depart from s and arrive at t; secondly, all users experience the same disutility for a given arrival time and travel time, consisting of two components:

- a **travel time** cost, precisely the time between the departure of the user from the origin and their arrival at the destination, scaled by a factor α (the value of time); and
- a **scheduling cost**, based on the arrival time of the user compared to a fixed desired arrival time T^*. More precisely, users arriving at a time $r \leq T^*$ experience a cost of $\beta(T^* - r)$, and users arriving at a time $r > T^*$ experience a cost of $\gamma(r - T^*)$. We will write r for the scheduling cost associated with arriving at time r.

When pricing is implemented, a (possibly time-varying) toll constitutes a third type of disutility. It is part of the generalized *price* of travel, but since it constitutes a monetary transfer, it does not make up a societal *cost* of travel.

The value of α should be strictly larger than β, so that for a fixed departure time, a user would always want to arrive as early as possible to minimize their cost, and no detouring to postpone arrival is induced. We will use $\alpha = 2, \beta = 1, \gamma = 3$ (which imply quite standard ratios between these shadow prices α, β and γ) throughout (our results can easily be reproduced for other reasonable choices).

Dynamic user equilibria A *dynamic user equilibrium* (or simply *equilibrium*) is a joint choice amongst all the users in the system of routes and departure times, with the property that no user has a unilateral deviation that strictly decreases their disutility. This results in a flow over time, which we will call the *equilibrium flow*. Since all users have the same strategy set (i.e., the same collection of routes and departure times) and are homogeneous, all users should experience the same disutility in a deterministic dynamic user equilibrium.

Let us introduce some notation in order to detail the structure of the equilibrium flow, and for later reference. Consider some flow over time $(f_e^+)_{e \in E}$. Let Q denote the total value of this flow; this is the total "mass" of users, measured in some arbitrary unit. The rate of users departing from s at time ξ is given by

$$\rho(\xi) := \nabla_s(\xi).$$

(Note that a user might "depart" s at some moment but then immediately be forced to wait on an arc.)

Now define the *earliest arrival functions* ℓ_v, for $v \neq s$, such that $\ell_v(\xi)$ is equal to the earliest time a particle can arrive at v, given that it leaves s at time ξ. Note that if $(f_e^+)_{e \in E}$ is an equilibrium flow, then no user departing at time ξ would arrive to a node v at any time later than $\ell_v(\xi)$ (otherwise, the user could take an alternate route to arrive at v, and hence t, earlier than it currently does). Let us be slightly more precise. Say that an arc $e = vw$ is *active* at entrance time ξ if $\ell_w(\xi) = T_e(\ell_v(\xi))$. This means that it is possible for a particle leaving s at time ξ to arrive at w as early as possible, namely at time $\ell_w(\xi)$, by a path that includes the arc e. Then the following condition is a requirement for a flow to be an equilibrium:

[4] One could modify the model to allow waiting, but since this would never be desirable for any individual user, as long as schedule delay costs are continuous over time as we assume, this would not impact equilibrium behavior. With step tolls though, users may have an incentive to wait before entering a link, until a downward step in a toll has been made.

Condition 1. for all $e \in E$ and $\xi \in \mathbb{R}$ for which $f_e^+(\xi) > 0$, e is active at entrance time ξ.

Condition 1 ensures that all users are satisfied with their choice of route, given their choice of departure time. For the setting of exogenously given demand, this is the entire description of an equilibrium; see (Cominetti et al., 2015). We must add to this a condition that ensures all users are satisfied with their choice of departure time as well. Thus we have

Condition 2. for any ξ and ξ', where in addition $\rho(\xi) > 0$, we have

$$C_{\text{sched}}(\ell_t(\xi)) + \alpha(\ell_t(\xi) - \xi) \leq C_{\text{sched}}(\ell_t(\xi')) + \alpha(\ell_t(\xi') - \xi').$$

Equilibria do always exist in this model, as long as $\beta < \alpha$[5]. Much more can be said about the structure of equilibrium flows, as well as computing them. We will explore this a little further in the next section.

Tolls To discuss throughput hypercongestion, it is helpful to consider optimal tolls as a means to decentralize the societal (system-)optimum outcome. Our tolls will be charged on a per-link basis, and moreover, we will allow them to be time-varying. However, all users traversing a link at the same moment incur the same toll on the link. The only other restriction we will have is that tolls must be nonnegative. Under our assumption of inelastic overall demand, this assumption is innocent (for optimal tolls with price-sensitive demand, it will still be satisfied).

So formally, we describe the tolling scheme on a link e by a function $\delta_e : \mathbb{R} \to \mathbb{R}_+$. Then $\delta_e(\xi)$ will be the toll charged to a flow particle *exiting* the link at time ξ.

The essential notion of a dynamic user equilibrium in the tolled setting remains unchanged, aside from the adjustment to the disutility of a user. No user should be able to change their route or departure time in a way that decreases their *generalized price*, which is the sum of the travel cost, scheduling cost, and tolls paid by the user. All users should experience the same generalized price in a tolled equilibrium. The *generalized cost* for a user refers to the sum of travel costs and scheduling costs only; note that this may differ between users in the presence of tolls. Optimal (first best) tolls are tolls which minimize the average generalized cost of the users, amongst all possible tolls. Note that this does not in general minimize the average generalized *price*.

Hypercongestion We now return to the two types of hypercongestion discussed earlier, and give precise formal definitions.

Speed-flow hypercongestion. Consider the evolution of the (untolled) equilibrium. Plot the inverse of the travel time $\ell_t(\xi) - \xi$, against $\nabla_t(\ell_t(\xi))$, the inflow rate into t at time $\ell_t(\xi)$, for all choices of ξ. The inverse travel time can be viewed as a measure of speed across the network as a whole.[6]

This plot is a macroscopic analogue of Fig. 1(d). Speed-flow hypercongestion is manifest by the appearance of a backward-bending curve in this figure; a region where the speed (inverse travel time) and the network throughput (flow rate into t) are both decreasing.[7]

Throughput hypercongestion. This is straightforward: if the generalized price at equilibrium under optimal tolls (recall this is identical for all users) is strictly smaller than the cost (including both travel and scheduling costs) experienced by all users in the untolled equilibrium, throughput hypercongestion is exhibited, reflecting a reduction in the peak duration and, equivalently, an increase in the time-averaged (over the full peak) arrival rate.

3. An instance exhibiting hypercongestion

In this section we present an instance (Fig. 2) where both forms of hypercongestion occur, and we describe the evolution of the dynamic equilibria both when no tolls are present, and when first-best pricing is applied.

Evolution without tolls For concreteness, we will fix the transit times, link capacities, and the total mass Q to fixed values (chosen for numerical convenience). Precisely the same qualitative behavior holds for quite a large set of parameter choices; we discuss this further in Section 4. Recall that we use $\alpha = 2, \beta = 1, \gamma = 3$. Let $Q = 1760, T^* = 75,$ and capacities and transit times as in Fig. 3.

The evolution of the dynamic equilibria when no tolls are present can be described through six *phases* (see Fig. 4). A single phase corresponds to a time interval where (1) the rate of users departing s is constant, and (2) these users all make the same aggregate route choices, i.e., the same fraction of users take any given path, for any given departure time in this interval. Within a phase, queue waiting times on any given arc change *linearly* over time. Of most interest to us will not be the rate of change of the queue length with respect to time, but *with respect to the time of departure from s*. Since a user departing from s at time ξ would arrive at a node v at time $\ell_v(\xi)$, the rate of interest for a queue on arc vw is precisely $\frac{1}{\nu_{vw}} \frac{dz_{vw}(\ell_v(\xi))}{d\xi}$. This quantity is shown for each queue in each phase in Fig. 4; when it is positive, the queue is growing, and when it is negative, it is depleting.

[5] We can prove this by slightly adjusting the existence proof by Cominetti et al. (2015) for the exogenous demand setting (meaning that agents cannot choose their departure time; the rate of flow leaving the source is thus a fixed function of time). We will not discuss the details in this paper.

[6] But note that it does not represent an averaged physical travel speed, since the actual distance travelled depends on the route the user takes through the network.

[7] One could also consider plotting some measure of density against network outflow, which would more directly correspond to the macroscopic fundamental diagram of Fig. 1(d). The main difficulty is providing a generally appropriate definition of density that can sensibly be compared to network throughput at a specific moment in time. The total number of users in the system at a particular time has little direct relationship to the inflow rate at t at the same moment in time. This is not an issue in the bathtub model, because of the assumed spatial homogeneity. While other definitions are possible, we prefer to stick to a single notion.

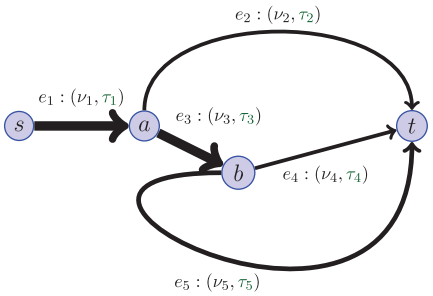


Fig. 2. Instance where both forms of hypercongestion occur; the label of an arc represent the capacity (first element) and the free transit time (second element). Thicker arcs represent links with relative high capacity, and longer arcs represent links with relative long transit time as assumed in the numerical example.

	e_1	e_2	e_3	e_4	e_5
Capacity (ν_e)	30	10	20	10	20
Travel time (τ_e)	0	5	0	0	25

Fig. 3. Remaining parameters for the instance considered in this section.

We now describe the six phases of the equilibrium.

(a). The first phase starts when the network is empty and the first agent departs, at time 6, taking the unique shortest path $e_1e_3e_4$. Per unit of time, 20 drivers take this path and hence a queue grows on e_4 at rate 1: this means that a particle of this phase appearing at s an ε amount of time later than the first one, experiences a queue waiting time of ε on e_4. Notice that the cost of these two particle is the same since the latter pays 2ε more in travel delay cost and, arriving 2ε time later, pays 2ε less in scheduling cost. (b). At time 11, when the queue waiting time on e_4 equals the free-flow transit time of e_2, the situation changes. If phase 1 were to continue, the travel time on e_4 would continue to increase, meaning that e_1e_2 would be a strictly shorter route from a to t. Instead, users start to take the path e_1e_2 in addition to the path $e_1e_3e_4$. At this point the inflow into the network also increases, from 20 to 40, with the effect that queues simultaneously grow on e_1, e_2 and e_4. Notice that, along these two paths, the total queue waiting time grows at rate 1 and hence all the particles of this phase incur the same cost, equal to that in phase (a). (c). The situation changes again at time 40.5. The agent that departs at this moment arrives at the destination at time T^*. Since the penalty for being late is non increasing over (arrival) time, the total travel time has to decrease swiftly for agents departing from this moment onward. As a consequence, the inflow into the network drops to 8. In this phase the queue on e_1 starts to deplete while the queues on e_2 and e_4 continue to grow. This happens since the capacity (and hence the outflow) of the arc e_1 is bigger than the capacity of the arc e_2 and e_4 combined.

From this phase onward, the total queue waiting time decreases at rate $\frac{3}{5}$: a particle departing an ε amount of time later pays $\frac{3}{5}\alpha\varepsilon = \frac{6}{5}\varepsilon$ less in travel delays and $\gamma \cdot \frac{2}{5}\varepsilon = \frac{6}{5}\varepsilon$ more in scheduling cost, keeping average cost constant as required for equilibrium.

The instance has been chosen so that due to the growing queue on e_4, the travel time on e_4 eventually equals the free-flow travel time on e_5, at which point the phase ends. (d). The next, fourth, phase is crucial and starts at time 43. The path $e_1e_3e_5$ starts being utilized, alongside all the paths that were used in the previous phase and thus the inflow rate increases to 12. All the drivers of this phase will split evenly among the three routes. At this point the queue lengths on e_2 and e_4 stay constant while the queue on e_1 continues to deplete. Note that this phase will cause an inflow rate into t particularly high; equal to 30, the capacity of e_1. The phase ends once the queue on e_1 empties. (e) and (f). In the final two phases, all remaining queues gradually deplete, and route choices consolidate to the shorter paths. Phase (e) starts at time $56 + \frac{1}{3}$ and phase (f) at $89 + \frac{2}{3}$. The final agent to depart arrives at t just at the moment when the queue on e_4 depletes, at time 98, thus experiencing again an empty network.

Fig. 5, displaying how average cost evolves over time for routes when they are and when they are not used, confirms that the patterns described jointly provide a dynamic equilibrium, in departure time *and* route choice.

The outflow over time is shown in figure Fig. 6(a). After T^* the outflow of the network does not behave monotonically: it first increases and then decreases. Since the schedule delay cost increases over time, after T^* the travel time of the agents arriving at t decreases over time. Hence, after T^*, we have an increase in the outflow with an increase of speed (inverse travel time), as shown in Fig. 6(b). Therefore, the instance exhibits speed-flow hypercongestion.

This result does not rely on our choice of measuring the passing flow at the sink of the network; it is in fact quite robust. For instance, we see exactly the same behavior if we compare the *inflow* into the network at a given moment with the speed (inverse transit time) of the agents departing at the same moment. The inflow over time is shown in Fig. 7(a); after time 40.5 (which is the departure time of the agent arriving at destination at time T^*), we again see that there is a

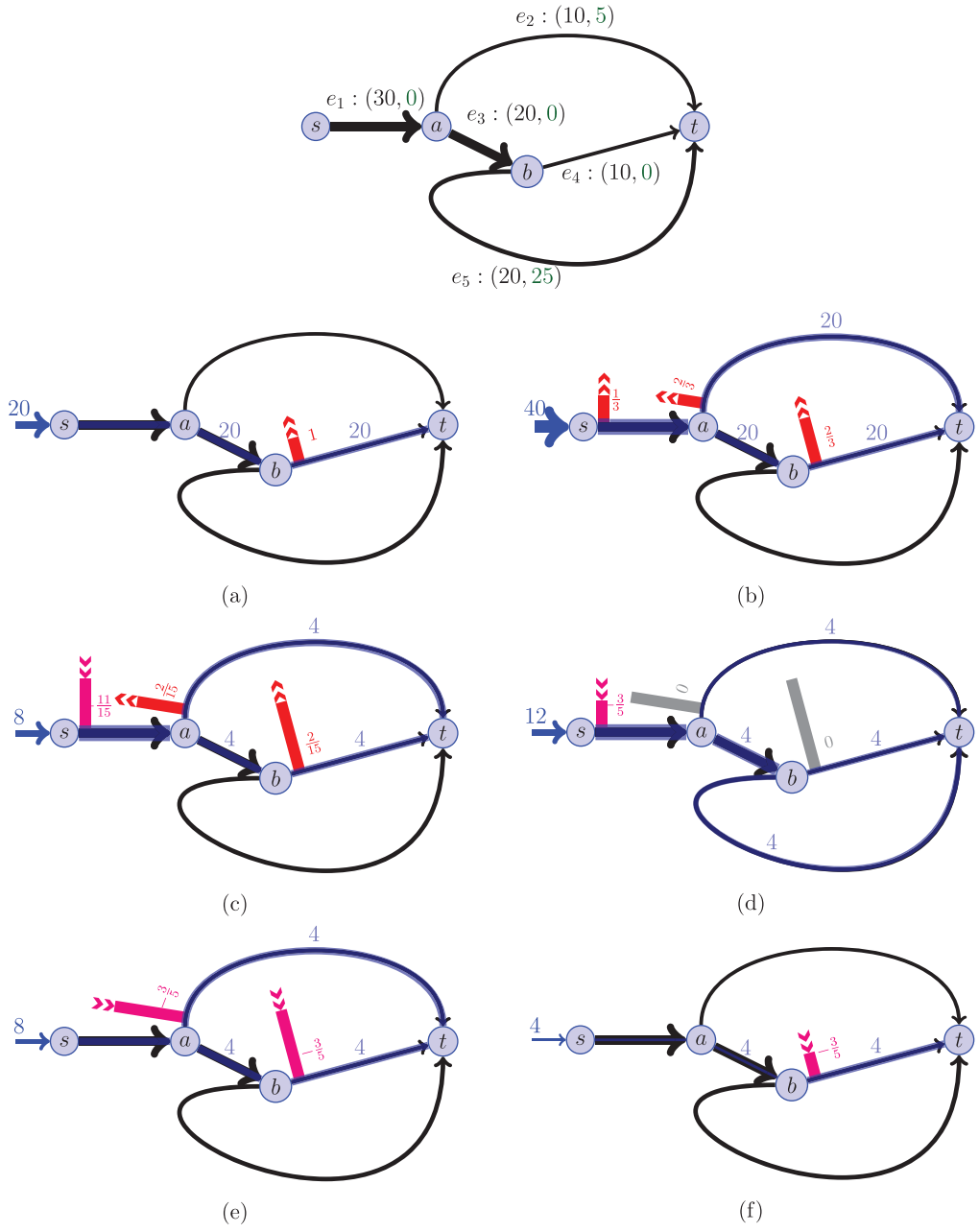


Fig. 4. Chronological evolution without tolls: in blue the inflow into the network, the routes that the drivers take and how these drivers split among these routes; in red, purple and gray the queue waiting times that change for consecutive drivers and their rate of change: in red the ones that increase, in purple the ones that decrease and in gray the ones that stay constant. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

moment where the inflow increases. Since the travel time (meaning the inverse speed) of agents departing from s decreases from time 40.5, we again observe an increase in the inflow with an increase of speed (Fig. 7(b)).

Remark. The textbook exposition of the backward-bending speed-flow relation (Fig. 1a) assigns at most one flow level to each positive speed. This is not the case for the effective speed-flow relation in our example. Speeds (inverse travel times) within an appropriate interval are seen twice; once before the desired arrival time and once after. For some speeds, the arrival rates at these two corresponding moments differ. This does not occur in the setting of a single link, or multiple parallel links, and represents another noteworthy feature of general networks from the conceptual viewpoint. Moreover, this could be one reason—among others—that empirical plots of flow versus speed display so much scatter.

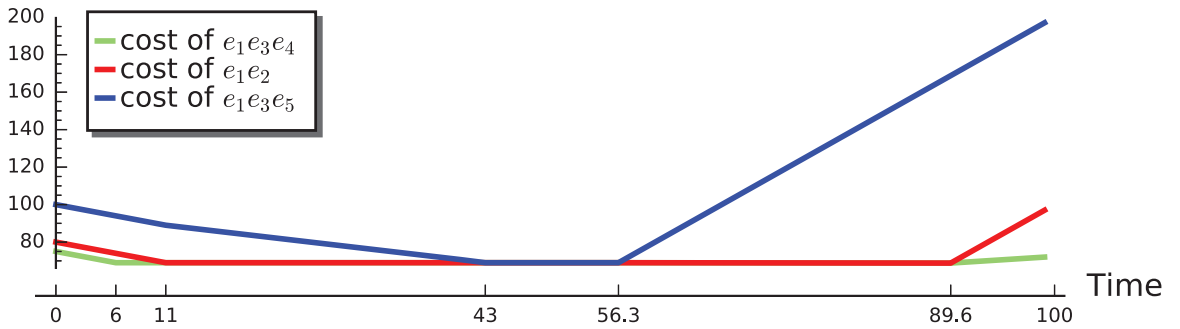


Fig. 5. Commuting cost of the different routes under the no-toll equilibrium. The horizontal axis represents the departing time. A route is active when its cost is equal to the minimum.

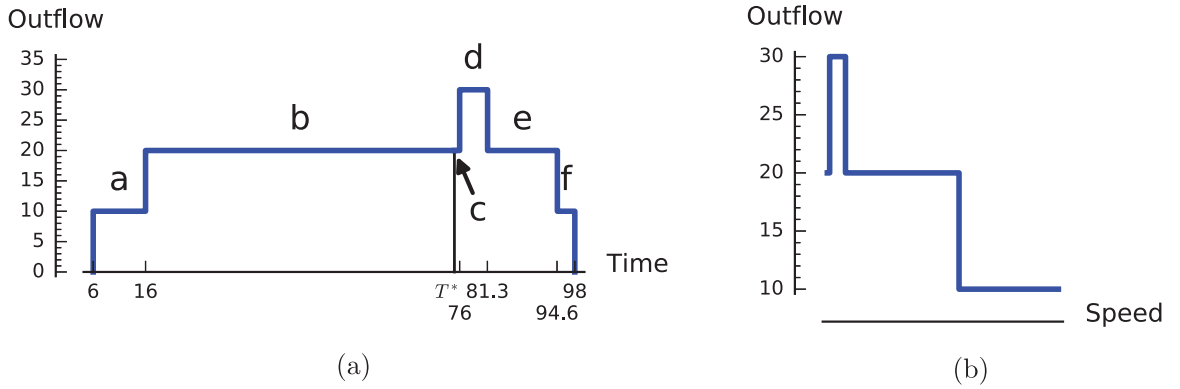


Fig. 6. On the left the plot showing the outflow of the network in any given time; the letters indicates the phase to which the outflow is associated to. On the right the plot indicating the relation between the outflow of the network and the speed (inverse travel time) for the agents arriving after T^*.

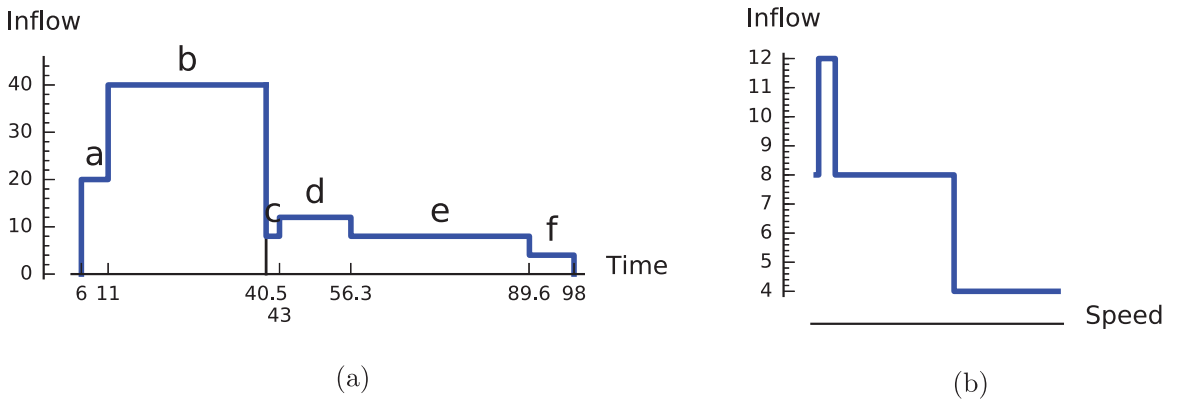


Fig. 7. On the left the plot showing the inflow of the network in any given time; the letters indicates the phase to which the outflow is associated to. On the right the plot indicating the relation between the inflow of the network and the speed (inverse travel time) for the agents departing after time 40.5.

Evolution with optimal tolls Under optimal pricing all three paths from s to t are used, and no queue is ever formed. An example of optimal pricing is obtained by tolling one arc per path and by increasing these tolls linearly at rate β until time T^* and then decreasing it at rate $-\gamma$ (see Fig. 8). This way each tolled arc is in a unique path, and that path is used continuously in the time interval where its arc is tolled. Chronologically, the evolution can be described through six phases. Each phase indicates a time interval in which the particles arriving at t take the same paths. The last three are mirrored but scaled version of the first three: the scaling is due to the ratio of β and γ. In the first and last phase (Fig. 9(a) and (f)) agents arrive at t only from the path $e_1e_3e_4$; in the second and fifth phase (Fig. 9(b) and (e)) they come from paths e_1e_2 and $e_1e_3e_4$ and in the third and fourth (Fig. 9(c) and (d)) they arrive from all the paths e_1e_2, $e_1e_3e_4$ and $e_1e_3e_5$.

In both the equilibria the first agent that arrives at t takes the same path $e_1e_3e_4$ and incurs no cost in tolls, but the arrival time is later when first-best pricing is applied: 6 under no toll and 11 under optimal pricing (Fig. 10). This implies

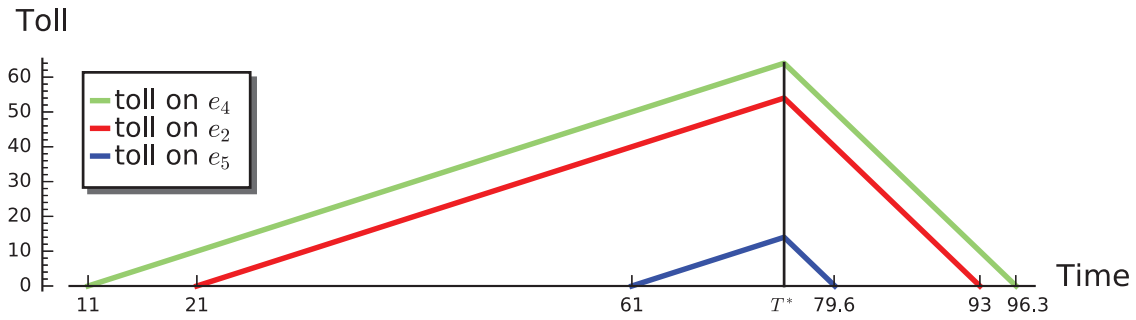


Fig. 8. Example of optimal tolls for the studied instance.

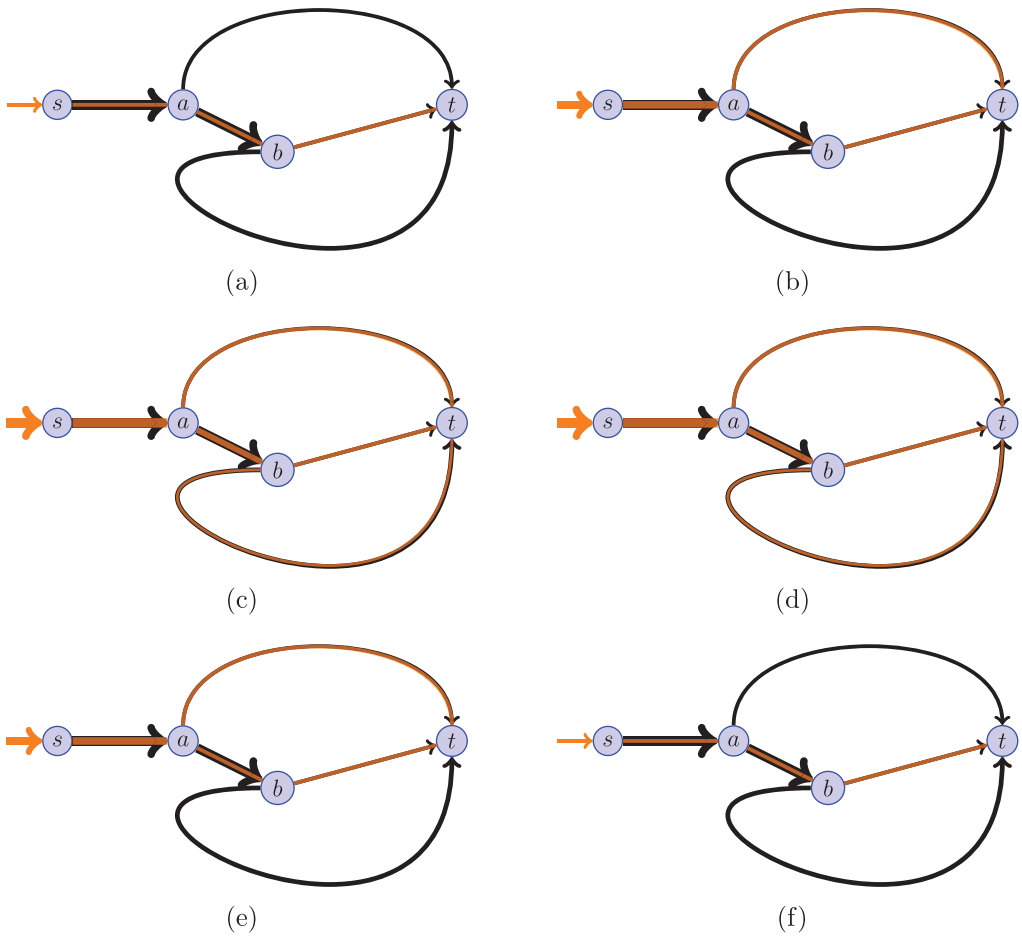


Fig. 9. Chronological evolution with tolls: in orange the routes that the agents arriving at t have taken.

that her general price, and hence that the general price of all the agents, is lower under first-best pricing. Similarly, the very last driver arrives at 98 under no toll and $96 + \frac{1}{3}$ under optimal pricing (Fig. 10). With optimal tolls, the peak duration is therefore shorter and, as a consequence, the time-averaged throughput is higher. This occurs because the path $e_1e_3e_5$ is active over a longer period with optimal tolling than without. Thus, the instance also exhibits throughput hypercongestion: the time averaged arrival flow is higher, the arrival window shorter and the generalized price is lower in the optimum compared to the no-toll equilibrium.

4. Sensitivity analysis

Consider again the network of Fig. 2, that we reproduce here for convenience:

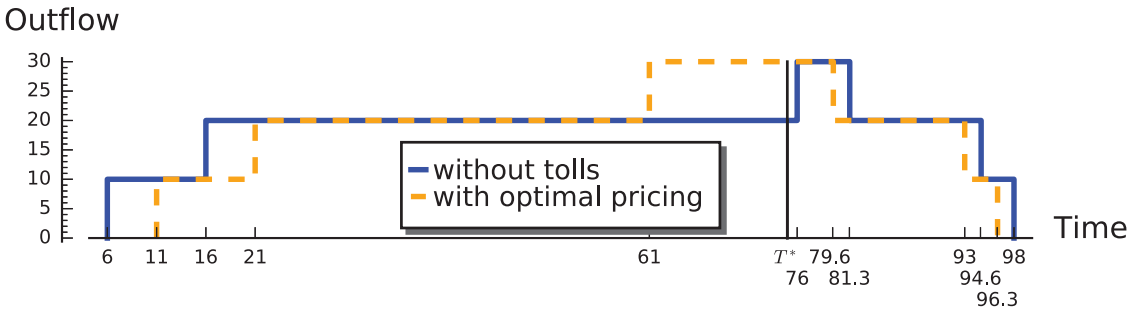


Fig. 10. Outflow over time of the no-toll and of the best-pricing equilibria.

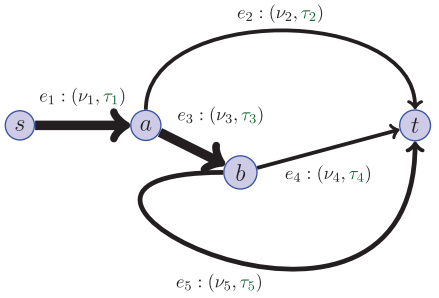


Fig. 11. Instance where both forms of hypercongestion occur.

The following theorem specifies a set of parameter choices for which both forms of hypercongestion occurs in this network.

Theorem 1. *Suppose all the following conditions are satisfied, for an instance of the network shape shown in Fig. 11:*

$$\tau_3 + \tau_4 < \tau_2 < \tau_5 \tag{4}$$

$$\frac{\alpha}{\alpha - \beta} \cdot \nu_4 \leq \min\{\nu_1, \nu_3\} \tag{5}$$

$$\nu_3 \geq \max\{\nu_1 - \nu_2, \frac{\alpha}{\alpha + \gamma}\nu_4\} \tag{6}$$

$$\nu_5 \geq \nu_3 - \nu_4 \tag{7}$$

$$\nu_2 + \nu_4 < \nu_1 < \frac{\alpha}{\alpha - \beta} \cdot (\nu_2 + \nu_4). \tag{8}$$

Then there exists a non-empty interval[8] *of total demand for which both speed-flow and throughput hypercongestion occur.*

Notice that constraint (4) says that $e_1e_3e_4$ is the shortest path and $e_1e_3e_5$ is the longest one; constraint (5) says that e_4 has relative small capacity while constraints (6) and (7) say that e_3 and e_5 cannot have a capacity that is too small; finally, constraint (8) gives a lower and an upper bound to the capacity of e_1.

Theorem 1 implies that after fixing the graph structure for the instance, there is a range of parameters that produce the two forms of hypercongestion, confirming that the occurrence is not a peculiarity that requires a very specific, fine-tuned combination of parameters.

From an efficiency viewpoint, throughput hypercongestion most clearly reflects the inefficiency involved. The fact that time-averaged throughput is in fact higher in the optimum, and the generalized price consequently lower, adds a clear second "dividend" to the implementation of pricing, on top of the well-known favorable impact on queues. It also suggests that social support for pricing might be higher than often thought, as also without recycling of tax revenues and also if bottleneck capacities exhibit no "drop" due to queuing, users may already benefit from the imposition of optimal tolls. It is therefore of interest to see how this type of hypercongestion varies with some key parameters of the model.

Fig. 12 shows the ratio, for the instance of Fig. 3, between the generalized price of the equilibrium without tolls and the generalized price of the equilibrium with optimal pricing as a function of the total users demand, using $\alpha = 2, \beta = 1, \gamma = 3$.

[8] Such an interval is (I_a, I_b) with $I_a = (\tau_2 - \tau_3 - \tau_4)\alpha\nu_4(\frac{1}{\beta} + \frac{1}{\gamma}) + (\tau_3 + \tau_5 - \tau_2)(\nu_2 + \nu_4) \cdot (\frac{\alpha}{\gamma} + \frac{\nu_1}{\nu_1 - \nu_2 - \nu_4})$ and $I_b = I_a + (\tau_3 + \tau_5 - \tau_2)\frac{\nu_1}{\gamma} \cdot \frac{\alpha\nu_2 + \alpha\nu_4 - \alpha\nu_1 + \beta\nu_1}{\nu_1 - \nu_2 - \nu_4}$.

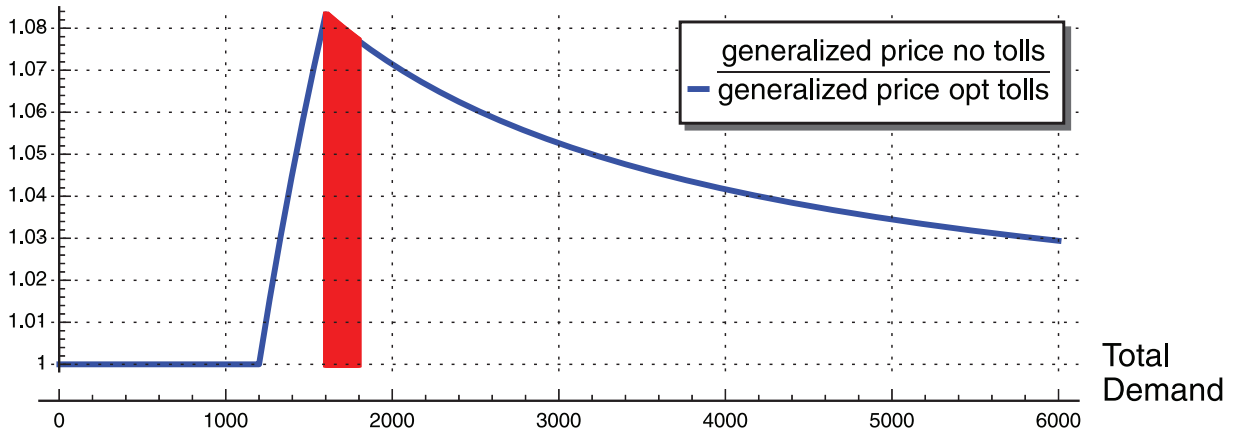


Fig. 12. Ratio between the generalized price of the equilibrium without tolls and the generalized price of the equilibrium with optimal pricing.

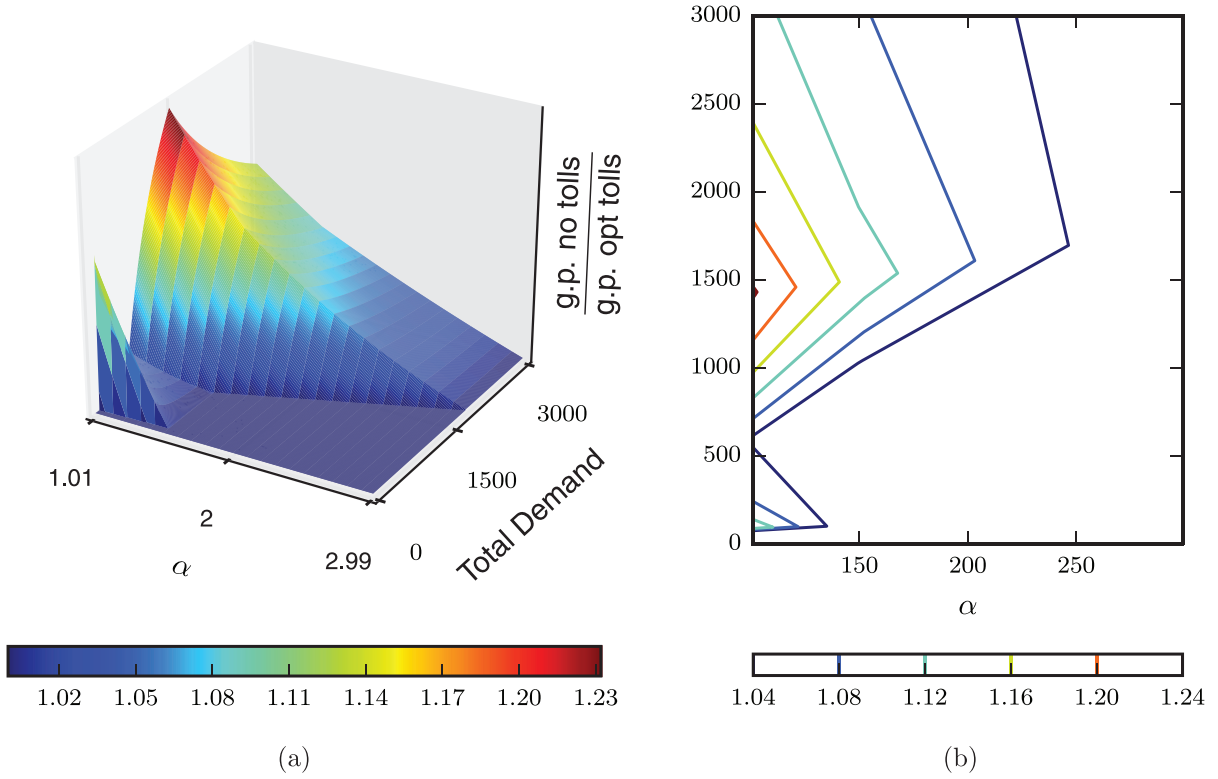


Fig. 13. The ratio between the generalized prices of the equilibrium without tolls and the equilibrium with optimal pricing as a function of α and the total demand; (b) shows a contour plot.

At first, the ratio is equal to 1. This scenario reproduces the behavior on a single bottleneck link, since all the agents takes only the shortest path. As the total demand increases beyond a critical value of around 1200, the generalized prices ratio first increases and then decreases, but it always remains strictly above 1. The red region demarcates the boundary of the range of choices for the total demand where speed-flow hypercongestion occurs. This shows that throughput hypercongestion can occur while speed-flow hypercongestion does not.

Fig. 13 shows, again referring to the instance of Fig. 3, the ratio between the generalized price of the equilibrium without tolls and the generalized price of the equilibrium with optimal pricing as a function of α and the total demand. Here the value of α varies in $(\beta, \gamma) = (1, 3)$. The ratio tends to increase when α becomes smaller, which is intuitive as a lower α would lead, ceteris paribus, to longer equilibrium queues. Given α, we observe a maximum ratio for some intermediate value of total demand, just as in Fig. 12.

	e_1	e_2	e_3	e_4	e_5
Capacity (ν_e)	9	4	6	4	4
Travel time (τ_e)	0	20	0	0	20

Fig. 14. Further parameters for an instance exhibiting speed-flow, but not throughput, hypercongestion.

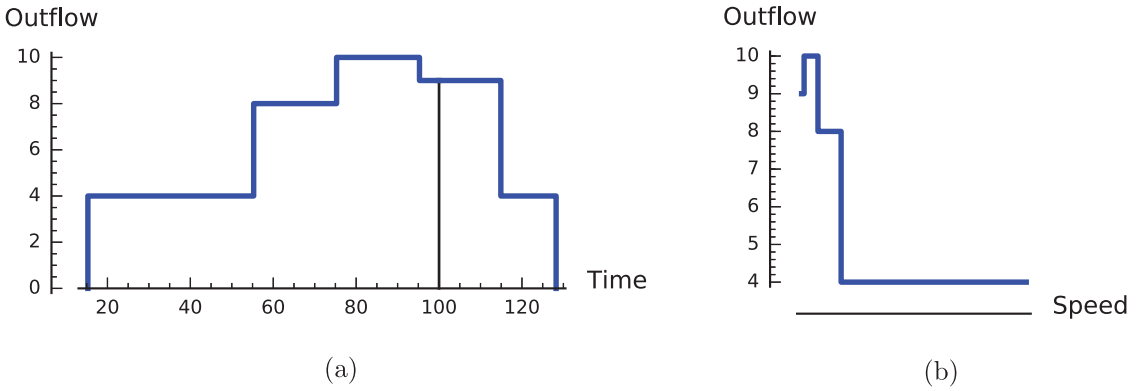


Fig. 15. Speed-flow hypercongestion of the instance described in Fig. 14: On the left the outflow in any given time; On the right the relation between the outflow and the speed (inverse travel time) for the agents arriving before T^*.

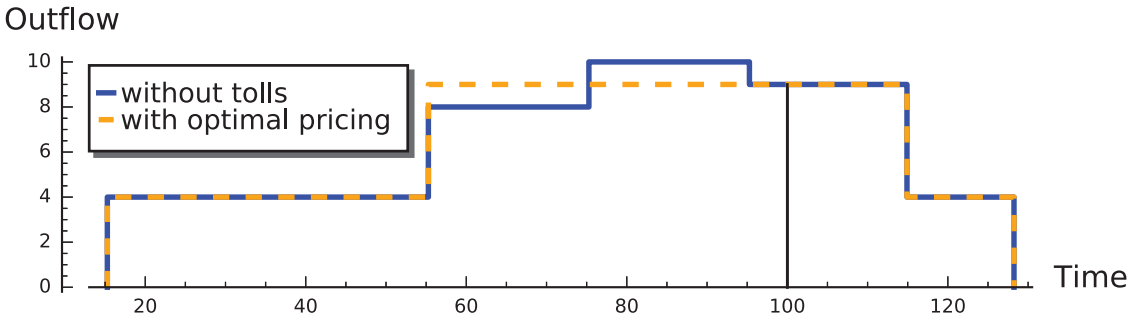


Fig. 16. Outflow over time of the equilibrium of the instance described in Fig. 14, both without tolls and under optimal pricing.

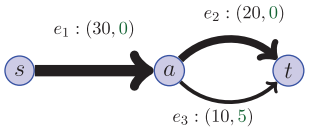


Fig. 17. The label of an arc represent the capacity (first element) and the free transit time (second element).

Separating the two forms of hypercongestion, and hypercongestion before the peak As already noted, the previous example shows that throughput hypercongestion can occur without the presence of speed-flow hypercongestion. The opposite can also happen: speed-flow hypercongestion can occur while throughput hypercongestion does not. An example is obtained on the network of Fig. 11 with $Q = 750$, $T^* = 100$ and capacities and transit times shown in Fig. 14.

Figs. 15 (a) and 15(b) shows that the instance exhibits speed-flow hypercongestion: there is a period of time before the desired arrival time where the throughput in the equilibrium decreases. Since, at this point, the travel time of the agents arriving at t increases over time, we have a decrease in the outflow with a decrease of speed (inverse travel time). Fig. 16 shows that the instance does not exhibits throughput hypercongestion.

This last example also demonstrates that speed-flow hypercongestion can occur *before* the peak (recall that our main example in Section 3 exhibited speed-flow hypercongestion after the peak). Since in the empirical literature on hypercongestion at the network level (e.g., Geroliminis and Daganzo (2008); Daganzo et al. (2011)), the effects do not seem to be restricted to occurring only after (or only before) the peak, this is important.

Simpler network topologies The examples discussed so far are not the smallest ones to exhibit hypercongestion. The following instance (Fig. 17) with three links also exhibits both forms of hypercongestion, with the usual choice of α, β, γ along with $Q = 400$. We have chosen $T^* = 15$.

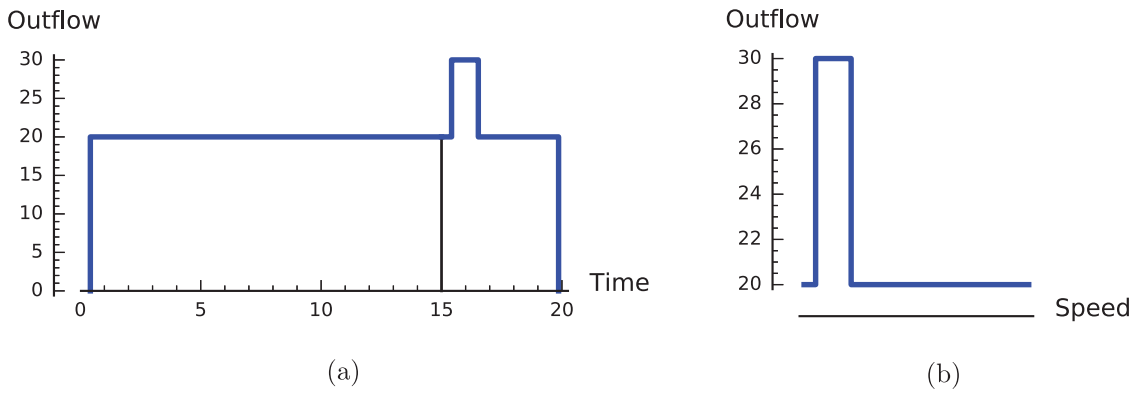


Fig. 18. Speed-flow hypercongestion of the instance described in Fig. 17: On the left the outflow of the network in any given time; On the right the relation between the outflow of the network and the speed (inverse travel time) for the agents arriving after T^*.

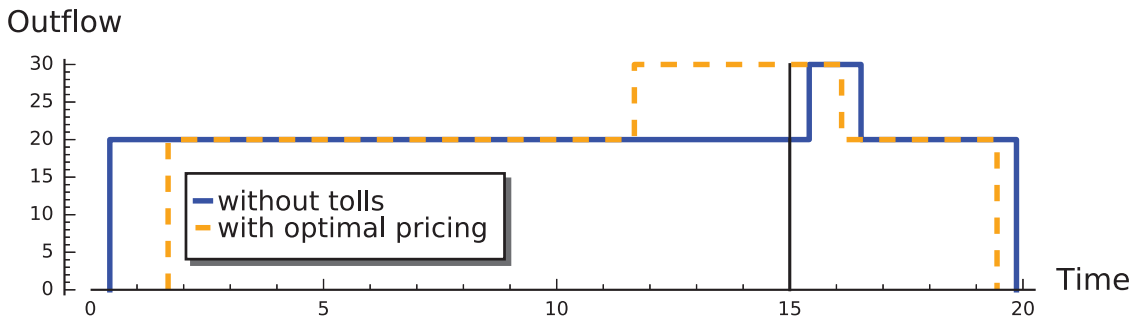


Fig. 19. Outflow over time of the equilibrium of instance described in Fig. 17, both without tolls and under optimal pricing.

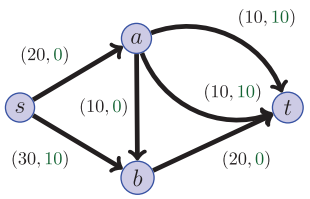


Fig. 20. Instance where optimal tolls increase the generalized price.

Figs. 18 (a) and 18(b) shows that the instance exhibits speed-flow hypercongestion, and Fig. 19 that it exhibits throughput hypercongestion.

The fact that hypercongestion occurs for such a simple network topology provides further evidence that hypercongestion is not rare for this type of model.

We remark that we have focused on the slightly larger five-arc instance in order to exhibit some phenomena beyond the existence of the two forms of hypercongestion. In particular, our observations that (i) that the two forms of hypercongestion are distinct: either form can occur without the other; and (ii) that speed-flow hypercongestion can occur *before* the peak rely on our larger example.

5. First-best pricing can strictly increase the generalized price

The fact the optimal tolling decreases the generalized price is remarkable, and runs counter intuition. It also is at odds with results from other dynamic models of traffic congestion, including the conventional single-link bottleneck model, where the generalized price does not change when an optimal toll is implemented, and models of flow congestion such as the one proposed by Chu (1995), where the generalized price would increase. That raises the question of whether, for other parameter combinations or networks, the current model could also produce instances where the generalized price rises instead of falls.

In this section we show through an illustration that first-best pricing does not always decrease the generalized price, and can in fact cause it to strictly increase. For this we consider the network of Fig. 20. Fig. 21 shows the ratio between the generalized price of the equilibrium without tolls and the generalized price of the equilibrium with optimal pricing, as a

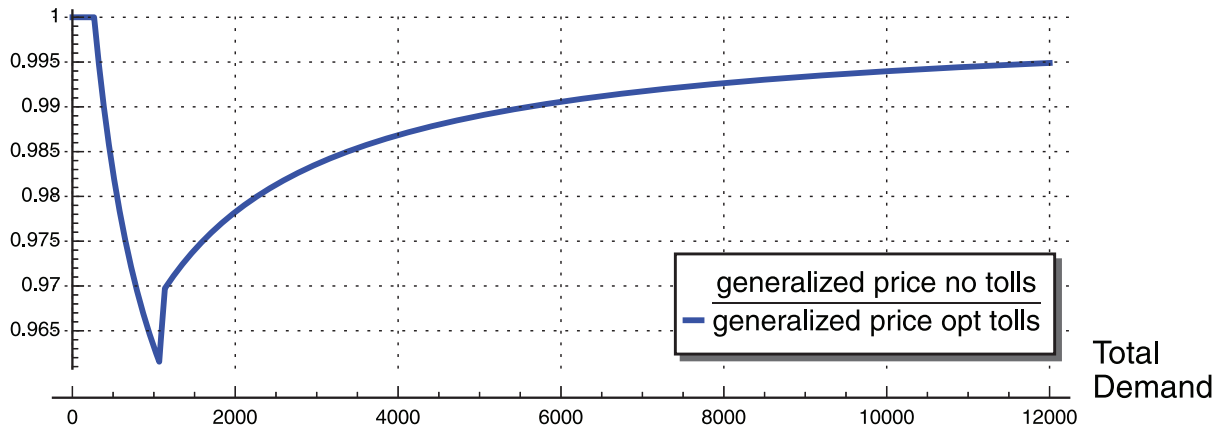


Fig. 21. Ratio between the generalized price of the equilibrium without tolls and the generalized price of the equilibrium with optimal pricing.

function of the total users demand (using $\alpha = 2, \beta = 1, \gamma = 3$ as before). At first, the ratio is equal to 1. This occurs since all the agents takes only the shortest path and hence we have the same behavior we would have on a single bottleneck link. As the total demand increases, the generalized price ratio first decreases sharply, and then increases asymptotically towards 1.

6. Conclusion

In this paper we showed that hypercongestion can occur as a purely emergent effect of the network interaction of multiple selfish users, all aiming to minimize their travel costs. For this we considered link congestion dynamics that do not exhibit hypercongestion and do not produce spillback effects (we considered Vickrey bottlenecks with spaceless vertical queues).

We distinguished two instructive aspects of hypercongestion that generally coincide for regular single-link models. One is related to the "macroscopic" versions of the fundamental traffic diagram and the other is exhibited when the imposition of optimal (first-best) pricing leads to an increase in the (time-averaged) arrival flow at the destination and therefore to a reduction in the generalized price, i.e., the total cost that each agent pays.

These findings are of interest because they show how the same flow can be reached at a higher-speed, lower-density configurations, thus eliminating queuing. Moreover, the second form of hypercongestion that we considered shows how optimal pricing may not only eliminate queuing but in addition decrease generalized prices before toll revenues are redistributed. As opposed to the bottleneck models for heterogeneous users where some users can gain from the imposition of optimal pricing to the detriment of other users' costs, in this paper we assumed homogeneous users and thus the reduction in travel price comes from an increase in the physical network usage. This is an important result for the political and social acceptability of pricing. However, we also showed that this is not always the case and that there are instances where optimal pricing can increase the generalized price. Finally, our pricing policy is not defined to be homogeneous over space, as in the case of other models that study hypercongestion, but it is differentiated over key links and bottlenecks in the whole network, as it is likely to be implemented in real applications.

Fully understanding the causes of hypercongestion as empirically observed remains a challenging task. It is unclear how the form of the network—for instance, it's topology—effects the prevalence of hypercongestion. It is well know that it does not occur in parallel link networks and our construction shows that it may occur in very simple networks, but little is known beyond that. It would be interesting to investigate what happens in the Vickrey bottleneck model on instances more indicative of real city networks, and with multiple origin-destination pairs. It is also natural to ask what impact other specifics of the model have on hypercongestion. For example, one can consider user heterogeneity, in the value of time or in scheduling preferences; or models where there is uncertainty in the delays experienced on links.

CRediT authorship contribution statement

Dario Frascaria: Methodology, Writing - original draft, Writing - review & editing, Formal analysis, Investigation, Visualization, Software. **Neil Olver:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Supervision, Funding acquisition. **Erik Verhoef:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing.

References

Arnott, R., 2013. A bathtub model of downtown traffic congestion. J. Urban Econ. 76, 110–121.

Arnott, R., Kokoza, A., Naji, M., 2016. Equilibrium traffic dynamics in a bathtub model: aspecial case. Econ. Transport. 7–8, 38–52.
Arnott, R., de Palma, A., Lindsey, R., 1990. Economics of a bottleneck. J. Urban Econ. 27 (1), 111–130.
van den Berg, V., Verhoef, E.T., 2011. Winning or losing from dynamic bottleneck congestion pricing? the distributional effects of road pricing with heterogeneity in values of time and schedule delay. J. Public Econ. 95 (7), 983–992.
Chu, X., 1995. Endogenous trip scheduling: the Henderson approach reformulated and compared with the Vickrey approach. J. Urban Econ. 37 (3), 324–343.
Cominetti, R., Correa, J., Larré, O., 2015. Dynamic equilibria in fluid queueing networks. Oper. Res. 63 (1), 21–34.
Daganzo, C.F., Gayah, V.V., Gonzales, E.J., 2011. Macroscopic relations of urban traffic variables: bifurcations, multivaluedness and instability. Transport. Res. Part B 45 (1), 278–288.
Fosgerau, M., 2015. Congestion in the bathtub. Econ. Transport. 4 (4), 241–255.
Friesz, T.L., Han, K., 2019. The mathematical foundations of dynamic user equilibrium. Transport. Res. Part B 126, 309–328.
Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. Transport. Res. Part B 42 (9), 759–770.
Koch, R., Skutella, M., 2011. Nash equilibria and the price of anarchy for flows over time. Theory Comput. Syst. 49 (1), 71–97.
Small, K.A., 2015. The bottleneck model: an assessment and interpretation. Econ. Transport. 4 (1), 110–117.
Small, K.A., Verhoef, E., 2007. The Economics of Urban Transportation. Routledge.
Verhoef, E.T., 2001. An integrated dynamic model of road traffic congestion based on simple car-following theory: exploring hypercongestion. J. Urban Econ. 49 (3), 505–542.
Verhoef, E.T., 2003. Inside the queue: hypercongestion and road pricing in continuous time continuous place model of traffic congestion. J. Urban Econ. 54 (3), 531–565.
Vickrey, W., 1969. Congestion theory and transport investment. Am. Econ. Rev. 59 (2), 251–260.