Modeling and control of congestion phenomena

Levering, N.A.C.

Publication date
2024

Chapter 1

Introduction

1.1 The rise of congestion

In the last fifty years, the world population has doubled to a size of more than eight billion people. The technological developments that took place in the same period of time allow these people to stay connected, both physically and virtually. However, as the networks facilitating these connections have limited capacity, the high connection demand has given rise to various congestion phenomena. Importantly, such congestion occurrences are typically extremely costly, in terms of delay, economic costs, and environmental impact. This has resulted in a key interest in the mathematical study of networks in which congestion plays a prominent role.

The notion of congestion is often associated with road traffic networks. Over the last decades, advances in the technology and production of cars have led to a high rate of car ownership. In combination with the increasing population density, this has made rising traffic congestion an inescapable condition. Especially during rush hours, when a high number of vehicles use the road network to commute from home to work (or vice versa), the capacity of certain roads may be insufficient, leading to the emergence of traffic jams. Congestion may also arise due to traffic events like accidents, as they may temporarily block a part of the road, thereby reducing the capacity of the road network.

In the last era, another application area of congestion control arose with the rapid developments in telecommunication networks. Examples of such networks include telephone networks, in which calls between telephones are routed through servers and controllers, and the Internet, in which packets of data (ranging from emails to big data sets) are sent from one device to another through communication and redistribution devices (e.g., routers, switches and hubs). Although the first mobile phone call was made just fifty years ago, a time in which the Internet was even non-existent, both communication networks are an indispensable component of modern daily life. However, with the high demand of virtual messaging, the bandwidth that exists to transfer messages between devices, or the capacity that devices have to process these messages, may get exceeded. This leads to high waiting times (i.e., a long time for messages to arrive at their destination), or even call or packet losses.

The rise of congestion phenomena in road and telecommunication networks has sparked the interest in the regulation of such networks. Mathematical descriptions for the events that cause congestion provide a useful tool in alleviating the impact of congestion in these net-

works, as they allow the study of different control measures. That is, traffic operators can use road traffic models to limit the effect of both rush hours and traffic incidents through real-time guidance to travelers about routing choices and departure times. Moreover, models for telecommunication networks allow policy makers to find efficient admission, routing, and processing procedures so as to prevent capacity exceedances in these networks.

An important aspect in the modeling of congestion – and a key challenge in the control of congestion – is formed by stochasticity. That is, road traffic dynamics do not follow deterministic laws, as traffic variables such as traffic flow and vehicle speed are influenced by the occurrence of (semi-)random events such as traffic incidents and bad weather conditions. Importantly, the occurrence or impact of these events may even be time-dependent, think of delays during rush hours typically being less severe during summer holidays, and incident probabilities typically being higher during rush hours. Second-order effects that govern unpredictability are due to the heterogeneity in human characteristics (e.g., different driving habits) and vehicle specifications (e.g., vehicle sizes). As there may be a significant impact of both types of randomness on the vehicle dynamics, in terms of optimal control, it is important to incorporate this randomness in road traffic models.

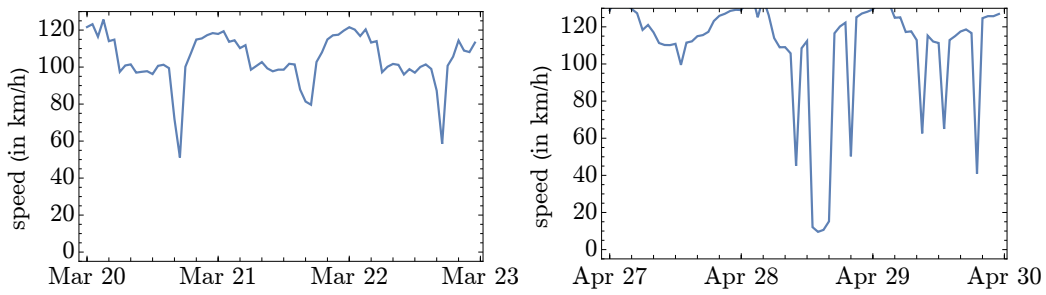
In telecommunications, there are also different processes in which uncertainty may play a role. First, there may be randomness in the arrival stream; for example, the times that packets of data are sent or telephone calls are made are typically irregular. Second, with the size of these packets or the lengths of these calls unknown beforehand, there may be additional uncertainty in the required service or processing times of the arriving jobs. Just as in the road traffic setting, time-dependence may play a role in this context as well. For example, the number of telephone calls is typically higher during the day than around midnight, resulting in a time-dependent arrival rate.

In this work, we focus on the modeling and control of congestion phenomena, thereby taking the impact of time-dependence and stochasticity into account. The first part of the thesis will specifically consider the road traffic setting, and will present stochastic road models and corresponding control problems in the areas of routing, optimal departure-time advice, and input rate control. The second part of the thesis will consider a classical congestion model for an element in a telecommunication network: a single-server queue. To guide control policies that aim to find a social optimum by levying tolls to arriving customers, we study the so-called *externalities*: the cost of additional arrivals to queues, in terms of the total extra waiting time.

The remainder of this chapter, which serves as an introduction to the terminology in the remainder of the thesis, has a similar outline as the thesis itself. First, Section 1.2 discusses the modeling and control of road traffic. Then, Section 1.3 introduces queueing models and the concept of externalities. Section 1.4 provides an overview of the contributions of this thesis and describes the organization of the chapters to come.

1.2 Congestion in road networks

Road traffic congestion has a big impact on today's world. The low speed levels that correspond to congestive settings inflict time and fuel waste, forming a direct cost for individual drivers experiencing these delays. In a 2021 report from the UK National Traffic Opera-



(a) Recurrent congestion.

(b) Non-recurrent congestion.

Figure 1.1: Speeds as measured by two loop detectors on the Dutch A9 highway in 2023. In (a), we observe both the time-dependence and randomness of rush hours. In (b), a speed reduction that is caused by an incident around the corresponding detector is displayed.

tions Centre, it was estimated that only the time lost to delays in the England highway network already costs around 3.5 billion pounds per year [104]. For the population, there are additional indirect costs, as congestion elevates the shipping prices in the transport sector, causing an increase in the consumer prices of many products. Another important consequence of traffic congestion – currently receiving much attention – is the increase of pollutant concentrations, and the resulting decrease in ambient air quality.

There are various ways to tackle congestion in road networks, so as to reduce the impact of the delays on society. For decades, countries have tried to cope with the rising demands by expanding their infrastructure. The Dutch government, for example, increased the total road length from 130.446 km in 2001 to 141.361 km in 2020 [29]. Even if such expansions would have proven effective, their future use is questionable, as they are extremely costly, and there is only limited room that can be used for land-based transportation purposes. Therefore, there is a need for policies that do not require building new infrastructure, but that better exploit existing resources.

Effective control policies for the road network should account for different types of congestion. In general, we distinguish two categories:

- *Recurrent congestion*: congestion caused by events that occur daily around the same time, such as congestion during peak hours (see Figure 1.1a). We refer to these events as *recurrent events*.

- *Non-recurrent congestion*: congestion that does not follow a daily or weekly pattern, but is caused by (semi-)random events such as incidents or bad weather conditions, referred to as *non-recurrent events* (see Figure 1.1b).

Notably, the focus of network expansion is mostly to reduce recurrent congestion, whereas the effect of non-recurrent events is significant as well. The latter is especially true in highway networks, in which incidents may lead to a big reduction in road capacity, and the number of routing options is limited.

An interesting development that offers new potential in the control of both types of congestion in highway networks is the increasing presence of Intelligent Transportation Systems (ITS). These days, most existing infrastructure is equipped with sensors like speed cameras and inductive loop detectors, that can provide real-time information about the

network dynamics. Moreover, via mobile devices and on-board navigation systems, there is the possibility for almost-instant communication with the vehicles on the road. This gives traffic managers the potential to learn the current state of the system, and to use this information to improve and dynamically update the advice given to the vehicles, e.g., the route to travel and the time to depart. This may reduce the time the vehicles spend in the system and, consequently, the total congestion in the network.

The goal of Part I of this thesis is to use the advances in ITS to model and reduce the impact of recurrent and non-recurrent congestion in highway networks, by providing vehicles with continual trip guidance based on the most recent state of the network. To this end, we need to address two major challenges. First, it should be observed that road traffic dynamics are inherently random. First-order (semi-)unpredictable events impacting the speeds vehicles are effectively able to drive are, e.g., traffic surges and incidents. Second, whereas the information load of ITS may be high, the computational costs of the algorithms that derive the control policy must be real-time, as advice regarding optimal routes or departures times must be readily available.

In the remainder of this section, we will introduce the traffic models and control techniques that have a central role in Part I of this thesis. Specifically, Section 1.2.1 highlights road traffic models that could account for the stochastic and dynamic nature of road traffic, and whose tractability and low-computational complexity allow application in congestion control. Then, Section 1.2.2 discusses three important control problems in road traffic: optimal routing, optimal departure times, and input rate control.

1.2.1 Operational models for congestion control

There is a broad literature on traffic models. These models differ at the level of detail (i.e., modeling the dynamics per car or the total dynamics per highway part), and adhere various assumptions on the time-dependence or stochasticity of the set of traffic variables. In this dissertation, as we aim to develop dynamic trip guidance frameworks that can be made operational with relatively low time-complexity, we do not consider models that track each vehicle in the network in a continuous way. Moreover, acknowledging the impact of uncertainty, we focus on stochastic models that either incorporate first-order random events such as incidents, or second-order effects that are caused by the driving style that individual drivers may have.

Specifically, in Part I of the dissertation, we work with three operational traffic models for congestion. First, in Chapters 2-4, we develop a Markov model that tracks the impact of recurrent and non-recurrent events on the speeds vehicles are able to drive. The Markovian nature makes the model tractable, and therefore suitable for control purposes. Second, in Chapters 5 and 6 we work with two stochastic models that are based on the concepts of the so-called *fundamental diagram* of traffic flow. In contrast to the Markov model, such a diagram provides a direct relation between road densities and vehicle speeds, giving a more detailed account of the impact of high demands.

Markov models

Markov processes are important types of stochastic processes that are frequently used in predictive modeling and decision analysis. The Markov process is characterized by the Markov property: the system transitions are only dependent on its current state, not on the prior history. This property often makes Markov models mathematically tractable. In the continuous case, this property is also referred to as the memoryless property, since the Markovian nature implies that the time between transitions follows the exponential distribution. More detail on the theory of Markov processes can be found in the works of Norris [169] and Ibe [109].

In road traffic, Markov processes have been used to track the conditions in a highway network, such as incident occurrences and network demand. As the network dynamics will be inherent to the evolution of these conditions, travel-time measures are modeled as functions of the state of this Markov process. The Markov process therefore serves as *Markov-modulated background process* or *environmental process*. A simple example would be to define a Markov process $M_a(t)$, for each arc a in the network, such that $M_a(t) = 2$ if arc a is congested at time t, and $M_a(t) = 1$ otherwise. Such a framework was suggested by [67, 92, 132, 133, 226], who let the expected travel time on arc a upon departure at time t be a direct function of $M_a(t)$.

In our work, we develop a Markov model that provides a more detailed account of the impact of congestion on the network dynamics. Similar to the Markov model presented by Kharoufeh and Gautam [129], the background process on an arc is considered to continuously model the velocities on that arc, i.e., the state of the Markov process uniquely determines the arc speeds. Then, instead of only a discrete and predefined set of values, realized travel times can take values in a complete range. In this framework, besides the expected travel time on arc a upon departure at time t, higher order travel-time moments may also be found, and are all based on the state of the system at time t. Notably, whereas Kharoufeh and Gautam introduced an elementary model on a single link, our model considers a complete network, is explicitly capable of incorporating both recurrent and non-recurrent congestion, and allows for correlation between the speeds on the network links.

Fundamental diagram of traffic flow

An important subject of study in traffic flow theory is the relation between the flow rate q (in veh/h) and the traffic density k (in veh/km), referred to as the (macroscopic) fundamental diagram of traffic flow. The first empirical relation between these two traffic variables was found in 1930, in a seminal paper by Greenshields [90], who represents the vehicle speed v (in km/h) as a simple linear function of the density, relating flow and density via $q = k \cdot v$ (see, e.g., Figure 1.2).

Under the model of Greenshields, the flow is a parabolic function of both the traffic density and the vehicle speed, categorizing two sets of road conditions. First, there is the *stable* regime, encoding pairs of low densities and high speeds. If we would take such a pair, deduce the corresponding flow rate, and let this rate increase gradually, we will reach the top of the parabolic functions. The flow and density value encoding this top are referred to as the critical flow and critical density, respectively. Increasing the density beyond this

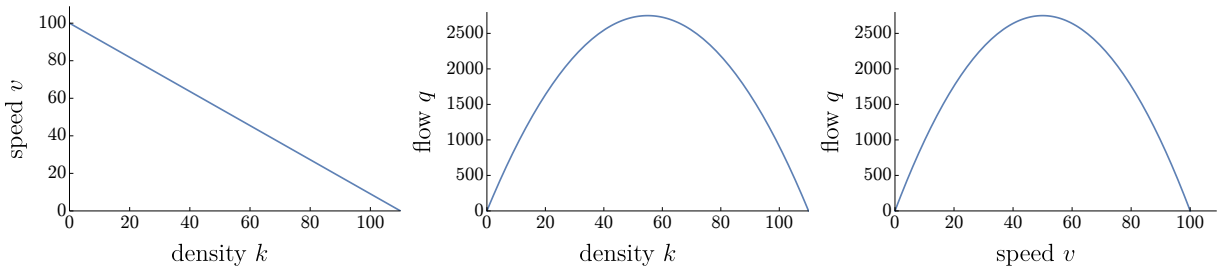


Figure 1.2: An example of the fundamental diagram of Greenshields [90].

critical point would yield values in an *unstable* regime with low vehicle speeds. The density in the unstable regime for which both the speed and the flow equal zero is called the jam density.

After Greenshields, numerous studies have developed a more detailed relation between the traffic variables v, q and k. Some important references are the fundamental diagrams of Greenberg [89] and Edie [61], and the trapezoid fundamental diagram [45]. Notably, these studies are mostly deterministic in nature, such that their diagrams describe the functional form of the relation between the mean flow rate and traffic density. Therefore, they do not account for the intrinsic fluctuations due to, e.g., driver-specific behavior and the various vehicle sizes.

Recently, there have been attempts to develop stochastic counterparts for the fundamental diagram. One line of literature, including the works of Wang et al. [215], Qu et al. [183], and Ni et al. [166], captures the fluctuations by a fundamental diagram in which the link's traffic density is not mapped to a specific velocity, but to a velocity distribution. Chapter 5 uses this modeling perspective. A different perspective, which we will consider in Chapter 6, is to let the road density be a random variable whose value depends on the number of vehicles on that road.

1.2.2 Regulation measures

ITS can learn the current state of the system, in terms of, e.g., the number of vehicles on the road, and the presence or absence of incidents. Traffic operators can use this information to take decisions about the amount of traffic they allow into the network, or to provide vehicles with advice regarding routing and departure time decisions. In Part I of this thesis, we study these control problems, and present corresponding real-time and (close-to-)optimal procedures. Application of these procedures will limit the impact of recurrent and non-recurrent congestion in highway networks.

Before introducing the three control problems in more detail, it is important to remark that procedures that take the dynamic and stochastic nature of highway traffic into account have a clear benefit over most state-of-the-art control policies. Typically, these policies assume demand to follow deterministic patterns, and are therefore unable to cope with the impact of first-order random events such as traffic incidents. Moreover, in terms of vehicle guidance, they often assume a procedure in which the vehicle receives trip advice once, whereas in reality, (almost) continuous updating is possible.

Routing

These days, physical maps have largely been replaced by on-board navigation systems and routing applications on mobile devices. Through an underlying routing algorithm, this software presents users with a path from their current location to a chosen destination. The first known and perhaps most famous routing algorithm is *Dijkstra's algorithm*, published by the Dutch computer scientist Edsger W. Dijkstra in 1959 [55]. This algorithm, which was later recognized to be a special case of a graph algorithm discovered earlier by both Prim and Jarník [206], initially assigns a label to each node in the graph. Starting from the origin, the algorithm then iteratively updates these labels, ending up with an optimal path in a deterministic graph with constant, deterministic, and positive link weights.

Although Dijkstra's algorithm yields a significant computational improvement over naively comparing the weights of all paths in the network, the current sizes of road networks make application of the algorithm still somewhat slow. Therefore, there is an extensive literature on speed-up versions of Dijkstra's algorithm. We specifically mention the A* algorithm [50, 76], that will be used in different chapters of the thesis. Just as Dijkstra, the A* algorithm is of label-updating nature. However, whereas Dijkstra sets a node label to be the shortest known distance between the origin and the node, A* aims to direct the search towards the destination by letting the node label be the sum of Dijkstra's node label and a lower-bound for the distance between this node and the destination. Naturally, the closer the lower bound comes to the true shortest remaining weight, the faster the A* algorithm executes.

Kaufman and Smith [123] have proven that Dijkstra's algorithm (and, as a consequence, the A* algorithm) is still applicable in networks whose link weights are time-dependent, as long as these networks satisfy the *first-in first-out* or *FIFO property*. The latter means that for a vehicle traversing a single network link, in terms of travel time, the arrival epoch is an increasing function of the departure epoch. However, determining the best path is no longer that transparent when the link weights are not only dynamic, but also stochastic, as will be true for real-life highway networks. Importantly, if randomness is involved, there is not one notion of an optimal path. Whereas some vehicles may consider the path with minimum expected travel time as optimal, other vehicles may prefer to travel via the path which maximizes the on-time arrival probability for a given arrival threshold, and other objectives can be thought of.

It has been proven that under some objectives and model assumptions, Dijkstra-based shortest path algorithms may even provide optimal results in the stochastic setting. An example is the minimum expectation objective, under the premise that the expected travel times satisfy the introduced FIFO property. Alternatively, under the same assumption there exist iterative schemes that output the correct path for the on-time-arrival objective.

In the above, the notion of optimal path concerns the setting in which a vehicle requests the path to travel at the origin, and does not deviate from this choice after departure. We refer to this as the *offline* or *static* setting. It is, however, important to observe that the realized link travel times during the trip, and the updates in network state the vehicle receives, may provide an incentive for the vehicle to switch to an alternative path midway. For example, if the vehicle learns that there is an incident further down the chosen path, blocking part of the path and thereby increasing the vehicle's travel time, it may be more beneficial to circumvent the known congestion and deviate from the path initially selected. The setting

in which vehicles are allowed to change their route anytime after their departure from the origin is referred to as the *online*, *adaptive* or *dynamic* setting.

In the online setting, a vehicle makes a routing decision upon arrival at each network node. Therefore, a corresponding routing procedure is no longer a single path, but a routing *policy*, which prescribes to which node to travel next under which set of traffic conditions. In Markov models as introduced in Section 1.2.1, for example, such a routing policy defines the next node for each location and background-state pair.

In general, finding an optimal routing policy is a much bigger challenge than finding an optimal path. Moreover, even if techniques to compute this policy are known, they may be computationally infeasible. Already in the Markovian framework, which has a finite state space and is relatively tractable, the optimal policy, following from dynamic programming procedures, cannot be deduced for highway networks of realistic size. In this thesis, we will therefore more than once propose the dynamic execution of a static shortest path algorithm as routing policy. Specifically, at each node in the network, given the current network state and the past observations, one determines the optimal path and travels to the first node on this path, to then repeat these steps. Remarkably, various numerical experiments that are presented in this thesis show the (close-to-)optimality and high efficiency of such procedures in different settings, making them extremely suitable for practical purposes.

Optimal departure times

In the routing setting above, after learning its optimal path or optimal routing policy, the vehicle is assumed to leave the origin instantly. However, in reality, when the objective is to arrive at a certain destination before a given time, one may want to get an indication of the time it takes to reach the destination beforehand, so as to know at what time to depart. On the one hand, if this time is chosen too far in the future, the probability of late arrival may exceed the risk the driver is willing to take. On the other hand, if the driver would depart too early, there is the cost of wasted time. Thus, there is a need for efficient algorithms that generate the latest departure time for which a user-specified on-time arrival probability can be guaranteed.

The main challenges are again posed by the time-dependent and stochastic nature of travel times. Based on the state of the network, one may have access to the current travel-time distributions, allowing one to compute the on-time arrival probability for various paths. However, to learn the on-time arrival probabilities at future departure times, one needs to derive the travel-time distributions that correspond to future travel times as well, where the evolution of the current network state needs to be taken into account.

For the optimal departure time, we can also distinguish both an offline and online variant. In the offline variant, the driver requests its optimal departure time once, to then depart at the advised time instant. In the online variant, the driver may receive departure time updates while still at the origin, to then only depart when the time itself and advised departure time match. To keep the procedure computationally feasible, we will again propose the dynamic execution of the offline algorithm as procedure for the online variant.

Input rate control

The two problems described above aim to reduce congestion in the highway network by providing individual travelers with tools that allow them to limit the time spent in the system. Whereas they may be able to reduce the impact of recurrent and non-recurrent events on individual drivers, application of these tools will not directly be successful in avoiding congestion altogether. That is, just like control mechanisms that perform flow management inside the road network (i.e., mechanisms that control each vehicle in the network), they are challenged by the worldwide increase in travel demand, which especially hinders the alleviation of recurrent congestion. To keep a better handle on congestive settings that arise through high demands, we consider the problem of input rate control at the boundary of the network.

As captured in the fundamental diagram of traffic flow, when the vehicle flow on a road exceeds its capacity, speeds will drop rapidly, and those speed drops will in turn propagate through space and time. The aim of input rate control is to fully avoid these capacity-violating scenarios, which deteriorate the performance of the network. To this end, we should again take the inherently random nature of vehicle traffic into account. Focusing on the impact of high demands, we should acknowledge that the capacity needs of individual vehicles suffer from randomness, due to the heterogeneity in, e.g., vehicle sizes and individual driving habits. By developing a control procedure that takes these random spikes in capacity needs into account, at the cost of a little waiting time at the network boundaries, we aim to ensure that the probability of high delays within the network is only small.

1.3 Queueing theory

After having introduced models and control procedures for congestion in the road traffic setting, we now consider regulation in what is perhaps the most canonical and general set of models for congestion: queueing models. The first queueing model was already introduced in 1909, in a paper of A.K. Erlang [65], who used this framework to model the delay in telephone systems. Since then, the field has developed rapidly, and queueing models have been used to describe the dynamics of congestion in a wide variety of applications. Examples include waiting lines in supermarkets, waiting rooms in hospitals, and the queues of data packets that must be processed by the devices that are part of the Internet network. Notably, whereas we have not considered them before, there is also a considerable amount of literature that applies queueing models to study traffic jams in road networks.

There are many ways in which the delay in queues can be controlled. In the supermarket, for example, one may increase the number of cashiers to reduce the waiting times of the customers. Note that this is a control decision on the supply side. Alternatively, one may aim to reduce the number of customers in the queue, i.e., the demand. To this end, a common practice is to consider *tolls*, a cost customers have to pay before joining the queue. As not all customers may be willing to pay such a fee, this will decrease the customer input rate. An important question is how to set the price of these tolls. That is, on the one hand, if this price is set too high, too many customers may choose to balk (i.e., not join), giving the queue a low throughput. On the other hand, a too low price may cause too

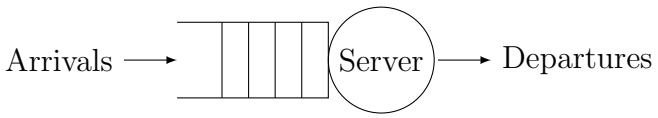


Figure 1.3: The single-server queue.

little customers to balk, yielding high waiting times for the ones that choose to join.

It has been suggested that in setting the correct price for such tolls one should levy the *externalities* that arriving customers impose, and in Part II of this thesis we will therefore study these externalities in a specific queueing model. Now, before explaining the concept of externalities in more detail, we will first briefly outline this model. More thorough introductions to the theory behind queueing models can, e.g., be found in the books of Cohen [42] and Kleinrock [134].

Queueing models

A queueing model is a mathematical description of a stochastic process in which arriving jobs need to wait until being processed by one of the servers in the network. In the context of a supermarket, the jobs represent the customers that arrive at the checkout with their goods, and the servers are the available cashiers. The randomness of a queueing model concerns the potentially stochastic process that governs the arrivals of the jobs, and the uncertainty in job sizes.

The most elementary queueing models are single-server queues (Figure 1.3), in which there is one waiting line to which jobs arrive, and one server to process these jobs. An example is the M/G/1 queue, a model to which many studies have been devoted, and which will also be under consideration in Part II of this thesis. In this queue, customers arrive according to a homogeneous Poisson process, indicated by the letter 'M' in the notation, which refers to the corresponding Markovian or memoryless nature of the inter-arrival times. Their job size or service requirement is the total time the server would need to fully process the job (given that, during its service, the server does not process other jobs as well). In the M/G/1 queue, it is assumed that the job sizes are independent samples of a general distribution, to which the letter 'G' in the notation refers. Alternatively, to emphasize the independence, this letter is sometimes replaced by 'GI'.

The notation 'M/G/1' follows the shorthand notation for a range of queueing models that was introduced by Kendall [128]. Specifically, Kendall proposed to denote queueing systems in the code A/S/c, where 'A' and 'S' respectively correspond to the inter-arrival and service time distribution, and 'c' is a number that specifies the amount of available servers. Typically, 'A' and 'S' are in the set {'G','M','D'}, with 'M' and 'G' as introduced above, and 'D' the degenerate distribution. Notably, Kendall's notation can be extended with additional letters to denote queues with finite waiting rooms or queues in which the customers arrive from a finite population.

In most real-life observable queueing systems, the server attends the jobs in their order of arrival. This service discipline is called first-come first-served (FCFS). Alternative service disciplines that are a frequent topic of study and that predominantly appear in the context of computer systems are, e.g., last-come first-served without preemption (LCFS-NP), last-

come first-served with preemption (LCFS-PR), and processor sharing (PS). Under the LCFS-NP discipline, new arrivals are placed at the second position in the queue, directly behind the customer in service (i.e., the queue consists of both the waiting customers and the customer in service). Then, at the time the server finishes a job, this job leaves, and the last-arrived job is taken into service. Under LCFS-PR, the queueing order is the same (i.e., reversed order of arrival), but with the additional property that a new arrival interrupts the processing of the job in service upon its arrival, this job now being the second job in the queue. Under PS, the order of arrival does not play a role, as the server processes all present jobs in parallel. That is, if there are k jobs in the queue at a given time instant, the speed at which these jobs are processed is $1/k$. Note that this service discipline cannot be offered by human servers, and that such service can only be approximated by computer systems.

With more than one hundred years of research since the first queueing model of Erlang, many fundamental results have been derived. In general, it is thereby often assumed that the queue is stable (i.e., the arrival rate is lower than the service rate) and in steady state. This allows for a mean value analysis, in which the following two properties have proven to be extremely useful:

- Little's law: founded by Little [150], this formula states that, in stationarity, the long-term average number of customers equals the product of the long-term arrival rate and the long-term average time that a customer spends in the system.
- PASTA property: being an acronym for 'Poisson arrivals see time averages', this property states that for queues with Poisson arrivals, the value of a certain quantity at arrival epochs (e.g., the number of customers) equals the long-term average value of that quantity in the system [219].

Regulation

One may aim to reduce the total delay in queueing models by increasing the total number of servers in the system. Note that employing such a measure resembles the efforts in limiting congestion in road traffic by network expansions. As we already noticed in that context, the future use of such supply changes may be limited and costly. Think also about the supermarket, in which there is a maximum number of check-out counters, and hospital waiting rooms, there being a limited number of doctors present.

Another measure that could potentially help to control the total delay would be a better choice of service discipline. For example, whereas FCFS is used in most physical queues, given that the server knows the service requirements of customers upon their arrival, the total waiting time is known to be minimal under the *shortest remaining processing time* discipline. However, changing to such a procedure is, in some practical applications, far from evident, as it is feared that such a policy is considered unfair, given that it does not attend customers in their order of arrival, and that it may produce a high variability in the waiting times of individual customers.

Instead of changing the queue on the supply side, in the 1960's, Naor [163] suggested to alternatively control the demand in the queue (i.e., the number of arrivals), by levying tolls. Specifically, the author considers a stable and stationary M/M/1 FCFS queue in which

arriving customers may choose to join or balk, based on the number of customers in the queue upon their arrival. If a customer joins, they receive a (positive) reward upon service completion, and they pay a (positive) cost per unit of time spent in the system. Naturally, optimizing their total gains (reward minus costs), there is a threshold n_0 such that, if arriving customers see n_0 or less customers in the queue, they will join, and otherwise balk. A similar threshold n_1 exists when aiming to maximize the expected sum of the net gains. Naor shows that $n_1 \leq n_0$, and that by levying tolls to individual customers, one may bridge this gap between self-optimization and overall optimization.

There are various pricing mechanisms that can be thought of, of which one is the natural notion of *externalities*. Externalities are the total waiting costs that a customer imposes on the system, i.e., the total additional waiting time of all customers due to the presence of this customer. By levying the externalities of a customer as toll, the customer directly pays for the delays they cause. Importantly, these delays are a function of their service requirement, such that, typically, customers with a large job pay higher tolls than customers with a small service requirement.

Due to its application in queue regulation, there is a need for formulations of externalities in different queueing systems. Haviv and Ritov [101] derive the expected externalities in a stationary M/G/1 queue, under the FCFS, LCFS-PR, LCFS-NP, and PS service disciplines. Their work is extended by Jacobovic and Mandjes [115], who consider externalities in the transient M/G/1 FCFS queue. In Chapter 7, we perform a similar analysis for the joint externalities in an M/G/1 queue under FCFS and LCFS-PR.

1.4 Contributions and outline

This thesis consists of two parts. In Part I, introduced in Section 1.2, we consider congestion phenomena in a specific application: highway networks. We develop stochastic models to capture the characteristics of congestion in these networks, and look at control policies that aim to limit the time vehicles spend in the system. In Part II, as introduced in Section 1.3, we shift our focus from the road-traffic application, and look at a more general class of congestion models: queueing models. Specifically, we consider the role of externalities in the regulation of an M/G/1 queue under different service disciplines, and study a corresponding mean-variance optimization problem in more generality.

Below, we provide, per part, a brief overview of the contributions of each chapter in that part. At the same time, this presents the outline of the thesis.

Part I

In Chapter 2 (based on [144]), our objective is to minimize the expected travel time of a vehicle in a highway network, assuming the vehicle may adapt the chosen route while driving. To this end, we first develop the Markovian velocity model (MVM), which uses a background process to track recurrent and non-recurrent traffic events that affect the speed of travelers, allowing for correlation between the speeds on different edges. The optimal routing policy being computationally intractable for realistic network sizes, we present the EDSGER* algorithm, a Dijkstra-like shortest path algorithm that can be used

dynamically with real-time response.

Modeling the vehicle speed uncertainties with the MVM, Chapter 3 considers the problem of determining a vehicle's optimal departure time. Specifically, we devise a computationally efficient algorithm that uses individual link travel-time distributions to obtain the optimal departure time for a given path or origin-destination pair. Since the conditions in the road network may change between the time of request and the advised time of departure, we present an online version of this procedure as well, in which the traveler receives departure time updates while still at the origin. The content of this chapter is based on [121].

Whereas the usefulness of a stochastic velocity framework follows from Chapters 2 and 3, calibration of such models is still crucial, as management centers and individual drivers need access to travel-time distribution estimates. In Chapter 4 (based on [143]), we therefore focus on operationalizing the MVM. First, we demonstrate how traffic data can be used to obtain an accurate description of the randomness of incidents. Then, we explicitly show how incidents and daily patterns can be incorporated into the background process of the MVM, and how their effect on vehicle speeds must be chosen to accurately reflect their impact on the travel time of a vehicle.

Chapter 5 (based on [120]) again considers optimal routing in stochastic road networks, but in a setting in which the current traffic densities are unknown, and we only have access to (i) vectors of velocity measurements on the different network links and (ii) historic routing probabilities. Using a stochastic fundamental diagram to model the uncertainty of the vehicle speeds as a function of the traffic density, the current link densities are estimated via a maximum likelihood procedure. We infer expected future density estimates by implementing a well-known discrete-time vehicle transmission scheme, and devise a highly efficient and close-to-optimal label-updating algorithm that aims to find the shortest path in a network whose weights depend on these time-stamped density estimates.

In Chapter 6 (based on [145]), we study the problem of the real-time control of input rates, with the aim to avoid any violations of link capacities. To account for the fact that, by the heterogeneity within and between traffic streams, the available capacity of a road suffers from randomness, we introduce a stochastic flow model that describes the impact of traffic input streams on the available road capacities. Exploiting similarities with traffic control in telecommunication networks, we propose a traffic rate control policy based on the concept of *effective bandwidths*.

Part II

The study of Chapter 7 (based on [114]) analyzes the joint distribution of the externalities created by the arrival of an additional customer c with service requirement x in a stable M/G/1 queue when the underlying service distribution is either LCFS-PR or FCFS. We first prove a decomposition of the externalities under the above-mentioned service disciplines. Then, this decomposition is used to derive several other results regarding the externalities: moments, asymptotic approximations as $x \to \infty$, asymptotics of the tail distribution, and a functional central limit theorem.

Motivated by a classical mean-variance analysis of the externalities in the M/G/1 LCFS-PR queue included in Chapter 7, Chapter 8 (based on [113]) studies the minimization of these externalities under a different set of constraints. That is, we consider a system

manager who is able to observe the service requirement of the additional customer c, and, upon arrival of c, both the total number of customers and the remaining service requirement of the customer in service. However, of the waiting customers, the manager only has access to their *total* remaining service demand. Under this constraint, we explicitly solve the minimal variance that can be attained for some families of parameterizations, and put forth a conjecture for the general case.