From Corridor to Network Macroscopic Fundamental Diagrams: A Semi-Analytical Approximation Approach

Gabriel Tilg, Lukas Ambühl, Sérgio Batista, Mónica Menéndez, Ludovic Leclercq, Fritz Busch

Submitted to *Transportation Science*
manuscript (Please, provide the manuscript number!)

From Corridor to Network Macroscopic Fundamental Diagrams: A Semi-analytical Estimation Approach

Gabriel Tilg
Chair of Traffic Engineering and Control, Department of Civil, Geo and Environmental Engineering, Technical University of Munich, Germany, gabriel.tilg@tum.de

Lukas Ambühl
Traffic Engineering Group, Institute for Transport Planning and Systems, ETH Zurich, Switzerland

Sérgio F. A. Batista, Mónica Menéndez
Division of Engineering, New York University Abu Dhabi, United Arab Emirates

Ludovic Leclercq
Univ. Gustave Eiffel, Univ. Lyon, ENTPE, LICIT, F-69518, Lyon, France

Fritz Busch
Chair of Traffic Engineering and Control, Department of Civil, Geo and Environmental Engineering, Technical University of Munich, Germany

The design of network-wide traffic management schemes or transport policies for urban areas requires computationally efficient traffic models. The macroscopic fundamental diagram (MFD) is a promising tool for such applications. Unfortunately, empirical MFDs are not always available, and semi-analytical estimation methods require a reduction of the network to a corridor which introduces substantial inaccuracies. We propose a semi-analytical methodology to estimate the MFD for realistic urban networks, without the information loss induced by the reduction of networks to corridors. The methodology is based on the method of cuts, but applies to networks with irregular topologies, accounts for different spatial demand patterns, and determines the upper bound of network flow. Thereby, we consider flow conservation and the effects of spillbacks, both at the network level. Our framework decomposes a given network into a set of corridors, creates a hypernetwork including the impacts of source terms, and then treats the dependencies across corridors (e.g. due to turning flows and spillbacks). Based on this hypernetwork, we derive the free-flow and capacity branch of the MFD. The congested branch is estimated by considering gridlock characteristics and utilizing recent advancements in MFD research. We showcase the applicability of the proposed methodology in a case study with a realistic setting based on the Sioux Falls network. We then compare the results to the original method of cuts, and a ground truth derived from the cell transmission model. The results reveal that our method is over five times more accurate than the state of the art in estimating the network-wide capacity and jam density. Moreover, they clearly indicate the MFD's dependency on spatial demand patterns. Compared to simulation-based MFD estimation approaches, the potential of the proposed framework lies in the modeling flexibility, explanatory value, and reduced computational cost.

Key words: macroscopic fundamental diagram, method of cuts, network modeling, traffic flow theory, variational theory

History:

1. Introduction

Population growth in large metropolitan areas leads to disruptions in the transportation system. Improving its efficiency essentially requires the design of appropriate traffic monitoring and control schemes. Aggregated traffic models based on the macroscopic fundamental diagram (MFD) (e.g Daganzo 2007, Geroliminis and Daganzo 2008) are powerful tools, consistent with traffic flow theory, for modeling system dynamics in large urban areas. These models require the definition of regions in the city network, where traffic conditions are approximately homogeneous, thus showing a clear relationship between travel production and the accumulation of vehicles. Such models have a wide range of applications, including the design of control strategies, pricing schemes, public transport priority and traffic management (see Johari et al. (2021) for an overview), or even the network-wide monitoring of emissions (Batista, Tilg, and Menéndez 2022).

The respective traffic models require the estimation of the MFD to describe aggregated traffic state dynamics. Microscopic simulations can be used in the absence of real data and for hypothetical

scenario analyses. Unfortunately, these have high computational costs, require laborsome calibration, and their results remain sensitive to the specific demand patterns. Alternatively, analytical solutions exist for estimating the MFD (e.g. Ambühl et al. 2020). However, limited or unavailable a-priori knowledge limits their application. Another possibility is to use semi-analytical approaches, such as the method of cuts (MC) (Daganzo and Geroliminis 2008, Leclercq and Geroliminis 2013, Laval and Castrillón 2015, Tilg, Amini, and Busch 2020). So far, however, these methods only explicitly model single corridors, which are defined as an ordered sequence of links. Hence, their application is limited to small and regular synthetic urban networks, where the difference between the MFD of an abstract corridor and that of the whole network is minor. Hitherto, the literature lacks a comprehensive framework to semi-analytically estimate MFDs for realistic, complex urban networks.

This can lead to major limitations as the topological mapping of a network into a single corridor is not necessarily accurate. Existing approaches implicitly assume that network supply characteristics are relatively regular and show low variability throughout the network (i.e. similar signal offsets, green-to-cycle ratios, block length, etc.). Thus, the average network characteristics can be mapped into a single corridor. This means, however, that for realistic network, the variability of the network is lost, leading to inaccuracies in the approximated MFD. For example, the capacity of a single corridor is typically constrained by the most restrictive intersection, but we know that networks are much more complex than that. Also, the aggregation of a network to one corridor simplifies demand-related aspects. This might be particularly problematic for general networks, since in many cases multiple routes connect the same origin-destination (OD) pair. Furthermore, existing methods implicitly assume that traffic dynamics evolve in a similar manner on all corridors in the network. This occurs rarely in urban networks with complex demand patterns. Furthermore, diverging flows at intersections inevitably lead to undersaturated conditions at downstream links. Neglecting this contradicts the network-wide flow conservation. Lastly, the effects of spillbacks may propagate to adjacent corridors. These aspects make the need for a valid approximation of the *network* MFD obvious.

We should note that in many cases, the observed macroscopic performance of an urban network is substantially lower than the *idealized* MFD which refers to a theoretical upper bound for traffic conditions in a corridor (Daganzo 2007). The reason for this lies in the observed traffic heterogeneity, non-stationary traffic states, and the path-flow distribution of the demand (Mahmassani, Saberi, and Zockaie 2013, Knoop, Van Lint, and Hoogendoorn 2015, Leclercq et al. 2015, Geroliminis and Sun 2011, Mazloumian, Geroliminis, and Helbing 2010). These observed aggregated traffic states are herein denoted as the *realized* MFD (Ambühl et al. 2021, Loder et al. 2019). Unfortunately, the two different MFD notions are currently used interchangeably. Both MFD types are very similar for a corridor, under steady-state conditions, and homogeneous demand. However, they might differ

significantly for general networks. Simply put, the idealized MFD corresponds to stationary flows resulting from a spatial demand pattern that perfectly utilizes the given supply. Thus, the ambiguity of results stemming from models that simplify the network to a corridor, as required by existing semi-analytical approaches, and neglecting the network-specific influence of demand-related aspects, becomes apparent. First approaches to tackle these issues are reported in Geroliminis and Boyacı (2012) and Xu, Yu, and Gayah (2020), which incorporate turning flows at intersections. The former study analyzes these effects on a simulation basis and focuses on incoming turns. The second one models the effects merely stochastically. Neither of them describes network-wide spillbacks nor models network conservation explicitly, which is important to identify the underlying mechanisms for congestion dynamics. Only such understanding enables the proper design of measures to mitigate congestion and increase urban network capacities. Our proposed methodology addresses this by considering the effects of turning ratios on the network MFD. The realized MFD is especially important because it replicates observed traffic states based on which further applications can be designed (e.g. traffic control, transport planning, analysis of local bottleneck effects at the network level). Additionally, as we will discuss later, it could potentially be extended to also estimate the idealized MFD, although this is left for future research.

In this paper, we propose a framework using semi-analytical methods to estimate the realized MFD at the network level. Thereby, we do not rely on extensive empirical traffic data sets or computationally expensive microscopic simulations. In contrast, we develop a framework based on semi-analytical methods, but do not reduce the network to a corridor. We focus on a single region and explicitly consider flow conservation as well as the effects of spillbacks. Thus, we can account for the effects of different demand patterns on the realized MFD. Hereafter, we refer with 'MFD' to the realized one unless explicitly stated otherwise. The contributions are threefold:

First, we develop a framework to semi-analytically estimate the network MFD. So far, existing methods were not able to account for spatial demand patterns and the related violation of the conservation of flows at intersections from the corridor perspective. Our framework proposes a way to address this challenge. We decompose a given network into a set of corridors and construct a hypernetwork in the spirit of the original MC. The decomposition maintains the connection of different corridors and thus, enables us to consider the network-wide effects of spillbacks. This is a major difference to current approaches, where multiple corridors are either reduced to a single one or modeled as being largely independent of each other, which may lead to inaccurate traffic state estimations.

Second, we propose three different approaches to derive the maximum flows at intersections, and two different approaches to approximate the network-wide jam density. To our best knowledge, no method exists so far to approximate the network-wide jam density. With the maximum flows at intersections and the network-wide jam density, the MFD is then estimated following the philosophy

of the original MC. The proposed approaches vary in terms of modeling complexity, computational cost, and estimation accuracy. Thus, they offer high flexibility and enable the choice of a case-specific MFD estimation. Moreover, they shed light on the effects of related assumptions and thereby provide additional explanatory value.

Third, we evaluate the proposed framework using a realistic network. Within this case study, we compare the proposed approaches to the state of the art for such networks and the cell transmission model (CTM) (Daganzo 1994, 1995) acting as a ground truth. Thereby, we provide proof of our concept and test the framework for a range of different input parameters.

The remainder of this paper is organized as follows. First, we provide a brief summary of the existing attempts to extend the original MC for realistic networks with turning flows and their drawbacks. Then, we present the framework to estimate the MFD at the network level. Subsequently, we conduct a case study for the realistic network of Sioux Falls. Finally, we draw conclusions and outline potential avenues for future work. In the appendix, we provide a nomenclature table (Appendix **??**), additional background on variational theory (VT) and the MC (Appendix **??**), a discussion of the framework's main assumptions (Appendix **??**), methodological details (Appendix **??** and **??**), and the results of a sensitivity study (Appendix **??**).

2. Extension attempts of the original method of cuts to the network level

This section briefly describes the deficiencies of current MC-based approaches to estimate the network MFD. The original MC is based on the concept of VT and was introduced by Daganzo and Geroliminis (2008). Vt itself was formulated by Daganzo (2005a,b) to solve complex and heterogeneous kinematic wave theory (KWT) problems. More background on both theories are provided in appendix **??**.

The MFD approximation by the MC is independent of the corridor demand since a single corridor without any turning flows only involves one OD pair. In this context, the *demand* only refers to the loading, and the idealized MFD covers all stationary loading levels. From the network perspective, however, demand distribution in the form of path flows can lead to heterogeneous link flows which can have a strong influence on the shape of both the realized and the idealized MFD. Applying MC to networks requires a reduction of the network to a single corridor. This inevitably introduces a bias with respect to the actual demand distribution and supply conditions.

Daganzo and Geroliminis (2008) originally proposed to identify a *representative* corridor of the network and apply the MC to find the network MFD. This represents the already mentioned reduction of the network's complexity to a single corridor. In Girault et al. (2016), the authors reduced a bidirectional grid network to four corridors representing each cardinal direction. They applied the MC to each corridor and subsequently averaged the resulting corridor MFDs. This approach neglects some

of the inter-dependencies such as spillback propagation between corridors. Moreover, the network reduction implicitly assumes that the most constraining intersection is saturated when the capacity is reached. However, this is not necessarily the case when turning flows occur, as some vehicles might turn before such an intersection, and thus reducing the maximum throughput. Aghamohammadi and Laval (2022) proposed a maximum likelihood approach for the method of Laval and Castrillón (2015) to improve the estimation of the idealized MFD for networks. While their results are promising, the methodology still builds upon a single corridor. Although it is not explicitly stated in these original papers, they implicitly aim at estimating the idealized MFD since they rely on the concept of the original MC.

Some initial attempts have been made to incorporate the effects of turning flows into MC. Since these approaches only consider a specific set of turning ratios, they implicitly focus on the realized MFD. Geroliminis and Boyacı (2012) modeled the effects of inflows for the estimation of a route-specific MFD, although they did not attempt to estimate network MFDs. They modeled inflows as bottlenecks but did not thoroughly incorporate them in the estimation procedure. Therefore, their resulting MFD describes traffic dynamics for a certain route and not for the full corridor. Recall that we defined a corridor as an ordered sequence of links. In distinction to that, we define a route as an ordered sequence of links a vehicle travels from its origin to its destination. Thus, certain routes might correspond to a corridor, others might not. Xu, Yu, and Gayah (2020) extended the stochastic model of Laval and Castrillón (2015) to account for turning flows. The authors proposed to estimate the effects of turning flows as a Markov process based on a two-ring model. They modeled turning flows with a global ratio which was kept constant across intersections in the network. Therefore, they again simplified the network to a single corridor. Other analytical approaches (Gan, Jin, and Gayah 2017, Jin, Gan, and Gayah 2013) are also based on the two-ring model and are therefore limited to very regular and simplified network topologies.

3. Network method of cuts

In an attempt to overcome the mentioned drawbacks of existing methods, we propose a framework based on the MC that applies to realistic networks with varying turning flows at each intersection. Thereby, we include network-wide traffic dynamics, do not reduce the entire network to a single corridor, and thus consider its original topology. Hereafter, we refer to this framework as 'nMC'. It consists of the following three steps:

- *Step 0* initializes the problem by defining required supply and demand characteristics. This consists of the definition of the network topology, signal control settings, and turning ratios. Note that this implies an indirect definition of routes, and thus reflects the observed spatial demand pattern for a specific case. These parameters describe the physical road network $\mathcal{N}$

composed by a set of directed links and intersections, as well as the demand that exists for that network.

- *Step 1* transforms $\mathcal{N}$ into a hypernetwork $\mathcal{G}$ which is the basis for deriving the MFD. Recall that a hyperlink represents a physical corridor in the original MC. Analogously, a hypernetwork represents the physical road network in time and space and can be seen as a multi-dimensional variational graph. To generate $\mathcal{G}$, we decompose $\mathcal{N}$ into a set of corridors, define the structure of the hypernetwork $\mathcal{G}$, and exogenously derive the costs for horizontal edges in $\mathcal{G}$. For this, we propose three different approaches with varying modeling complexity, computational cost, and estimation accuracy. These approaches are labeled as 'network variational theory' (nVT), which is based on an extension of VT to networks (Tilg et al. 2021), 'full spillbacks' (FS), and 'limited spillbacks' (LS). They are further explained in Section 4.
- *Step 2* derives the cuts based on which the free-flow branch, the capacity, as well as the congested branch of the MFD are estimated. In the spirit of the original MC, moving observers travel through $\mathcal{G}$ in order to estimate stationary traffic states. This step includes the approximation of the network-wide jam density and the utilization of symmetries in the propagation of free-flow and congested traffic states across the network. Again, two different approaches are proposed, namely the network variational theory (nVT) and the queue propagation (QP) approach, which are explained in Section 5.3.

Figure 1 summarizes the overall framework.

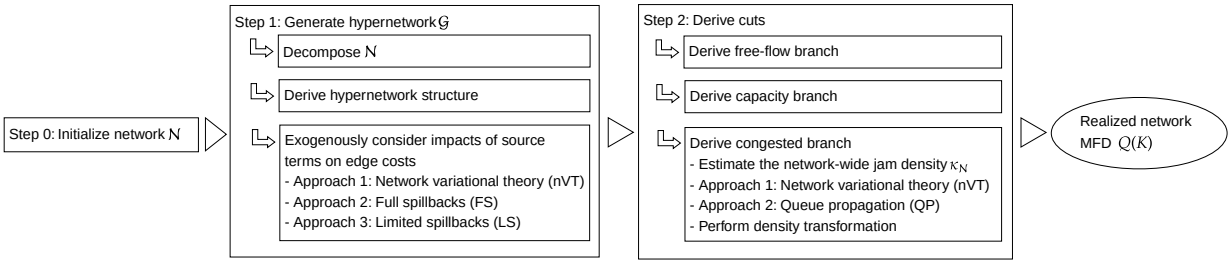


Figure 1 Flowchart of the proposed framework, the nMC.

4. Generation of a hypernetwork

The representation of $\mathcal{N}$ as a hypernetwork $\mathcal{G}$ is key to our framework. It allows us to derive the cuts similar to the original MC (see Appendix **??**) while accounting for source terms and network-wide spillback propagation. These cuts are the basis to find the network MFD. In the following, we list our main assumptions. A more detailed discussion about these assumptions is provided in the Appendix **??**:

- We only consider unimodal networks.

- We assume that stationary states exist and can be reached for any spatial demand pattern if the temporal change is slow.
- We only consider signalized intersections without modeling conflicting streams explicitly.
- We assume that vehicles follow a FIFO discipline on all links and at diverges (Newell 1993).
- Our framework only applies to cases where vehicles at intersections can either remain on the main corridor or change to a single adjacent corridor.
- We assume turning ratios to be constant across time.

We generate a hypernetwork $\mathcal{G}$ representing a physical road network $\mathcal{N}$ including signal settings and turning flows. First, $\mathcal{N}$ is decomposed into a set of corridors $\mathcal{C}$. Second, the structure of a hypernetwork $\mathcal{G}$ is defined. Third, the cost of edges in $\mathcal{G}$ are specified. The costs of horizontal edges in the hypernetwork, which represent maximum flows, are derived depending on the proposed methods applied, i.e. nVT, FS, or LS. This concludes the generation of the hypernetwork $\mathcal{G}$. The resulting maximum flows in $\mathcal{G}$ represent the network traffic at the capacity level.

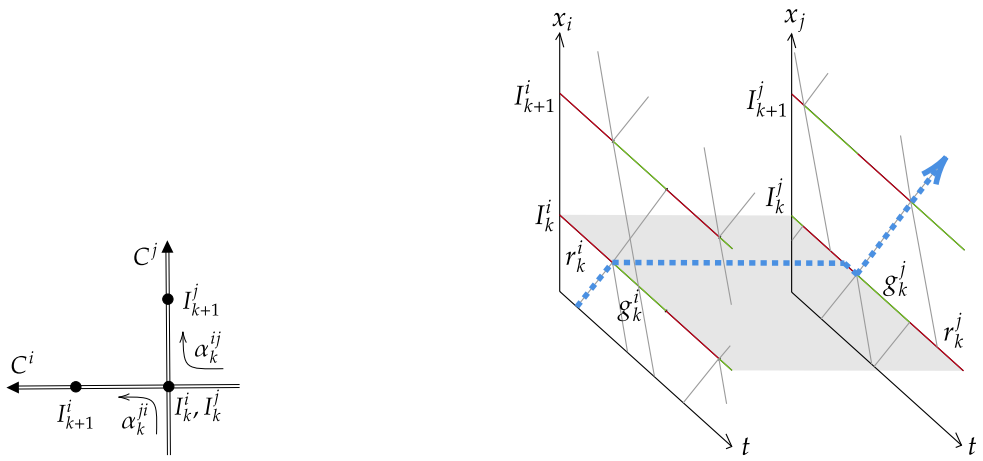
4.1. Network decomposition

The decomposition of the physical road network $\mathcal{N}$ into a set of non-overlapping corridors facilitates the generation of $\mathcal{G}$. Furthermore, this set of corridors is an input to the MFD estimation (see Section 5.1). Tilg et al. (2021) described such a decomposition in the context of VT. Following their methodology is advantageous for our framework, as their proposed network VT solution method, the 'nVT', can be utilized to derive the hypernetwork (see Section 4.3.1). We provide a brief summary of the decomposition method in Appendix **??**.

4.2. Structure of the hypernetwork

The set of corridors $\mathcal{C}$ allows us to generate a hypernetwork $\mathcal{G}$ representing the maximum traffic flows in the entire network $\mathcal{N}$. This includes the generation of a hyperlink for each corridor along with intersections and corresponding turning ratios α according to the physical network topology. Hence, in contrast to the original MC where only single corridors could be modeled, $\mathcal{G}$ represents networks where multiple edges downstream and upstream of an intersection might exist. This expansion of the concept of the corridor-specific hyperlink to a network-wide hypernetwork $\mathcal{G}$ represents one of the major cornerstones of our framework.

Figure 2a illustrates an example of a decomposed network based on which we further introduce our notation. It shows two corridors C connected at an intersection I. The indices denoted as superscripts i, j refer to a specific corridor. The index k as subscript refers to an intersection and the link directly upstream of it. The subscript is unique in combination with the corridor-related index. For example, the intersection I_k^i is the k^{th} intersection on corridor C^i. Similar to that, subscripts of maximum flows $\tilde{q}$ indicate intersections, and superscripts the corridor to which the flow refers. For example, the



a) Physical road network $\mathcal{N}$. **b)** Hypernetwork $\mathcal{G}$.

Figure 2 Example of a network with two corridors and a corresponding hypernetwork excerpt.

variable $\tilde{q}^i_k$ describes the maximum cycle-based average flow on corridor C^i upstream of intersection I^i_k. By maximum flow, we do not refer to the capacity, but the maximum flow that can be sustained on average during one cycle as a function of both the demand and the supply. The indices of turning ratios α denote the origin and destination corridor of the corresponding flows at a given intersection. For example, the turning ratio α^{ij}_k describes the ratio of the flow aiming to change from C^i to C^j at intersection $I^i_k = I^j_k$.

The network $\mathcal{N}$ can be translated into a hypernetwork $\mathcal{G}$, see Figure 2b. For a single corridor, the graph $\mathcal{G}$ consists of horizontal and slanted edges, the latter with slopes equal to u or w. The duration of green and red phases is denoted with g and r, respectively. Subscripts denote the intersection to which these phases refer, while superscripts refer to the corresponding corridors. For example, g^i_k is the duration of the green phase at intersection I^i_k on corridor C^i. The horizontal edges have associated costs z equal to the maximum flow $\tilde{q}$ passing a moving observer traveling along the respective edge. In the original MC, this flow is zero during red times, i.e. $\tilde{q} = 0$, and equal to the link capacity during green times, i.e. $\tilde{q} = q_{max}$ (see Appendix **??**). However, the occurrence of inflows and outflows at intersections might affect the observed maximum flow, and consequently the related costs z. For example, the maximum flow $\tilde{q}^i_k$ during a green phase at an intersection I^i_k might be reduced when a net outflow occurs at the upstream intersection I^i_{k-1}. Also, if the inflow at an intersection I^i_k exceeds the capacity of the downstream intersection I^i_{k+1}, spillbacks occur after a while, potentially affecting $\tilde{q}^i_k$.

These examples highlight the importance of the inclusion of effects of turning flows and indicate that such impacts can be represented via $\tilde{q}$. Thus, we aim at modifying the costs z of horizontal edges to represent the effects of source terms. As explained before, these effects need to be considered

exogenously. Therefore, we propose three different methods to exogenously determine the effects of source terms on the maximum flows $\tilde{q}$ at intersections, and adapt the costs of related edges in $\mathcal{G}$.

4.3. Exogenous consideration of source terms

To complete the generation of the hypernetwork $\mathcal{G}$, we exogenously account for the effects of source terms at intersections which represent inflows and outflows. In particular, we modify the costs of horizontal edges in $\mathcal{G}$ representing green phases. Thereby, we include both demand- and supply-related aspects, such as undersaturated intersection approaches and the occurrence of spillbacks coming from the downstream intersection. Note that while the edge costs in $\mathcal{G}$ shall represent maximum flows, similar to the original MC, the hypernetwork allows deriving an upper bound for flows during free-flow traffic states as well. For this purpose, we present three different approaches of decreasing modeling complexity and computational cost at the expense of lower accuracy:

1. *Approach 1 - Network variational theory (nVT)*: Based on the nVT proposed by Tilg et al. (2021), this approach allows to accurately consider spillback propagation including circular dependencies at the network scale according to the KWT.
2. *Approach 2 - Full spillbacks (FS)*: This approach builds on analytical formulations and includes spillback propagation throughout the network and thus related circular dependencies.
3. *Approach 3 - Limited spillbacks (LS)*: Here, we only consider spillback propagation within links. The capacity reduction from downstream intersections will affect the upstream ones but will not further propagate into the network. Therefore, circular dependencies cannot be considered with this approach.

In the following subsections, we first describe the utilization of nVT within the framework and highlight the synergies between nVT and nMC. Then, we introduce the mathematical framework for the approximate approaches FS and LS. Finally, we set up an optimization problem to derive the maximum flows at each intersection, which correspond to the network capacity state, and thus the costs for the hypernetwork.

4.3.1. Approach 1 - Network variational theory (nVT) The 'nVT' as proposed by Tilg et al. (2021) enables one to describe the evolution of traffic states in a signalized network with precision and efficiency. Its advantage compared to other macroscopic traffic models is the ability to model complex intra-link bottlenecks. Its advantage compared to microscopic simulation models is the numerical efficiency. Moreover, nVT requires the same inputs as nMC and builds upon a numerical grid similar to a hypernetwork. This leads to further synergies, as the hypernetwork $\mathcal{G}$ can be utilized for such a numerical grid. While it promises a high accuracy compared to the approximate approaches (FS and LS), it requires the highest modeling and computational efforts to derive the costs for horizontal edges in $\mathcal{G}$.

Cumulative count The numerical grid is defined by the time-step length Δt, the spatial step length $\Delta x = u\Delta t$, the length of each corridor $C \in \mathcal{C}$, and the length of the analysis period (see Appendix **??**). The latter has to be chosen such that stationary traffic states are reached on each link $L \in \mathcal{L}$, since the MFD is based on such stationary traffic states. In other words, the total computation time of the numerical grid needs to be sufficiently long such that stationary states can be reached. Lastly, turning ratios $\alpha \in A$ are input parameters. Note that this grid has some overlap with the hypernetwork $\mathcal{G}$ (e.g. at intersections) but is generally more fine-meshed as grid nodes also cover space within links.

The nVT finds the cumulative count N at each node $P(x,t)$ in the numerical grid. For a given node $P(x,t)$ within links, the classical VT formulation applies (see Appendix **??**). Furthermore, Tilg et al. (2021) proposed to exogenously account for the effects of source terms at the points P which are exactly at, and right downstream of an intersection. The corresponding formulations ensure the correct propagation of free-flow and congested traffic states across intersections despite the existing discontinuities in the Moskowitz surface at such locations. The result is the precise cumulative count N at each point in the numerical grid. For further details on the methodology, please refer to Tilg et al. (2021).

Edge costs The cumulative count $N(x,t)$ found by nVT lets us now derive the costs for horizontal edges z_k^i in the hypernetwork $\mathcal{G}$. Recall that these costs correspond to the flow which would pass a moving observer traveling along such edges. Such moving observers always travel until the termination of a red phase in the original MC, which includes the entire green phase. Hence, it is sufficient to derive the green phase-specific average flows from the nVT solution to specify the costs of horizontal edges in $\mathcal{G}$. We measure the difference in N which occurs during the beginning and the end of green phases at all intersections to calculate the costs z. These costs consider both, the fact that intersections can be undersaturated due to net outflow at the upstream intersection, and spillbacks which block certain parts of the green phase. In both cases, the costs of the corresponding edge are effectively reduced. More formally, we can determine the costs z_k^i related to a green phase at intersection I_k^i as:

$$z_k^i = \frac{N_{k,s}^i - N_{k,e}^i}{g_k^i}. \tag{1}$$

where $N_{k,s}^i$ and $N_{k,e}^i$ denote the cumulative count at intersection I_k^i at the start and end of a green phase, respectively. Note that the numerator is always positive, as N is a monotonic increasing function in t. Moreover, it is necessary to derive the spillback duration. Again, such durations can be extracted from the nVT solution by analyzing the Moskowitz surface.

The nVT approach can deliver a high accuracy of flows within the network in accordance with KWT. However, it involves a comparably high modeling complexity and computational cost. In order to reduce both, we develop a conceptually simpler and computationally more efficient approach to derive the costs for the hypernetwork as shown in the following subsection.

4.3.2. Approach 2 - Full spillbacks (FS) The FS approach enables one to consider congestion propagation throughout the network. Thus, we label this approach 'Full spillbacks'. Compared to the nVT approach it simplifies traffic dynamics, thus it requires a lower computational effort and modelling complexity. It applies to use cases where a fast and simple MFD estimation is essential.

This section formulates a framework to approximate maximum flows $\tilde{q}$ considering the inter-dependencies between different intersections $I \in \mathcal{I}$ in the network $\mathcal{N}$. First, we consider demand-related aspects and the network-wide conservation of flows. Second, we account for supply reductions due to spillbacks. The main assumptions are a triangular link FD and that traffic signals at all intersections I have a common cycle length.

Demand The demand-induced reductions of maximum flows $\tilde{q}_{k,d}$ due to source terms need to be represented in $\mathcal{G}$. This is of particular importance for the estimation of the capacity branch of the MFD. The index d indicates that this maximum flow is constrained by the demand. Note that such effects were implicitly included in the original MC, e.g. undersaturated intersections due to upstream bottlenecks. However, source terms eliminate the possibility for such an implicit consideration. The following example lets us derive the key formulation to represent demand-related effects in our model.

Consider a case where net outflows at the intersection I^i_{k-1} lead to flows at the downstream link L^i_k and intersection I^i_k that are lower than the link capacity q_{max}. This is illustrated in an excerpt of the time-space diagram corresponding to C^i in Figure 3.

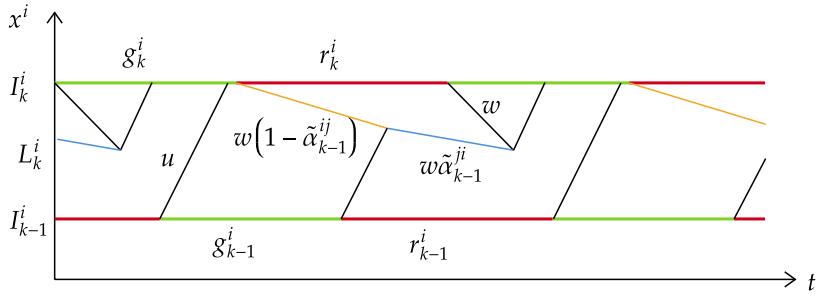


Figure 3 Influence of undersaturated links.

The figure shows a link L^i_k between two intersections, I^i_{k-1} and I^i_k, on corridor C^i. The red and green lines represent the signal phases at both intersections. Shock waves with the slopes u and w are shown as black lines. We assume turning ratios such that the total flow during g^i_k is undersaturated, i.e. $\tilde{q}^i_k < q_{max}(g^i_k/c^i_k)$ where c^i_k denotes the cycle length at I^i_k. The shock waves illustrating the propagation of the queue from I_k are shown as orange and blue lines. The orange shock wave results from flows staying on C^i, i.e. retaining flows, which join the queue. Thus, the queue grows with speed $w(1-\tilde{\alpha}^{ij}_{k-1})$. The blue shock wave results from turning flows from C^j which join the queue. Thus,

the queue grows with $w\tilde{\alpha}^{ji}_{k-1}$. Assuming a triangular FD, the parameter $\tilde{\alpha}$ can be derived as shown in Appendix **??** and written as:

$$\tilde{\alpha} = \alpha \frac{u}{u + w(1-\alpha)}. \tag{2}$$

This example shows that the maximum flow $\tilde{q}^i_k$ can be undersaturated since the flow which discharges from the queue does not last for the entire green phase g^i_k. Accounting for intersection-specific signal settings, we can describe the demand $\tilde{q}^i_{k,d}$ for an entire cycle as:

$$\tilde{q}^i_{k,d} = \tilde{q}^i_{k-1}(1-\alpha^{ij}_{k-1})\frac{g^i_{k-1}}{g^i_k} + \tilde{q}^j_{k-1}\alpha^{ji}_{k-1}\frac{r^i_{k-1}}{g^i_k}. \tag{3}$$

This equation describes the maximum flow $\tilde{q}^i_{k,d}$ which can occur considering demand only. Note that if a net inflow exists, i.e. $\tilde{q}^i_{k,d} > q_{max}$, the demand exceeds the link capacity q_{max}. To address such cases, the required supply-related constraints are introduced in the following.

Supply Next, we model the effects of spillbacks on the maximum flows, i.e. $\tilde{q}_{k,s}$, denoted by the additional subscript s. Note that such effects are implicitly considered in the original MC as flow conservation is kept. However, the occurrence of inflows and outflows at intersections might lead to a violation of flow conservation at the corridor level even though it is still satisfied at the network level. Therefore, in order to consider the effects of spillbacks, we propose to find the most constraining spillback duration σ_k exogenously as explained in the following. Note that the assumption of a FIFO diverge implies that a spillback occurring on a single link downstream of I_k affects all other outgoing links of I_k, too, if there is turning demand. Below, we derive the spillback duration and incorporate that into $\tilde{q}_{k,s}$ by means of an example.

Recall that flow conservation applies at the link level despite the existence of source terms as so far we have assumed those source terms are located at intersections. Thus, we can derive the duration σ_k during which a spillback in a link affects the upstream intersection I_k by applying traditional concepts of traffic flow theory.

Figure 4 displays an example of such a spillback occurrence for the stationary case. It shows an excerpt of a time-space diagram for a link L^i_{k+1} between two intersections, I^i_k and I^i_{k+1}. The red and green lines represent the signal phases at both intersections. Shock waves with the slopes u and w are shown as black lines. For this case, we assume turning ratios such that the total flow during g^i_{k+1} is oversaturated, i.e. the demand exceeds the intersection capacity $\tilde{q}^i_{k+1,d} > q_{max}(g^i_{k+1}/c^i_{k+1})$. The queue grows with speed w while saturation flows reach the queue. Once vehicles from upstream of I^i_k join the queue, its propagation speed changes. This is highlighted by the blue and orange curves in the figure. The blue curve represents a queue growth due to inflows from the adjacent corridor (i.e. turning vehicles coming from corridor C^j to C^i), similar to Figure 3. The orange one represents a queue growth due to the straight flows (i.e. vehicles staying on C^i). Once the orange shock wave

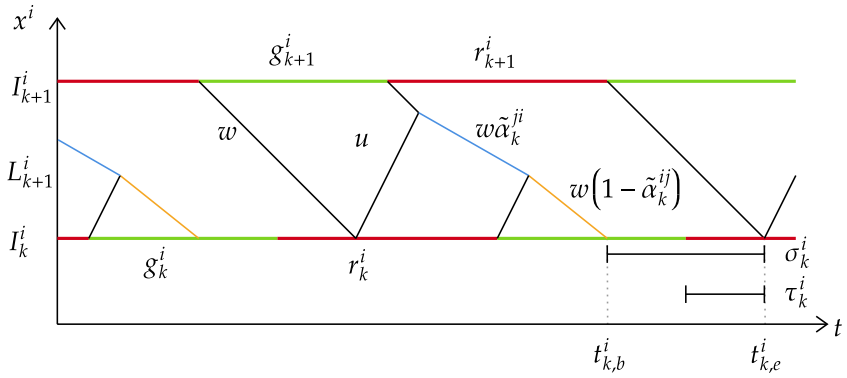


Figure 4 Spillback occurrence in a time-space diagram.

reaches I^i_k the flow of the corresponding upstream link is blocked, and the queue spills over to this link. This marks the beginning of the spillback duration, $t^i_{k,b}$. The spillback ends once the queue at I^i_k dissolves again at the time instant $t^i_{k,e}$. The total duration of the spillback impact is denoted as σ^i_k. This blockage affects the capacity of upstream links and the demand for downstream links on all adjacent corridors. Note that $t^i_{k,e}$ occurs during a red phase in the shown example, and the following considerations refer to that case. Nevertheless, the concept is valid for $t^i_{k,e}$ occurring during a green phase as well. The equations can be derived analogously.

First, we calculate the spillback impact ending time $t^i_{k,e}$ which can be found from the time-space diagram:

$$t^i_{k,e} = r^i_{k+1} + \frac{l^i_k}{w}. \tag{4}$$

Then, we derive the parameter τ^i_k, which denotes the time between the beginning of the active phase and $t^i_{k,e}$. Considering the offset o^i_k, we derive the parameter as follows:

$$\tau^i_k = mod(t^i_{k,e}, c^i_k) - o^i_k. \tag{5}$$

where *mod* is the modulo operator which calculates the remainder of the division of $t^i_{k,e}$ and c^i_k.

To facilitate the estimation of σ^i_k, we shift our perspective from the time-space diagram to a cumulative plot, as illustrated in Figure 5. We then resort to the illustrative and well-established technique of comparing cumulative counts N at different locations introduced as the 3-detector problem by Newell (1993). As described in Tilg et al. (2021), a discontinuity in the Moskowitz function N occurs at intersections with source terms. Thus, it is important to note that the following explanations always refer to a link segment where the flow conservation applies and no discontinuity exists. Correspondingly, we derive the cumulative count N^i_{k+1} at the downstream intersection I^i_{k+1}, and $N^i_{k+\delta x}$ at an infinitesimally small distance downstream of I^i_k, denoted as $I^i_{k+\delta x}$. Furthermore, we assume the spillback duration at I^i_k and $I^i_{k+\delta x}$ to be the same, $\sigma^i_k = \sigma^i_{k+\delta x}$. The count N^i_{k+1} represents

the cumulative curve corresponding to the supply, and $N^i_{k+\delta x}$ the demand. Since we are interested in deriving the spillback impact duration at $I^i_{k+\delta x}$, we shift N^i_{k+1} by l^i_k/w in the temporal dimension, and $l^i_k\,\kappa_{max}$ in the N-dimension, where κ_{max} is the link jam density. Figure 5 illustrates the shifted

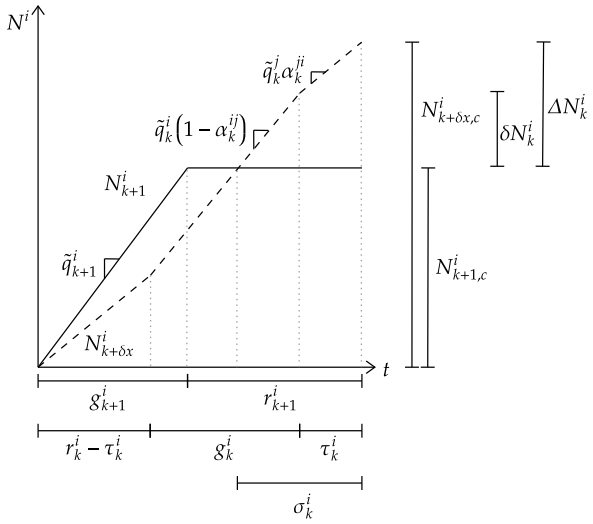


Figure 5 Spillback occurrence in a cumulative plot.

cumulative curves, where N^i_{k+1} and $N^i_{k+\delta x}$ are depicted as a solid and as a dashed curve, respectively. Both N-curves refer to a single cycle length c. This is sufficient, as we assume a common cycle length for all intersections, and analyze stationary states which occur periodically with such cycle length. The maximum cumulative count during the respective cycle is denoted with the subscript c. Based on both cumulative curves, we can derive the spillback impact duration σ^i_k as follows.

The cumulative count $N^i_{k+1,c}$ is equal to the product of the maximum flow $\tilde{q}^i_{k+1,s}$ at the downstream intersection I^i_{k+1} and the corresponding green phase:

$$N^i_{k+1,c} = g^i_{k+1}\tilde{q}^i_{k+1}. \tag{6}$$

Note that this formula already contains the flow $\tilde{q}^i_{k+1}$ which is calculated based on eq.(13). That is, this flow can itself be impacted by spillbacks at the respective intersection. This dependency shows the recursive nature of this approach, and it allows us to model the propagation of spillbacks downstream of I^i_k. Since $\tilde{q}^i_{k+1}$ can be reduced by spillback from *all* links downstream of I^i_k, this equation allows to account for circular dependencies within the network. Nevertheless, we assume a uniform spread of the effects of any constraint within the green phase.

The cumulative count related to the demand $N^i_{k+\delta x,c}$ equals the sum of the inflows for one cycle. Building upon eq.(3), we can write:

$$N^i_{k+\delta x,c} = \tilde{q}^i_k(1-\alpha^{ij}_k)g^i_k + \tilde{q}^j_k\alpha^{ji}_k r^i_k. \tag{7}$$

The comparison of both N-curves lets us determine the spillback duration σ_k^i at intersection I_k^i. More specifically, we can determine the actual cumulative plot at I_k^i as the lower envelope of both N-curves. The spillback begins once the flow is constrained by the cumulative curve at the downstream intersection, I_{k+1}^i. During that period, the flow is zero. We determine the difference between the demand and the capacity accounting for the spillback:

$$\Delta N_k^i = N_{k+\delta x,c}^i - N_{k+1,c}^i = \tilde{q}^i(1-\alpha^{ij})g_k^i + \tilde{q}^j\alpha^{ji}r_k^i - g_{k+1}^i\tilde{q}_{k+1}^i. \tag{8}$$

Next, we determine the part of ΔN_k^i which remains after subtracting the blocked vehicles during τ_k^i, denoted as δN_k^i. This simplifies the notation to derive σ_k^i.

$$\delta N_k^i = \Delta N_k^i - \tau_k^i\tilde{q}_k^j\alpha_k^{ji}. \tag{9}$$

Note that a positive δN_k^i corresponds to a case where $\sigma_k^i > \tau_k^i$, i.e. the spillback spans both phases. This case is illustrated in Figure 5. On the other hand, a negative δN_k^i corresponds to the case where $\sigma_k^i < \tau_k^i$, i.e. where the spillback only occurs during the red phase.

To derive the total spillback impact duration σ_k^i, we distinguish between the two cases referring to the sign of δN_k^i:

$$\sigma_k^i = \begin{cases} \tau_k^i + \frac{\delta N_k^i}{\tilde{q}_k^i(1-\alpha^{ij})} & \text{, if } \delta N_k^i > 0 \\ \tau_k^i + \frac{\delta N_k^i}{\tilde{q}_k^j\alpha^{ji}} & \text{, otherwise.} \end{cases} \tag{10}$$

With this equation, one can calculate the spillback duration due to a queue on C^i, which can be relevant for both corridors C^i and C^j at the intersection I_k^i. To approximately account for a FIFO diverge behavior, we propose to calculate ΔN_k for all outgoing links, and subsequently consider the maximum ΔN_k to compute σ_k. During a stationary state, the inflow and outflow in a link are balanced by the spillback duration. Thus, the maximum ΔN_k is required to ensure all flows are stationary. Note that this is an approximation, as we neglect the fact that spillbacks can also affect inflows on adjacent links. However, this approach does not aim at an extremely precise estimation of σ_k, but at estimating the spillback duration sufficiently precise for an accurate MFD estimation at the network level.

The total spillback duration σ_k^i is further divided into the corridor-specific blockage times, which are denoted as $\sigma_k^{i\rightarrow i}$ and $\sigma_k^{i\rightarrow j}$. These depend on the length of σ_k^i, i.e. whether the spillback lasts only during one signal phase, or spans two phases. Note that σ_k^i can occur during two phases at a maximum, as we assume a common cycle length for all intersections and therefore the theoretically maximum blockage duration equals to the red phase at the downstream intersection which cannot exceed a cycle length. As we assume that $t_{k,e}^i$ occurs during a red phase, $\sigma_k^{i\rightarrow j}$ is either τ_k^i or σ_k^i, if

$\sigma_k^i < \tau_k^i$. Similar to that, $\sigma_k^{i\rightarrow i}$ is the difference of the total spillback duration and τ_k^i, or equal to zero in $\sigma_k^i < \tau_k^i$. To cover both cases, we write:

$$\begin{aligned} \sigma_k^{i\rightarrow i} &= \max(0, \sigma_k^i - \tau_k^i), \\ \sigma_k^{i\rightarrow j} &= \min(\tau_k^i, \sigma_k^i). \end{aligned} \tag{11}$$

It is left to the interested reader to verify that the framework can be derived analogously for the other case, where $t_{k,e}^i$ occurs during an active green phase.

Eventually, the corridor-specific spillback impacts allow us to derive the respective capacity constraints. Therefore, we can formulate the reduction of maximum flows on corridor C^i and corridor C^j due to spillbacks on corridor C^i:

$$\begin{aligned} \tilde{q}_{k,s}^{i\rightarrow i} &= q_{max}(1 - \frac{\sigma_k^{i\rightarrow i}}{g_k^i}), \\ \tilde{q}_{k,s}^{i\rightarrow j} &= q_{max}(1 - \frac{\sigma_k^{i\rightarrow j}}{r_k^i}). \end{aligned} \tag{12}$$

Maximum flows The maximum flow on each link is then the minimum of the demand (eq.(3)) and the supply (eq.(12)):

$$\tilde{q}_k^i = \min\left(\tilde{q}_{k,d}^i, \tilde{q}_{k,s}^{i\rightarrow i}, \tilde{q}_{k,s}^{j\rightarrow i}\right). \tag{13}$$

This equation is formulated for each link and results in a non-linear system of equations. Its structure assembles widely applied concepts such as the CTM or VT. For network entry and exit links, the demand and supply parts have to be treated separately. For example, there are no demand-related restrictions for links on which flows enter the network. Similarly, there are no capacity constraints for links on which flows exit the network. These links represent origin and destination nodes within the network.

The inter-dependencies of flows $\tilde{q}_k^i$ can make the framework complex. Hence, to further simplify the description of network-wide traffic dynamics, we propose another approach which is introduced in the following section.

4.3.3. Approach 3 - Limited spillbacks (LS) The 'Limited spillback' (LS) approximate approach is similar to the FS, with the difference that we further simplify the consideration of spillbacks at downstream intersections. More specifically, the effects of spillbacks are only considered at the intersection level and not across entire corridors. To that end, we need to adapt eq.(6) as explained in the following.

We modify the way to estimate the cumulative count $N_{k+1,c}$ representing the maximum number of vehicles passing the downstream intersection I_{k+1} during a cycle. In this approach, $N_{k+1,c}$ is equal to the product of the link capacity q_{max} at the downstream intersection I_{k+1} and the corresponding green phase:

$$N_{k+1,c}^i = g_{k+1}^i \tilde{q}_{max}^i. \tag{14}$$

Note that this represents the essential assumption of the LS approach, i.e. queues blocking I^i_{k+1} are not considered in deriving σ^i_k. In other words, spillbacks propagating from links downstream are not considered here. This assumption might decrease the accuracy of the method for certain cases but further simplifies the overall modelling complexity.

The modification of the method to derive $N_{k+1,c}$ further affects the computation of ΔN^i_k. Based on eq.(8) and eq.(14) above, we write:

$$\Delta N^i_k = N^i_{k+\delta x,c} - N^i_{k+1,c} = \tilde{q}^i(1-\alpha^{ij})g^i_k + \tilde{q}^j\alpha^{ji}r^i_k - g^i_{k+1}\tilde{q}^i_{max}. \tag{15}$$

The difference in the demand and the capacity ΔN^i_k can then be put into eq.(9), and thus enables one to follow the procedure described for the previous approach to eventually calculate the flows $\tilde{q}^i_k$.

4.3.4. Maximum network flows To finally specify the costs for the hypernetwork, the flows $\tilde{q}^i_k$ must be chosen such that the network-wide average flow is maximized. The network-wide maximum does not necessarily result from maximum inflows at origin links, since these can lead to congestion within the network. Turning ratios are critical to define local capacity reductions that trigger spillbacks. These internal bottlenecks might become active at network inflows lower than q_{max} and the corresponding congestion can limit the maximum flows inside the network. Therefore, the maximum network-wide average flow is reached at the highest network inflows where the effects of activated internal bottlenecks still remain low. To treat this problem adequately within the FS and LS approach, we formulate an optimization problem based on eq.(13). The capacity of a network is defined by the maximum stationary flows possible. We define the inflows at origin links as decision variables. Therefore, the equivalent of eq.(13) for origin links, i.e. the link index $k=1$, can be written as follows:

$$\tilde{q}^i_1 = \min\left(\lambda^i q_{max}, \tilde{q}^{i\rightarrow i}_{1,s}, \tilde{q}^{j\rightarrow i}_{1,s}\right). \tag{16}$$

where the parameter λ is a factor to determine the demand at origin links as the share of q_{max}. This factor is found within an optimization problem and corresponds to the highest demand level possible where the effects of internal bottlenecks are low such that the network-wide average flows are maximized. The overall optimization problem for all links in the network can be formulated as:

$$\max_{\lambda^i} \quad \sum_{i=1}^{n} \tilde{q}^i_k \tag{17a}$$

$$\text{s.t.} \quad 0 \le \lambda^i \le 1 \tag{17b}$$

since all constraints are effectively included in eq.(13). For the nVT approach, we can also find the capacity state of the network with eq.(17). However, the maximum flows $\tilde{q}$ are obtained based on eq.(1), since the cost of horizontal edges in $\mathcal{G}$ corresponds to these flows.

5. Derivation of cuts

The previous section describes three approaches with different levels of modelling complexity to define the costs of horizontal edges in $\mathcal{G}$ representing green phases, which include the effects of source terms on network-wide traffic state evolution for the capacity state. Building upon $\mathcal{G}$, in this section, we propose a method to estimate the realized network MFD. Thereby, we apply methods from the original MC and utilize recent findings in MFD-related research. First, we estimate the free-flow branch by deriving cuts from $\mathcal{G}$. Similar to Leclercq and Geroliminis (2013), moving observers are free to choose any path in $\mathcal{G}$, and are not restricted to one direction as in Daganzo and Geroliminis (2008). Second, we approximate the capacity branch also based on $\mathcal{G}$. Finally, we estimate the network-wide jam density, and exploit the symmetry of the link FD indicated in Laval and Castrillón (2015) and Daganzo and Knoop (2016) to derive the congested branch.

5.1. Free-flow branch

Similar to the original MC, we rely on the concept of (virtual) moving observers that travel through the network. We aim at estimating the free-flow branch of each route MFD to be able to aggregate them to the free-flow branch of the network MFD. However, flow conservation within a route might be violated when source terms exist, and thus MC cannot be directly applied (see Appendix **??**). To address this, we propose an approximate approach. It is based on the commonly known notion of path flows, which we will refer to as *route flows* to avoid any confusion with the term *path* as used in Section 2.

Route flows consist of vehicles which share a common route from an origin to a destination, and flow conservation applies. Thus, we can apply the original MC by deriving hyperlinks for these routes from the hypernetwork $\mathcal{G}$. Spillback effects are accounted for as the costs are derived from $\mathcal{G}$, and demand-related effects are not considered as flow conservation applies. The general cost of an edge representing a green phase relates to the share of the route flow to the total flow at the specific intersection approach. By deriving all route MFDs, and subsequently taking the route length-weighted average, one can approximate the network MFD. Note that this corresponds to the aggregation of traffic states according to Edie's definitions (Edie 1963).

Unfortunately, the number of routes connecting each OD pair can become intractable for realistic networks. Hence, to reduce the number of evaluated routes, we conjecture that the network MFD is estimated sufficiently well when the network variability is covered by the evaluated hyperlinks. The corresponding set of routes is non-overlapping and includes each intersection approach. Note that such a set is given by the set of corridors $\mathcal{C}$ defined by the network decomposition in Section 4.1. We derive the cuts corresponding to forward-moving observers as described in the original MC for each hyperlink. Eventually, we average the route-specific free-flow branches to derive the one corresponding to the network MFD.

5.2. Capacity branch

To derive the capacity branch of the network MFD, we fully utilize the hypernetwork $\mathcal{G}$. It already includes estimates for the maximum flows on each link, $\tilde{q}_k$. The computation of these flows considered both decreased demand due to a net outflow, and spillback induced capacity reductions. We take the link-weighted average to derive the network-wide capacity, as proposed in Geroliminis and Daganzo (2008). Note that phenomena such as the short-blocks problem (e.g. Daganzo and Lehe 2016) are taken into account in our methods to define the hypernetwork as we consider related spillback effects explicitly.

5.3. Congested branch

To conclude the estimation of the MFD we approximate its congested branch as follows. First, we estimate the network jam density K_{max}. Then, we exploit a symmetry in the evolvement of free-flow and congested traffic states considering the estimation for K_{max}. This symmetry was shown by Laval and Castrillón (2015), Daganzo and Knoop (2016) concerning the FD which exists when a canonical density transformation is applied. This aspect reflects the symmetric propagation of free-flow and congested traffic states throughout a link. They utilized this symmetry also for the MFD. Our main assumption here is that the symmetry still holds when the network jam density is reduced. We discuss this assumption in the Appendix **??**.

5.3.1. Network jam density To our best knowledge, no explicit attempts to estimate the network-wide jam density in the context of MFD approximation exist. The network jam density differs from the link jam density because the network is not necessarily fully jammed in case of gridlock (Daganzo, Gayah, and Gonzales 2011, Mahmassani, Saberi, and Zockaie 2013, Mazloumian, Geroliminis, and Helbing 2010).

First, we let the traffic in the physical network reach the capacity state. Then, we define virtual links at the destination nodes in the network, hereafter referred to as 'destination links', and simultaneously set their capacity to zero. This results in queues originating on those links which propagate through the network until a gridlock state is reached. By approximating the queue growth throughout the network, we can derive the network jam density K_{max}. Note that gridlocks in reality might occur due to an interplay of many different aspects such as outflow reductions due to spillbacks, driver heterogeneity, restrictive internal bottlenecks such as traffic lights, and additional demand generation inside or at the fringe of the given network. Nevertheless, we assume that reducing the outflow capacities at destination links leads to the highest number of jammed links in the network. Therefore, it results in the highest value for the network jam density which is preferable as we look for the MFD as an upper bound.

Approach 1 - Network variational theory (nVT) Similar to estimating the cost of horizontal edges in $\mathcal{G}$, we propose an approach based on nVT to estimate the network-wide jam density K_{max}. It has the advantage of being accurate as few assumptions are involved. On the other hand, this approach has a comparably high modeling complexity and computational cost. Alternatively, any other numerical solution method to solve network KWT problems applies. However, there exist synergies between nVT and the nMC framework proposed in this paper which we intend to utilize (see Section 4.3.1).

To estimate K_{max}, we let the network reach the capacity state as in the estimation of costs for $\mathcal{G}$, and then block the destination links as described above. This leads to queues being propagated throughout the entire network. Once the system reaches a stationary state, the link density can be derived by evaluating the difference of the Moskowitz function in the spatial dimension for each link, and subsequently dividing it by the link length:

$$K_k^i = \frac{N_{k,x=0}^i - N_{k,x=l_k^i}^i}{l_k^i}. \tag{18}$$

Then, the network-wide jam density is calculated as:

$$K_{max} = \frac{\sum_{i,k} K_k^i}{\sum_{i,k} l_k^i}, \tag{19}$$

Approach 2 - Queue propagation (QP) We propose a second approach which simplifies traffic dynamics compared to nVT. This leads to a reduction in estimation accuracy, but also decreases the computational cost involved.

We assume the capacity of destination links at the boundary of $\mathcal{N}$ to be zero and trace the propagation of queues throughout the network. Thus, we label this approach as 'queue propagation' (QP). Once a stationary gridlock state is reached, we derive the number of vehicles per link, and subsequently, take the link-weighted average to derive K_{max}.

For each link L_k^i we derive the time instant $t_{k,dq}^i$ when the queue starts to grow at the link's downstream end, and approximate the time when it reaches the link's upstream end, $t_{k,uq}^i$. Note that the latter time instant equals to $t_{k-1,dq}^i$, i.e. the time when the queue reaches the upstream link L_{k-1}^i, if no earlier blockage from an adjacent link occurs. To derive $t_{k,uq}^i$, we distinguish between two cases.

In the first case, we focus on links where no spillback occurred when the network was at capacity, i.e. $\tilde{q}_k^i < q_{k,s}^i$. Such a case is displayed in Figure 6. It shows a time-space diagram for the link L_k^i between two intersections I_{k-1}^i and I_k^i.

The left part of the figure reflects the traffic conditions for the network capacity state (see Section 4.3.2). At time $t_{k,dq}^i$ a queue reaches the intersection I_k^i from downstream. Note that $t_{k,dq}^i$ can result from downstream links from both corridors, C^i and C^j. However, the earlier queue is decisive for the further propagation, as we assume FIFO diverging behavior. If the earliest queue arrives from

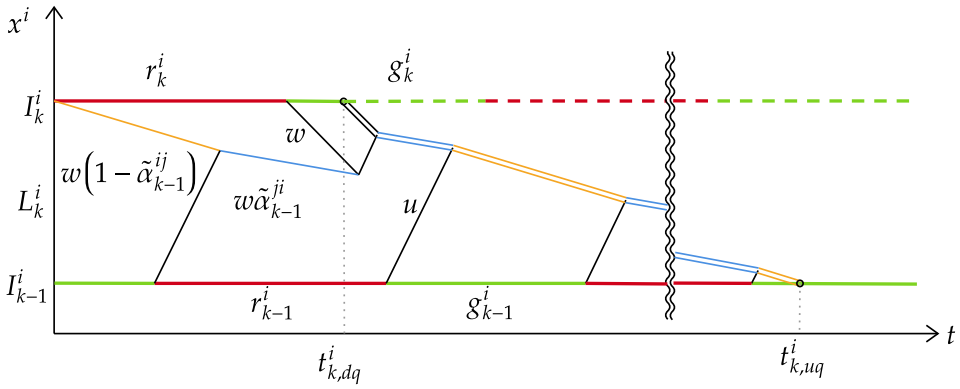


Figure 6 **Queue growth for undersaturated links.**

C^j, it occurs during a red phase on C^i. The queue on link L_k^i starts to grow with the start of the respective red phase, as vehicles stopping due to a red light will not be able to continue their travel once the green phase starts due to the existing queue downstream. Formally, the time when the queue starts to grow at the downstream end of link L_k^i can be written as:

$$t_{k,dq}^i = \min\left(\lfloor\frac{t_{k+1,uq}^j}{c_k^i}\rfloor c_k^i, t_{k+1,uq}^i\right), \tag{20}$$

where the $\lfloor . \rfloor$ is the *floor* function, which ensures $t_{k,dq}^i$ starts with a cycle and therefore with the red phase if the decisive queue comes from C^j.

From this time onward, the capacity at I_k^i is reduced to zero and the queue starts to grow on link L_k^i. During the occurrence of discharging flows, the queue grows with speed w as shown in the figure. In case that flows from the upstream intersection I_{k-1}^i reach the queue, it grows with speeds $w(1-\tilde{\alpha}_{k-1}^{ij})$ and $w\tilde{\alpha}_{k-1}^{ji}$. The growth of the queue is displayed as a double line in the figure. For the sake of simplicity, we neglect the impact of discharging flows which are supposed to be small in the considered case, and approximate the growth as the cycle-based average of the blue and orange shock waves:

$$t_{k,uq}^i = t_{k,dq}^i + \frac{l_k^i}{w}\frac{c_k^i}{r_k^i\tilde{\alpha}_{k-1}^{ji} + g_k^i(1-\tilde{\alpha}_{k-1}^{ij})}. \tag{21}$$

In the second case, we focus on an intersection I_{k+1}^i that is oversaturated when the network is at capacity state, and therefore a spillback occurs. Thus, the queue grows faster as in the first case since incoming flows are typically close to saturation.

This is shown in Figure 7 as a time-space diagram for link L_{k+1}^i between the intersections I_k^i and I_{k+1}^i. Note that a spillback occurs at I_k^i denoted as σ_k^i. In the figure, a queue reaches I_{k+1}^i at $t_{k+1,dq}^i$ from downstream. The queue grows with speed w as long as saturation flow reaches it. Closer to I_k^i, the queue grows depending on the inflows with speed $w(1-\tilde{\alpha}_{k-1}^{ij})$ and $w\tilde{\alpha}_{k-1}^{ji}$, and reaches I_k^i at

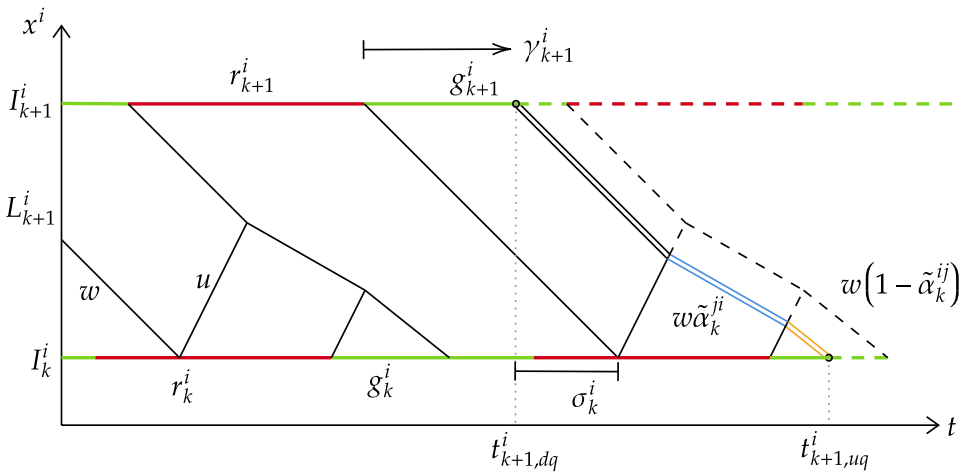


Figure 7 Queue growth for oversaturated links.

$t^i_{k+1,uq}$. If $t^i_{k+1,dq}$ equals the beginning of the red phase r^i_{k+1}, we can calculate $t^i_{k+1,uq}$ based on eq.(10) in Section 4.3.2, i.e. by deriving the spillback time σ^i_k. Contrarily, if $t^i_{k+1,dq}$ is close to the beginning of the green phase g^i_{k+1}, the queue grows with speed w nearly all the time. However, if $t^i_{k+1,dq}$ is exactly equal to the beginning of the green time, the queue starts growing at the beginning of the previous red phase. To describe this formally, we introduce the parameter $\gamma = [0,1]$, as shown in the figure, which is 1 if $t^i_{k,dq} = r^i_k$, and close to 0 if $t^i_{k,dq}$ is close to g^i_k. To approximate all cases in between, we linearly interpolate between $\gamma = (0,1]$. The case of $\gamma = 0$ is treated separately.

$$t^i_{k+1,uq} = \begin{cases} t^i_{k+1,dq} + \frac{l^i_{k+1}}{w} + \gamma^i_{k+1}(r^i_{k+1} - \sigma^i_k) & \text{, if } \gamma^i_{k+1} > 0 \\ t^i_{k+1,dq} + \frac{l^i_{k+1}}{w} - \sigma^i_k & \text{, otherwise.} \end{cases} \tag{22}$$

With equations (20)-(22) we approximate congestion propagation throughout the network. However, a blocked intersection due to a queue on C^i also blocks flows to C^j due to the FIFO diverging behavior. This leads to reduced inflows to the respective link. In the extremal case, the entire link on C_j might become empty due to such a blockage. This is essentially the reason for K_{max} being smaller than the link jam density κ_{max}. To account for such cases, we trace the last vehicle which enters and exits a link. A corresponding example is illustrated in Figure 8. In this example, a blockage occurs at intersection I^i_{k-1} at $t^i_{k-1,dq}$. This is due to a queue reaching the intersection from the adjacent corridor C^j, which is not shown in the figure. The remaining vehicles on link L^i_k join the existing queue at I^i_k, and leave the link at $t^i_{k,dq}$ during the respective green phase. In this example, no queue from downstream of I^i_k occurs at $t^i_{k,dq}$ and therefore the link L^i_k becomes empty. In contrast to the cases described above, the flows at I^i_k are reduced to zero due to a lack of demand and not due to a queue-induced blockage.

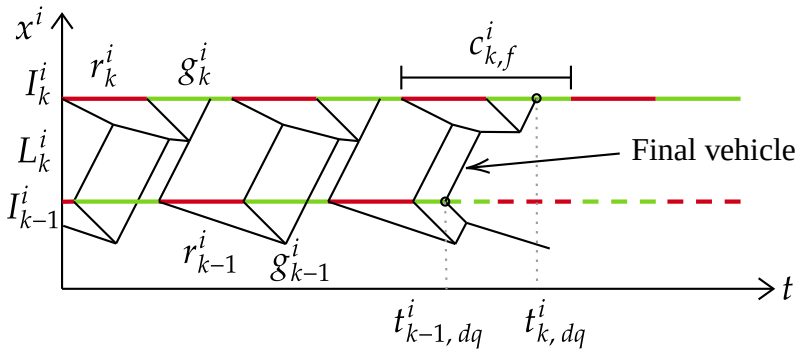


Figure 8 Downstream effects of spillbacks.

To describe this formally, let $c_{k,f}^i$ denote the beginning of the cycle in which the final vehicle reaches the existing queue at the downstream end of the link.

$$c_{k,f}^i = c_k^i \lfloor \frac{t_{k-1,dq}^i + \frac{l_k^i}{u}}{c_k^i} \rfloor, \tag{23}$$

The remaining flow relevant to $c_{k,f}^i$ is denoted as $q_{k,f}^i$. It can be derived based on inflows at the upstream intersection. The ratio of the remaining flow to the saturation flow q_{max} corresponds to the part of the green phase where flows larger than zero occur at I_k^i. Thus, the time $t_{k,dq}^i$ when $q = 0$ is equal to the red phase r_k^i of the cycle $c_{k,f}^i$, and the part of the green time until the intersection is cleared.

$$t_{k,f}^i = c_{k,f}^i + r_k^i + g_k^i \frac{q_{k,f}^i}{q_{max}}, \tag{24}$$

To account for this, we rewrite eq.(20):

$$t_{k,dq}^i = \min \left(\lfloor \frac{t_{k+1,uq}^j}{c_k^i} \rfloor c_k^i, t_{k+1,uq}^i, t_{k,f}^i \right). \tag{25}$$

This completes the set of equations required to track congestion propagation throughout the network for the gridlock case. The equations (21), (22), and (25) are applied for each link L_k^i on a corridor C^i of the set $\mathcal{C}$. However, the queues from an adjacent corridor might reach an intersection earlier than the current one being evaluated with the equations. Thus, an iterative approach needs to be employed until the resulting times $t_{k,uq}^i$ and $t_{k,dq}^i$ converge. This further ensures the consideration of circular dependencies.

Subsequently, we estimate the maximum density per link K_k^i during the gridlock state. Thereby, we consider that a link might not fully load due to a blockage from an adjacent corridor at the upstream intersection. The time span needed to fully load the link is $t_{k,uq}^i - (t_{k,dq}^i - l_k^i/u)$, which is based on the equations above. However, the actual loading duration for a link is $\max(t_{k-1,dq}^i - t_{k,dq}^i + l_k^i/u, 0)$,

since $t^i_{k-1,dq}$ also accounts for blockages from adjacent corridors. The maximum operator is applied to exclude infeasible values. We approximate the actual maximum density per link as follows:

$$K^i_k = \kappa_{max} \frac{\max(t^i_{k-1,dq} - t^i_{k,dq} + l^i_k/u, 0)}{t^i_{k,uq} - t^i_{l,dq} + l^i_k/u}, \tag{26}$$

Based on K^i_k for each link, we eventually apply eq.(19) to derive K_{max}.

5.3.2. Density transformation In order to finish the estimation of the congested branch of the network MFD, we utilize the symmetry between the free-flow and congested branch, as suggested by Laval and Castrillón (2015), Daganzo and Knoop (2016).

More specifically, we transform densities of the free-flow and capacity MFD branch:

$$\kappa' = \kappa - \frac{1}{2}\left(1 + q\frac{u+w}{uw}\right). \tag{27}$$

where κ' denotes the transformed density. Note that this equation is based on a normalized FD with $q_{max} = 1$ and $\kappa_{max} = 1$. Based on the estimation of the free-flow and the capacity branch in Sections 5.1 and 5.2, we know the corresponding (κ, q) values. As noted in Laval and Castrillón (2015), the respective transformed densities range between $[-0.5, 0]$.

6. Case study

In this section, we conduct a case study to test our framework for the network of Sioux Falls. Thereby, we apply the nMC to estimate the network-wide realized MFD, the current state-of-the-art methods, and the CTM to derive a ground truth.

6.1. Case study design

6.1.1. Network initialization The case study is conducted based on the bi-directional Sioux Falls network as shown in Figure 9. For the sake of simplicity, we alter the original Sioux Falls network slightly by only considering intersections with four or fewer legs, and we disallow left-turns at intersections. In total, the network consists of 23 intersections, shown as filled and empty circles, connected by 36 bi-directional links, as highlighted by the arrows. All intersections are controlled by a fixed signal control scheme with a cycle time of 90 s, and a green and red phase each of 45 s. All offsets are set to zero. We assume an identical FD for all links, characterized by $q_{max} = 1800$ veh/h, $u = 10$ m/s, and $w = -5$ m/s.

We define origin and destination nodes as those with less than four legs. They are shown as filled circles in the figure. At these nodes, we add virtual links which serve as origin and destination links and thereby allow to consider demand generation within the network. Additionally, we randomly select reasonable but different turning ratios for each intersection between $0.25 \leq \alpha \leq 0.75$. These limits are chosen to exclude extremal turning ratios which lead to special and rather unrealistic

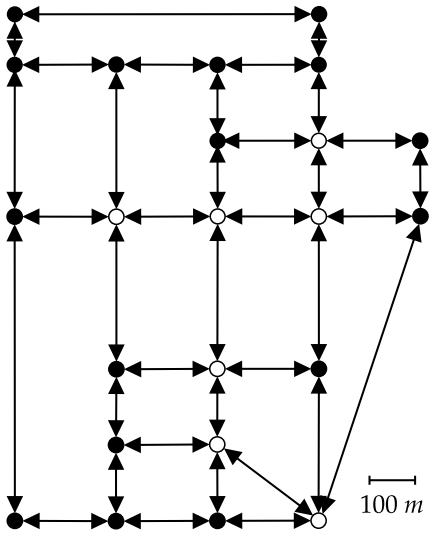


Figure 9 **Sioux Falls network for the case study.**

spillback occurrences in the network, as numerical investigations have shown. More importantly, the chosen turning ratios lead to an inter-dependency of traffic dynamics across the entire network. This includes the propagation of spillbacks within and across corridors, i.e. circular dependencies exist in our experiment. This becomes clear when considering that turning ratios at all intersections are non-zero and that the network is bidirectional. Thus, circular flows can occur between any set of at least four intersections forming a rectangle within the network.

6.1.2. Ground truth In order to further evaluate our MFD estimation results, we compare it to a ground truth derived from the CTM which is a macroscopic, time- and space-discrete traffic model. We implement the same network topology and the total simulation time to 2 h. We set the simulation time step to $\Delta t = 0.25$ s, and the spatial step size to $\Delta x = u\Delta t$ to satisfy the Courant-Friedrichs-Lewy condition (Courant, Friedrichs, and Lewy 1928). Setting a lower time-step was not feasible due to the high computational cost.

To derive the full range of the MFD with the CTM, we define several scenarios which are characterised by a constant inflow level and an exogenous reduction of the capacity at destination links. The former aspect is required in order to fully load the network and to reach capacity. The latter measure is applied to create significant congestion spreading throughout the network and affecting the stationary network-wide traffic state accordingly. While heavy congestion can generally occur in networks also without the explicit reduction of outflows, e.g. due to very restrictive intersection control, this procedure facilitates the replicability of the experiments and the reproducibility of the corresponding results. We consider multiple inflow and exogenous capacity reduction levels for the CTM simulation. More specifically, we specify ten different inflow levels between 0 and 900 veh/h per origin link, and 20 different capacity reduction factors between 0 and 1 at destination links. In

each scenario, the network is initially empty and the same level of inflow and capacity reduction applies to all origin and destination links, respectively. Subsequently, the network is loaded with the specified inflow, and traffic dynamics are simulated until a stationary state is reached. If a congested traffic state is to be obtained in a scenario, the respective exogenous capacity reduction at destination links applies once the network is stationary at the capacity level. This ensures the transition from a stationary state. To plot the related MFD, we aggregate flow and density values for one cycle length once stationary states are reached at the end of the simulation. We verify the existence of a network-wide stationary traffic state by ensuring that no differences exist between the average densities of the last two cycles in each simulation run.

Note that the CTM includes a certain error in the calculated density and flow values due to the numerical diffusion. This error converges to zero with a decreasing time step Δt. The chosen time-step represents a reasonable trade-off between accuracy and computational cost. Nevertheless, the ground truth results should be treated accordingly when being interpreted.

6.1.3. Proposed approaches The network geometry and topology as well as the triangular fundamental diagram are the main input data for the proposed approaches to derive the hypernetwork and estimate the network-wide jam density. The approximate approaches do not require any further data. The nVT, however, necessitates the specification of a numerical grid. Furthermore, the total computation period is set to 2 h to reach stationary traffic states, and the time-step is set to $\Delta t = 0.1$ s. This small time-step highlights the lower computational cost and consequently increased accuracy of nVT compared to the CTM. In order to estimate K_{max}, the capacity of destination links is set to zero after 1.5 h. This ensures that the stationary traffic states related to the network capacity are reached before the gridlock state occurs.

6.1.4. State of the art We choose the MC by Daganzo and Geroliminis (2008), by Leclercq and Geroliminis (2013), and the stochastic MC by Laval and Castrillón (2015) as state of the art for this case study. Hereafter, we abbreviate them with 'Dag', 'Lec', and 'Lav'. Note that none of these methods explicitly account for the effects of source terms, i.e. the exchange of flows across corridors. Furthermore, these methods implicitly focus on the idealized MFD for corridors. Still, we consider them as state of the art for estimating the realized MFD as no better alternatives exist to our best knowledge. More specifically, methods that estimate the MFD while accounting for turning flows in realistic networks do not exist since the reported studies all apply to simplistic networks such as regular grids or two-ring networks.

6.2. Results and discussion

6.2.1. Impact of hypernetwork generation method We compare the MFDs estimated by our proposed framework with the one derived by the 'Dag', 'Lec', and 'Lav' methods, as well as with

the CTM ground truth. The results are displayed in Figure 10. The x-axis displays the network-wide average density K in veh/km, and the y-axis the network-wide average flow Q in veh/h. The CTM results are shown as grey diamonds. The MFDs from the proposed framework are based on the nVT approach, as well as on the FS and LS approaches to derive $\mathcal{G}$ combined with the QP approximation for the network jam density. They are displayed as a solid black curve, a dashed black curve, and a dotted black curve, respectively. The MFD resulting from the state-of-the-art methods are shown as solid, dashed, and dotted grey curves.

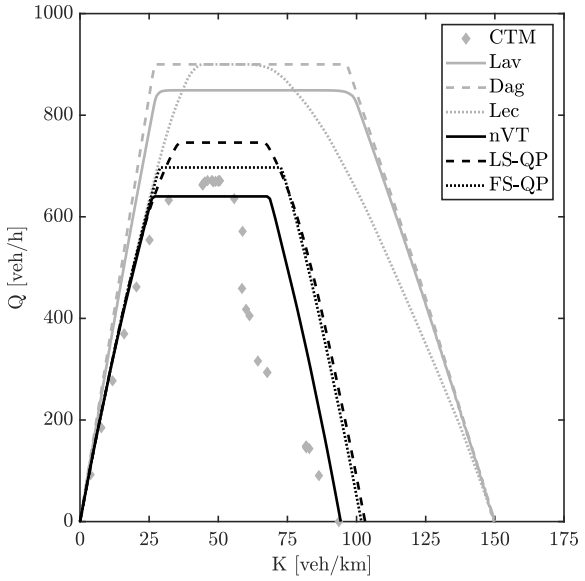


Figure 10 **Resulting MFDs from the proposed framework based on the nVT, the FS-QP, and the LS-QP approach, the state-of-the-art methods represented by the original method of cuts by Daganzo and Geroliminis (2008) ('Dag') and by Leclercq and Geroliminis (2013) ('Lec'), the stochastic approximation by Laval and Castrillón (2015) ('Lav'), and the CTM ground truth.**

The ground truth MFD reaches a capacity of ca. 670 veh/h, and a network-wide jam density of ca. 94 veh/km.

The nVT approach to derive the hypernetwork and estimate the network-wide jam density results in the most accurate MFD estimation as indicated in the figure. The capacity is ca. 25 veh/h lower than the one of the CTM, while the network-wide jam density is the same. The reason for this is that nVT itself solves the KWT problem at the network level accurately, and its integration into the hypernetwork generation, as well as into the estimation of K_{max}, let the overall framework profit from such accuracy. As expected, the estimated MFDs resulting from the hypernetwork built based on the approximate approaches are less accurate. The first approach, abbreviated as 'FS-QP', which includes network-wide spillback propagation, estimates Q_{max} as accurately as the nVT approach,

with an error of 25 veh/h. However, here the difference results from an overestimation rather than an underestimation. The error in the jam density estimation is ca. 8 veh/h. The second one, abbreviated as 'LS-QP', which does not consider network-wide spillback propagation overestimates the capacity by ca. 75 veh/h and the jam density by ca. 10 veh/km. In reality, an accurate capacity branch estimation is more important as the congested branch is rarely observed (e.g. Loder et al. 2019). Thus, we adjudge the increased modeling complexity of the FS approach as valuable for the presented case.

The differences in the network-wide capacity and the jam density estimation by the nVT approach and the approximate ones are due to the assumptions made within the approximations. In particular, averaging the downstream spillback impact uniformly on the downstream capacity neglects the temporal component of the spillback occurrence. However, such temporal aspects impact the further propagation of the spillback, i.e. to what share it is propagated to corridors C^i and C^j. Similar to that, we assume the inflow into a link as the equally weighted mean of both retaining and turning flows to derive K_{max}. The shorter a link is the more likely these assumptions are violated.

The results from the state-of-the-art methods substantially differ from the ground truth. Both the capacity and the jam density are significantly overestimated. However, this is expected, as these methods cannot account for the effects of turning flows on the network MFD. In other words, they cannot account for the effects of spatial demand patterns which is important to derive an estimate for the realized MFD as our results clearly show.

Given the error included in the ground truth and the magnitude of improvement compared to the results of the state-of-the-art methods, the differences in Q_{max} and K_{max} appear to be small for all MFDs resulting from our proposed framework. Moreover, these first results indicate that indeed the overall estimation accuracy is the highest for the nVT approach and the consideration of spillback propagation throughout the network in the FS approach is valuable.

6.2.2. Computational costs The modeling complexity of the nVT approach is clearly the highest compared to the approximate ones. This is not surprising, as it numerically solves the underlying KWT problem at the network level. For the case study settings, the nVT takes 4 h 50 min to derive the MFD. The FS-QP approach consists of more decision variables in the optimization problem than the LS-QP one and thus has the higher complexity of the two. Both approaches take 1.2 min for the case study settings and are thus substantially cheaper than the nVT from a computational perspective. The computational cost of the nVT strongly scales with the chosen time-step and the spatial extent of the network (Tilg et al. 2021). For example, if the time-step is set to $\Delta t = 1$ s, the computation time is only 14.3 minutes for the case study. The other two methods scale primarily with the number of intersection legs in the network.

The computation times for these three approaches are of the magnitude of some minutes, while the CTM requires several days to derive the MFD for this case study. The methods were evaluated on a computer with an Intel(R) Xeon(R) W-2145 CPU with 3.70 GHz and 64 GB RAM. This high computational time for the ground truth results from the small time-step (even though it is still 2.5 times larger than that used in the nVT approach), the resulting large number of cells, the necessity to reach stationary states and to replicate the simulation for multiple demand levels. Similar or even larger computational requirements could be equally expected from other simulations. Furthermore, note that the optimization procedure integrated in the nMC framework ensures that highest network-wide flows are found, while the CTM-related MFD only represents a point-wise evaluation of traffic states which can miss the true capacity.

6.2.3. Sensitivity analysis Here, we further analyze the nMC framework with respect to two specific aspects. First, we evaluate how the selection of routes impacts the estimation of the free-flow branch. Second, we investigate the impact of turning ratios on the estimation of the MFD. Below, we describe the main results from the sensitivity analysis. The details are provided in Appendix **??**.

- **Route selection:** As described in Section 5.1, the free-flow branch is estimated based on a specific set of routes. This is done to reduce modeling complexity and computational burden while providing a reasonable approximation of the network MFD. The sensitivity analysis reveals that indeed the estimation's accuracy is further increased by considering more routes in the process. In the end, the selection of routes of the MFD estimation is use-case dependent and the choice represents a trade-off between computational burden and estimation accuracy.
- **Turning ratios:** In order to assess the impact that spatial demand patterns have on the MFD, we first analyze the variability of the resulting MFDs associated with different demand patterns compared to the ground truth. The significant variability in the results indicates that the MFD indeed depends on the spatial demand pattern in the network. This further shows that a framework to estimate the MFD for specific spatial demand patterns is beneficial. Despite the assumption regarding the existence of stationary states, which might be violated in reality, the approximated MFDs can be useful. Furthermore, we compare the results from our proposed framework to the state of the art and the ground truth for several cases of different sets of turning ratios to assess the quality of our approach. The results clearly show the substantial improvement in the estimation accuracy regarding the network-wide capacity and jam density of all three proposed methods. On average, the estimated capacity and jam density are more than five times closer to the ground truth value than those estimated with the state of the art methods.

7. Conclusion

In this paper, we propose a framework, the nMC, to estimate the network-wide realized MFD without relying on extensive empirical data or microscopic simulations. Its flexibility allows to trade-off modeling complexity and computational cost against estimation accuracy. We base our framework on the well-established original MC (Daganzo and Geroliminis 2008, Leclercq and Geroliminis 2013) and follow its philosophy while exogenously accounting for the effects of source terms at intersections. The framework consists of two main steps after the problem initialization. First, we build a hypernetwork that reflects the network flows in the capacity state. For this purpose, we propose three different methods, a precise one based on the nVT (Tilg et al. 2021) and two approximate ones based on analytical formulations. The hypernetwork is the basis to derive the network MFD. Similar to the original MC, we derive cuts to approximate the free-flow branch of the MFD. Moreover, the hypernetwork allows us to efficiently estimate the network capacity. Last, we estimate the congested branch of the network MFD by approximating the maximum density in the network during gridlock, and utilizing symmetries of the underlying link FD. We thoroughly investigate the proposed framework in a case study. This includes the comparison to a ground truth provided by the CTM, as well as to the MC (Daganzo and Geroliminis 2008, Leclercq and Geroliminis 2013, Laval and Castrillón 2015) as state-of-the-art for realistic networks. Furthermore, we analyze our framework's robustness to different spatial demand patterns. The results demonstrate that the proposed framework enables one to estimate the realized network MFD sufficiently well for a realistic network. Moreover, they evidently show the improvement of the estimation accuracy of the network-wide capacity and jam density by our proposed framework compared to existing methods.

Our methodology allows estimating the realized MFD for realistic urban networks without the information loss induced by the reduction of networks to a single corridor as implicitly done by existing methods. Therefore, our framework is able to account for different spatial demand patterns and to determine the upper bound of the network-wide average flow. The potential of this methodology lies in its flexibility, and reduced computational cost compared to simulation-based studies. Moreover, it sheds light on the impact of specific assumptions on the realized MFD estimation. Note that existing approximation methods for realistic networks neglect the effects of demand patterns. Following the empirical study of Ambühl et al. (2021) and the theoretical work by Leclercq and Paipuri (2019) which establish the demand-dependency of the MFD from a temporal perspective, our framework is methodological proof that the effects of the spatial aspects of demand on the realized MFD cannot be neglected as well. One field of application is the calibration of aggregate traffic models for large metropolitan regions (Mariotte et al. 2020).

Future work includes the derivation of the idealized MFD based on our proposed framework. This can be achieved by integrating the nMC into an optimization framework with turning ratios as

decision variables and the maximum network-wide flow as the objective value. This would lead to the maximum MFD while considering the effects of spatial demand patterns on network-wide traffic states as well as spatial dependencies of traffic dynamics. Moreover, multi-modal aspects can be incorporated by modeling the effects of buses on traffic flow and integrating a passenger model. The incorporation of the nVT facilitates the modeling of such aspects as moving or stationary intra-link bottlenecks. This can be modeled as such a type of bottleneck, and therefore can be interesting for estimating the three-dimensional MFD (Geroliminis, Zheng, and Ampountolas 2014). Similar extensions of the MC exist already for the corridor level (Dakic et al. 2020, Chiabaut 2015), and are therefore promising starting points for this research direction. Last, by integrating continuous source terms (Laval, Costeseque, and Chilukuri 2016) in nVT, the effects of generated and finished trips along links can potentially be modeled. This could then be incorporated in the nMC framework to offer the user multiple demand generation methods.

Acknowledgments

G. Tilg acknowledges support from the German Federal Ministry of Transport and Digital Infrastructure (BMVI) for the funding of the project LSS (Capacity increase of urban networks). S. F. A. Batista and M. Menéndez acknowledge the support by the NYUAD Center for Interacting Urban Networks (CITIES), funded by Tamkeen under the NYUAD Research Institute Award CG001. L. Ambühl acknowledges the support by the ETH Research Grant ETH-27 16-1 under the project name SPEED. L. Leclercq acknowledges funding by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 646592 - MAGnUM project).

References

Aghamohammadi R, Laval JA, 2022 *Parameter estimation of the macroscopic fundamental diagram: A maximum likelihood approach. Transportation Research Part C: Emerging Technologies* 140:103678.

Ambühl L, Loder A, Bliemer MC, Menendez M, Axhausen KW, 2020 *A functional form with a physical meaning for the macroscopic fundamental diagram. Transportation Research Part B: Methodological* 137:119–132.

Ambühl L, Loder A, Leclercq L, Menendez M, 2021 *Disentangling the city traffic rhythms: A longitudinal analysis of mfd patterns over a year. Transportation Research Part C: Emerging Technologies* 126:103065.

Batista S, Tilg G, Menéndez M, 2022 *Exploring the potential of aggregated traffic models for estimating network-wide emissions. Transportation Research Part D: Transport and Environment* 109:103354.

Chiabaut N, 2015 *Evaluation of a multimodal urban arterial: The passenger macroscopic fundamental diagram. Transportation Research Part B: Methodological* 81:410–420.

Courant R, Friedrichs K, Lewy H, 1928 *Über die partiellen differenzengleichungen der mathematischen physik. Mathematische annalen* 100(1):32–74.

Daganzo CF, 1994 *The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. Transportation Research Part B: Methodological* 28(4):269–287.

Daganzo CF, 1995 *The cell transmission model, part ii: network traffic. Transportation Research Part B: Methodological* 29(2):79–93.

Daganzo CF, 2005a *A variational formulation of kinematic waves: basic theory and complex boundary conditions. Transportation Research Part B: Methodological* 39(2):187–196.

Daganzo CF, 2005b *A variational formulation of kinematic waves: Solution methods. Transportation Research Part B: Methodological* 39(10):934–950.

Daganzo CF, 2007 *Urban gridlock: Macroscopic modeling and mitigation approaches. Transportation Research Part B: Methodological* 41(1):49–62.

Daganzo CF, Gayah VV, Gonzales EJ, 2011 *Macroscopic relations of urban traffic variables: Bifurcations, multivaluedness and instability. Transportation Research Part B: Methodological* 45(1):278–288.

Daganzo CF, Geroliminis N, 2008 *An analytical approximation for the macroscopic fundamental diagram of urban traffic. Transportation Research Part B: Methodological* (42):771–781.

Daganzo CF, Knoop VL, 2016 *Traffic flow on pedestrianized streets. Transportation Research Part B: Methodological* 86:211–222.

Daganzo CF, Lehe LJ, 2016 *Traffic flow on signalized streets. Transportation Research Part B: Methodological* 90:56–69.

Dakic I, Ambühl L, Schümperlin O, Menendez M, 2020 *On the modeling of passenger mobility for stochastic bi-modal urban corridors. Transportation Research Part C: Emerging Technologies* 113(November 2018):146–163.

Edie L, 1963 *Discussion of traffic stream measurements and definitions.* Almond J, ed., *Proceedings of the 2nd International Symposium on the Theory of Traffic Flow*, 139–154 (Paris, France: OECD).

Gan QJ, Jin WL, Gayah VV, 2017 *Analysis of traffic statics and dynamics in signalized networks: a poincaré map approach. Transportation science* 51(3):1009–1029.

Geroliminis N, Boyacı B, 2012 *The effect of variability of urban systems characteristics in the network capacity. Transportation Research Part B: Methodological* 46(10):1607–1623.

Geroliminis N, Daganzo CF, 2008 *Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. Transportation Research Part B: Methodological* 42(9):759–770.

Geroliminis N, Sun J, 2011 *Properties of a well-defined macroscopic fundamental diagram for urban traffic. Transportation Research Part B: Methodological* 45(3):605–617.

Geroliminis N, Zheng N, Ampountolas K, 2014 *A three-dimensional macroscopic fundamental diagram for mixed bi-modal urban networks. Transportation Research Part C: Emerging Technologies* 42:168–181.

Girault JT, Gayah VV, Guler I, Menendez M, 2016 *Exploratory analysis of signal coordination impacts on macroscopic fundamental diagram. Transportation Research Record: Journal of the Transportation Research Board* 2560(2560):36–46.

Jin WL, Gan QJ, Gayah VV, 2013 *A kinematic wave approach to traffic statics and dynamics in a double-ring network. Transportation Research Part B: Methodological* 57:114–131.

Johari M, Keyvan-Ekbatani M, Leclercq L, Ngoduy D, Mahmassani HS, 2021 *Macroscopic network-level traffic models: Bridging fifty years of development toward the next era. Transportation Research Part C: Emerging Technologies* 131(June):103334.

Knoop VL, Van Lint H, Hoogendoorn SP, 2015 *Traffic dynamics: Its impact on the macroscopic fundamental diagram. Physica A: Statistical Mechanics and its Applications* 438:236–250.

Laval JA, Castrillón F, 2015 *Stochastic approximations for the macroscopic fundamental diagram of urban networks. Transportation Research Part B: Methodological* 81:904–916.

Laval JA, Costeseque G, Chilukuri B, 2016 *The impact of source terms in the variational representation of traffic flow. Transportation Research Part B: Methodological* 94:204–216.

Leclercq L, Geroliminis N, 2013 *Estimating mfds in simple networks with route choice. Transportation Research Part B: Methodological* 57:468–484.

Leclercq L, Paipuri M, 2019 *Macroscopic Traffic Dynamics Under Fast-Varying Demand. Transportation Science* 53(6):1526–1545.

Leclercq L, Parzani C, Knoop VL, Amourette J, Hoogendoorn SP, 2015 *Macroscopic traffic dynamics with heterogeneous route patterns. Transportation Research Part C: Emerging Technologies* 59:292–307.

Loder A, Ambühl L, Menendez M, Axhausen KW, 2019 *Understanding traffic capacity of urban networks. Scientific Reports* 9(1):16283.

Mahmassani HS, Saberi M, Zockaie A, 2013 *Urban network gridlock: Theory, characteristics, and dynamics. Transportation Research Part C: Emerging Technologies* 36:480–497.

Mariotte G, Leclercq L, Batista S, Krug J, Paipuri M, 2020 *Calibration and validation of multi-reservoir mfd models: A case study in lyon. Transportation Research Part B: Methodological* 136:62–86.

Mazloumian A, Geroliminis N, Helbing D, 2010 *The spatial variability of vehicle densities as determinant of urban network capacity. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368(1928):4627–4647.

Newell GF, 1993 *A simplified theory of kinematic waves in highway traffic, part i: General theory, part ii: Queueing at freeway bottlenecks, part iii: Multi-destination flows. Transportation Research Part B: Methodological* 27(4):281–313.

Tilg G, Ambühl L, Batista SF, Menendez M, Busch F, 2021 *On the application of variational theory to urban networks. Transportation Research Part B: Methodological* 150:435–456.

930 Tilg G, Amini S, Busch F, 2020 *Evaluation of analytical approximation methods for the macroscopic funda-*
931 *mental diagram. Transportation Research Part C: Emerging Technologies* 114:1 – 19.

932 Xu G, Yu Z, Gayah VV, 2020 *Analytical method to approximate the impact of turning on the macroscopic*
933 *fundamental diagram. Transportation Research Record* 2674(9):933–947.