Failure of classical traffic and transportation theory: The maximization of the network throughput maintaining free flow conditions in network

Boris S. Kerner[1]

[1] *Physics of Transport and Traffic, University Duisburg-Essen, 47048 Duisburg, Germany*

We have revealed general physical conditions for the *maximization* of the network throughput at which free flow conditions are ensured, i.e., traffic breakdown cannot occur in the whole traffic or transportation network. A physical measure of the network – *network capacity* is introduced that characterizes general features of the network with respect to the maximization of the network throughput. The network capacity allows us also to make a general proof of the deterioration of traffic system occurring when dynamic traffic assignment is performed in a network based on the classical Wardrop' user equilibrium (UE) and system optimum (SO) equilibrium.

I. INTRODUCTION

To find traffic optima and organize dynamic traffic assignment and control in traffic and transportation networks, a huge number of traffic theories have been introduced (see, e.g., [1–23] and references there). Some of traffic models developed recently are devoted to the development of optimal dynamic feedback strategies in the networks (e.g., [15–20]) and search algorithms based on stochastic processes which find local optima with asymmetric look-ahead potentials (e.g., [21–23]).

The most famous approach to an analysis of traffic and transportation networks is based on the Wardrop's user equilibrium (UE) and system optimum (SO) equilibrium introduced in 1952 [1]. *Wardrop's UE*: traffic on a network distributes itself in such a way that the travel times on all routes used from any origin to any destination are equal, while all unused routes have equal or greater travel times. *Wardrop's SO*: the network-wide travel time should be a minimum. Wardrop's principles reflect either the wish of drivers to reach their destinations as soon as possible (UE) or the wish of network operators to reach the minimum network-wide travel time (SO). During last 50 years based on the Wardrop's equilibria a huge number of theoretical works to dynamic traffic assignment and control in the networks have been made by several generations of traffic and transportation researchers (see, e.g., [2–14] and references there).

In particular, in accordance with the Wardrop's equilibria, travel times or/and other travel costs on network links are considered to be self-evident traffic characteristics for objective functions used for optimization transportation problems [2–14]. The main aim of associated classical approaches is to minimize travel times or/and other travel costs in a traffic or transportation network (see, e.g., [2–17]). The classical traffic and transportation theories based on Wardrop's equilibria have made a great impact on the understanding of many traffic phenomena. However, network optimization approaches based on these theories have failed by their applications in the real world (see explanations of these critical statements and references in critical reviews [24, 25]). Even several

decades of a very intensive effort to improve and validate network optimization models have had no success. Indeed, there can be found no examples where on-line implementations of the network optimization models based on these theories could reduce congestion in real traffic and transportation networks.

As explained in recent reviews [24, 25], assumptions about traffic features used for the development of classical traffic flow models are not consistent with the empirical nature of traffic breakdown at network bottlenecks. The empirical nature of traffic breakdown explained in three-phase traffic theory [26, 27] is as follows: Traffic breakdown is a phase transition from free flow (F) to synchronized flow (S) that occurs in a *metastable state of free flow* at a network bottleneck. Each of network bottlenecks is characterized by a minimum capacity $C_{\min}$ that separates stable and metastable states of free flow at the bottleneck [26, 27]: At

$$q_{\text{sum}} < C_{\min}, \tag{1}$$

free flow is stable at the bottleneck, where q_{sum} is the flow rate in free flow at the bottleneck [24, 26]. Therefore, under condition (1) no traffic breakdown can occur at the bottleneck. Contrarily, at

$$q_{\text{sum}} \geq C_{\min} \tag{2}$$

free flow is metastable with respect to traffic breakdown at the bottleneck. Therefore, under condition (2) traffic breakdown can occur at the bottleneck.

In this paper, we reveal an approach to the maximization of the network throughput ensuring free flow conditions at which traffic breakdown cannot occur in the whole network. We call this approach *network throughput maximization approach*. We introduce a physical measure of a traffic or transportation network called *network capacity*. The network capacity characterizes general features of the network with respect to the maximization of the network throughput at which free flow conditions are ensured in the networks. One of the consequences of the general analysis of traffic and transportation networks made in the article is a general proof of the deterioration

of traffic system occurring when dynamic traffic assignment is performed in a network based on the Wardrop' equilibria. This can explain the failure of the application of the Wardrop' equilibria for the prevention of traffic congestion in real traffic and transportation networks.

The article is organized as follows. In Sec. II, we introduce the network throughput maximization approach for the prevention of breakdown in networks at the maximum network throughput as well as the network capacity. In discussion Sec. III we make a general analysis of the deterioration of the traffic system through the use of Wardrop's equilibria, illustrate this general analysis by numerical simulations, discuss how appropriate information to the drivers can be given to organize the maximization of the network throughput as well as consider a connection between the network throughput maximization approach introduced in this article and the breakdown minimization (BM) principle of Ref. [28, 29].

II. THE MAXIMIZATION OF THE NETWORK THROUGHPUT ENSURING FREE FLOW CONDITIONS IN NETWORK

A. Network model

In known real field (empirical) traffic data, traffic breakdowns occur usually at the same road locations of a traffic network called network bottlenecks (e.g., [26, 27, 30–32]). Network bottlenecks are caused, for example, by on- and off-ramps, road gradients, road-works, a decrease in the number of road lines (in the flow direction), traffic signals in city traffic, etc. A bottleneck introduces a speed disturbance localized in a neighborhood of the bottleneck. As a result, in the empirical data at the same flow rate the probability of traffic breakdown at a bottleneck on a network link is considerably larger than on the link outside the bottleneck.

We consider a traffic or transportation network with N network bottlenecks, where $N > 1$. We assume that there are M network links (where $M > 1$) for which the inflow rates q_m, $m = 1, 2, \ldots, M$ can be adjusted; q_m is the link inflow rate for a link with index m. At the network boundaries, there are I links for the network inflow rates $q_i^{(o)}(t)$, $i = 1, 2, \ldots, I$ (called "origins", for short O_i, $i = 1, 2, \ldots, I$), where $I \geq 1$. At the network boundaries, there are also J links for the network outflow rates $q_j^{(d)}(t)$, $j = 1, 2, \ldots, J$ (called "destinations", for short D_j, $j = 1, 2, \ldots, J$), where $J \geq 1$. The network inflow rates $q_i^{(o)}(t)$ and the total network inflow rate $Q(t)$ are determined by the network inflow rates $q_{ij}^{(o)}(t)$ of vehicles moving from origin O_i to destination D_j (called as origin-destination pair O_i–D_j of the network) through the well-known formulas

$$q_i^{(o)}(t) = \sum_{j=1}^{J} q_{ij}^{(o)}(t), \quad Q(t) = \sum_{i=1}^{I} q_i^{(o)}(t). \qquad (3)$$

In this article, the network inflow rates $q_{ij}^{(o)}(t)$ (origin-destination matrix) are assumed to be known time-functions. Each of the network bottlenecks $k = 1, 2, \ldots, N$ is characterized by a minimum highway capacity $C_{\min}^{(k)} = C_{\min}^{(k)}(\alpha_k, \mathbf{R}_k)$, α_k is the set of control parameters of bottleneck k [33]; $\mathbf{R}_k$ is a matrix of percentages of vehicles with different vehicle (and/or driver) characteristics that takes into account that dynamic assignment is possible to perform individually for each of the vehicles [34]; the flow rate in free flow at bottleneck k will be denoted by $q_{\text{sum}}^{(k)}$.

In contrast with classical traffic and transportation theories (see, e.g., [2–14]), our approach for the maximization of the network throughput does not include travel times (or other "travel costs") on different network routes. To avoid the use of network routes with considerably longer travel times in comparison with the shortest routes, the network throughput maximization approach to dynamic traffic assignment and network control is applied for some "alternative network routes (paths)" (for short, alternative routes) *only*. This means that there is a constrain "alternative network routes (paths)" by the application of the network throughput maximization approach.

We define the constrain "alternative network routes (paths)" as follows. The alternative routes are possible different routes from origin O_i to destination D_j (where $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$) for which the maximum difference between travel times in free flow conditions does not exceed a *given value* denoted by ΔT_{ij} (values ΔT_{ij} can be different for different origin-destination pairs O_i–D_j of the network, where $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$). Before the network throughput maximization approach is applied to a large traffic or transportation network, for each of the origin-destination pairs O_i–D_j of the network a set of the alternative routes should be found.

B. Network throughput maximization approach: Prevention of breakdown in networks at the maximum network throughput

In real traffic and transportation networks the network inflow rate Q (3) and, therefore, the flow rates $q_{\text{sum}}^{(k)}$, $k = 1, 2, \ldots, N$ in the network increase from very small values (at night) to large values during daytime. At small initial values Q and $q_{\text{sum}}^{(k)}$ condition (1) is valid for each of the network bottlenecks:

$$q_{\text{sum}}^{(k)} < C_{\min}^{(k)}, \quad k = 1, 2, \ldots, N. \qquad (4)$$

In accordance with condition (1), conditions (4) mean that free flow is stable with respect to traffic breakdown at each of the network bottlenecks. Therefore, no traffic breakdown can occur in the network. Due to the increase in Q, the flow rate $q_{\text{sum}}^{(k)}$ at least for one of the network

bottlenecks $k = k_1^{(1)}$ becomes very close to $C_{\min}^{(k)}$. Therefore, in formula

$$q_{\text{sum}}^{(k)} + \epsilon^{(k)} = C_{\min}^{(k)} \text{ for } k = k_1^{(1)} \qquad (5)$$

a positive value $\epsilon^{(k)} = C_{\min}^{(k)} - q_{\text{sum}}^{(k)}$ becomes very small:

$$\epsilon^{(k)}/C_{\min}^{(k)} \ll 1 \text{ for } k = k_1^{(1)}. \qquad (6)$$

The approach to the maximization of the network throughput at which free flow conditions are ensured in the whole traffic or transportation network (network throughput maximization approach) is as follows. We maintain condition (5) at the expense of the increase in q_m on other alternative routes. As a result, condition $q_{\text{sum}}^{(k)} + \epsilon^{(k)} = C_{\min}^{(k)}$ at $\epsilon^{(k)}/C_{\min}^{(k)} \ll 1$ is satisfied for another bottleneck $k = k_2^{(1)}$. When Q increases further, we repeat the above procedure for other network bottlenecks. Consequently, we get

$$q_{\text{sum}}^{(k)} + \epsilon^{(k)} = C_{\min}^{(k)} \text{ for } k = k_z^{(1)} \qquad (7)$$
$$z = 1, 2, \ldots, Z_1, \text{ where } Z_1 \geq 1, \ Z_1 \leq N,$$

where $\epsilon^{(k)}/C_{\min}^{(k)} \ll 1$. Thus, when Q increases, we maintain conditions (7) at the expense of the increase in q_m on other alternative routes.

However, due to the constrain "alternative routes", the number Z_1 of bottlenecks satisfying (7) is limited by some value Z (where $Z \leq N$). All values $\epsilon^{(k)}$ in (7) are positive ones. Therefore, conditions (7) are equivalent to (4). Thus, due to the application of the network throughput maximization approach free flow remains to be stable with respect to traffic breakdown (F→S transition) at each of the network bottlenecks. For this reason, no traffic breakdown can occur in the network. This means that conditions (7) in the limit case $Z_1 = Z$ and $\epsilon^{(k)} \to 0$ are related to *the maximum possible network throughput* at which conditions (4) for the stability of free flow in the whole network are still satisfied. This is because at the subsequent increase in the network inflow rate Q the constrain "alternative routes" does not allow us to maintain conditions (7) at the expense of the increase in q_m on other alternative routes: At least for one of the network bottlenecks conditions (4) cannot be satisfied any more.

The latter means that when the maximum network throughput determined by conditions (7) in the limit case $Z_1 = Z$ and $\epsilon^{(k)} \to 0$ is exceeded, then at least for one of the network bottlenecks condition (2) is valid. Under condition (2), free flow is in a metastable state with respect to traffic breakdown (F→S transition) at a network bottleneck. Therefore, traffic breakdown can occur in the network.

C. A physical measure of traffic and transportation networks – Network capacity

Within a steady-state analysis of traffic and transportation networks [35], the limit case $Z_1 = Z$ in conditions (7) allows us to define a network measure (or "metric") – *network capacity* denoted by C_{net} as follows. The network capacity C_{net} is the maximum total network inflow rate Q at which conditions [36, 37]

$$q_{\text{sum}}^{(k)} = C_{\min}^{(k)} \text{ for } k = k_z^{(1)}, \qquad (8)$$
$$q_{\text{sum}}^{(k)} < C_{\min}^{(k)} \text{ for } k \neq k_z^{(1)},$$
$$z = 1, 2, \ldots, Z; \ Z \geq 1, \ Z \leq N; \ k = 1, 2, \ldots, N$$

are satisfied. From a comparison of conditions (7) and (8) we can see that the total network inflow rate reaches the network capacity [38]

$$Q = C_{\text{net}} \qquad (9)$$

when the limit case $Z_1 = Z$ in (7) is realized *and* all values $\epsilon^{(k)}$ in (7) are set to zero: $\epsilon^{(k)} = 0$. In accordance with the definition of the network capacity (8) and condition (2), at $Q = C_{\text{net}}$ (9) free flow becomes the metastable one with respect to traffic breakdown at network bottleneck(s) $k = k_z^{(1)}$ (8). This means that under conditions (8) traffic breakdown can occur in the network.

When

$$Q < C_{\text{net}}, \qquad (10)$$

as explained above, conditions (7) are satisfied. Under conditions (7), free flow is stable with respect to traffic breakdown (F→S transition) in the whole network. Therefore, no traffic breakdown can occur in the network. This explains the physical sense of the network capacity C_{net}: Dynamic traffic assignment in the network in accordance with the network throughput maximization approach maximizes the network throughput at which traffic breakdown cannot occur in the whole network.

Respectively, when

$$Q > C_{\text{net}}, \qquad (11)$$

due to the constrain "alternative routes" of the network throughput maximization approach at least for one of the network bottlenecks the flow rate in free flow at the bottleneck becomes larger than the minimum bottleneck capacity: $q_{\text{sum}}^{(k)} > C_{\min}^{(k)}$. For this reason, rather than conditions (8) under condition (11) we get [41]

$$q_{\text{sum}}^{(k)} > C_{\min}^{(k)} \text{ at } k = k_w^{(2)}, \qquad (12)$$
$$q_{\text{sum}}^{(k)} \leq C_{\min}^{(k)} \text{ at } k \neq k_w^{(2)},$$
$$w = 1, 2, \ldots, W; \ W \geq 1, \ W \leq N; \ k = 1, 2, \ldots, N.$$

Therefore, in accordance with (2), free flow is the metastable one with respect to traffic breakdown at the related network bottleneck(s). This means that under conditions (12) traffic breakdown can occur in the network.

D. The maximization of the network throughput in non-steady state of network

As above-mentioned, the definition of the network capacity (8), (9) is valid only within a steady-state analysis

of the networks. Contrarily, conditions (7) at $Z_1 = Z$ and $\epsilon^{(k)} \to 0$, for which free flow is stable with respect to traffic breakdown at each of the network bottlenecks, are applicable even when free flow distribution in the network cannot be considered a steady one. Indeed, in conditions (7) only local flow rates $q_{\text{sum}}^{(k)}$ in free flow at network bottlenecks are used.

Therefore, conditions (7) at which free flow conditions persist in the whole network can be used for dynamic traffic assignment in real traffic and transportation networks under real non-steady state conditions, i.e., without involving an analysis of the network capacity.

We can see that the basic objective of the approach of the maximization of the network throughput introduced in the paper is to guarantee that condition (2) is satisfied at *none* of the network bottlenecks.

III. DISCUSSION

A. The network throughput maximization approach versus Wardrop's equilibria

From results of Ref. [28] in which a different application of the BM principle has been studied (see Sec. III B), one can assume that the assignment procedure with the network throughput maximization approach (5)–(7) introduced in this article should give a better performance than assignment procedures designed through the use of Wardrop's UE or SO. However, the following questions arise that could not be answered in [28]:

(i) Even when the network throughput maximization approach (7) exhibits the better performance in comparison with the Wardrop's equilibria, whether does lead to *large enough benefits* that justify to use this approach instead of the Wardrop's equilibria? Indeed, it seems that the use of the network throughput maximization approach (7) exhibits considerable disadvantages: Some of the drivers should use longer routes.

(ii) Is there a general measure for a comparison of the performance of different assignment procedures for an arbitrary traffic network?

Both a general analysis and simulations of the network throughput maximization approach versus Wardrop's equilibria are made below *only* for the case $Q < C_{\text{net}}$ (10), when the application of the network throughput maximization approach (7) ensures that traffic breakdown cannot occur in the network. One of the benefits of this analysis is that we can find a general explanation of the deterioration of traffic system through application of Wardrop's equilibria that is independent on network characteristics.

1. Deterioration of traffic system through application of Wardrop's equilibria: General analysis

A real traffic or transportation network consists of alternative routes with very different lengths. Therefore, at small enough flow rates q_m, there are routes with short travel times ("short routes") and routes with longer travel times ("long routes"). When Q and, consequently, values q_m increase, the minimization of travel times in the network with the use of dynamic traffic assignment based on Wardrop's UE or SO leads to considerably larger increases in the flow rates on short routes in comparison with increases in the flow rates on long routes (e.g., [2–14]).

Therefore, at $Q < C_{\text{net}}$ (10), rather than conditions (7), any application of Wardrop's UE or SO leads to conditions $q_{\text{sum}}^{(k)} < C_{\text{min}}^{(k)}$ for some of the bottlenecks on long routes, whereas for some of the bottlenecks on short routes, we get $q_{\text{sum}}^{(k)} > C_{\text{min}}^{(k)}$. Therefore, in accordance with condition (2), traffic breakdown can occur at network bottlenecks on the short routes.

One of the consequences of this general conclusion is that already at relatively small total network inflow rates $Q < C_{\text{net}}$ the application of the Wardrop's equilibria leads to the occurrence of congestion in urban networks. An improvement of the performance of the Wardrop's equilibria with respect to the prevention of traffic breakdown in a network could *not* be achieved, even if the metastability of free flow with respect to traffic breakdown at network bottlenecks has been taken into account in travel time costs [42]. Indeed, due to the metastability that is realized under condition (2), travel time costs can exhibit discontinuities when traffic breakdown has occurred. However, possible discontinuities in travel time costs cannot change the above conclusion that the application of the Wardrop's equilibria at $Q \to C_{\text{net}}$ does result in the metastability of free flow at some of the network bottlenecks: Under free flow conditions, in accordance with the Wardrop's UE or SO the flow rate on a short route should be larger than the flow rate on a long route. This is independent on values $C_{\text{min}}^{(k)}$ for bottlenecks on the route. Contrarily, in accordance with conditions (7) resulting from the network throughput maximization approach, the flow rate at any network bottleneck $q_{\text{sum}}^{(k)}$ should be smaller than the minimum capacity $C_{\text{min}}^{(k)}$ of the bottleneck. This does not depend on whether a bottleneck is on a short route or it is on a long route.

The above conclusion that at $Q \to C_{\text{net}}$ any application of the Wardrop's equilibria results in the occurrence of the metastable free flow at some of the network bottlenecks can additionally be explained through a consideration of the following hypothetical network: In the network with different lengths of alternative routes, values of minimum capacity C_{min} are assumed the same for any bottleneck. We assume that our statement about the metastability of free flow at some of the network bottlenecks might be incorrect: An application of Wardrop's

UE or SO might result in conditions (7) at which free flow is stable with respect to traffic breakdown at each of the network bottlenecks. However, at $C_{\min}^{(k)} = C_{\min}$ and $\epsilon^{(k)} \to 0$ in (7), on routes with different lengths the flow rates $q_{\text{sum}}^{(k)}$ must be the same for all bottlenecks for which conditions (7) are satisfied. This contradicts the sense of any application of Wardrop's UE or SO: On average, the flow rates should be larger on short routes then those on long routes.

Thus, due to the application of Wardrop's UE or SO, already at $Q < C_{\text{net}}$ (10) it turns out that for some of the network bottlenecks $q_{\text{sum}}^{(k)} > C_{\min}^{(k)}$. Therefore, traffic breakdown can occur at these bottlenecks. In this case, it is not possible to predict the time instant at which traffic breakdown occurs at these bottlenecks. This is because the time delay to the breakdown $T^{(B)}$ is a random value [24, 26]. We can apply the Wardrop's UE or SO for dynamic traffic assignment without any delay after the breakdown has occurred. This is possible because the speed decrease due to the breakdown can be measured. However, after the assignment has been made, there is always a time delay in the change of traffic variables at the bottleneck location. This time delay is caused by travel time from the beginning of a link to a bottleneck location on the link. Therefore, it is not possible to avoid congested traffic occurring due to the breakdown. A control method can only effect on a spatiotemporal distribution of congestion in the network.

From this general analysis, we can see that the main benefit of the network throughput maximization approach in comparison with the Wardrop's equilibria is as follows: Even when the total network inflow rate Q approaches the network capacity, i.e., at $Q \to C_{\text{net}}$ the network throughput maximization approach does ensure free flow conditions at which traffic breakdown cannot occur in the whole network. Contrarily, at $Q \to C_{\text{net}}$ any application of the Wardrop's equilibria does lead to congested traffic in the network.

The physics of traffic breakdown in traffic and transportation networks revealed in Secs. II B and II C shows that the wish of humans to use shortest routes of a network contradicts fundamentally the physical nature of traffic breakdown in the network. Therefore, in contrast to the human wish, the use of the classical Wardrop's equilibria results basically in the occurrence of congestion in urban networks. This can explain why approaches to dynamic traffic assignment based on the classical Wardrop's UE or SO equilibria (e.g., [2–14]) have failed by their applications in the real world.

2. Numerical simulations

The Wardrop's UE or SO and the network throughput maximization approach are devoted to the optimization of large, complex vehicular traffic networks. However, to illustrate the general conclusions made above, it is
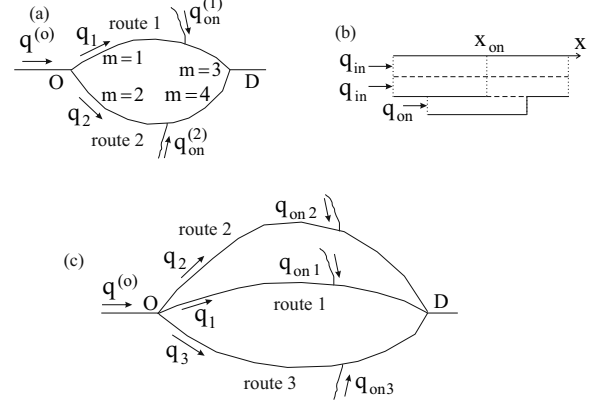

FIG. 1: Model of two-route and three-route networks: (a, b) Sketch of a simple network with two routes and two on-ramp bottlenecks with $L_1 = 20$ km, $L_2 = 25$ km (a) and model of on-ramp bottleneck on a two-lane road (b); the beginning of the on-ramp merging region is at $x_{\text{on}} = 15$ km (b); $q_{\text{on}}^{(1)} = 400$ vehicles/h, $q_{\text{on}}^{(2)} = 700$ vehicles/h; calculated values $C_{\min}^{(k)} = 3980$ for $k = 1$, 3760 vehicles/h for $k = 2$. (c) Sketch of network with three routes and three on-ramp bottlenecks that model is shown in (b); $L_1 = 20$ km, $L_2 = 25$ km, $L_3 = 22.5$ km; given on-ramp inflow rates $q_{on}^{(k)} = 300$ vehicles/h for route 1 ($k = 1$), 800 vehicles/h for route 2 ($k = 2$), and 500 vehicles/h for route 3 ($k = 3$); calculated values $C_{\min}^{(k)} = 4000$ for $k = 1$, 3740 for $k = 2$, and 3850 vehicles/h for $k = 3$.

sufficient to simulate simple models of traffic networks, two-route network with two on-ramp bottlenecks (Fig. 1 (a, b)) [43] and three-route network with with three on-ramp bottlenecks (Fig. 1 (c)) [44]. In both networks, vehicles move from origin O to destination D (Figs. 1 (a, c)). Each of the network routes is a two-lane road with an on-ramp bottleneck (Figs. 1 (a–c)). In two-route network (Fig. 1 (a)), there are two alternative routes 1 and 2 with lengths L_1 and L_2 (with $L_2 > L_1$). In three-route network (Fig. 1 (c)), there are three alternative routes 1, 2, and 3 with lengths L_1, L_2, and L_3 (with $L_2 > L_3 > L_1$). The total network inflow rate is equal to $Q = q^{(o)} + \sum_{k=1}^{N} q_{on}^{(k)}$, where $N = 2$ and 3, respectively,
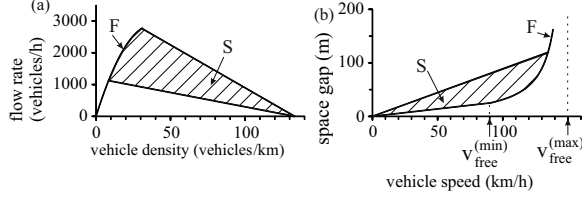
FIG. 2: Model steady states in Kerner-Klenov model in the flow-density plane (a) and in the space-gap-speed plane (b); F – free flow, S – synchronized flow; maximum free flow speed is given by formula (13).

for two-route and three-route networks (Figs. 1 (a, c)). In free flow, due to a complex dynamics of permanent speed disturbances at the bottlenecks as well as a decreasing-dependence of the vehicle speed on the vehicle density, route travel times depend considerably on the link inflow rates q_m ($m = 1, 2$ for Fig. 1 (a) and $m = 1, 2, 3$ for Fig. 1 (c)). The flow rates downstream of the bottlenecks are $q_{sum}^{(k)} = q_k + q_{on}^{(k)}$, where on-ramp inflow rates $q_{on}^{(k)}$ are given constants (see caption of Fig. 1), $q_k = q_m$ and $k = m$ ($m = 1, 2$ in Fig. 1 (a, c)).

For simulations, we use the Kerner-Klenov stochastic microscopic three-phase traffic flow model [45, 46], in which the maximum speed of vehicles in free flow is a function of the space gap g between vehicles (Fig. 2). Dependence of the maximum free flow speed on the space gap g between vehicles is given by formula:

$$v_{free}(g) = \max(v_{free}^{(min)}, \; v_{free}^{(max)}(1 - k_v d/(d + g))), \quad (13)$$

where $v_{free}^{(min)} = 90$ km/h, $v_{free}^{(max)} = 150$ km/h, $d = 7.5$ m (vehicle length), and $k_v = 1.73$. The model incorporates the metastability of free flow with respect to an F→S transition (traffic breakdown) at network bottle-

necks [26, 27, 45, 46]. Simulations show that this stochastic model allows us to take into account the metastability in travel time costs. With the use of simulations, this model feature illustrates a general conclusion made in Sec. III A 1 that the incorporation of the metastability in travel time costs does not prevent the occurrence of congestion by the use of the Wardrop's equilibria at $Q < C_{net}$ (10). Because the physics of this model has already been studied in details [26, 27, 45, 46], the model is given in Appendix A.

Simulations show that for these network models the network capacity

$$C_{net} = \sum_{k=1}^{N} C_{min}^{(k)}. \quad (14)$$

At chosen network parameters (see captions to Fig. 1), $C_{net} = 7740$ vehicles/h for two-route and to $C_{net} = 11590$ vehicle/h for three-route networks, respectively. As long as $Q < C_{net}$, when the network throughput maximization approach is applied, then no traffic breakdown occurs in both networks.

The Wardrop's UE for the two-route network shown in Fig. 1 (a) results in

$$T_1(q_1, q_{on}^{(1)}) = T_2(q_2, q_{on}^{(2)}), \; q^{(o)} = \sum_{m=1}^{2} q_m, \quad (15)$$

where T_r, $r = 1, 2$ are travel times at route $r = 1$ and $r = 2$, respectively. For the three-route network (Fig. 1 (c)) under conditions $q_m > 0$, $m = 1, 2, 3$ that are realized in simulations, we get

$$T_1(q_1, q_{on}^{(1)}) = T_2(q_2, q_{on}^{(2)}) = T_3(q_3, q_{on}^{(3)}), \; q^{(o)} = \sum_{m=1}^{3} q_m, \quad (16)$$

where T_r, $r = 1, 2, 3$ are travel times at route 1, 2, and 3, respectively. To disclose the physics of the dynamic traffic assignment with Wardrop's UE under condition (10), we consider and compare both a hypothetical case of a time-independent inflow rate $q^{(o)} = \sum_{m=1}^{M} q_m$ at the origin of the network (Figs. 3 and 4) with a more realistic case of a time-dependent inflow rate $q^{(o)}(t)$ (Figs. 5–7). In Figs. 5 and 6, in accordance with almost all empirical observations we simulate a morning (or evening) rush hour in which the inflow rate $q^{(o)}(t)$ firstly increases and then decreases over daytime (Figs. 5 (a) and 6 (a)) [47].

For two-route network (Fig. 1 (a)), route 1 is shorter than route 2. To satisfy (15), the flow rate q_1 should be larger than the flow rate q_2. For this reason, already at $Q = 7000 < C_{min}^{(net)} = 7740$ vehicles/h, the probability of traffic breakdown at bottleneck 1 is equal to 0.59. Therefore, under application of the *static* dynamic assignment with Wardrop's UE (flow rates q_m that satisfy (15) at $t = 0$ do not depend on time) we have found a random time-delayed traffic breakdown (F→S transition) at bottleneck 1 leading to traffic congestion (Fig. 3).
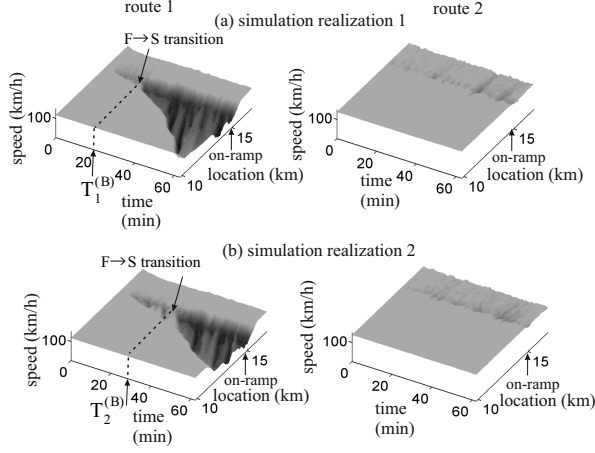
FIG. 3: Simulations of static traffic assignment with the Wardrop's UE for network in Fig. 1 (a). (a–d) Speed in space and time on routes 1 (left panel) and 2 (right panel) for two simulation realizations with time delays of the breakdown $T^{(B)} = 20$ min (realization 1) and $T^{(B)} = 29$ min (realization 2). $Q = 7000$ vehicles/h. In accordance with (15), $q_1 = 4060$ vehicles/h, $q_2 = 1840$ vehicles/h. Arrows F→S show time instants of F→S transitions.
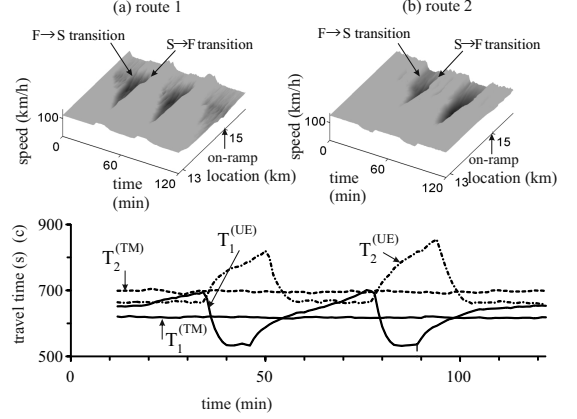
FIG. 4: Simulations of dynamic traffic assignment for two-route network (Fig. 1 (a)) at time-independent inflow rate $q^{(o)}$. (a, b) Speed in space and time on routes 1 (a) and 2 (b) under application of Wardrop's UE. (c) Travel times $T_1 = T_1^{(UE)}(t)$, $T_2 = T_2^{(UE)}(t)$ for the Wardrop's UE principle, $\Delta q = 2000$ vehicles/h [48]. Travel times $T_1 = T_1^{(TM)}$, $T_2 = T_2^{(TM)}$ are related to application of the network throughput maximization approach for which $q_1 = 3560$ vehicles/h, $q_2 = 2340$ vehicles/h. $Q = 7000$ vehicles/h.

The application of the Wardrop's UE for *dynamic* traffic assignment [48] results in a *random process* of the congested pattern emergence due to an F→S transition with the subsequent dissolution of the pattern due to a return S→F transition, and so on (Fig. 4): In different simulation realizations, we have found different sequences of the congested pattern emergence and dissolution. As in many other applications of the Wardrop's UE [15–17], this random process leads to large oscillations of travel times $T_1^{(UE)}$ and $T_2^{(UE)}$ (Fig. 4 (c)), whereas travel times $T_1^{(TM)}$ and $T_2^{(TM)}$ under the use of the network throughput maximization approach are time-independent [49].

In Fig. 5, the total network inflow rate $Q(t)$ depends on time and it does not exceed the network capacity. We can see that even in this case under application of the Wardrop's UE we get qualitatively the same random process of sequences of the congested pattern emergence and dissolution with large oscillations of travel times. Due to the application of the network throughput maximiza-

tion approach, condition (5) is satisfied for bottleneck 1; therefore, the increase in $q^{(o)}(t)$ leads to the increase in q_2 on the alternative route 2 only. For this reason, from Fig. 5 (d, g) we can see that as long as the inflow rate $q^{(o)}$ increases over time the flow rate q_1 and travel time $T_1^{(TM)}$ does not depend on time, whereas q_2 and $T_2^{(TM)}$ increase over time.

In Fig. 6 (a), the total network inflow rate $Q(t)$ depends also on time and it does not exceed the network capacity for three-route network (Fig. 1 (c)). A more complex network structure results in a more sophisticated spatiotemporal distribution of congestion between different routes (Fig. 7 (a–c)). Nevertheless, we find qualitatively the same random process of sequences of the congested pattern emergence and dissolution on different routes with oscillations of route travel times (Fig. 7 (d)) as that in a simpler two-route network (Fig. 5). A differ-
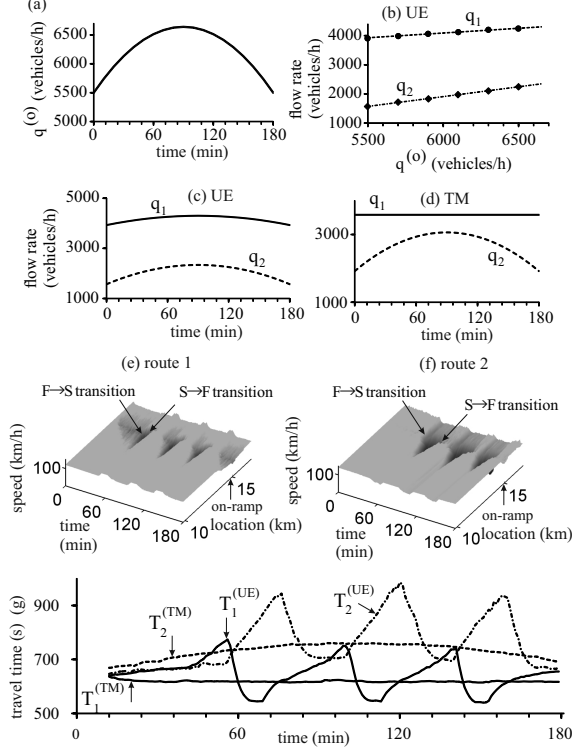
FIG. 5: Simulations of dynamic traffic assignment for two-route network (Fig. 1 (a)) at time-dependent inflow rate $q^{(o)}(t)$ (a); the maximum value of this flow rate $q_{\max}^{(o)} = 6636$ vehicles/h satisfies condition (10). (b, c) Flow rates q_m, $m = 1, 2$ for Wardrop's UE as functions of $q^{(o)}$ (b) and on time (c) calculated in accordance with (15) [47]; in (b), calculated points are well fitted by lines $q_1 = 0.33q^{(o)} + 2170$ and $q_2 = 0.67q^{(o)} - 2170$ vehicles/h. (d) Time-dependent flow rates $q_m(t)$ $m = 1, 2$ for the network throughput maximization approach. (e, f) Speed in space and time on routes 1 (e) and 2 (f) under application of Wardrop's UE. (g) $T_r = T_r^{(\mathrm{UE})}(t)$ on different routes $r = 1, 2$ for the Wardrop's UE principle, $\Delta q = 2000$ vehicles/h [48]. In (g), travel times $T_r = T_r^{(\mathrm{TM})}(t)$ on different routes $r = 1, 2$ are related to application of the network throughput maximization approach at which no traffic breakdown occurs at the bottlenecks.
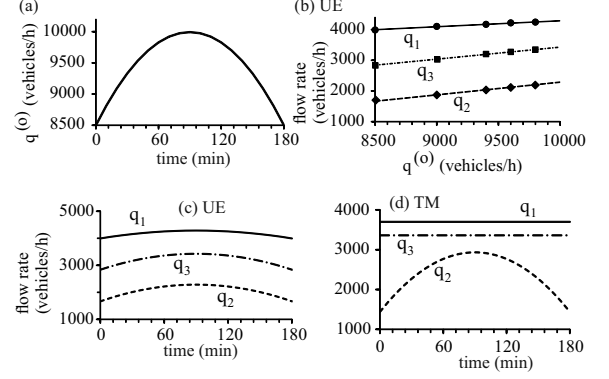
FIG. 6: Simulations of distribution of flow rates between different routes through dynamic traffic assignment in three-route network (Fig. 1 (c)) at time-dependent inflow rate $q^{(o)}(t)$ (a); the maximum value of this flow rate $q_{\max}^{(o)} = 9994$ vehicles/h satisfies condition (10). (b, c) Flow rates q_m, $m = 1, 2, 3$ for Wardrop's UE as functions of $q^{(o)}$ (b) and on time (c) calculated in accordance with (16) [47]; in (b), calculated points are well fitted by lines $q_1 = 0.19q^{(o)} + 2345$, $q_3 = 0.41q^{(o)} - 1850$, and $q_3 = 0.39q^{(o)} - 495$ vehicles/h. (d) Time-dependent flow rates $q_m(t)$ $m = 1, 2, 3$ for the network throughput maximization approach.

3. About application of network throughput maximization approach for real traffic and transportation networks

When the network throughput maximization approach is applied, route travel times determine only the constrain "alternative routes" that prevents the use of too long routes. After these long routes have been excluded from dynamic traffic assignment with the network throughput maximization approach, the assignment is determined by conditions (7) as explained in Sec. II B. Thus it seems that the use of the network throughput maximization approach in real traffic and transportation networks has considerable disadvantages in comparison with applications of the Wardrop's equilibria: Some of the drives should use longer routes to avoid congestion in the network.

ence with two-route network is also found by the application of the network throughput maximization approach: For the flow rate $q^{(o)}$ used in Fig. 6 (a) for three-route network, conditions (7) are satisfied for bottlenecks 1 and 3. Therefore, the whole increase in $q^{(o)}(t)$ leads to the increase in q_2 on the alternative route 2 only (Fig. 6 (d)).
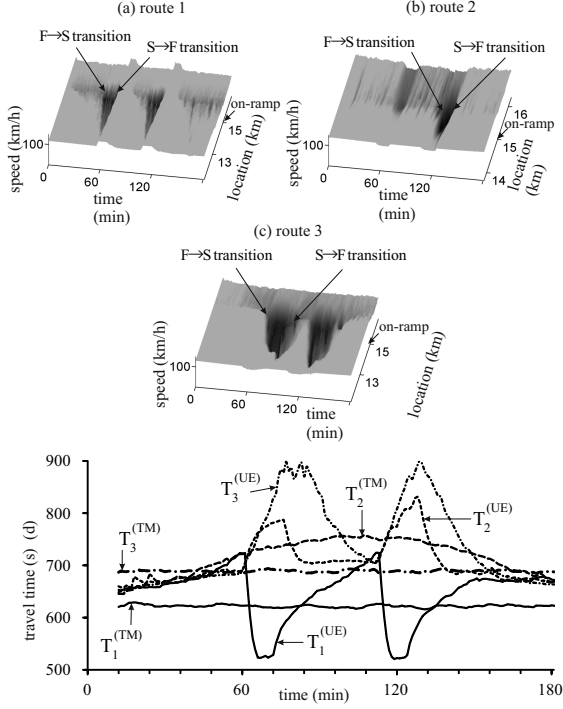
FIG. 7: Simulations of dynamic traffic assignment in three-route network accordingly to the flow rates on different routes that calculations are shown in Fig. 6: (a–c) Speed in space and time on routes 1 (a), 2 (b), and 3 (c) under application of Wardrop's UE. (d) $T_r = T_r^{(\mathrm{UE})}(t)$ on different routes ($r = 1, 2, 3$) for the Wardrop's UE principle, $\Delta q = 2400$ vehicles/h [48]. In (d), travel times $T_r = T_r^{(\mathrm{TM})}(t)$ ($r = 1, 2, 3$) are related to application of the network throughput maximization approach at which no traffic breakdown occurs at the bottlenecks.

However, possible benefits of the application of the network throughput maximization approach for the future organization of traffic and transportation networks can overcome these disadvantages. Indeed, empirical analysis of real field microscopic traffic data shows that fuel consumption and, therefore, CO_2 emission can be about 3–3.5 times larger in congested traffic than those in free flow (e.g., [50] and references there). Thus the maintenance of free flow conditions in urban areas of industrial countries through the application of the network throughput maximization approach as explained in this article can contribute to the environment protection against the increase in CO_2 emission in the world.

Moreover, through the use of short routes in applications of the Wardrop's equilibria congested traffic in a network occurs. Travel times can considerably increase due to congestion on short routes (Figs. 4, 5, and 7).

Therefore, the use of short routes does not necessarily lead to the reduction of travel time costs in real networks.

To maintain free flow conditions in urban networks for such an environment protection, already now there are technical possibilities [51]. Through communication of GPS vehicle data to a traffic control center, the center can provide appropriate information to the drivers as well as organize an efficient network organization with the network throughput maximization approach as introduced Secs. II B and II D: (i) Traffic center can store characteristics of traffic breakdown at network bottlenecks found from measurements of traffic variables with road detectors, video cameras and/or GPS probe vehicles (FCD – floating car data). (ii) Traffic center can inform drivers individually about an eco-route calculated with the approach of Secs. II B and II D. (iii) Electronic road charge systems based on GPS vehicle data can facilitate the use of alternative routes associated with the network throughput maximization approach for the maintenance of conditions (7). For example, as very small network inflow rates Q, when condition (5) is not satisfied, vehicles can freely choose routes without any (or the same) road charge. At larger values of Q, when conditions (7) are satisfied for some of the network bottlenecks, the vehicles using alternative (longer) routes should pay considerably smaller road charge (or no charge at all) than those using the routes via network bottlenecks for which conditions (7) are satisfied.

B. Network throughput maximization approach as an application of breakdown minimization (BM) principle

In [28, 29], a breakdown minimization (BM) principle for dynamic traffic assignment and control in traffic and transportation networks has been introduced. The BM principle states that the optimum of a traffic or transportation network with N bottlenecks is reached, when dynamic traffic assignment, optimization and/or control are performed in the network in such a way that the probability P_{net} for the occurrence of traffic breakdown in at least one of the network bottlenecks during a given observation time interval T_{ob} reaches the minimum possible value. The BM principle reads as follows [28, 29]:

$$\min_{q_1,q_2,...,q_{\mathrm{M}},\mathbf{R}_1,\mathbf{R}_2,...,\mathbf{R}_{\mathrm{N}},\alpha_1,\alpha_2,...,\alpha_N} \{P_{\mathrm{net}}(q_1, q_2, ..., q_{\mathrm{M}},$$
$$\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_{\mathrm{N}}, \alpha_2, ..., \alpha_N)\}, (17)$$

where is assumed that the probability P_{net} depends on variables q_m, $m = 1, 2, \ldots, M$ as well as $\mathbf{R}_k$ and α_k, $k = 1, 2, \ldots, N$.

In [28, 29, 39], it has been assumed that the probability of traffic breakdown in the network

$$P_{\mathrm{net}} = 1 - \prod_{k=1}^{N}(1 - P^{(\mathrm{B},k)}). \qquad (18)$$

This means that we have assumed that different traffic breakdowns occurring at different network bottlenecks can be considered independent events. In (18), $P^{(\mathrm{B},k)} = P^{(\mathrm{B},k)}(q_{\mathrm{sum}}^{(k)}, \alpha_k, \mathbf{R}_k)$ is the probability that during a time interval for observing traffic flow T_{ob} traffic breakdown occurs at bottleneck k, where $k = 1, 2, \ldots, N$.

An analysis of traffic assignment in a network with the BM principle (17), (18) made in [28] permits the *distinct* assignment of network link inflow rates q_m for the case of relatively large values Q, when, after the application of the BM principle (17), (18), free flow remains in a metastable state (2) at least at one of the network bottlenecks. In this case, although the application of the BM principle (17), (18) reduces the probability of traffic breakdown in the network to some minimum possible value $P_{\mathrm{net}} = P_{\mathrm{net}}^{(\mathrm{min})}$, this minimum value of the probability is larger than zero:

$$P_{\mathrm{net}} = P_{\mathrm{net}}^{(\mathrm{min})} > 0. \tag{19}$$

No application of the BM principle for the distinct assignment of network link inflow rates q_m for the case

$$P_{\mathrm{net}} = 0 \tag{20}$$

has been made in [28]. The network throughput maximization approach resulting in formula (7) at $Z_1 = Z$ as introduced in this article is the application of the BM principle (17) for "zero breakdown probability" (20). The approach permits the distinct assignment of network link inflow rates q_m under condition (20).

Thus, there can be two different applications of the BM principle:

(i) The application of the BM principle (17), (18) related to the case (19). The distinct assignment of network link inflow rates q_m for this case has been made in [28].

(ii) The application of the BM principle (17) for "zero breakdown probability" (20). This application made in this article is the network throughput maximization approach. As shown in Sec. II B, this approach resulting in conditions (7) at $Z_1 = Z$ permits the distinct assignment of network link inflow rates q_m for the case (20).

When the value Q is a relative small one, conditions (7) for the network throughput maximization can be applied. When the value Q becomes a relative large one, specifically, when conditions (12) are satisfied, then the application of the BM principle (17), (18) of Ref. [28] to the distinct assignment of network link inflow rates q_m related to the case (19) can be applied.

We can conclude that the network throughput maximization approach is applied to guarantee that condition (2) is satisfied at *none* of the network bottlenecks. This is possible as long as by the increase in the network inflow rate Q conditions (7) can be satisfied. When under subsequent increase in the total network inflow rate Q conditions (12) are satisfied, free flow is in a metastable state at some of the network bottlenecks. In this case, the network throughput maximization approach cannot be applied any more. Instead, to minimize the probability of traffic breakdown, the application of the BM principle (17), (18) of Ref. [28] to the distinct assignment of network link inflow rates q_m related to the case (19) can be applied. An analysis of the sequence of these different applications of the BM principle is out of scope of this paper; this could be interesting task for future investigations.

C. Conclusions

1. We have revealed a network throughput maximization approach that ensures free flow conditions at which traffic breakdown cannot occur in a traffic or transportation network.

2. A physical measure of a network – network capacity is introduced that characterizes general features of the network with respect to the maximization of the network throughput. As long as the total network inflow rate is smaller than the network capacity no traffic breakdown can occur in the network.

3. Based on the physics of the network capacity, we have shown that the classical Wardrop's UE or SO equilibrium deteriorates basically the traffic system: Even when the total network inflow rate is smaller than the network capacity, the dynamic traffic assignment with the Wardrop's equilibria leads to the occurrence of traffic congestion in traffic and transportation networks.

Appendix A: Kerner-Klenov model for two-lane road with on-ramp bottleneck

In this Appendix, we present a discrete version of the Kerner-Klenov stochastic three-phase traffic flow model for single-lane road with on-ramp bottleneck [46]. In the model, index n corresponds to the discrete time $t_{\mathrm{n}} = \tau n$, $n = 0, 1, \ldots$, v_n is the vehicle speed at time step n, a is the maximum acceleration, $\tilde{v}_n$ is the vehicle speed without speed fluctuations, the lower index ℓ marks variables related to the preceding vehicle, $v_{\mathrm{s},n}$ is a safe speed at time step n, $v_{\mathrm{free}} = v_{\mathrm{free}}(g_n)$ is the maximum speed in free flow which is assumed to be a function of space gap g_n, ξ_n describes speed fluctuations; $v_{\mathrm{c},n}$ is a desired speed; all vehicles have the same length d that includes the mean space gap between vehicles within a wide moving jam where the speed is zero. In the model, discretized space coordinate with a small enough value of the discretization cell δx is used. Consequently, the vehicle speed and acceleration (deceleration) discretization intervals are $\delta v = \delta x / \tau$ and $\delta a = \delta v / \tau$, respectively, where time step $\tau = 1$ s. Because in the discrete model version discrete (and dimensionless) values of space coordinate, speed and acceleration are used, which are measured respectively in values δx, δv and δa, and time is measured in values of τ, value τ in all formulas below is assumed to be the dimensionless value $\tau = 1$. In the

TABLE I: Discrete stochastic model [46]

$$v_{n+1} = \max(0, \min(v_{\text{free}}, \tilde{v}_{n+1} + \xi_n, v_n + a\tau, v_{s,n})),$$
$$x_{n+1} = x_n + v_{n+1}\tau,$$
$$\tilde{v}_{n+1} = \min(v_{\text{free}}, v_{s,n}, v_{c,n}),$$
$$v_{c,n} = \begin{cases} v_n + \Delta_n & \text{at } g_n \leq G_n, \\ v_n + a_n\tau & \text{at } g_n > G_n, \end{cases}$$
$$\Delta_n = \max(-b_n\tau, \min(a_n\tau, \ v_{\ell,n} - v_n)),$$
$$g_n = x_{\ell,n} - x_n - d,$$
a, d, and τ are constants,
$v_{\text{free}} = v_{\text{free}}(g_n)$ is function of space gap g_n.

TABLE II: Functions in discrete stochastic model I: Stochastic time delay of acceleration and deceleration

$$a_n = a\Theta(P_0 - r_1), \ \ b_n = a\Theta(P_1 - r_1),$$
$$P_0 = \begin{cases} p_0 & \text{if } S_n \neq 1 \\ 1 & \text{if } S_n = 1, \end{cases} \quad P_1 = \begin{cases} p_1 & \text{if } S_n \neq -1 \\ p_2 & \text{if } S_n = -1, \end{cases}$$
$$S_{n+1} = \begin{cases} -1 & \text{if } \tilde{v}_{n+1} < v_n \\ 1 & \text{if } \tilde{v}_{n+1} > v_n \\ 0 & \text{if } \tilde{v}_{n+1} = v_n, \end{cases}$$
$r_1 = \text{rand}(0,1)$, $\Theta(z) = 0$ at $z < 0$ and $\Theta(z) = 1$ at $z \geq 0$;
$p_0 = p_0(v_n)$, $p_2 = p_2(v_n)$ are speed functions, p_1 is constant.

model of an on-ramp bottleneck (Table VI; see explanations of model parameters in Fig. 16.2 (a) of [26]), superscripts + and − in variables, parameters, and functions denote the preceding vehicle and the trailing vehicle in the target lane during the lane changing on the main road or during the vehicle merging from the on-ramp lane. Initial and boundary conditions are the same as that explained in Sec. 16.3.9 of [26]. Model parameters are presented in Tables VII and IX.

TABLE III: Functions in discrete stochastic model II: Model speed fluctuations

$$\xi_n = \begin{cases} \xi_a & \text{if } S_{n+1} = 1 \\ -\xi_b & \text{if } S_{n+1} = -1 \\ \xi^{(0)} & \text{if } S_{n+1} = 0, \end{cases}$$
$$\xi_a = a^{(a)}\tau\Theta(p_a - r), \ \ \xi_b = a^{(b)}\tau\Theta(p_b - r),$$
$$\xi^{(0)} = a^{(0)}\tau \begin{cases} -1 & \text{if } r \leq p^{(0)} \\ 1 & \text{if } p^{(0)} < r \leq 2p^{(0)} \text{ and } v_n > 0 \\ 0 & \text{otherwise}, \end{cases}$$
$r = \text{rand}(0,1)$; p_a, p_b, $p^{(0)}$, $a^{(0)}$, $a^{(a)}$, $a^{(b)}$ are constants,

TABLE IV: Functions in discrete stochastic model III: Maximum speed v_{free}, synchronization gap G_n and safe speed $v_{s,n}$

$$v_{\text{free}} = v_{\text{free}}(g_n),$$
$$v_{\text{free}}(g_n) = \max(v_{\text{free}}^{(\min)}, v_{\text{free}}^{(\max)}(1 - \kappa d/(d + g_n))),$$
where $v_{\text{free}}^{(\min)}$, $v_{\text{free}}^{(\max)}$ and κ are constants.
$$G_n = G(v_n, v_{\ell,n}),$$
$$G(u, w) = \max(0, \lfloor k\tau u + a^{-1}u(u - w) \rfloor),$$
$k > 1$ is constant.
$$v_{s,n} = \min(v_n^{(\text{safe})}, g_n/\tau + v_\ell^{(a)}),$$
$$v_\ell^{(a)} = \max(0, \min(v_{\ell,n}^{(\text{safe})}, v_{\ell,n}, g_{\ell,n}/\tau) - a\tau),$$
$$v_n^{(\text{safe})} = \lfloor v^{(\text{safe})}(g_n, \ v_{\ell,n}) \rfloor,$$
$v^{(\text{safe})}(g_n, \ v_{\ell,n})$ is taken as that in [53],
which is a solution of the Gipps's equation [52]
$$v^{(\text{safe})}\tau_{\text{safe}} + X_d(v^{(\text{safe})}) = g_n + X_d(v_{\ell,n}),$$
where τ_{safe} is a safe time gap,
$$X_d(u) = b\tau^2\left(\alpha\beta + \frac{\alpha(\alpha-1)}{2}\right),$$
$$\alpha = \lfloor u/b\tau \rfloor \text{ and } \beta = u/b\tau - \alpha$$
are the integer and fractional parts of $u/b\tau$, respectively; b is constant.

TABLE V: Lane changing rules from the right lane to the left lane ($R \rightarrow L$) and from the left lane to the right lane ($L \rightarrow R$) and safety conditions for lane changing

$$R \rightarrow L: v_n^+ \geq v_{\ell,n} + \delta_1 \text{ and } v_n \geq v_{\ell,n},$$
$$L \rightarrow R: v_n^+ > v_{\ell,n} + \delta_1 \text{ or } v_n^+ > v_n + \delta_1.$$
$$\text{Safety conditions:}$$
$$g_n^+ > \min(v_n\tau, \ G_n^+),$$
$$g_n^- > \min(v_n^-\tau, \ G_n^-),$$
$$G_n^+ = G(v_n, v_n^+), \ G_n^- = G(v_n^-, v_n),$$
$G(u, w)$ is given in Table IV;
lane changing occurs with probability p_c.

TABLE VI: Models of vehicle merging at on-ramp bottleneck that occurs when a safety rule $(*)$ *or* a safety rule $(**)$ is satisfied

Safety rule $(*)$:
$g_n^+ > \min(\hat{v}_n\tau,\ G(\hat{v}_n, v_n^+)),$ $g_n^- > \min(v_n^-\tau,\ G(v_n^-, \hat{v}_n)),$ $\hat{v}_n = \min(v_n^+,\ v_n + \Delta v_r^{(1)}),$ $\Delta v_r^{(1)} > 0$ is constant.
Safety rule $(**)$: $x_n^+ - x_n^- - d > \lfloor \lambda_b v_n^+ + d \rfloor,$ $x_{n-1} < x_{n-1}^{(m)}$ and $x_n \geq x_n^{(m)}$ or $x_{n-1} \geq x_{n-1}^{(m)}$ and $x_n < x_n^{(m)},$ $x_n^{(m)} = \lfloor (x_n^+ + x_n^-)/2 \rfloor,$ λ_b is constant.
Parameters after vehicle merging: $v_n = \hat{v}_n,$ under the rule $(*)$: x_n maintains the same, under the rule $(**)$: $x_n = x_n^{(m)}.$
Speed adaptation before vehicle merging $v_{c,n} = \begin{cases} v_n + \Delta_n^+ & \text{at } g_n^+ \leq G(v_n, \hat{v}_n^+), \\ v_n + a_n\tau & \text{at } g_n^+ > G(v_n, \hat{v}_n^+), \end{cases}$ $\Delta_n^+ = \max(-b_n\tau, \min(a_n\tau,\ \hat{v}_n^+ - v_n)),$ $\hat{v}_n^+ = \max(0, \min(v_{\text{free}}^{(\text{max})},\ v_n^+ + \Delta v_r^{(2)})),$ $\Delta v_r^{(2)}$ is constant.

TABLE VII: Model parameters: Vehicle motion in road lane

$\tau_{\text{safe}} = \tau = 1$, $d = 7.5$ m$/\delta x$, $\delta x = 0.01$ m, $\delta v = 0.01$ ms^{-1}, $\delta a = 0.01$ ms^{-2}, $v_{\text{free}}^{(\text{min})} = 25$ ms$^{-1}/\delta v$, $v_{\text{free}}^{(\text{max})} = 41.67$ ms$^{-1}/\delta v$, $\kappa = 1.73$, $b = 1$ ms$^{-2}/\delta a$, $a = 0.5$ ms$^{-2}/\delta a$, $k = 3$, $p_1 = 0.3$, $p_b = 0.1$, $p_a = 0.17$, $p^{(0)} = 0.005$, $p_2(v_n) = 0.48 + 0.32\Theta(v_n - v_{21})$, $v_{01} = 10$ ms$^{-1}/\delta v$, $v_{21} = 15$ ms$^{-1}/\delta v$, $p_0(v_n) = 0.575 + 0.125\min(1, v_n/v_{01})$, $a^{(0)} = 0.2a$, $a^{(a)} = a$, $a^{(b)}(v_n) = 0.2a + 0.8a\max(0, \min(1, (v_{22} - v_n)/\Delta v_{22})$, $v_{22} = 12.5$ ms$^{-1}/\delta v$, $\Delta v_{22} = 2.778$ ms$^{-1}/\delta v$.

TABLE VIII: Model parameters: Lane changing

$\delta_1 = 1$ ms$^{-1}/\delta v$, $p_c = 0.2$.

TABLE IX: Parameters of model of on-ramp bottleneck

$\lambda_b = 0.75$, $v_{\text{free on}} = 22.2$ ms$^{-1}/\delta v$, $\Delta v_r^{(2)} = 5$ ms$^{-1}/\delta v$, $L_r = 1$ km$/\delta x$, $\Delta v_r^{(1)} = 10$ ms$^{-1}/\delta v$, $L_m = 0.3$ km$/\delta x$.

[1] J.G. Wardrop, in: Proc. of Inst. of Civil Eng. II. 1 (1952) 325–378.

[2] Y. Sheffi, Urban transportation networks: Equilibrium analysis with mathematical programming methods. New Jersey: Prentice-Hall, 1984.

[3] M.G.H. Bell, Y. Iida, Transportation network analysis, John Wiley & Sons, Incorporated, Hoboken, NJ 07030-6000, USA, 1997.

[4] S. Peeta, A.K. Ziliaskopoulos, Networks and Spatial Economics **1** 233–265 (2001).

[5] H.S. Mahmassani, Networks and Spatial Economics **1** 267–292 (2001).

[6] H. Rakha, A. Tawfik, in *Encyclopedia of Complexity and System Science*, ed. by R.A. Meyers. (Springer, Berlin, 2009), pp. 9429–9470.

[7] C.F. Daganzo, Y. Sheffi, Transp. Sci. **11** 253–274 (1977).

[8] D.K. Merchant, G.L. Nemhauser, Transp. Sci. **12** 187–199 (1978).

[9] D.K. Merchant, G.L. Nemhauser, Transp. Sci. **12** (1978).

[10] M.G.H. Bell, Transp. Res. B **26** 303–313 (1992).

[11] M.G.H. Bell, Transp. Res. B **36** 671–681 (2002).

[12] H.S. Mahmassani, G.L. Chang. Transp. Sci. **21** 89–99 (1987).

[13] T.L. Friesz, K. Han, P.A. Neto, A. Meimand, T. Yao, Transp. Res. B **47** 102–126 (2013).

[14] L. Lu, Y. Xu, C. Antoniou, M. Ben-Akiva Transport. Res. C **51** 149–166 (2015).

[15] J. Wahle, A.L.C. Bazzan, F. Klugl, M. Schreckenberg, Physica A **287** 669 (2000).

[16] L.C. Davis, Physica A **388** 4459 (2009).

[17] L.C. Davis, Physica A **389** 3588 (2010).

[18] Wen-Xu Wang, Bing-Hong Wang, Wen-Chen Zheng, Chuan-Yang Yin, and Tao Zhou Phys. Rev. E **72** 066702 (2005).

[19] R. Claes, T. Holvoet, and D. Weyns, IEEE Trans. on ITS **12** 364–373 (2011).

[20] Chuan-Fei Donga, Xu Mab, Bing-Hong Wang, Phys. Lett. A **374**, 1326–1331 (2010).

[21] A. Sopasakis, M.A. Katsoulakis, SIAM J. Appl. Math. **66**, 921–944 (2006).

[22] Cory Hauck, Yi Sun, and Ilya Timofeyev, Stoch. Dyn. **14**, 1350022 (2014).

[23] Yi Sun and Ilya Timofeyev Phys. Rev. E **89**, 052810 (2014).

[24] B.S. Kerner, Physica A **392** 5261–5282 (2013).

[25] B.S. Kerner, Physica A **450** 700–747 (2016).

[26] B.S. Kerner, *The Physics of Traffic* (Springer, Berlin, 2004).

[27] B.S. Kerner, *Introduction to Modern Traffic Flow Theory and Control.* (Springer, Berlin, 2009).

[28] B.S. Kerner, J. Phys. A: Math. Theor. **44** 092001 (2011).

[29] B.S. Kerner, Phys. Rev. E **84**, 045102(R) (2011).

[30] L. Elefteriadou, An Introduction to Traffic Flow Theory. (Springer, Berlin 2014).

[31] A.D. May, Traffic Flow Fundamentals. (Prentice-Hall, Inc., New Jersey, 1990).

[32] D. Helbing, Rev. Mod. Phys. **73** 1067–1141 (2001); D. Chowdhury, L. Santen, A. Schadschneider, Phys. Rep. **329** 199 (2000); M. Treiber, A. Kesting, Traffic Flow Dynamics (Springer, Berlin, 2013). Springer Optimization and Its Applications, Vol. 84, Springer, Berlin (2014).

[33] Examples of possible control parameters of network bottlenecks are the on-ramp inflow $q_{\text{on},k}$ for the case when bottleneck k is an on-ramp bottleneck as well as the flow rate of vehicles $q_{\text{off},k}$ leaving the main road to an off-ramp for the case when bottleneck k is an off-ramp bottleneck.

[34] For example, different assignment can be done for long vehicles (trucks) and passenger vehicles as well as for usual vehicles and electric vehicles.

[35] Usually, the network inflow rate $Q(t)$ changes in a network over time very slowly in comparison with any characteristic times of dynamic traffic effects at network bottlenecks under free flow conditions in the network. For this reason, as usually assumed in theories of traffic and transportation networks (see, e.g., [1–6]), at any given Q free flow distribution in the network can be considered as a steady state (steady-state analysis of traffic and transportation networks). This means that $Q = Q_{\text{out}}$, where $Q_{\text{out}}(t) = \sum_{j=1}^{J} q_j^{(\text{d})}(t)$ is the total network outflow rate.

[36] To explain the term *maximum total network inflow rate* in the definition of the network capacity C_{net} in more details, we note that in some cases the minimum capacity C_{min} of a network bottleneck can be a function of flow rates q_m. For example, for an on-ramp bottleneck C_{min} can be a function of the on-ramp inflow rate q_{on}. For this reason, to find C_{net}, the assignment of the flow rates q_m as well as control of bottleneck parameters should ensure the maximum total network inflow rate Q at which conditions (8) are satisfied. A development of a general procedure for the calculation of the network capacity for traffic networks is out of scope of this paper. Two examples of calculations C_{net} for simple network models are presented in Sec. III A 2.

[37] Note that the network capacity C_{net} depends on the network inflow rates $q_{ij}^{(\text{o})}(t)$ and the set of the alternative routes, which are assumed to be given. For the same day, there can be also different values C_{net} at different time instants.

[38] In city traffic, there is a hypothetical case of "red wave", when all vehicles approach a traffic signal during the red signal phase *only*. In this case, the definition of the network capacity C_{net} through conditions (8) remains. However, when at one of the network bottlenecks due to the signal "red wave" is realized, then the minimum capacity of the signal C_{min} is equal to the classical signal capacity C_{cl} [39, 40]. Therefore, in (8) the value of C_{min} for this signal should be replaced by C_{cl}. Moreover, when the signal is one of the network bottlenecks $k = k_w^{(2)}$ in (12), then over-saturated (congested) traffic does occur at this signal. To explain this, we note that in the case of "red wave" the classical theory of traffic at the signal is a special case of the three-phase theory [39, 40]: When the average arrival flow rate (flow rate at a bottleneck due to the signal) exceeds $C_{\text{min}} = C_{\text{cl}}$, then traffic breakdown, i.e., the transition from under-saturated traffic to over-saturated (congested) traffic occurs at the signal without time delay. The network capacity C_{net} follows also from the above application of the network throughput maximization approach, if in all formulas of Sec. II C we replace $C_{\text{min}}^{(k)}$ by $C_{\text{cl}}^{(k)}$. In particular, C_{net} is determined

from conditions (8) as follows:

$$q_{\text{sum}}^{(k)} = C_{\text{cl}}^{(k)} \text{ for } k = k_z^{(1)}, \quad \text{(A1)}$$
$$q_{\text{sum}}^{(k)} < C_{\text{cl}}^{(k)} \text{ for } k \neq k_z^{(1)},$$
$$z = 1, 2, \ldots, Z; \ Z \geq 1, \ Z \leq N; \ k = 1, 2, \ldots, N,$$

where bottlenecks $k = k_z^{(1)}$ and value Z are found in accordance with the constrain "alternative routes" as described in Sec. II B. Formula (A1) is also applicable in the framework of the classical traffic flow theories in which is assumed that there is a particular value of capacity for any network bottleneck. If we denote capacity of free flow at network bottleneck k by $C_{\text{cl}}^{(k)}$, formula (A1) remains for any classical traffic flow theory. In other words, the application of the network throughput maximization approach leading to formula (A1) does not depend on a traffic flow theory applied for the explanation of the physical nature of highway capacity. In particular, the measure "network capacity" has the same sense in the classical theory as that in the three-phase traffic theory. In a steady state analysis of a network [35], condition $Q \to C_{\text{net}}$ provides the maximum possible network throughput at which free flow conditions are ensured at which traffic breakdown cannot occur in the whole network.

[39] B.S. Kerner, Physica A **397**, 76–110 (2014).
[40] B. S. Kerner, S.L. Klenov, M. Schreckenberg, J. of Stat. Mech., P03001 (2014).
[41] As the total flow rate Q in (11) is subsequently increased, the number W of bottlenecks for which conditions (12) are satisfied, can change.
[42] In numerical simulations of the application of the Wardrop's equilibria to dynamic traffic assignment in two simple models of networks (Sec. III A 2), the metastability of free flow with respect to an F→S transition (traffic breakdown) at network bottlenecks has been taken into account in travel time costs through the use of the Kerner-Klenov microscopic stochastic three-phase model.
[43] Two-route models are often used for studies of other different traffic phenomena (e.g., [15–17]).
[44] Numerical simulations of more complex networks are out of scope of this article. Simulations for specific networks could be interesting task for further investigations.
[45] B.S. Kerner, S.L. Klenov, J. Phys. A: Math. Gen. **35**, L31–L43 (2002); B.S. Kerner, S.L. Klenov, Phys. Rev. E **68** 036130 (2003); B.S. Kerner and S.L. Klenov. J. Phys. A: Math. Gen. **37** 8753–8788 (2004).
[46] B.S. Kerner, S.L. Klenov, Phys. Rev. E **80** 056101 (2009).
[47] The dynamic traffic assignment of the link inflow rates q_m ($m = 1, 2$ in Fig. 1 (a) and $m = 1, 2, 3$ in Fig. 1 (c)) under time-dependent flow rate $q^{(o)}(t)$ is performed with the use of standard methods [2–6]: For some flow rates $q^{(o)}$, values q_m have been found in accordance with the Wardrop's UE (points in Figs. 5 (b) and 6 (b)). Simulations show that these points are well fitted with lines given in captions to Figs. 5 and 6. These lines have further been used for calculations of q_m.
[48] We have proven different control methods to decrease congestion in the network occurring due to dynamic traffic assignment with Wardrop's UE or SO (as above-explained it is not possible to avoid congestion fully). Some of the best results shown in Figs. 4–7 are achieved with "congested pattern control method" [26], in which bottleneck control starts *only* after traffic breakdown (F→S transition) is registered on the shortest route (route 1) through the use of a road detector installed 500 m upstream of bottleneck location $x_{\text{on}} = 15$ km. An F→S transition is registered when the speed at the detector $v \leq 80$ km/h (for two-route network) and $v \leq 70$ km/h (for three route network) during 3 min; a return S→F transition is registered when the speed at the detector $v \geq 92$ km/h during 2 min. After the F→S transition has been registered on bottleneck 1, the flow rate on route 1 decreases on a value Δq and the flow rate on route 2 increases on the same value Δq (Figs. 4 and 5). For three-route network (Fig. 7), the increase in the flow rate Δq has been divided (in accordance with Eq. (16) for free flow conditions) between two other routes (routes 2 and 3). After a return S→F transition has been registered at route 1, the flow rates q_k return to their initial values (at $t = 0$).
[49] Qualitative the same results as shown in Figs. 3 and 4 have been derived for Wardrop's SO at $Q = 7300 < C_{\min}^{(\text{net})} = 7740$ vehicles/h.
[50] P. Hemmerle, M. Koller, H. Rehborn, B. S. Kerner, M. Schreckenberg, IET Intell. Transp. Syst., 1–8 (2015), doi: 10.1049/iet-its.2015.0014.
[51] While working at the Daimler Company, the author was lucky to take part in the development of systems for traffic information used in vehicle routing systems, which are on the market.
[52] P.G. Gipps, Trans. Res. B 15 (1981) 105–111.
[53] S. Krauß, P. Wagner, C. Gawron, Phys. Rev. E 55 (1997) 5597–5602.