Equilibrium analysis of macroscopic traffic oscillations

Yu (Marco) Nie
Department of Civil and Environmental Engineering, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, USA

Article history:
Received 16 December 2008
Received in revised form 24 May 2009
Accepted 4 June 2009

Keywords:
Traffic oscillations
Wardrop equilibrium
Boston traffic equilibrium
Freeway bottlenecks

Abstract
Using a simple network model with two parallel links connecting a diverge and a merge, this paper studies under what conditions traffic oscillations may be initiated and propagated in a traffic stream, specially at freeway bottlenecks. Drivers are assumed to minimize either the experienced or instantaneous travel times, and in doing so, they settle at a Wardrop (day-to-day) equilibrium or a Boston (within-day) traffic equilibrium, respectively. We prove that the path travel time function in our model is not monotone, and show that this property leads to multiple Wardrop equilibria, of which only one is both stable and efficient. The paper shows that periodic traffic oscillations do not arise from Wardrop equilibria. Trivial oscillations exist at Boston equilibria, which are caused by drivers’ overreaction to traffic conditions. However, periodic oscillations are likely to emerge when (1) transitions between stable and unstable equilibria take place, and more importantly, (2) drivers make decisions based on out-of-date information of traffic conditions. The latter finding is useful in guiding control practice at freeway bottlenecks and work zones to prevent traffic oscillations.

1. Introduction

Traffic oscillations often arise in congested traffic flow, such as vehicular queues induced by freeway bottlenecks (for recent empirical evidences, see e.g., Smilowitz et al., 1999; Mauch and Cassidy, 2002; Ahn and Cassidy, 2007). In the past, this frustrating stop-and-go motion is often explained using car-following behavior (e.g., Chandler et al., 1958; Herman et al., 1959; Treiterer and Myers, 1974), lane-changing maneuvers (e.g., Gazis et al., 1962; Munjal and Pipes, 1971; Daganzo, 2002a,b), and the instability predicted by higher-order traffic flow models (e.g., Kerner and Konhauser, 1994; Jin and Zhang, 2003).

Traffic oscillations may also be triggered by macroscopic mechanisms such as queue interactions (Jin, 2003). Recently, Nie and Zhang (2008) and Jin (2009) characterized this type of oscillatory traffic pattern using a two-route network with a diverge and a merge (hereafter referred to as the D–M model) (see Fig. 1). These studies employ the traffic flow model of Lighthill and Whitham (1955) and Richards (1956) whereas the merge and diverge traffic follow the models of Jin and Zhang (2001) and Daganzo (1995), respectively. Periodic oscillations may occur in this model when queues formed at the merge spill back to the diverge, thereby reducing the discharging capacity of the diverging branches due to the first-in–first-out (FIFO) discipline imposed at the diverge. Noticeably, key features of such oscillatory traffic patterns appear to agree with empirical evidence such as reported in Mauch and Cassidy (2002) and Ahn and Cassidy (2007). This coincidence raises an interesting question of whether such a model can be used to explain, if not predict, traffic oscillations often observed at freeway bottlenecks. Valid though that question may seem, it should be noted that the original
D–M model ignores drivers’ route choice behavior, namely, oscillations occur only when flow distribution at the diverge is fixed within a certain range. Obviously, unless an effective control device is implemented in its favor, drivers may not follow such a fixed route choice.

Would traffic oscillations emerge when drivers’ behavior is reasonably taken into consideration? The present paper is intended to address this question. Well-embraced behavioral assumptions state that drivers tend to make travel choices (departure time, routes, etc.) to maximize their utility. For the purpose of this study, it suffices to focus on route choice and assume that travel time is the only factor at work in that choice. It is well-known that the traffic assignment problem (the problem of assigning traffic to shortest routes) with such behavioral assumptions can be formulated as a Nash–Cournot non-cooperative game, whose solutions are characterized by a set of traffic equilibrium states. Since the temporal evolution of traffic flow has to be considered in order to study oscillations, our equilibrium analysis falls into a class of dynamic traffic assignment (DTA) models. The reader is referred to Peeta and Ziliaskopoulos (2001) for a comprehensive review of the DTA literature.

Our goal is to study the properties of dynamic equilibrium solutions to the diverge–merge network model illustrated in Fig. 1a. Such a simple model allows us to derive analytical solutions that promise useful insights. More importantly, the D–M model reasonably represents a bottleneck situation where lane drop may cause vehicular queues and subsequently traffic oscillations, as shown in Fig. 1b. Assumptions necessary to make such a connection are:

- Drivers are informed of the bottleneck by a traffic sign upstream of the lane drop. In response to this event, drivers will make a lane-changing decision at a point near the sign. That point corresponds to the diverging junction in Fig. 1a. The ratio of drivers who select the shoulder lanes (i.e., link 1 in Fig. 1a) is denoted as $r$.
- Once passing the imaginary diverge, drivers will not change lane until they arrive at the actual lane-drop location, which constitutes the merging junction in Fig. 1a.

The settings in Fig. 1 will be frequently used hereafter. Particularly, we emphasize that links 1 and 2 in Fig. 1a refer to the shoulder lanes and through lanes in Fig. 1b, respectively.

Two different behavioral assumptions, which lead to different equilibrium states, are considered. In the first, drivers want to minimize their experienced travel times. By learning from and adjusting according to daily travel experience, drivers will settle at the so-called day-to-day equilibrium, which is a dynamic extension of the Wardrop equilibrium (Wardrop, 1952) and is widely used for long-term travel forecasting (e.g. Smith, 1993; Friesz et al., 1993; Ran et al., 1996). However, lane-changing maneuvers as those triggered by a lane drop in Fig. 1b may be too minor to be predicted from such a day-to-day equilibrium. It is more likely that drivers would make those lane-changing decisions en-route according to local traffic conditions. This assumption drives the system to a Boston traffic equilibrium (Friesz et al., 1993), in which drivers minimize their instantaneous travel times. The focus of the paper is, therefore, to obtain both Wardrop and Boston traffic equilibria of the D–M model and reveal their analytical properties, particularly those pertinent to oscillations. Numerical experiments will be conducted when it is difficult to get simple analytical solutions.

This paper is organized as follows. Section 2 briefly reviews the oscillatory traffic patterns yielded from the D–M model when the route choice is fixed. Sections 3 and 4 discuss the Wardrop and Boston traffic equilibria, respectively. Section 5 concludes the paper.
2. Oscillatory traffic pattern with fixed route choice

Nie and Zhang (2008)\(^1\) provides analytical solutions of dynamic traffic flows to the D–M model depicted in Fig. 1a, when diversion ratio \(r\) is treated as a fixed exogenous variable. It is shown that the evolution of the system is uniquely determined by two factors: the ratio of capacities of links 1 and 2 \((k)\, and the diversion ratio \(r \in (0,1)\). A synopsis of their results follows.

It is useful to first recall that the LWR model (Lighthill and Whitham, 1955; Richards, 1956), the diverge model of Daganzo (1995) and the merge model of Jin and Zhang (2001)\(^2\) are used to describe traffic flow. An identical triangle fundamental diagram is employed to describe the flow–density relationship on all links. Assuming that the total demand is \((1 + k)c\) (i.e., the capacity of link 3 in Fig. 1a), the outflow and inflow at the diverge are

\[
v_{o3} = \min \left\{ (1 + k)c, \frac{kc}{r}, \frac{c}{1 - r} \right\}, \quad v_{i1} = rv_{o3}, \quad v_{i2} = (1 - r)v_{o3}
\]

where \(v_{o3}\) denotes the outflow of link 3, and \(v_{i1}\) and \(v_{i2}\) are the inflows of links 1 and 2, respectively. The outflows of links 1 and 2 are determined by

\[
v_{o4} = \min \{c, D_1 + D_2\}, \quad v_{o3} = v_{o4} \frac{D_j}{D_1 + D_2}, \quad j = 1, 2
\]

where \(v_{o4}\) is the inflow of link 4 and \(D_j\) is the demand for link \(j\) (i.e., the maximum flow rate allowed to leave a link). Traffic congestion in the system may be initiated either at the merge due to the insufficient downstream supply, or at the diverge due to an imbalanced diversion ratio. The interplay of the two effects will lead the system to different terminal states depending on \(k\) and \(r\), as summarized in Fig. 2 (the reader is referred to Section 3.4 (Nie and Zhang, 2008) for more detailed analysis which is deemed unnecessary to repeat herein). Note that, due to the insufficient capacity, link 3 will become congested as soon as the queues reach the diverge, regardless of traffic condition on links 1 and 2. Specifically, when \(r > k\) (Region I), the traffic flow pattern is always stable; when \(\frac{k}{1 + k} < r < \min(0.5, k)\) (Region III), three different periodic oscillation patterns occur at different \(k - r\) combinations; otherwise (Regions II and IV), the system will oscillate at the beginning but eventually converge to a stable solution conforming to the initial diversion ratio, i.e., \(v_{i1} = v_{o3} = cr\) while \(v_{i2} = v_{o2} = c(1 - r)\).

Nie and Zhang (2008) noted that some periodic oscillatory traffic patterns from the above analysis (Region III in Fig. 2) well coincide with the empirical observations of traffic oscillations (e.g. Mauch and Cassidy, 2002). However, these oscillations are not realistic in the sense that they do not reflect drivers’ route choice behavior. Perhaps the most obvious flaw is that all periodic oscillations occur only when the diversion ratio is such set that link 2 (through lanes) is consistently uncongested. Clearly those oscillations may not happen in reality since no driver would rather wait to enter a congested link 1 when link 2 is under free flow condition. The above observation necessitates the analysis that takes drivers’ behavior into account.

3. Analytical Wardrop equilibrium solutions

If drivers are free to choose among links 1 and 2 (cf. Fig. 1a), they would choose such that neither link is faster than the other, which will drive the system to the well-known Wardrop equilibrium (Wardrop, 1952). Following the convention, we shall call such a pattern a user-equilibrium (UE). Let \([0, T]\) be an analysis period during which departure may take place, and \(c_i(t,f)\) denote the travel time on path \(i\) departing at a time \(t \in [0, T]\), which is a function of the path flow vector \(f = [f_1,f_2]^T\). If

\[\text{Fig. 2. Traffic evolution of the diverge–merge system.}\]
When any of the Conditions (I)–(III) is met, because the bottleneck (the merge point) always operates at its capacity
Proof. could reduce travel delay), and thus identical. Further, because these cases imply equilibrium, travelers should experience
the total travel times experienced by travelers departing at any time must be the lowest (note that no other flow pattern
can take an arbitrary value between (I) implies that \( \nu_{32} = \nu_{34} = \nu_{12} = \nu_{14} = c \) and \( \nu_{11} = 0 \). Since \( \nu_{11} \), \( \nu_{24} \), \( \nu_{21} \), \( \nu_{22} \), \( \nu_{23} \), \( \nu_{31} \), \( \nu_{32} \), \( \nu_{33} \), \( \nu_{34} \), \( \nu_{41} \), \( \nu_{42} \), \( \nu_{43} \), \( \nu_{44} \) are identical and attain the same
speed), and have the same density. Consequently, the travel time on either link will increase, but at the same pace.
will develop on these links because \( \nu_{32} + \nu_{24} = c \) and \( \nu_{14} = \nu_{42} = c \). After the traffic arrives at the merge, queues
will develop on these links because \( \nu_{31} + \nu_{21} > c \). The flow distribution (2) yields \( \nu_{31} = kc/(1 + k) \). \( \nu_{21} = c/k \) + 1/k. Noting
that the incoming flow rate is at capacity on both links, the queues will grow at the same speed (equals the wave
speed), and have the same density. Consequently, the travel time on either link will increase, but at the same pace.
When the traffic reaches the merge, the total outflow of link 3 is reduced from \( (1 + k)c \) to \( c \), with \( \nu_{31} \) reduced to
\( kc/(1 + k) \). In this case, both links 1 and 2 are congested, but the travel times on them remain the same \( \forall t \).
(IV) When \( f_1(t) > kq(t) \), the reader may verify that the total inflow into links is smaller than \( c \), using the diverge formula
(1). Therefore, links 1 and 2 never become congested and travel times on them are always free flow travel time. □

We note that Condition IV actually contains infinite number of solutions, because \( f_1(t) \) can take an arbitrary value between \( kq(t) \) and \( q(t) \) for any \( t \) without violating the UE condition. Consequently, the D–M model demonstrates that multiple
user equilibria corresponding to different link flow patterns exist in the DTA problems with physical queues. Multiple equilibria are a 
result of the lack of monotonicity of the cost function \( c(t, f) \) in the VI problem. In the static traffic assignment problem (Beckmann et al., 1956; Sheffi, 1985), for example, multiple equilibria may occur if non-separable cost functions (such as those involving intersection delays) renders the failure of monotonicity (Smith, 1982). It is known that monotonicity of path travel times may not hold in the DTA problems even when physical queues are not considered. For example, Mounce (2003) showed, based on a point-queue traffic model, that path travel times may not be monotone when a path contains more than one bottleneck. As demonstrated in the following, non-monotonicity in our model arises not due to the interaction between bottlenecks, but rather from the link interactions imposed through the merge and diverge models. Daganzo (1998) showed that DTA solutions may be chaotic when queue spillover is permitted, because a small perturbation of network inputs can override an over-saturated equilibrium with an under-saturated one, or vice versa. The multiple-equilibria phenomenon revealed from the D–M model is along the line of Daganzo’s discovery (in the sense that they both have to do with interactions between links), but has a different cause.

Let us first present the following lemma which characterizes the relationship between the path travel time and diversion ratio in the D–M model.

Lemma 1. When any of the Conditions (I)–(III) in Proposition 1 is satisfied, \( c_1(t) \) and \( c_2(t) \) are identical and attain the same
minimum value \( c(t) \). ∀t. If \( 0 < f_1(t) < kc/(1 + k) \). ∀t, \( c_1(t) < c(t) < c_2(t) \) almost everywhere in \( t \in [0, T] \); if \( kc/(1 + k) < f_1(t) < kc \). ∀t, \( c_1(t) > c(t) > c_2(t) \) almost everywhere in \( t \in [0, T] \).

Proof. When any of the Conditions (I)–(III) is met, because the bottleneck (the merge point) always operates at its capacity, the total travel times experienced by travelers departing at any time must be the lowest (note that no other flow pattern could reduce travel delay), and thus identical. Further, because these cases imply equilibrium, travelers should experience

\[ q(t) = \sum_{i=1}^{2} f_i(t) \] is the departure rate of drivers at time \( t \), a UE flow pattern \( f^* \) can be characterized by the following equilibrium conditions

\[ f_i'(t) > 0 \Rightarrow c_i'(t) = \mu_i(t); \quad f_i'(t) = 0 \Rightarrow c_i'(t) \geq \mu_i(t), \quad \forall i, t \in [0, T] \]

where \( \mu_i(t) = \min(c_i'(t), i = 1, 2) \). ∀t, \( f^* \) may be obtained by solving an infinite-dimension variational inequality (Friesz et al., 1993)

\[ \frac{2}{2} \sum_{i=1}^{2} \int_{0}^{T} c_i(w, f^*) f_i(w) - f_i^*(w) dw \geq 0, \quad \forall f^* \text{ such that } f_i(t) \geq 0, \quad i = 1, 2, \quad \sum_{i=1}^{2} f_i(t) = q(t), \quad \forall t \]

For a problem as simple as the D–M model, however, the UE flows may be obtained analytically.

Proposition 1. When \( q(t) = (1 + k)c \). ∀t \in [0, T] in the D–M model, a flow pattern \( f^* \) satisfies the UE condition (3) if it meets any of the 
following conditions: (I) \( f_1(t) = 0 \), ∀t; (II) \( f_1(t) = kq(t) \), ∀t, (III) \( f_1(t) = k \), ∀t, and (IV) \( f_1(t) > kq(t) \), ∀t, where \( \sum_{i=1}^{2} f_i(t) = q(t) \), ∀t.

Proof. We only need to show that, when any of the conditions is met, the travel time on links 1 and 2 are the same at any time \( t \), which leads to \( c_1(t) = c_2(t) \). ∀t.

(I) In this case, \( r = 0 \), so \( \nu_{13} = \nu_{12} = \nu_{23} = \nu_{24} = c \) and \( \nu_{11} = 0 \). Since \( \nu_{11} \), \( \nu_{24} \), \( \nu_{21} \), \( \nu_{22} \), \( \nu_{23} \), \( \nu_{31} \), \( \nu_{32} \), \( \nu_{33} \), \( \nu_{34} \), \( \nu_{41} \), \( \nu_{42} \), \( \nu_{43} \), \( \nu_{44} \) are identical and attain the same
speed), and have the same density. Consequently, the travel time on either link will increase, but at the same pace.
When the traffic reaches the merge, the total outflow of link 3 is reduced from \( (1 + k)c \) to \( c \), with \( \nu_{31} \) reduced to
\( kc/(1 + k) \). In this case, both links 1 and 2 are congested, but the travel times on them remain the same \( \forall t \).
(IV) When \( f_1(t) > kq(t) \), the reader may verify that the total inflow into links is smaller than \( c \), using the diverge formula
(1). Therefore, links 1 and 2 never become congested and travel times on them are always free flow travel time. □

This condition is introduced to simplify the analysis. One may argue that \( q(t) \) equals the capacity of link 3 only when the queue does not reach the origin. We note that this need not undermine the analysis if traffic is allowed to queue at the origin.
the same travel time whether they take path 1 or 2. Therefore, \( c_1(t) \) and \( c_2(t) \) correspond to the minimum possible travel time \( c(t) \).

When \( 0 < f_1(t) < kc/(1 + k) \), \( \forall t \). Link 2 is always congested as soon as traffic reaches the merge, because \( v_{x2} < c = v_{x1} \). Although link 1 may also become congested\(^4\) before the system converges to a stable solution, the density is always lower and accordingly the queue grows more slowly. Thus, the travel time on link 1 is lower than that on link 2 except at instants with zero Lebesgue measure when they are identical such as when \( t = 0 \) and \( t = T \). This implies \( c_1(t) < c_2(t) \) almost everywhere. Moreover, note that the bottleneck still operates at capacity in this case. Therefore, the total travel time experienced by travelers departing at any time remains to be the lowest possible. That is, \( c_1(t) < c_2(t) < c(t) \) almost everywhere. This in turn implies \( c_1(t) < c(t) \) and \( c_2(t) > c(t) \). □

Fig. 3 visualizes the result of Lemma 1, illustrating how the travel times on paths 1 and 2 associated with a departure time \( 0 < t < T \) may change with the diversion ratio \( r \).

We are now ready to show that \( c_1(t, f) \) is not a monotone mapping. This is formally stated as follows.

**Proposition 2.** The cost function \( c_1(t, f) \) in the VI problem (4) is not monotone if traffic flow is described by the LWR model, the diverge model (1) and merge model (2).

**Proof.** Monotonicity of \( c(t, f) \) implies that for any feasible \( f \) and \( g \), the following inequality always holds

\[
\sum_{i=1}^{2} \int_{0}^{T} [c_i(w, f) - c_i(w, g)] [f_i(w) - g_i(w)] dw \geq 0
\]

(5)

We only need to find one case in which the above does not hold. Let \( f_1(t) = kc \), \( \forall t \), and \( g_1(t) = r_0c \), \( \forall t \), \( k/(k + 1) < r_0 < k \). Thus \( f_1(t) > g_1(t) \) and \( f_2(t) < g_2(t) \). According to Lemma 1, however, \( c_1(t, f) < c_1(t, g) \) and \( c_2(t, f) > c_2(t, f) \) almost everywhere. Thus \( c_1(t, f) - c_1(t, g) |f_1(t) - g_1(t)| < 0 \) and \( c_2(t, f) - c_2(t, g) |f_2(t) - g_2(t)| < 0 \) almost everywhere. So the inequality (5) does not hold for this \( f \) and \( g \). □

The existence of multiple equilibria is a bad news for the purpose of travel forecasting since one has to determine which equilibrium is more likely to prevail in reality. Nevertheless, different equilibria may be distinguished by some properties, of which the most important are stability and efficiency.

**Definition 1 (Stability).** An equilibrium path flow pattern \( f^* \) is stable if the following two conditions are met when an arbitrarily small number of traveler switch from one path to another.

1. Travelers who stay on their original paths do not have a better alternative after the perturbation.
2. The travel time experienced by any traveler who changes their paths is strictly longer.

Intuitively, the above definition ensures that travelers who have changed their paths have the incentive to return immediately to the previous equilibrium. A more rigorous definition of stability requires that all eigenvalues of the Jacobian matrix \( c(t, f) \) have positive real parts (e.g. Pappalardo and Passacantando, 2002). Since the Jacobian of \( c(t, f) \) usually does not have an analytical form, it is difficult to establish the stability of the system using its Jacobian.

---

\(^4\) Link 1 becomes congested in this case only when \( f_1(t) > kc/(1 + 2k) \). The reader is referred to Fig. 4 in Nie and Zhang (2008) and related discussions.
Definition 2 (Efficiency). An equilibrium path flow pattern $f^*$ is efficient if there is no other equilibrium in which at least one driver has a lower travel time and no driver has a higher travel time.

Now we are able to characterize the four equilibria given in Proposition 1 using these properties.

Proposition 3. Of the four equilibria described in Proposition 1, Equilibria I and II are efficient but unstable; Equilibrium IV is neither stable nor efficient; Equilibrium III is both efficient and stable.

Proof. That equilibria I, II and III are all efficient directly follows from Lemma 1. Equilibrium IV is inefficient because the bottleneck always operates below the capacity $c$. At Equilibrium I, if a tiny number of drivers switches from path 2 to path 1, these drivers will experience a lower travel time, as shown in Fig. 3. Similarly, a driver switch from path 1 to path 2 at Equilibrium II may reduce its travel time. This violates the second condition for stability. The stability of Equilibrium III is easy to verify using Lemma 1, noting that any small change of path flow at that equilibrium will change path travel times in the same direction. Finally, Equilibrium IV is unstable because the second condition in Definition 1 is violated when flow slightly shifts from path 1 to 2, and hence increases the bottleneck capacity and reduces travel time. □

While multiple equilibria exist for the D–M model, only Equilibrium III is likely to be observed in the long run. The unstable Equilibria I, II and IV are short-lived, if not unobservable, because any small perturbation in the proper direction may initiate an irreversible transition from any of the unstable equilibria to Equilibrium III. Suppose, for example, that all drivers initially merge to the through lane (link 2) and leave the shoulder lane (link 1) empty. This corresponds to Equilibrium I. However, some aggressive drivers may wish to cut in line by taking advantage of the shoulder lane. Because these “pioneers” do improve their travel times, they will be followed by more drivers who are now at a disadvantage by staying the course. Consequently, this trend will continue until link 1 becomes as congested as link 2. From a behavioral point of view, Equilibrium III is likely to become a dominant pattern through drivers’ day-to-day learning process. In this process, drivers may gradually recognize that switching to the main line only at the real merge point is the best strategy to choose in dense traffic. However, oscillatory traffic will inevitably appear when transitions between equilibria take place. Particularly, a non-equilibrium state rest between Equilibria II and III is associated with periodic oscillations.

The existence of multiple equilibria in the D–M model suggests that such a situation is bound to arise in more general DTA models which can only be solved numerically. Since forecasting traffic is a major application of DTA models, it is important to ensure that numerical solutions yielded from such models are both stable and efficient. Unfortunately, unlike its static counterparts, most general DTA models can only be solved with heuristic algorithms that do not guarantee convergence, owing to mostly the lack of monotonicity. In practice, these algorithms are typically unable to tightly converge to an equilibrium solution. While an approximated equilibrium may be acceptable for many practical purposes, it is a considerable challenge to verify its stability and efficiency, which is critical to its realism.

4. En-route decision and oscillatory traffic patterns

The previous section suggests that traffic oscillations described in Nie and Zhang (2008) may not manifest in the D–M model should drivers observe any of the Wardrop equilibria. In fact, any oscillatory traffic pattern in the D–M model necessarily implies different travel times on links 1 and 2 which cannot be satisfied at a Wardrop equilibrium. As shown in Fig. 3, in the oscillation region $(k/(1 + k) < r < k$ and $0 < r < 1/(1 + k))$, travel times on the two paths never equilibrate with each other.

However, we caution that Wardrop equilibria may not accurately predict driver’s behavior when a bottleneck necessitates lane-changing maneuvers. For one thing, drivers may have to travel through a number of congested freeway bottlenecks in their journey. Therefore, where and how to perform lane-changing at each bottleneck is a rather minor choice compared to other choices such as routes and departure times. Moreover, due to limited availability of data, the inherent randomness in both supplies and demands, and the individual’s perception error and capacity of short-term memory, it is unlikely that day-to-day equilibrium could predict those micro choices with acceptable accuracy. Consequently, it is more reasonable to assume that drivers would make such lane-changing decisions “en-route” based on local conditions. Such an assumption leads to the so-called Boston traffic equilibrium (BTE), in which “network users do not settle into a day-to-day equilibrium flow pattern, but rather try to optimize their route and departure time choices based on current information” (Friesz et al., 1993). At a BTE, drivers constantly seek to minimize their journey times through en-route rerouting. They are assumed to know the accurate and up-to-date network-wide traffic information in order to make those rerouting decisions. Although drivers seem to make the best choice at each instant of time in BTE, they are myopic in that they do not properly account for other drivers’ decisions. Consequently, drivers may experience a journey time totally different from what they anticipate at each decision point.

It seems that oscillations may happen to a BTE flow pattern because BTE could produce differences in experienced travel times. In the following, we obtain the BTE solution for the D–M model based on the concept of reactive traffic assignment. Suppose that link travel times are updated according to real-time traffic conditions. Then, for each diverge point (where a rerouting decision is required), shortest paths leading to the destination are computed. Flows approaching a diverge are assigned to the up-to-date shortest path. Note that the above procedure is just a sequence of all-or-nothing assignments decomposed in time and space, which does not guarantee to yield a BTE. Nevertheless, like the incremental assignment
for static problems, reactive assignment is generally adequate to get a rough approximation, provided that traffic states and shortest paths are updated frequently enough. After all, obtaining a highly converged BTE solution is neither intended in this study nor essential to the following analysis.

Fig. 4 illustrates how reactive assignment is performed at a diverge in our analysis. As shown in the figure, the actual diversion is assumed to take place upstream of the physical diversion point. Once entering the ending cell, they join one of the vehicular queues (cf. Fig. 4) and cannot switch to other queues afterwards. Note that diverge model (1) still applies, with the diversion ratio \( r \) determined by the volumes of the vehicular queues. Moreover, drivers choose the next link to visit as soon as they receive the information of current traffic conditions on links 1 and 2 through, for example, a variable message sign (VMS). The location where they receive such information and thereby make the decision may or may not be the actual diversion point, where diversion is actually executed. The distance between the decision-making and decision-execution points is denoted as \( x \).

When approaching the bottleneck, drivers will select a less congested lane because it may provide a higher travel speed. Thus, we postulate that instantaneous link travel time in reactive assignment can be estimated from a speed–density relationship. To be consistent with the triangle fundamental diagram employed in previous analysis, the following speed–density function is used:

\[
\begin{align*}
  u(t) &= \left\{ \begin{array}{ll}
    u_f & \text{if } \rho(t) \leq \rho_c, \\
    \alpha \left( \frac{1}{\rho_t} - \frac{1}{\rho_j} \right) & \text{if } \rho(t) > \rho_c
  \end{array} \right.
\end{align*}
\]

where \( u(t) \) is the average speed at time \( t \), \( u_f \) is the free flow speed, \( \rho_c \) and \( \rho_j \) are critical and jam densities, respectively, and \( \alpha \) is a parameter. The density \( \rho(t) \) can be calculated from the volume \( v(t) \), the distance \( d \) and the number of lanes \( n_l \), by \( \rho(t) = v(t)/d/n_l \).

We are now ready to examine the BTE flow patterns of the D–M model resulted from the reactive assignment. We first assume that drivers make decisions at the actual diversion point, i.e., \( x = 0 \). Let \( \eta \) be the proportion of drivers who are reactive to the traffic condition, and for simplicity assume that the non-reactive drivers all take link 2 (through lanes). Obviously, the influence of such collective en-route decisions is determined by the value of \( \eta \).

When \( \eta < k/(1 + k) \), reactive drivers may choose either link 1 or 2 at the beginning since both links have identical travel times. However, once some drivers go for link 1, they will induce congestion (hence density increases) on link 2, making it unattractive to the rest of reactive drivers. Thus, only when all reactive drivers select link 1 is the system stable. This coincides with the decaying oscillatory traffic pattern obtained from fixed route choice (Region IV in Fig. 2).

When \( \eta > k/(1 + k) \), reactive drivers can alter the composition of traffic heading to links 1 and 2 in the ending cell (hence the diversion ratio \( r \)) so as to equilibrate the instantaneous travel times on those links. According to the diverge model (1), \( r \) also dictates the total throughput of the system. If \( k/(1 + k) < \eta < k \), the choice of reactive drivers will never reduce the total throughput below the bottleneck capacity \( c \). Thus, a BTE pattern similar to Equilibrium III in Proposition 1 is expected to emerge in the end. In this case, queues on links 1 and 2 will grow at a roughly same pace until they both reach the diverge.

If \( \eta > k \), the instant throughput of the system (given by \( kc/r \)) is smaller than \( c \) when all reactive drivers choose link 1, and larger than \( c \) when link 2 is preferred. Clearly, reactive assignment leads to “overreaction” in this case because of its inability to take into account the first-in–first-out constraint imposed at the diverge. Since the system throughput oscillates, oscillatory traffic patterns are expected. However, compared to those generated from the fixed route choice case, these oscillations should have much shorter periods and smaller amplitudes. Due to this oscillation, it may take longer for the system to stabilize at a state similar to Equilibrium III.

To verify the above analysis, numerical experiments are conducted on a dynamic network loading platform (Nie et al., 2008). The platform implements a numerical scheme of the LWR model similar to the cell transmission model (CTM, cf. Daganzo, 1994; Daganzo, 1995), but adopts a different merge model (2). In our experiments, link 1 has one lane, link 3 has three lanes, links 2 and 4 each have two lanes. For all the links, per lane capacity is 2000 vehicle/h/lane, jam density 240 vehicle/
mile/lane, free flow speed 50 mile/h, and the jam wave speed –10 mile/h. The distance between the diverge and merge points is 0.5 mile. The upstream inflow rate is 6000 vehicle/h and loaded uniformly within a simulation period (the assignment horizon) of 30 min (assumed to start at 6:00 am), which is divided into 300 assignment intervals (6 s each). The total number of vehicles that depart in each assignment interval is 10. The simulation interval (or loading interval) is 2 s. Accordingly, traffic conditions as well as shortest paths are updated every 2 s. In the implementation, the distance between the physical and actual diverge points (cf. Fig. 4) is always larger than or equal to the product of the free flow speed and the simulation interval, and smaller than two times of that product. This restriction is not a reflection of realism, but rather for the modeling convenience.

Fig. 5 shows the instant travel times on links 1 and 2 at BTE patterns corresponding to different \( \eta \) values. These travel times were computed using Eq. (6) and recorded during simulation. Recall that in our example, \( k = 0.5 \), so \( k/(1 + k) = 1/3 \). When \( \eta = 0.2 \), travel time on link 2 oscillated at first but quickly stabilized. Link 1 was never congested because there are no enough reactive drivers. As a result, reactive drivers always experience a shorter travel time. When \( \eta = 0.4 \), travel times on both links increased at the same rate and reached the same value of about 2.3–2.4 min after about 6 min, implying that BTE is maintained throughout the simulation. When \( \eta = 0.6 \), it took about 20 min for the travel times on both links to achieve a similar value (2.3–2.4 min). Although the high frequency oscillations in travel times at the beginning indicate that the reactive assignment have caused sharp changes in the diversion ratio in the ending cell, the system nonetheless reached an equilibrium state in the end. Interestingly, when \( \eta = 0.8 \), neither link ever got fully congested. The instant travel times on these links jumped between the free flow (0.6 min) and a slightly congested value (0.8–0.9 min), indicating that a short queue was sustained on either link. A close look reveals that the average system throughput is well below \( c \) in this case.

In a nutshell, nontrivial traffic oscillations do not seem to arise from the D–M model when drivers have access to the up-to-date traffic information and actively seek better alternative routes (i.e., \( x = 0 \)). The oscillations observed from a BTE flow pattern either are decaying or have insignificant amplitudes and periods, which mainly result from overreaction rooted in reactive assignment. However, more pronounced oscillations may emerge if \( x > 0 \), as demonstrated in the following numerical experiments. In this case, reactive drivers will select a link immediately after they read the message from VMS, and observe the decision without recognizing that the information will be out of date by the time they get to the actual diversion point. Therefore, the information provided through VMS provokes a short-term proactive decision.

Fig. 6 depicts the resulting oscillatory traffic patterns when \( \eta = 0.4 \) and \( x = 0.3 \) (miles). Fig. 6 shows that link 2 was always congested while link 1 periodically oscillated between free flow and congested conditions. The period of the oscillation is approximately 10 min. The instant travel times on the two links are not identical for most of the time. Thus, reactive assignment did not achieve BTE, which is not a surprise since drivers actually make route choices using lagged travel times. What is unexpected here is that the information lag produced periodic oscillations. Fig. 6 shows that the outflow rate of link 3 (the one upstream of the diverge) oscillates between about 2800–5200 vehicle/h, with a regular period of about 10 min. When the instant travel time on link 1 is lower than that of link 2, all reactive drivers prefer link 1, which creates a peak value (approximately \( kc/r = 2000/0.4 = 5000 \)) in the outflow rate of link 3 lagged by

---

**Fig. 5.** Instant travel times on links 1 and 2 for various values of \( \eta|x = 0 \).
the travel time from the decision-making to decision-execution points. Similarly, when link 2 has a lower travel time, the outflow rate of link 3 reaches its valley value. The pattern of the oscillatory traffic may be better revealed from the density contour plot in Fig. 6.

Fig. 7 demonstrates the impact of $x$ on the oscillatory traffic patterns. When $x$ is decreased to 0.1 mile, the lag of information access is reduced. In this case, link 1 becomes more congested, while its travel time is still lower than that of link 2 for most of the time. The period of oscillations is reduced from 10 min to about 2.5–3 min. As expected, Fig. 7 looks more
similar to the case of $\eta = 0.4$ in Fig. 5 than Fig. 6. When $x$ is 0.5 mile, the instant travel time pattern is generally similar to that depicted in Fig. 6, except that link 2 seems less congested. Although the free-flow periods on link 1 are slightly enlarged, the period of oscillatory traffic remains to be about 10 min. We do not present results for larger $x$ because oscillations in those cases becomes too chaotic to analyze. However, our experiments indicate that, for large $x$ (i.e., larger than 1.5 mile), the D–M system of the above setting will operate below capacity, with links 1 and 2 congested alternatively. In a nutshell, the above findings suggest that providing out-of-date traffic information may generate unnecessary oscillations and even reduce the total system throughput.

Remark 1. Mauch and Cassidy (2002) found that the oscillations in traffic flow at different ramps have an amplitude around 1500–2000 vph and a period of several minutes. Moreover, the oscillations propagated upstream as waves at a nearly constant speed, and their amplitudes grew slightly in the upstream direction. Ahn and Cassidy (2007) confirmed these findings and suggested a possible linkage between lane-changing maneuvers and the oscillations. We note that the oscillations associated with lagged information access discussed in this section can be considered as being caused by bottleneck-triggered lane-changing maneuvers. Moreover, when parameters (information access point and reactive ratio) are properly selected, the Boston-equilibrium results for lagged information case are qualitatively comparable to these field observations. For example, Fig. 6b and c showed that the oscillation propagates upstream at a nearly constant speed, with a period about 10 min (note that the period decreases when the information access point is closer to the actual diverge point).

The above comparison does not imply that our “theory” can predict stop-and-go oscillations to any level of accuracy that would match field observations. While network effects and drivers’ access and response to information has a clear role in forging traffic oscillations, other factors, such as irregularities and instabilities in drivers’ car-following behavior, must be at work as well. Consequently, a partial model that does not consider all factors is unlikely to make any reasonable predictions about the puzzling stop-and-go phenomenon. Nonetheless, the model discussed herein offers one plausible explanation to the cause of such oscillation and, therefore, may be useful in its prevention.

5. Concluding remarks

It remains an open question under what conditions traffic oscillations may be initiated and propagated in a traffic stream. We attempt to address this question from a macroscopic perspective, using a simple network model with two parallel links connecting a diverge and a merge. Such a simple network model is employed because (1) it allows us to derive analytical solutions, (2) it is a building block of more general networks, and (3) it reasonably represents a bottleneck situation where lane drop may cause vehicular queues and traffic oscillations. The model proposed in this paper explicitly considers drivers’ route choice behavior. Depending on whether drivers want to minimize the experienced or instantaneous travel times, the system may settle at a Wardrop (day-to-day) equilibrium or a Boston (within-day) traffic equilibrium.

We proved that the route travel time function in the diverge–merge (D–M) model is not monotone owing to the interactions between physical queues. As a consequence, the D–M model has multiple Wardrop equilibria, of which only one is both stable and efficient. Unstable and/or inefficient equilibria are unlikely to be observed in the real-world because any small perturbation may cause the flow pattern to depart from them irreversibly. However, the existence of unstable and inefficient equilibria presents challenges for the applications of general DTA models for the purpose of traffic forecasting. In particular, multiple equilibria raise several concerns about the existing heuristic DTA algorithms. First, they may well be trapped at unrealistic equilibria. Thus, one should check the stability and efficiency of a solution obtained from a heuristic algorithm before using it. Although it is beyond the scope of the paper to discuss how such check can be effectively performed, a simple method may be to perturb the equilibrium solution and compare the path travel times against the two criteria proposed in Definition 1. Second, which equilibrium is attained may largely depend on the initial solution and algorithmic factors. Thus, it may be a useful practice to solve a given problem multiple times from a sample of initial solutions.

Our analysis does not reveal any oscillatory traffic patterns at Wardrop equilibria. Nevertheless, one may consider the oscillatory traffic induced by a bottleneck as being non-equilibrium, i.e., a pattern that has not yet come to an equilibrium, or a pattern that is approaching from an unstable equilibrium to a stable one. This explanation is, however, of limited value since a transitional pattern may be chaotic and thus difficult to predict.

We showed that traffic oscillations may appear in a Boston traffic equilibrium (BTE), which is likely to be resulted from a combination of drivers’ overreaction to traffic information and the first-in–first-out constraint imposed at the diverge. However, theses BTE oscillations are trivial, in the sense that they are either decaying or similar to a white-noise signal with high frequency and small amplitudes. However, nontrivial periodic oscillations emerge when information access is subject to a time lag and thereby prevents reactive assignment from achieving a BTE state. Although they are typically more complicated, such oscillations are similar to those obtained from the fixed route choice model of Nie and Zhang (2008) in that they have a stable period and a significant amplitude (in term of density and flow rate of the link upstream of the diverge, cf. Fig. 6).

It is unclear to what extent the D–M model with a lagged information access may explain the traffic oscillation observed in the real world. Most likely many other factors may be at work and hence no single theory could predict this puzzling phenomenon with reasonable accuracy. Nevertheless, the finding may be used in bottleneck management practice to prevent traffic oscillations. One example is the dynamic late merge system (e.g., MDOT, 2004) often implemented at work zones, which intends to avoid breakdown and balance lane use by providing local traffic conditions and/or merge guidance. As re-
sults in this paper suggest, such a system has to be implemented properly to ensure that the guidance is provided at the actual diverge point. Because the actual diverge point may be influenced by control strategies, achieving this goal in practice is not an easy task which require further investigation.

Acknowledgements

Constructive comments from two anonymous referees are appreciated. Professors David Boyce and Hani Mahmassani at Northwestern University kindly offered comments on an earlier draft of the paper. The views are those of the author’s alone.

References


