TRANSPORTATION AND TRAFFIC THEORY

Proceedings of the 13th International Symposium on Transportation and Traffic Theory
Lyon, France, 24–26 July, 1996

edited by

JEAN-BAPTISTE LESORT

Laboratoire d'Ingenierie Circulation Transports (LICIT)
Institut National de Recherche sur les Transports et leur Sécurité (INRETS) and
École Nationale des Travaux Publics de l'État (ENTPE), Lyon, FRANCE

PERGAMON

THE NATURE OF FREEWAY GRIDLOCK AND HOW TO PREVENT IT

Carlos F. Daganzo

Department of Civil and Environmental Engineering and
Institute of Transportation Studies
University of California, Berkeley, CA 94720 USA

ABSTRACT

This paper presents a continuum model of merge operations and shows how it can be used to describe traffic dynamics on a closed-loop freeway (a "beltway"). The paper shows that it is possible for traffic on a beltway to start a process of self-destruction from which it cannot recover without outside intervention. If allowed to continue indefinitely, the end result of such a process is a gridlocked traffic stream without any motion. A related phenomenon sometimes occurs in roundabouts when priority is given to the entering traffic.

Our result only hinges on three critical assumptions: (i) if a merge is congested, merging vehicles will "force" gaps and trickle into the freeway in a given ratio with freeway vehicles, (ii) vehicles stay on the freeway until they reach their specific destinations, and (iii) when on the freeway, vehicles take up some space that may depend on flow.

The paper then describes the collapse process and shows that there is a "relaxation time" or "half-life" during which the freeway flow declines by a factor of 1/2. The half-life can be estimated by a simple formula; it can be considerably less than 1 hour for 3-lane freeways. Because speed is very sensitive to flow when flow is reduced from its maximum level, the decline in speed is much more sudden: we estimate that speeds drop by a factor of six during the first half-life.

Although a state of complete gridlock should never be reached if there are alternative routes, because drivers would begin to exit before reaching their destinations, the low flows and speeds that would result would still negate the advantage of the beltway over the surface streets. The good news is that the collapse process can be reversed, or better prevented altogether, by restricting the ratio of input flow to mainline flow below a critical level. The paper also identifies such level and describes the recovery process.

1. INTRODUCTION

There are three possible reasons for congestion on any freeway: (1) some of the freeway exits cannot accommodate the demand and the resulting slow-moving queues block it; (2) the freeway has a natural bottleneck which generates queues behind it; and (3) there is too much inflow which gets in the way of through traffic.

It is well known by practicing engineers that if some of the exiting traffic in scenario (1) or the short trips using the bottleneck in scenario (2) can be diverted from the freeway to other underutilized routes, then the delay to the remaining freeway users can be reduced more (often considerably more) than it is increased to those diverted. This is the main rationale for ramp metering or closing. (In as much as delay can be reduced by eliminating a link this effect is a good illustration of what planners call Braess' "paradox", although it is hard to see why an obvious phenomenon should be called a paradox.) Comprehensive analyses of the topic, including comparisons of ramp metering versus ramp closing, can be found in Allen and Newell (1976) and Newell (1977).

Perhaps less well known is the fact that metering can be beneficial even when route choice is not an issue; it may be highly so even if there are no alternative routes whatsoever. The typical situation occurs on freeway sections upstream of a bottleneck when a significant fraction of the upstream traffic wishes to exit before the bottleneck; e.g. as might occur on beltways bypassing city centers. In these instances the rate at which vehicles are removed from the system is the sum of the flows through the bottleneck and through all the upstream exit ramps. Thus, it should be intuitive that it is inefficient to allow the bottleneck queue to grow past an exit ramp that carries a significant amount of demand because this would starve the exit ramp for flow. Note that the validity of this observation is independent of whether the bottleneck is of type (1), (2) or (3).

Metering can alleviate this type of problem. In particular, for type (3) bottlenecks the queue can be eliminated without knowing the origin-destination (O/D) table simply by monitoring and limiting merging activity.[1] As our first step in this line of research we will examine in detail the case of a closed-loop beltway with a fixed number of lanes in which the bottlenecks are only caused by merging traffic. A closed-loop is chosen as our scenario because in these instances lack of metering can be catastrophic, whereas for freeway sections containing just one or two ramps it may be merely inefficient.

[1]For other bottleneck types O/D information is needed. The goal in these cases should be to release just enough upstream traffic to keep the bottleneck at saturation. Priority should then be given to the on-ramps that send the least traffic through the bottleneck as this allows a maximum number of trips to leave the system per unit time. In general, the benefit obtained by this form of metering is highly dependent on the O/D table; and it is negligible if almost all the traffic goes through the bottleneck.

We have found that if the O/D table on a beltway involves long trips then a steady state solution with positive flow may not exist. The phenomenon arises when merging traffic is allowed to force its way into the beltway queues, slowing traffic down and choking the rate at which circulating vehicles can exit. On a closed loop the effects can lead to a catastrophic situation where more traffic enters than leaves the freeway at all times (!) and the result is a drift toward gridlock that only abates once rush hour queues have dissipated. [This is known to happen on a smaller scale at roundabouts, where gridlock is the only feasible steady state for certain O/D tables if absolute priority is not given to the circulating traffic.]

Our objective here is not to set forth general conditions for existence of a solution with positive flow—this should be done at a later date—but to demonstrate that gridlock must arise even in very simple (partially symmetric) situations, and also to study the decay and recovery process from arbitrary initial conditions. Such a basic understanding will help us find simple methods for gridlock diagnosis and prevention. In this spirit, it will be assumed for the most part of this paper that the beltway is homogeneous (no bottlenecks) and that the trip length distribution is rotationally symmetric, although some on-ramps may have more flow than others and their geometry may also be different. Generalizations of these assumptions are discussed qualitatively.

Section 2 of this paper describes the model for the merge. Then, for rotationally symmetric problems, Sec. 3 examines the stationary flow states that can arise on the beltway and Sec. 4 the transient effects. Section 5 extends the results to unequal input flows. The paper ends with a brief discussion.

2. Theory Of The Merge: The Merge Diagram

A merge consists of an upstream freeway segment (u), an incident on-ramp (r), and a downstream segment (d). The on-ramp could be another freeway, but the adopted terminology is more convenient. Because we are considering a homogeneous freeway, we assume that (u) and (d) have the same capacity, q^f_{max}. We also assume that the capacity of the ramp is q^r_{max} and that input and output flows through a merge cannot exceed the corresponding capacities. The set of feasible input flows defined by this rule is the trapezoid enclosed by the coordinate axes and lines ab and bc of Fig. 1. The slanted side of the trapezoid corresponds to the condition $q^d \le q^f_{max}$, since conservation requires that $q^d = q^u + q^r$. Note that the abscissa reached by a line with slope -1 passing through a point in the diagram is the downstream flow for the corresponding input flows; e.g. the abscissa of point c' for input conditions "e".[2]

[2]If q^f_{max} were to depend on the ratio of entering flow, the outer boundary of the trapezoid could be slanted and/or curved. The results in Secs. 2 and 3 of this paper also hold for boundaries of this type. We choose to present the less general results because they are more intuitive and easier to explain.

We note at this point that the definition of q^f_{max} will soon be generalized to represent the maximum flow that can be admitted by the downstream section of the merge under current traffic conditions. As such, it may vary with time. If for example a downstream queue completely blocks entry after a certain time, then $q^f_{max} = 0$ until the queue dissipates.

We assume that for every flow below capacity traffic can be either congested or uncongested, and that the congested (or queued) state corresponds to lower travel speeds. Therefore, to fully characterize the state of the merge we need to specify whether each one of its three components is "uncongested" or "queued" (U or Q).

Only the following state combinations are allowed:

No Q	(NQ):	no queues anywhere; flow at or below capacity.
Ramp Q	(RQ):	ramp queue; no queue on upstream freeway; downstream flow at capacity or queued.
Freeway Q	(FQ):	freeway queue; no queue on ramp; downstream flow at capacity or queued.
Both Q	(BQ):	queues on both upstream segments; downstream flow at capacity or queued.

Note that we do not allow upstream queues to exist when the output flow is below capacity and unqueued since the upstream queue(s) would quickly restore the flow to its maximum possible value; i.e. to the appropriate value on the outer boundary of Fig. 1. We also exclude the situation with a downstream queue but no upstream queues since that situation can only persist in the unlikely event that the sum of the arriving flows exactly matches the downstream flow, and in that case we can simply imagine that the existing queues are infinitesimal.

In the spirit of informal remarks made by pioneering Caltrans engineer K. Moskowitz to G.F. Newell (Newell, 1992), the feasible states are further restricted by assuming that if queues exist on both approaches, vehicles enter the freeway in a given ratio: α ramp vehicles for each upstream vehicle. A rough approximation for the merge of a 1-lane ramp and an L-lane freeway is $\alpha = 1/(2L-1)$, as if ramp vehicles and those on the outer freeway lane alternated advancing into the merge. The Moskowitz assumption is in agreement with a limited set of empirical observations (Lin, 1996). Thus, the BQ states are assumed to lie on a ray emanating from the origin, as shown in Fig. 1, which we call the "priority line". In principle, one could allow the priority line to curve, but this is premature in view of the available evidence.

This partition of the feasible region is logical because if the freeway carries too little traffic and there is an oversupply of ramp vehicles then these will grab the unused capacity without generating freeway queues; clearly, any such queues (e.g. arising from fluctuations) would

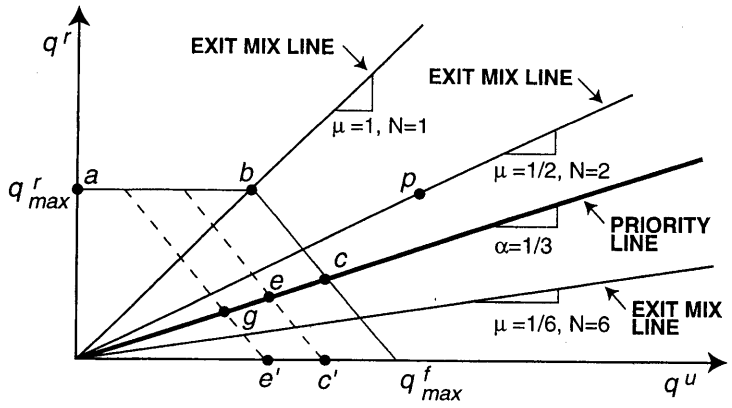

Figure 1. The merge diagram

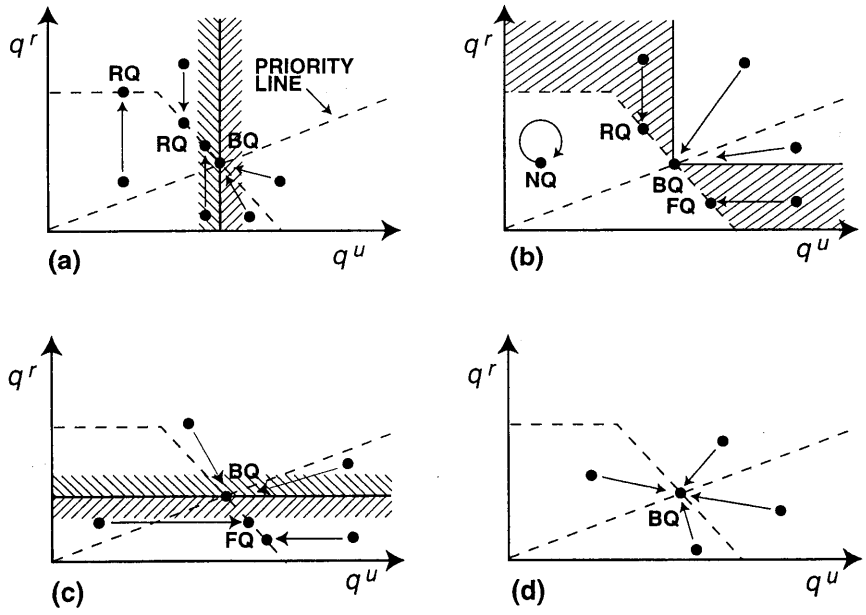

Figure 2. Merge transitions when initial merge state is: (a) RQ, (b) NQ, (c) FQ and (d)BQ

quickly dissipate by the BQ flows that would ensue. Thus, the merge would operate in a RQ state above the priority line. If the ramp carries too little traffic relative to the freeway, a similar argument reveals that the merge would operate in an FQ state below the priority line.

It should be intuitive that the information embodied in the merge diagram is sufficient to predict the cumulative flows advancing through the merge for any given set of desired cumulative inputs, even if one specifies some time-dependent upper bound to the downstream flow, $q^{d*}(t)$. Then the flows advancing on each approach at any given time t are the largest possible consistent with the condition $q^u(t) + q^r(t) \leq q^{d*}(t)$, and with the requirement not to exceed the desired input flow of an approach on which there is no queue.

The recipe can be described completely with the four merge diagrams of Fig. 2, where dotted lines have been used for the capacity constraints and the priority line to emphasize that conditions could change with time; e.g. the slanted dotted line with slope -1 corresponds to the downstream restriction, $q^{d*}(t)$. The change in conditions could arise from slow-moving traffic downstream, an incident, a change in the priority rule, or in the arrival rate(s). [Of course, the latter is noticed at time t only when there is no queue on the approach with the changed arrivals.] Each diagram corresponds to a different initial merge state, at time t. The arrows of the figure denote the change in advancing flows (and merge state) when the constraints at time t become those displayed. Note that changing conditions can only change the merge state by creating queues: RQ—>BQ, FQ—>BQ, NQ—>(RQ,FQ,BQ). This does not mean that queues cannot dissipate; they do so spontaneously when the cumulative departures calculated from Fig. 2 catch up with the curve of cumulative desired arrivals, which can be determined with the standard deterministic queueing diagrams. Figure 2 can then be used to determine the proper discharge rate for each of the approaches. The reader is encouraged to consult Newell (1982) which discusses from a slightly different perspective the case without downstream queues, and includes an example in which desired input curves leading to oversaturation are specified from t = 0 to t = T.

In order to analyze the beltway, we also need to define what happens at an exit, but since we are assuming that exit queues do not develop the answer is simple: flow past the exit is decreased by the flow of vehicles leaving there. To do this properly, one needs to keep track of the mix of destinations in the traffic stream; e.g. as explained in Newell (1993).

When entrances and exits are close together one can expect weaving effects to play some role, but these are ignored in our analysis.

3. STATIONARY BELTWAY FLOWS

We consider now a rotationally symmetric O/D table, where P_n denotes the fraction of entering traffic at any ramp that bypasses n on-ramps (n = 0, 1, 2, ...), and let N denote the

average number of on-ramps bypassed by an entering vehicle. [We recall at this point that $N = P_1 + P_2 + P_3 + ...$] Then, in a stationary and rotationally symmetric solution with entering flow q^r at each ramp, the freeway flow upstream of any on-ramp is the sum of the flows sent by it by all the other ramps. Since on-ramps and off-ramps alternate along our loop we find:

$$q^u = q^r[P_1 + P_2 + P_3 + ...] = q^rN. \tag{1a}$$

Then,

$$q^d = q^r(N+1). \tag{1b}$$

Equation (1a) defines a ray through the origin of the merge diagram with slope $\mu = 1/N$, as shown in Fig. 1. This ray is called the "exit mix line" because in a stationary solution the exiting and entering flows must balance out, i.e. $q^e = q^r$, and the ratio of the ordinate to the abscissa intercepted by a line with slope -1 is the fraction of vehicles taking the exit: q^e/q^d. In view of this, we see that a stationary solution to our problem, defined by a point on the merge diagram with coordinates q^u and q^r (and with $q^d = q^u + q^r$), must be on the exit mix line.

If the desired input flow, Q^r, generated upstream flows $Q^u = NQ^r$ that satisfy the capacity constraints, then no queues would be generated; the point on the exit line with coordinates (Q^u, Q^r) would be inside the feasible region and would be our solution.

If point (Q^u,Q^r) lies outside the feasible region, then less flow $q^r < Q^r$ can enter the system in a steady state. The solution (q^u, q^r) must still be on the exit mix line but shifted closer to the origin. Let us examine this in more detail.

If the exit line is above the priority line, as occurs for the cases with N=1 and N=2 in the figure, then flows in the interior of the feasible region (of RQ or BQ type) are unstable; as seen from Figs. 2a and 2d they would drift toward the boundary. The only stable solution is on the boundary itself; e.g. point b of Fig. 1 for the case with N=1. Physically, this means that the freeway flow past the on-ramps meters the entering traffic, and does so in the best possible way (keeping the downstream flows at capacity if the ramps can supply enough flow). Traffic self-regulates.

Similar considerations applied to the case where the exit line is below the priority line reveal that the only stable solution can be at the origin! With the exit mix line below the priority line, our merges must be in an NQ or FQ state if there is an equilibrium. However, these solutions are not allowed because we have specified that Q^r is so high that it generates a ramp queue. Figures 2a and 2d show that if there was a ramp queue, then on-ramp flows would increase. Thus, the only stable solution is gridlock, since a stopped traffic stream is nature's way of blocking the tendency for on-ramp flows to increase. In a world of "point queues" where gridlock is impossible, there would be no stationary solution at all.

In closing this section we note that the result we have obtained is independent of drivers behavior while overcoming distance; it is an exclusive property of the merging strategy. The transient process by which the system approaches the gridlocked state does depend on the spacings that drivers select while in a queue, and this is explored in the next sections.

4. TRANSIENTS

It is assumed in this section that there is a reproducible relation $q = F(k)$ between the flow and density on the freeway, indicating that drivers in a queue decrease their (average) spacing when flow declines. This is the central assumption of the kinematic wave model of Lighthill and Whitam (1955) and Richards (1956), which is the simplest continuum model recognizing in a physically meaningful way that vehicles take up space. We note in this regard that the merge rules of Sec. 2 define well-posed boundary conditions for the kinematic wave network problem; see Daganzo (1995). Furthermore, as suggested by the work of Hall et.al (1986), Banks (1989) and Newell (1993), we shall make the additional assumption that the F(k) relations on the freeway and the ramp are similar triangles; see Fig. 3. This special form of the kinematic wave model is attractive because it leads to a solution without expanding waves that can be depicted pictorially very clearly.

We recall that the slope of the left side of the triangle represents the free-flow speed, and the slope of the right side the velocity of flow disturbances, -w. It appears from a limited set of experiments that the two slopes should be approximately as shown in the figure, in the ratio 1:5, to fit the data best (Lin and Ahanotu, 1995). This ratio also seems sensible from casual observation of waves in queued traffic which appear to move somewhere between 10 and 15 MPH. [We note that the particular form of the F(k) relation on the ramp does not influence the results of our examples in any way; the curve is only used to display the prevailing traffic state on the ramp.]

4.1 A Description of the Collapse

This subsection describes graphically how an initially empty and symmetric system evolves toward and through the first stages of collapse. It then presents a simple formula for the collapse rate, which is independent of the initial conditions. Section 4.2 describes the recovery process, which takes place when input flows are restricted.

Let us consider Fig. 4. It displays the time-space evolution of the traffic states on the road segment between on-ramps for a rotationally symmetric beltway obeying the merge diagram of Fig. 1 for $\mu = 0$ (no exiting traffic) and $\mu = 1/6$. Although the first case is an unrealistic scenario it is introduced because its physical evolution is rather obvious, and this should help the reader interpret the results. Only a small step is then needed to understand the general case with $\mu < \alpha$, which is shown in Fig. 4b. We assume for both illustrations that a flow

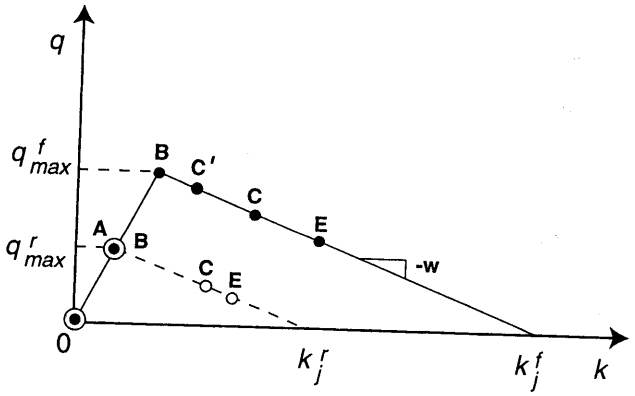

Figure 3. The "fundamental diagram"

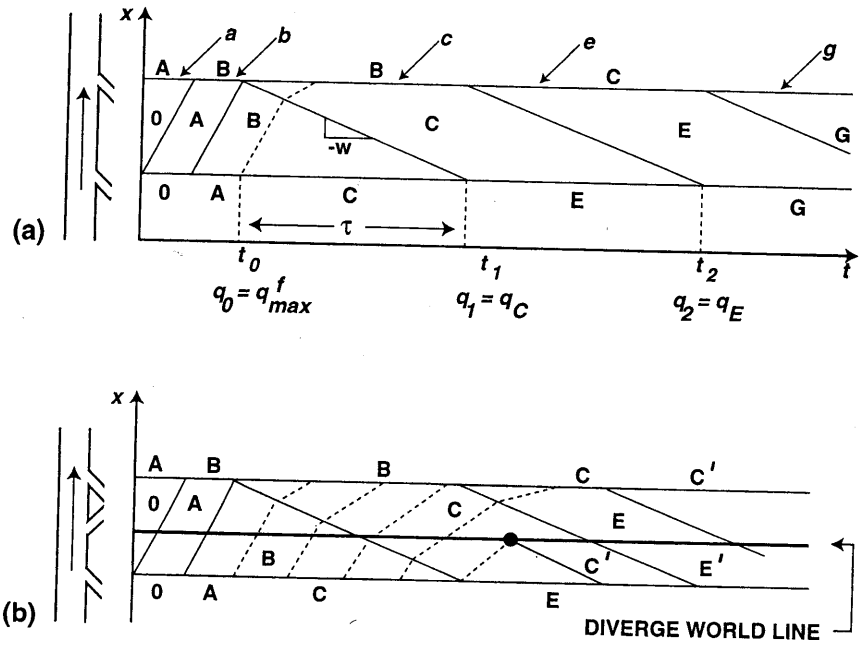

Figure 4. The beginning of the collapse process: (a) no exit flow (b) some exit flow

$Q^r = q^r_{max}$ is poured from every ramp into an initially empty beltway, and note that none of the results change if we imagine that the ramp capacity is greater than Q^r.

The horizontal lines of Fig. 4 represent the "world lines" of the ramps, which double up as interfaces between different traffic states. The remaining interfaces are the slanted solid lines. The lower case letters pointing to the merge world line (in part a) indicate the state of the merge as per the diagram of Fig. 1. The capital letters identify the traffic state prevailing in each portion of the time-space plane; they correspond to those displayed in Fig. 3. Dashed lines represent vehicle trajectories; only one of these has been included in part a to avoid clutter. The reader familiar with kinematic wave theory should have no difficulty verifying the validity of our solutions in the usual way. (In doing this one should remember that the solution for all x is periodic, with the pattern of the figure repeating itself for all inter-ramp intervals; we also note that Figs. 3 and 4 have been drawn approximately on a scale where lines with the same slope correspond to the same velocity.) Let us now examine what the figures are saying.

Figure 4a indicates that shortly before $t = t_0$ the beltway is approaching saturation and the merges are operating in state "b" of Fig. 1. When the interface separating freeway state A from state B (the freeway capacity) reaches the merge, the traffic demand suddenly jumps from point "b" to point "p" in Fig. 1; with the merge still in state NQ. According to Figure 2b, the merge shifts immediately into state BQ, which is point "c" of Fig. 1. Physically this means that the ramp flow is cut in half by the freeway surge and that the competition reduces the freeway flow by a like amount; thus, queues begin to develop on both the freeway and the ramp. The resulting upstream freeway state, labeled C, is shown in Fig. 3, together with the on-ramp state (open-circle). The flow restriction generates a shockwave between states B and C that travels backward (as shown in the figure). Similar shocks are issued from all the on-ramps. Symmetry implies that they will hit their neighboring upstream ramp simultaneously at a time t_1, and this will restrict the flow through all the merges even further; the freeway will be in a queued state everywhere. On the merge diagram, the consequences of this event are calculated by shifting the line with -1 slope down until it intersects the q^u axis for $q^u = q_C$ (the flow of state C). This is the abscissa of point c in Fig. 1; i.e. point c'. Because we are in merge state BQ we see from Fig. 2d that the new operating point of Fig. 1 after time t_1 should be "e". As before, this corresponds to a more congested upstream state E which is also shown on Fig. 3. New shocks are issued from all the ramps once again, and the ratcheting down process continues.

We see from Fig. 4a that in the time τ that it takes for the shocks to travel the distance between ramps, the system flow is reduced from q^f_{max} to q_C then to q_E, etc. [Since the shock travels at speed w, this time should be about 5 times larger than the free-flow vehicular trip time, yielding $\tau \approx 5$ min if ramps are spaced at 1 mile intervals.] We see from the geometry of Fig. 1 that flows are reduced by a constant factor $(1+\alpha)^{-1}$ at each step. Thus, at time $t_0 + t$, the flow is reduced by the factor $(1+\alpha)^{-t/\tau}$, if t is an integer multiple of τ. This

means that the flow has a half-life equal to $\tau\ln(2)/\ln(1+\alpha)$, which approximately reduces to $\tau\ln(2)/\alpha$ when α is small compared with 1 (as should be the case for freeways with more than two lanes).

Once the beltway is oversaturated, the collapse is rather quick. The half life for a three lane freeway with $\alpha \approx 1/5$ and $\tau \approx 5$ min. is about 20 minutes, and trip times are doubled much sooner than that. This occurs because speed is very sensitive to flow[3] and can also be verified directly from the trajectory of a vehicle entering the freeway at $t = t^0$. The figure shows that the travel time of such a vehicle increases by about 50% on the first leg of its trip and that its delay on succeeding legs becomes much greater. These qualitative remarks, of course assume that all the on-ramps are oversaturated when the freeway is at capacity and that there is no exiting traffic; longer half-lives can be expected if this is not the case. The remainder of this subsection and Sec. 5 relax these assumptions.

If there is exiting traffic the solution is similar. We assume that all the vehicles bypass the same number of on-ramps (N=6) because the presentation is easiest. (More complicated trip length distributions are considered later in this section with similar results.) The main difference appearing in Fig. 4b is that at the time when vehicles begin to leave the beltway a recovery wave is issued from the diverge world line, at the thick dot on the figure. This event introduces a less congested state C' with higher flow upstream of the diverge. The flow of the new state (shown on Fig. 3) satisfies $q_{C'} = q_C(N+1)/N = q_C(1+\mu)$ since 1 out of every N+1 vehicles leave. The displayed solution will not be valid if $q_{C'} > q_B \equiv q^f_{max}$; i.e. if $1+\mu > 1+\alpha$. However, in this case traffic would not be expected to collapse.[4] Because states with higher flows are introduced at the diverge world line in part (b) of the figure, the flows at the merge world line are higher than those of part (a); it should be intuitive that the drift toward gridlock is retarded.

An exact formula for the half-life can be obtained for arbitrary initial conditions if the trip length distribution is such that the fraction of exiting vehicles directly upstream of a diverge, $ß = q^r/q^d$, is fixed. An approximation that applies to asymmetric input flows is given in Sec. 5. We note that the constant ß is related to the slope of the exit mix line in Fig. 1 by $ß = \mu/(\mu+1)$, or by $1-ß = 1/(1+\mu)$.

Once the freeway is entirely in a queued state the kinematic wave equations governing its evolution become linear, with wave speed w. This means that any disturbances (large and small) will propagate with these speeds, independent of everything else going on on the

[3]The "elasticity" of speed with respect to flow in the queued regime is: $[dv/v]/[dq/q] = (1+v/w) \approx 6$

[4]If traffic did not collapse a clearing shock would be issued from the thick dot in the figure, in addition to the recovery wave. The state upstream of the diverge would then be B instead of C', and the state downstream would be uncongested with flow $q_B/(1+\mu)$. As shown in Sec. 4.2, this gets the solution on the way to recovery.

system. This is true for any set of (queued) initial conditions. Therefore an observer moving backward with velocity w would measure an unchanging flow relative to the ground when traveling between ramps. (Note that this statement is true in Fig. 4). On crossing an on-ramp, the flow measured by such an observer would decline by a factor of $(1+\alpha)^{-1}$; on crossing an off-ramp it would go up by a factor of $(1-ß)^{-1} \equiv (1+\mu)$. In the time τ that it takes to travel between two on-ramps the observer flow would decline by a factor

$$f = [(1-ß)(1+\alpha)]^{-1} = (1+\mu)/(1+\alpha). \qquad (2)$$

As expected, this factor is less than 1 if a stable solution on Fig. 1 does not exist. Once the observer has traveled around the loop once, the measured reduction will coincide with the one recorded by a stationary observer located at the same initial position. Although flows do not decline smoothly, the fractional decline in flow observed at any location after a time large compared with the wave trip time around the loop must be approximately $f^{t/\tau}$. Thus, the flow half-life can be expressed as:

$$\text{half-life} \approx \tau \ln(2)/\ln[(1-ß)(1+\alpha)]. \qquad (3)$$

for $t \longrightarrow \infty$. This expression is only marginally useful for predicting flows at a single location because the wave cycle may last a few hours for long beltways. Curiously, however, the expression can be shown to be very accurate on a much finer time scale when applied to the space-mean flow on the loop; i.e. to the integral of the flows around the loop divided by the length of the loop. The reason is that the sum of the flows relative to the ground seen by evenly and infinitesimally closely spaced observers on the loop coincides with the flow seen by an equivalent set of stationary observers at all times. Therefore, in as much as the factor (2) applies to the average flow measured by the moving observers every τ time units, the expression also applies to the space-mean flow on the loop every τ time units; i.e. every few minutes for typical freeways. The space-mean flow is relevant because it is linearly related to the total number of vehicles in the loop, since F(k) is linear. Thus, we are led to the conclusion that this grand total, like the space-mean flow, approaches the maximum possible occupancy in an exponential and rather smooth way.

If τ = 5 min, α = 1/5 and μ = 1/6 (ß = 1/7) the half life is about 2 hrs, with trip times doubled in about 20 min. The same trip lengths on a two lane freeway ($\alpha \approx 1/3$) would reduce the half life to about 25 min. This illustrates that the danger of collapse is largest on narrow expressways (high α) with closely spaced ramps (low τ and low ß); the extreme case being a roundabout without strict priority to the circulating traffic.

4.2 Recovery

The progress toward collapse can be reversed by reducing the input flow or augmenting the output. The latter is likely to occur automatically as drivers bail out of the system before their exit once the going gets to be too slow.

As long as the freeway is in a congested state and there are queues on all the ramps the problem continues to be governed by linear equations. Thus, Eqs.(2) and (3) continue to hold if ß is constant, independent of the initial conditions. However, (2) will now be a rate of increase and (3) will be the time needed to double the flow. Here we illustrate the process by tracing all the way to the uncongested equilibrium an idealized case in which the initial flows on each freeway segment are constant. This should also help us gain further practice with the transition diagrams of Fig. 2.

Figure 5 displays the complete solution to such a problem when the merges are initially in state (BQ) and the priority line is shifted (at t = 0) to the shown position. The rationale for the figure is explained below.

Initially the merge operating point moves from d to e, as per the rule of Fig. 2d; i.e., a state E' is introduced upstream without changing the downstream flow. When the interface between C and E' reaches the diverge, the increased flow must be matched by an increased flow upstream of the diverge (in the ratio 1-ß); this amount can be identified graphically by going up from point E' on the merge diagram to the exit mix line and coming back down with slope -1; the result is point G. When the interface between E and G reaches the next merge at $t = \tau$ the downstream flows at the merge increase to level G from level E and the rules of Fig. 2d (again) introduce merge operating point g. The graphical process is then continued in an obvious way until the interface between states G' and H' reaches the diverge world line (at the thick dot). We note from the geometrical construction that the flows so generated {E,G,H} and {E'G'H'} are geometric sequences with growth ratios $f = (1+\mu)/(1+\alpha)$ as claimed.

We now examine how the beltway reaches its final uncongested equilibrium state, recognizing that the prior construction procedure cannot be continued past the thick dot because the next flow in the sequence {E,G,H} would exceed capacity. As explained in footnote 4, the stable solution at the diverge is one where both a backward moving recovery wave and a forward-moving clearing shock are issued from the thick dot. The recovery wave introduces the capacity state (B) upstream of the diverge and the recovery shock an uncongested state (A) with $q_A = q_B(1-ß) = q_B/(1+\mu)$ downstream of the diverge.

This condition means that if these two flows are located on the abscissa of the merge diagram, then a vertical line passing through A and a slanted line with slope -1 passing through B must meet on the exit mix line (as shown on the figure). When the interface HB meets the merge world line the downstream flow at the merge increases, and on the merge

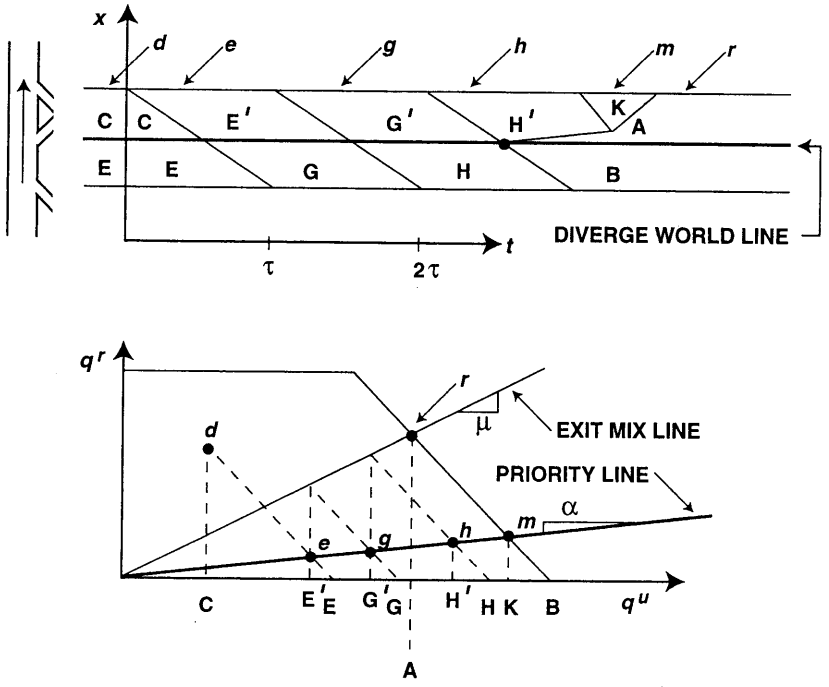

Figure 5. The recovery process

diagram we must translate the slanted line passing through H until it passes through B. Since we are still in state BQ, we see from Fig. 2d that the new merge state must be point m. Thus, a state K with flow $q_K = q_B/(1+\alpha)$ must be introduced upstream, as shown.

When the interface between H' and K meets the recovery shock, they coalesce into a single shock that will move with a positive speed since $q_A < q_K$. When this shock reaches the merge, the flow upstream would decrease to (uncongested) level A. Thus, the point depicting the desired flows on the merge diagram (not shown on the figure) would be at the intersection of a horizontal line passing through m and the vertical line passing through A. Because the merge is in state RQ at that time, Fig. 2a indicates that the operating point must be on the boundary, directly above the point corresponding to the desired flows; i.e. at point r. This is reassuring because point r is the stable equilibrium predicted by the theory of Sec. 2.

5. Asymmetric On-Ramps And Input Flows

It is unreasonable to expect all the on-ramps on a beltway to have the same α's or the same desired input flows, and this section relaxes these assumptions.

Let j index the various on-ramps (j = 1, 2,... R), as well as the two roadway segments on either side (u and d) and the off-ramp upstream. We start by imagining that the α_j are different but fixed in time, and that there are queues on every ramp. The distribution of trip lengths is geometric and rotationally symmetric so that ß is fixed, as before.

Since the kinematic wave equations are still linear, the arguments of Sec. 4.1 regarding moving observers still hold. Now the observer's flow is reduced by the factor:

$$f_j = [(1-ß)(1+\alpha_j)]^{-1}. \tag{4}$$

for every pair of ramps. By considering many evenly spaced observers, as before, it is easy to show that the rate of decay should be proportional to the geometric mean of the f_j. For small α_j's and ß's this is approximately $<\alpha>-\mu$, where $<\alpha>$ is the arithmetic average of the α_j's.

If not all the on-ramps are queued the solution is more complicated because the drop in flow measured by the observer after passing an on-ramp without a queue, q^r_j, is not a fixed fraction of the freeway flow. Since $q^d_j = q^d_{j-1}(1-ß) + q^r_j$, the observer's flow changes by the factor $f_j = q^d_{j-1}/q^d_j = (1-q^r_j/q^d_j)/(1-ß)$. Figure 2c shows that $q^r_j = \min\{Q^r_j, (\alpha_j/(1+\alpha_j))q^d_j\}$ and, thus, we can write:

$$f_j = (1-ß)^{-1}[1 - \min\{Q^r_j/q^d_j, \alpha_j/(1+\alpha_j)\}] \tag{5}$$

for ramps without queues.

We see from this expression that if the flow downstream of an on-ramp without a queue, q^d_j, is so small that the minimum function in (5) takes on the value of its second argument, then (5) coincides with (4); and a ramp queue is generated. Otherwise the reduction in flow seen by the observer is less (f_j is greater) than if there had been a queue. Thus, a conservative value for the reduction in the space-mean flow can be obtained by neglecting the ramp queues; i.e. by taking the geometric mean of (5) across all j.

A crude estimate independent of the detailed flow distribution around the loop uses the current space-mean flow, q, instead of q^d_j in (5). If we denote the resulting approximate geometric mean by E[f(q)], then the logarithmic rate of decay is approximately: $-(1/\tau)\ln(E[f(q)])$. Finally we note that (5) increases with q^d_j. Thus E[f(q)] declines as the

beltway flows drop, meaning that the decay process is predicted to accelerate. For sufficiently small q the decay rate is predicted to match that from (4) as one would expect.

6. DISCUSSION

Although we have not explored the effects of asymmetric O/D tables thoroughly and have not examined the effects of an inhomogeneous beltway (i.e. one in which there is a bottleneck) we believe that the qualitative description of the collapse presented in this paper also applies to these cases. After all, once the beltway is in a queued state everywhere and the drift toward collapse begins, the system of partial differential equations that govern the flow evolution on all the links is still linear. This means that a solution of the form (4) still exists in the general case. Furthermore, although further research is needed to extend the conditions for collapse presented in Sec. 3 to the general case, the results we have derived already suggest how the collapse can be avoided. This is explained below.

Beltway inhomogeneities mainly influence the way in which the point of no return is reached. For example, if there is a major bottleneck at a certain point, perhaps caused by insufficient exit flow on a major off-ramp, and there are no problems anywhere else, then the critical instant will be reached when/if the bottleneck queue grows all the way around the beltway and closes the loop. If other upstream ramps contribute to the flow the process is accelerated. It may be quite rapid (and comparable to the wave trip time between ramps) if the rush hour begins with a system that is already close to saturation. On the other hand, we repeat that the existence of attractive alternative routes will minimize the chances of lock-up. What we may observe in that case is a completely queued freeway on which traffic crawls at the speed of the surface streets surrounding it, and not benefiting its users.

What is remarkable is that gridlock/crawling conditions can be improved by controlling the input flows temporarily without knowing the O/D table. This is an important difference between the type of metering that is needed to correct gridlock problems and that needed to increase the flow upstream of bottlenecks; see Sec. 1. To avoid gridlock it suffices to make sure that some of the beltway links are in a non-queued state, which can be accomplished simply by restricting freeway entries directly upstream of problematic queues. Of course in a practical situation we probably should go further and attempt to eliminate all queues, as in roundabouts, by preventing entering traffic from disrupting the beltway traffic.

This is not to say that the O/D table is not important. Even in a seemingly ideal situation where the beltway is fully utilized with all the links at capacity, the combined exit flow on all the off-ramps may not be well-defined. It will usually depend on the priority given to the various on-ramps and on the O/D table. This is illustrated by Fig. 6 which depicts two possible equilibrium patterns for a one-lane roundabout with priority to the circulating traffic. We assume that all 4 approaches send a flow equal to the capacity of the roundabout: A

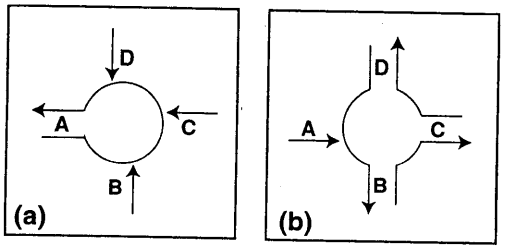

Figure 6. Equilibrium patterns for a roundabout

sends it to A, D to B, B to C, and C to D. Although both patterns exhibit the same vehicle miles of travel on the roundabout (and use it fully), pattern b is clearly preferred (equity issues aside) since it discharges the queues on the approaches three times faster. An area deserving continued research is the development of efficient time-dependent metering strategies given accurate O/D information.

We finally note that for a system that is on the verge of reaching the critical point of collapse on most days, there can be days when a fluctuation (e.g. more peaked demand) could send it over the edge. It is thus conceivable that, even with the same number of trips per day, the system would operate smoothly on some days and show a tendency toward gridlock on others... and this for no apparent reason!

REFERENCES

Allen, B. L. and Newell, G. F. (1976). Some issues relating to metering or closing of freeway ramps. Part I. Control of a single Ramp. Part II. Translationally symmetric corridor. *Trans. Sci.*, **10** (3), 227-242 and 243-268.

Banks, J. H. (1989). Freeway speed-flow-concentration relationships: more evidence and interpretations. *Trans. Res. Rec.*, **1225**, 53-60.

Daganzo, C. F. (1995). The cell transmission model: Part II. Network Traffic., *Trans. Res.*, **23B**, 79-94.

Hall, F. L., B. L. Allen and M. A. Gunter (1986). Empirical analysis of freeway flow-density relationships. *Trans. Res.*, **20A** (3), 197-210.

Lighthill, M. J. and J. B. Whitham (1955). On kinematic waves. I Flow movement in long rivers. II A Theory of traffic flow on long crowded roads. *Proc. Royal Soc.*, **A 229**, 281-345.

Lin, W. H. and Ahanotu, D. (1995). Validating the basic cell transmission model on a single freeway link. PATH Technical Note 95-3, *Institute of Transportation Studies, U. of California,* Berkeley, CA.

Lin, W. H. (1996). Validation of the cell-transmission merge model. *(in progress).*

Newell, G. F. (1977). The effect of queues on the traffic assignment to freeways. *Proc. 7th Int. Symp. on Transportation and Traffic Theory*, pp. 311-340, Kyoto, Japan.

Newell, G. F. (1982). *Applications of queueing theory* (2nd edition). Chapman Hall, London.

Newell, G. F. (1992). Private communication.

Newell, G. F. (1993). A simplified theory of kinematic waves in highway traffic, Part I: General theory; Part II: Queueing at freeway bottlenecks; Part III: Multi-destination flows. *Trans. Res.*, **27B** (4), 281-314.

Richards, P.I. (1956). Shockwaves on the highway. *Opns. Res.*, **4**, 42-51.