ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 00 (2010) 000–000

Procedia Social and Behavioral Sciences

www.elsevier.com/locate/procedia

19th International Symposium of Transportation and Traffic Theory

Extended Bottlenecks, the Fundamental Relationship, and Capacity Drop on Freeways

Benjamin Coifman[a,b]*, Seoungbum Kim[a]

[a]*The Ohio State University, Dept. of Civil and Environmental Engineering and Geodetic Science, Hitchcock Hall 470, Columbus, OH 43210*
[b]*The Ohio State University, Dept. of Electrical and Computer Engineering, Dreese Labs 205, Columbus, OH 43210*

Abstract

This paper presents evidence that the commonly used point bottleneck model is too simplistic for freeway bottlenecks, the actual mechanism appears to occur over an extended distance. We find evidence of subtle flow limiting and speed reducing phenomena more than a mile downstream of a lane drop bottleneck. These phenomena impact the fundamental relationship, FD. Close to the lane drop the free flow regime appears to come from a "parabolic" FD, but further downstream the relationship straightens to a "triangular" FD and throughput increases. We develop a theory to explain the underlying mechanisms. These insights should help resolve the decades long debate about the shape of the FD. The phenomena also provide a mechanism that may contribute to the empirically observed capacity drop often seen at bottlenecks. Although we study a lane drop, this work should be transferable to other bottlenecks where the capacity restriction persists for an extended distance, e.g., a corridor with a fixed number of lanes and an on-ramp bottleneck.

Keywords: bottlenecks; fundamental relationship; fundamental diagram; flow-density; freeway traffic; capacity; lane change maneuvers

1. Introduction

A bottleneck becomes active whenever demand exceeds capacity. There are many different freeway features that can cause bottlenecks, e.g., grades, lane drops, merge and weaving sections (Banks, 1990). Conventionally an active bottleneck is modeled as if occurring at a discrete point or short distance along the freeway, e.g., as Bertini and Leal (2005) succinctly put it, an active freeway bottleneck is, "a point on the network upstream of which one finds a queue and downstream of which one finds freely flowing traffic." Point bottleneck models are widely used and are well established in traffic flow theory. Contrary to such conventional wisdom, in this paper we present evidence that modeling the bottleneck mechanism as if it occurs at a single point along the road is too simplistic and we show that the mechanism appears to occur over an extended distance for some bottlenecks. A few previous papers have made a similar argument for various reasons (e.g., Coifman, et al, 2003; Koohong and Cassidy, 2004; Ogut and Banks, 2005). The present work highlights additional features that are obscured by the single point assumption. We adopt

* Corresponding author. Tel.: (614) 292-4282.
E-mail address: Coifman.1@osu.edu.

the term "apparent-point-bottleneck", APB, to specify the location where one would place the point bottleneck model while underscoring our belief that the actual bottleneck mechanism occurs over an extended distance.

A comprehensive understanding of the traffic conditions over space that give rise to the bottleneck remains elusive. Many earlier works have studied traffic evolution in the queue, upstream of freeway bottlenecks using loop detectors, film/video, etc.. Few of these studies consider conditions downstream of the APB beyond looking for free flow conditions to ensure that the bottleneck is active.[2] In this study we take the opposite approach and focus almost entirely on traffic downstream of a recurring APB that arises from a lane drop. Like earlier studies we employ conventional loop detectors, but we then go further and use over a hundred probe vehicle tours through the corridor. We find evidence of subtle flow limiting and speed reducing phenomena more than a mile downstream of the APB when the bottleneck is active. As discussed in Section 2, both data sets tell the same story: when the bottleneck is active it takes several miles downstream of the APB before drivers attain full free speed for the given flows.

The first major finding of this work is empirical and arises from the long distance before attaining free speed. Close to the APB the free flow regime of the fundamental relationship, FD, between flow and concentration (density or occupancy) tends slightly towards lower speeds at higher flows (a "parabolic" FD, e.g., as found in the current Highway Capacity Manual, TRB, 2000 and dating back to Greenshields, 1935), but further away from the APB the relationship straightens, exhibiting nearly constant free speed over the entire range of flows (a "triangular" FD, e.g., as found in Munjal, et al, 1971[3]; Hall, et al, 1986; Banks, 1989). The FD is fundamental to most traffic flow theory, yet debate about its shape remains unresolved after decades of research (e.g., Drake, et al, 1967; Hall, et al, 1986; Hsu and Banks, 1993; Banks, 2002; Jin and Ran, 2009; Del Castillo, 2010). As such, the present work could have far reaching impacts by explaining the different empirically observed shapes of the FD reported in the literature.

The second major finding of this work is also empirical, namely that the conventionally defined free flow regime of the parabolic FD data in this corridor do not appear to be purely free flowing, thus potentially leading to an underestimate of bottleneck capacity if care is not taken. When the bottleneck is active, immediately downstream of the APB flow increases monotonically with occupancy and speeds are within 85% of the free flow speeds observed when the bottleneck is inactive. By conventional practice one would measure capacity immediately downstream of the APB. Yet when the bottleneck is active we observed the throughput increasing further downstream of the APB due to ramp flows. The number of lanes remains constant in this segment, so even though this segment seems to be in the free flow regime, the subsequent ramp flows appear to consume some of the capacity that would otherwise be available immediately downstream of the APB.

The third major finding of this work comes from our theory developed in Section 3 to explain the empirically observed findings. Namely, the segment downstream of the lane drop (i.e., the APB) is unique, exhibiting properties of both the free flow regime and the congested regime (holes propagating downstream, and delay propagating upstream, respectively). In aggregate, the costs of entering vehicles propagate upstream and reduce the throughput of the APB while the benefits of exiting vehicles propagate downstream and are lost. Conventional data aggregation is too coarse to see these microscopic disturbances; but as will be shown, the microscopic disturbances provide a possible explanation as to why some researchers have found parabolic FD while others have found triangular FD. The presence of the disturbances also provides a mechanism that may contribute to the so-called "capacity drop," where the bottleneck throughput drops once it becomes active. This capacity drop is often attributed to lane change maneuvers in the vicinity of the APB because in general when a vehicle changes lanes it temporarily occupies two lanes at the same time and thus, consumes more than a single vehicle worth of capacity (e.g., Coifman, et al, 2006). If a vehicle changes lanes while passing through the assumed-point-bottleneck it will reduce the throughput (e.g., Laval and Daganzo, 2006; Duret, et al, 2010). But the literature does not discuss the impacts of lane change maneuvers or other minor flow limiting events further downstream of the APB.

We use the term *roadway capacity*, RCap, to denote the local capacity at any specific point along the freeway if the flow at that point were not constrained by the upstream and downstream roadway demands and capacities. The RCap at any given point in the corridor can be larger than the *bottleneck capacity*, which characterizes the maximum

---

[2] The most notable exception being Hall et al (1992), which assimilates findings from Hall and Gunter (1986), Persaud and Hurdle (1988) and Banks (1989) to show how the observable flows change relative to the bottleneck location in a fashion many now take for granted.

[3] Munjal, et al provide no empirical justification, instead attributing the triangular shape to Drake, et al, (1967), but the triangular shape does not follow directly from Drake, et al. Since Drake and his coauthors published three papers with the same title between 1965 and 1967, it is possible that the triangle came from one of the non-cited variants.

sustainable throughput of the bottleneck. As such, the present work is applicable to bottlenecks where the RCap remains roughly constant for an extended distance beyond the initial drop in RCap or increase in demand, e.g., a similar near-RCap situation can arise at an on-ramp bottleneck. So the results herein will likely apply to corridors with one or more on-ramp bottlenecks if the number of lanes remains constant (e.g., Ogut and Banks, 2005). However, as will become evident in Section 3, this work is not applicable in cases where the RCap is diminished for only a short distance (e.g., an incident blocking one lane or an off-ramp queue spilling into the mainline).

There have been surprisingly few studies of lane drop bottlenecks. Among these some are not directly relevant to the present work either because the bottleneck was not active (e.g., Munjal, et al, 1971; Goodwin and Lawrence, 1972; Hsu and Banks, 1993) or the work looked strictly upstream of the lane drop (e.g., Zhang and Shen, 2009). Only a few studies had both an active bottleneck and the researchers included the segment downstream. Among these, Persaud and Hurdle (1988) studied a lane drop bottleneck and unlike most studies, they explicitly denote their subject bottleneck extending 1 km. But this spatial terminology is to simply indicate that the physical restriction persists for this distance, the bottleneck mechanism itself was assumed to reside at the upstream edge (as per Hurdle and Datta, 1983). The authors attribute most of the distance to be in discharge, with traffic accelerating downstream of a queue. The main point of their effort was to illustrate that discharge flows will exhibit constant flow but speeds will increase as one moves downstream, thereby explaining why the empirical speed-flow curves of the day, sampled at many discrete locations, seemed to drop off abruptly at some *critical flow*. Bertini and Leal (2005) examined two different lane drop bottlenecks using cumulative arrival curves. They observed that once the bottlenecks became active the discharge flow drops by 7% compared to conditions just prior to activation. Laval and Daganzo (2006) propose a theoretical model to explain the drop in discharge rate at bottlenecks, noted above, and then simulate the lane drop from Bertini and Leal to validate their model.

Finally, a few studies have begun to move away from an assumption of a point bottleneck, e.g., Coifman, et al, (2003), Koohong and Cassidy (2004), and Ogut and Banks (2005). Of note, Laval (2006, 2009) use moving bottlenecks to stretch out a conventional point bottleneck model over a finite distance (e.g., trucks on an upgrade). But like a point bottleneck model, queues only originate at the upstream end of this finite region. The papers then define capacity to be the average of two or more distinct traffic states, i.e., whether or not a given period is impacted by a moving bottleneck. While this capacity definition certainly has practical applications for describing the maximum throughput, as noted in Laval (2009) it is not the apex of the FD, i.e., RCap. Laval and Leclercq (2010) present a macroscopic model to capture location dependent bottleneck formation assuming a continuum of ramps along a corridor. The authors show how demand to/from ramps could cause queues to form in different locations or the head of the queue to move in response to ramp demands. The work is predicated on the assumption that queuing forms at the point where traffic reaches a critical density. Laval and Leclercq's model does not capture either of the empirically observed findings of our research noted above.

The remainder of this paper is organized as follows. Section 2 presents empirical results demonstrating the anomalous relationships downstream of an APB, Section 3 presents our theory to resolve the anomalies, and Section 4 presents the discussion and conclusions.

2. Anomalous Relationships Downstream of an Apparent-Point-Bottleneck- Empirical Results

Consider the roughly 2 mi corridor along northbound I-71 in Columbus, Ohio, as shown in Fig 1d. This segment includes a lane drop around mile 3.9, as lane 4 exits to an off-ramp. While some vehicles exit the freeway at this ramp, most of the lane 4 traffic moves to the three through lanes. The segment is immediately downstream of the I-670 interchange, ensuring that demand on the four-lane section can exceed the capacity of the three-lanes past the drop. This lane drop is an APB that is often active during afternoon peak periods. Using a differential global positioning system (DGPS) equipped probe vehicle, we collected 134 passes through the roughly 2 mi corridor, from approximately 67 days during morning and evening peak periods, between 2005 and 2007 (Coifman, et al, 2003; Coifman, 2006). The driver was instructed to travel in lane 2 except when overtaking another vehicle. After snapping each pass to a common reference, a given pass was sorted into one of the following three groups:

- Free flow inactive- speeds remain above 45 mph over the entire pass - the 87 such passes are shown in Fig 1a.
- Active- speeds are below 45 mph upstream of the lane drop and above 45 mph downstream - the 29 such passes are shown in Fig 1b.
- Congested inactive- speeds stay below 45 mph downstream of the lane drop - the 18 such passes are not shown.

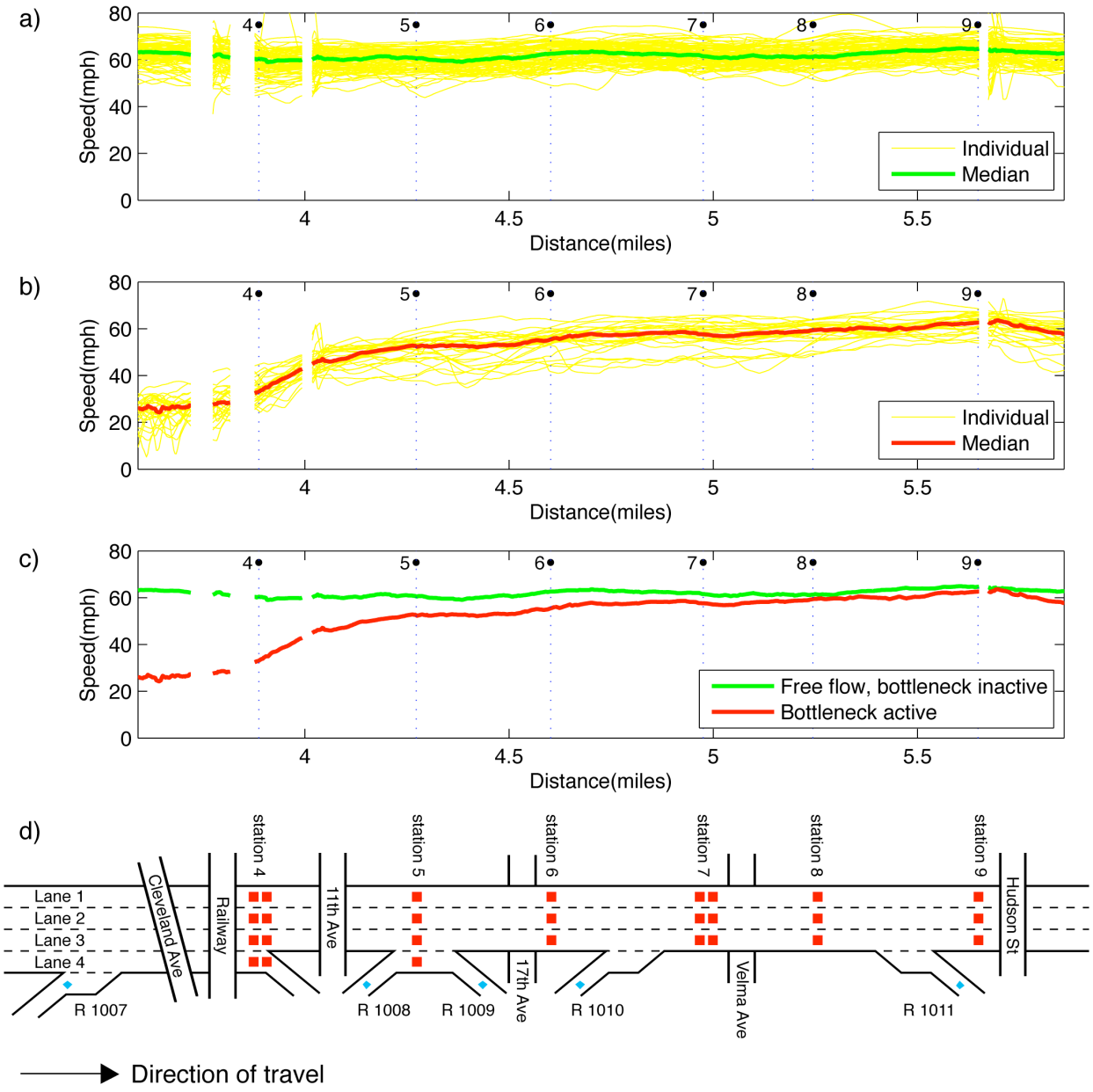

Figure 1, Probe vehicle speed as a function of distance for (a) 87 runs when the bottleneck was free flow inactive and the group median, (b) 29 runs when the bottleneck was active and the group median, (c) both medians together, and (d) a schematic of the corridor approximately to scale. Note that the exact location of detector stations 4-9 are shown in parts (a)-(c).

The congested inactive runs are excluded from further study because they indicate a queue from a more restrictive bottleneck downstream precludes the subject bottleneck from activating. Henceforth, for brevity we use "inactive" to refer to the *free flow inactive* group.

The median speed at each location is shown in Fig 1a-b for the given group. There are a few locations where the DGPS is obstructed by overpasses and no data is presented. As expected, Fig 1a shows relatively constant speeds throughout. Fig 1b shows congested conditions upstream of the APB. The median speed increases sharply past the lane drop and then begins to level off around mile 4.2, but it continues to slowly increase all the way to mile 5.6. Fig 1c compares the two median speed curves, and shows that when the bottleneck is active the speed remains below

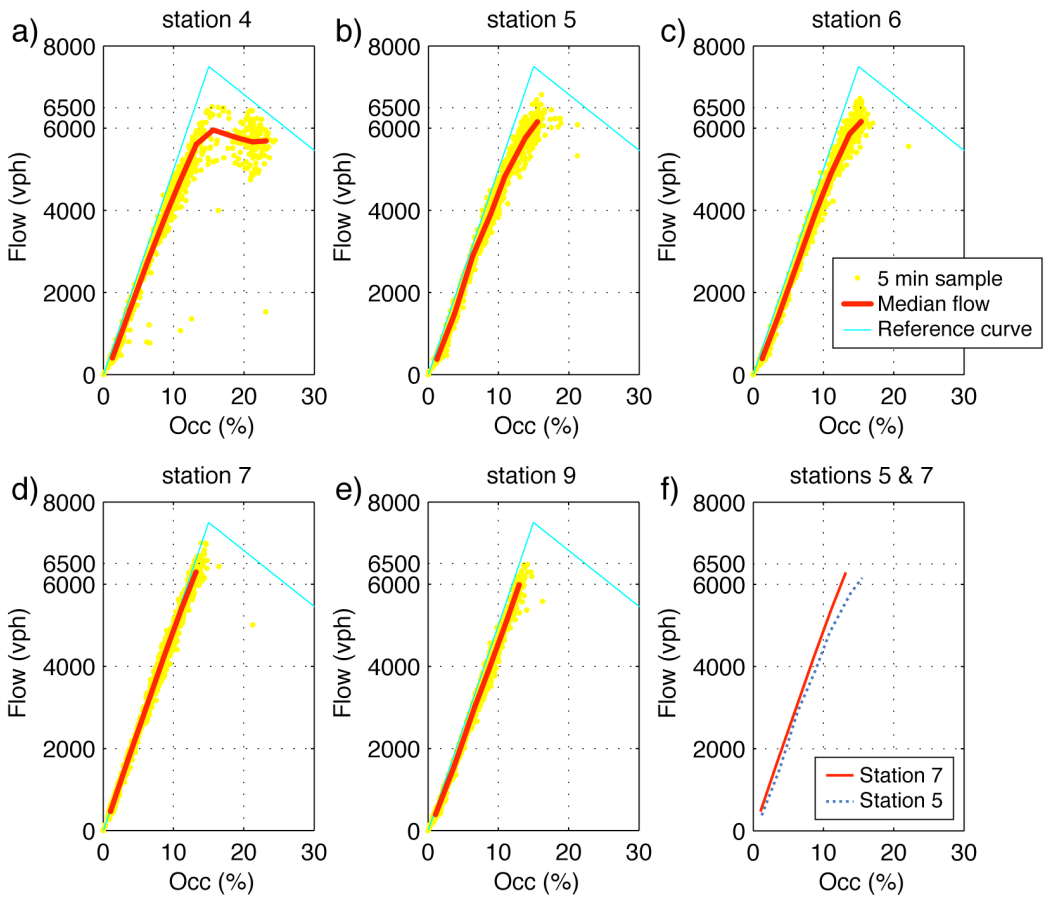

Figure 2, Flow versus occupancy, 5 min samples, over 6 weekdays at (a) station 4, (b) station 5, (c) station 6, (d) station 7, and (e) station 9. Each plot includes the median flow curve (by occupancy bin) and for comparison, a common reference curve. (f) A direct comparison of the medians from station 5 and station 7.

that of when the bottleneck is inactive until mile 5.2, if not further. So while the APB occurs around mile 3.9, the impacts of the restriction are evident in the probe vehicle speed for over a mile downstream. An active bottleneck should be discharging vehicles at capacity. It is often assumed that capacity flow occurs at a speed below free speed, this assumption is implicit in a parabolic FD (e.g., the current Highway Capacity Manual, TRB, 2000) and most conventional level of service measures. So the speed difference between when the bottleneck is active and inactive does not in and of itself suggest there was a problem with selecting the specific location of the APB. However, the fact that a measurable difference exists so far downstream of the lane drop and that the difference diminishes the further downstream one goes is counter to the idea that the phenomena can be modeled as occurring at a single point in space.

There are also six loop detector stations in the study area, two dual loop detector stations (stations 4 and 7) and four single loop detector stations, as shown with rectangles in Fig 1d and vertical dashed lines on the plots in Fig 1a-c. Station 4 is immediately downstream of the painted gore point on the pavement. Station 8 suffers from a chronic error and has never provided useable data, thus, station 8 is not employed in this study. The points in Fig 2 show flow versus occupancy using 5 min samples summed across lanes 1-3 at the five operational stations, from six

Table 1, The total number of 5 min observations with flow above 6,000 vph over the six days at each station. The maximum throughput increases from station 4 to station 7 then drops at station 9 after an off-ramp.

	Between 6,000 and 6,500 vph	Above 6,500 vph	Total
Station 4	40	2	42
Station 5	60	7	67
Station 6	65	9	74
Station 7	88	28	116
Station 9	42	0	42

weekdays in which the bottleneck was active in October 2009.[4] The dates were chosen arbitrarily, with the stipulation that the bottleneck had to be active for some portion of the given date. Fig 2a-e show the sampled data with points, the median flow sampled by occupancy bin with a bold curve, and a common reference consisting of two thinner line segments that meet at 15% occupancy. This latter reference curve is only meant to facilitate comparison across the plots at different stations. All of the stations in Fig 2 are in the same corridor, all of the plots use data from the same set of days, and the vast majority of vehicles passing any one of the stations will pass all of the stations in the figure.

These stations report per vehicle data, and for the present study we first employ the data cleaning processes presented in Lee and Coifman (in-review-A, in-review-B, in-review-C) to all of the detector stations. The results are similar when applied to the uncleaned data, with one exception that we found in an earlier data set as follows. This analysis was initially conducted using data from 5 days in June 2005. During the summer of 2009 we diagnosed a hardware fault at station 9 and the operating agency corrected it. Prior to this date the sensor cards at this station were set too low and the detectors often dropped out in the middle of long vehicles (for more details, see Lee and Coifman, in-review-C). Although our cleaning corrected the problem, to ensure consistency with the uncleaned data, for this paper we only present loop detector data collected after the hardware fix at station 9, but note that these results are consistent with the 2005 data set as well.

Fig 2a shows that some of the station 4 data fall in the congested regime, while station 3 (located just beyond the left edge of Fig 1d, data not shown) exhibited much more congested data over the study period, consistent with Fig 1b. More importantly, Fig 2b shows that the station 5 data appear to come strictly from the free flow regime. In the context of the point bottleneck model, one would likely conclude that station 5 is downstream of any active bottleneck on the study days used in the plot. Furthermore, the station 5 data appear to be consistent with a parabolic FD. If the bottleneck mechanism were actually confined to a point in space near the lane drop at station 4, then ordinarily the station 5 data should include the capacity flow of the active bottleneck. But station 5 is located on link with an auxiliary lane and as noted earlier, the auxiliary lane is not included in the flow-occupancy measurements. Generally the ramp flows are small, neither the on nor the off-ramp exceeded 600 vph over any 5 min sample on the study days (roughly 1/10th of the maximum observed throughput) and the ramp flows were usually smaller than 480 vph in any given sample. The on-ramps are metered, which generally prevents pulses of demand during the peak periods. During the evening peak period when the bottleneck was active the inflow from the 11th Ave on-ramp usually exceeds the outflow to the 17th Ave off-ramp. Table 1 shows that the through lanes at station 5 had an average of 55 min per day with flow in excess of 6,000 vph (67 samples over six days) while station 4 only had an average of 35 min per day with flow in excess of 6,000 vph (42 samples over six days). Fig 2c shows station 6, and is very similar to station 5 in terms of the median curve while Table 1 shows station 6 had slightly more samples with flow in excess of 6,000 vph.

[4] Following conventional practice we use occupancy as a proxy for density since few alternatives are available for single loop detector stations. It can be shown that in a given sample occupancy and density are related by the average effective vehicle length. Even if the effective vehicle length varies from sample to sample, it should impact all of the stations similarly since the vast majority of vehicles passing any one of these stations will pass all of these stations.

But something changes by the time the traffic reaches station 7. Fig 2d shows the flow-occupancy relationship straightens out in the free flow regime at station 7, being indicative of a triangular FD. Table 1 shows the maximum throughput continues to grow at station 7, with an average of 96 min per day with flow in excess of 6,000 vph. In fact station 7 has more than three times as many samples above 6,500 vph compared to station 6, and more than ten times as many compared to station 4. After passing a major off-ramp to Hudson St. the throughput drops at station 9 (Fig 2e), but the flow-occupancy relationship in the free flow regime at station 9 remains straight and indicative of a triangular flow-occupancy relationship.

Fig 2 highlights two anomalies downstream of the APB at the lane drop. As one travels downstream over the span of 0.7 miles from station 5 to station 7, first, although all of the data from the stations appear to come from the free flow regime, the FD straightens from parabolic to triangular in shape (Fig 2f), and second, as quantified in Table 1, the throughput progressively increases even though the number of through lanes remains constant. This relationship holds whether using median flow by occupancy bin (as shown) or median occupancy by flow bin (not shown). The analysis was repeated by individual lane, as shown in Appendix A. Each lane at station 5 and 6 exhibited the curvature. Lane 1 at station 7 exhibited a slight curvature, the FD is straight in the remaining lanes at station 7 and all lanes at station 9. Because the curvature was seen in all lanes, it cannot be explained by ramp flows alone. Likewise, few long vehicles travel in lane 1, so it is not likely that trucks could be the sole explanation for the empirically observed curvature.

3. A Theory to Resolve the Anomalies

While the commonly accepted definition of an active bottleneck calls for free flow conditions downstream, the interpretation of what constitutes free flow is vague and subjective. It is also commonly accepted that speeds decrease as flow nears capacity, i.e., a parabolic FD, but these small drops in speed may be due to previously unaccounted for traffic features. Whenever demand approaching the lane drop at station 4 exceeds the capacity of the three-lane section downstream, it ensures that all available throughput at the start of the three-lane section will be consumed, i.e., the lanes are at or near their RCap. This near RCap state can extend for many miles if the following three conditions are met. First, the geometry does not change significantly, e.g., the number of lanes remains roughly constant. Second, no queue from a more restrictive bottleneck backs up into the section. Third, the ramp inflow is roughly equal to or greater than the ramp outflow, i.e., vehicles departing at off-ramps are generally replaced by vehicles entering from on-ramps.

When a lane is near its RCap we can no longer assume the traffic will be able to accommodate entering vehicles over a negligible distance or even over a finite distance. Yet all too often the literature on bottleneck operation implicitly assumes that the lane change accommodation distance is negligible. Consider the implications, first Fig 3a shows vehicle trajectories in a single lane that is near RCap. At some time and location, denoted with a star, a vehicle enters the lane. Assuming the vehicles are close enough together, those immediately behind the lane change maneuver must briefly slow down to accommodate the headway of the entered vehicle. Borrowing the discrete traffic state methodology from Wang and Coifman (2008) and using the flow-density relationship in Fig 3c,[5] if traffic is initially at free flow state A and the transient queue behind the entered vehicle can achieve a higher flow at congested state D, the disturbance will propagate downstream, as shown in Fig 3d. Once more the star denotes the time and location of the maneuver. This discretized traffic state diagram is adopted to clearly show the key events, but as a result, it is impossible to capture the transient continuum of short headways experienced by the vehicle that actually changed lanes or the vehicle immediately behind it. Generally these states will be off of the flow-density curve, falling to the right of the congested regime; but the specific evolution is inconsequential to the present work, so the region M is used to denote the transient states experienced by these two vehicles. Outside of region M, all vehicles are assumed to follow the LWR (Lighthill and Whitham, 1955; Richards, 1956) theory for kinematic waves. Downstream of the lane change maneuver, the traffic now exhibits a brief platoon of vehicles at the RCap for the lane, C. If the initial flow is higher, e.g., state B in Fig 3e, flow must momentarily drop to state D when a vehicle enters the lane, and thus, a transient queue propagates upstream. The queue eventually dissipates and downstream of the lane change maneuver again there is a platoon of vehicles at state C, but this time longer than that in Fig 3d.

[5] As per conventional practice, q denotes flow and k denotes density.

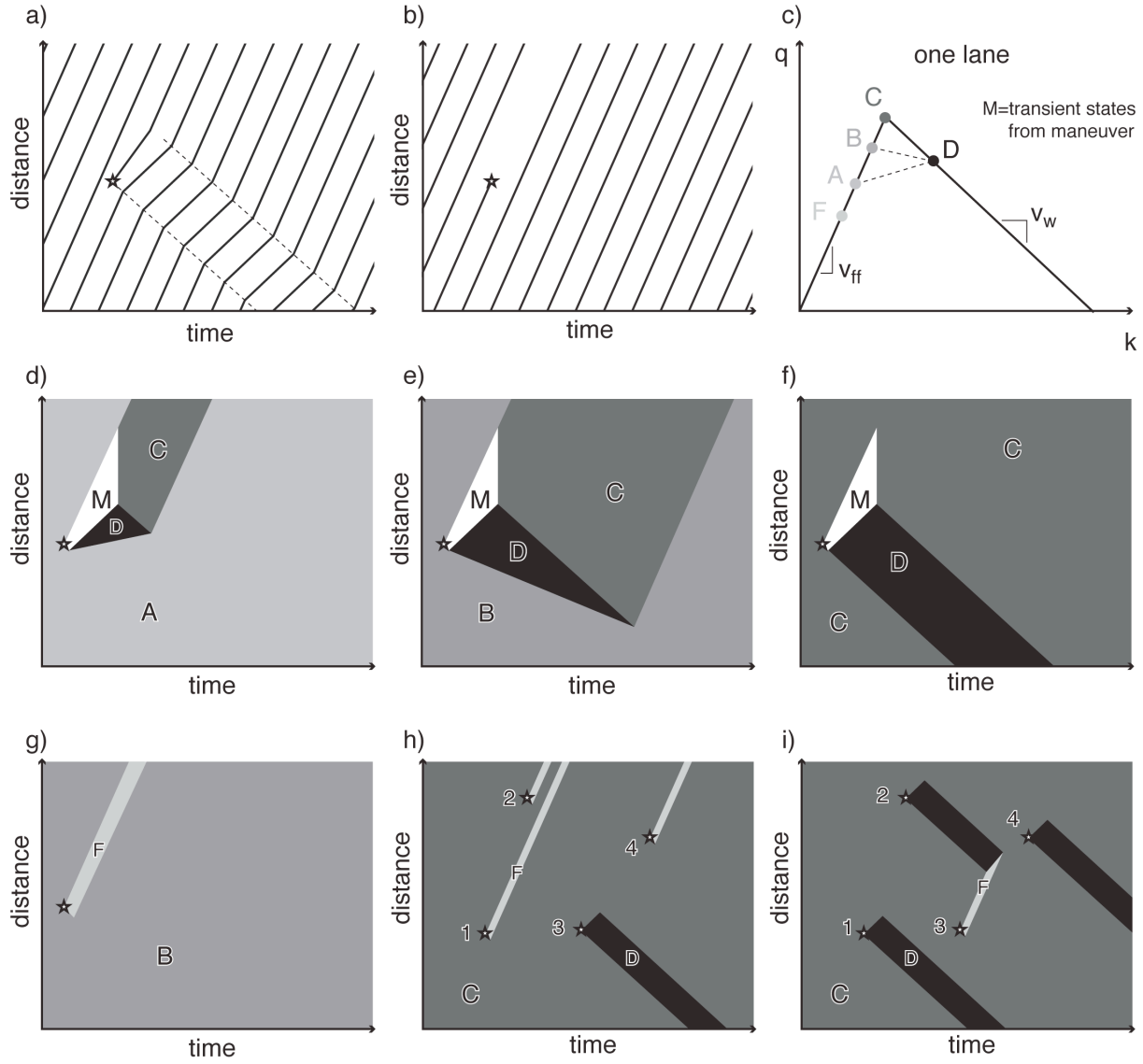

Figure 3, Vehicle trajectories in a single lane that is near roadway capacity, (a) when a vehicle enters at the star, (b) when a vehicle exits at the star. (c) Flow-density relationship for one lane. The state diagram in an entered lane when (d) initial flow is moderate and the flow behind an entering vehicle can climb to a higher congested state, (e) initial flows are higher and the flow behind an entering vehicle must drop to a lower flow, (f) initial flows are at roadway capacity so there is no spare capacity to absorb the disturbance and it will propagate upstream indefinitely. (g) The state diagram in an exited lane, because the traffic is at free speed the following traffic cannot close the gap and it will propagate downstream. Finally, the combined impacts on a two lane road from four lane change maneuvers in lane (h) and lane (i).

If the traffic is initially at state C, the traffic is still traveling at free speed, v_{ff}. But as Fig 3f shows, there is no unused capacity to absorb the impact of the entering vehicle and the resulting disturbance propagates upstream indefinitely, delaying all drivers upstream by one headway. Since traffic is already at capacity, no platoon is evident downstream of the maneuver (i.e., there is no memory of the event downstream in this lane). Now consider the progression behind an exiting vehicle. Fig 3b shows vehicle trajectories in a single lane. At some time and location, denoted with a star, a vehicle departs the lane. But since the traffic is already at free speed, the following vehicle cannot consume the gap. The resulting state diagram is shown in Fig 3g assuming traffic is initially at state B. The evolution remains strictly in the free flow regime, so a momentary drop in flow to state F propagates downstream

with the traffic, in the gap that the exited vehicle left behind. The figure would look the same if the initial ambient state were A, B or C, though the specific state F will depend on the ambient state: the flow will be half the ambient flow while the duration of state F will be twice the ambient headway.

In the case of a lane change maneuver one would see an entrance in one lane and an exit in another lane, with the respective disturbances in the two lanes. But the disturbance in a given lane is independent of where the vehicle came from or is going to. Thus, if a vehicle enters via an on-ramp, one could see Fig 3f in one lane without observing Fig 3g in another (or vice versa for an off-ramp). When a segment is at or near state C, the asymmetry between Fig 3f and 3g gives rise to unexpected phenomena. First, consider a hypothetical two-lane freeway segment, without ramps, initially at state C in both lanes. A vehicle that departs lane h (lane change maneuvers 1, 2, and 4 in Fig 3h) must enter lane i (Fig 3i) and vice versa (lane change maneuver 3). In this example both lanes experience delays from entering vehicles that propagate to the upstream edge of the plot. For three of the four maneuvers any benefit from departing vehicles propagates past the downstream end of the segment. The resulting gap can consume the delay behind an entrance if the two disturbances collide, e.g., exit 3 and entrance 2 in Fig 3i. Although many such disturbances may be consumed in this fashion, in general it is impossible for all of the delays to be absorbed without ramps generating excess holes and even then it is unlikely, e.g., the delay arising from the upstream-most maneuver cannot be consumed. Of course if the signals are in different lanes, they will pass without collision, e.g., the disturbance from exit 1 in Fig 3h simply passes the disturbance from entrance 2 in Fig 3i. Next, suppose the two plots did not include maneuver 3, then there would strictly be downstream moving holes in Fig 3h and upstream moving disturbances in Fig 3i. If one took the average flow across the top edge of Fig 3h or the bottom edge of Fig 3i, the average flow would be less than that of state C. The lane change maneuvers are consuming the RCap even though they far downstream of the APB.[6]

Now consider what happens as the delay disturbance of Fig 3f propagates further upstream to the APB. If demand exceeds capacity of the three-lane section, the lane drop ensures the three-lane section is initially operating at state C and a queue exists on the four-lane section. Conservation of flow dictates that the queue in the four-lane section must have the same flow as the three-lane section. The brief disturbance to state D behind the entering vehicle is more restrictive than state C and when this disturbance reaches the lane drop, the four-lane section must also briefly drop to the lower flow. The disturbance then continues propagating upstream of the lane drop. Herein lies a conflict: a disturbance from downstream propagates through the lane drop, something that should not happen if the bottleneck mechanism strictly occurs at single point in space near the lane drop. Or, in other words, if one applies the point bottleneck model at the lane drop, then the model momentarily becomes inactive as the disturbance passes. The magnitude of this disturbance should be small, making it easy to miss. For example, if Δv_3 drops 5 mph in the three lane section[7], then Δq would drop by only 1% and the new state would be so close to C that it would be almost impossible to distinguish the difference in aggregate flow and occupancy data. Meanwhile, Δv_4 in the four lane section would drop by less than 1 mph, but it would still drop due to downstream conditions and thus, it would still momentarily deactivate an otherwise active point bottleneck. In general, as long as the duration of the disturbance does not change as it crosses the lane drop, Δq and Δk will be the same on both sides of the drop, on the other hand, Δv_3 will be larger than Δv_4 both in relative and absolute terms. While the speed drop is attenuated when the disturbance enters the four-lane section, only a fraction of this attenuation is realized in delay savings by the individual drivers because although Δv_4 is smaller than Δv_3, the slower moving vehicles in the four-lane section will take more time to traverse the disturbance.

In the case of an on-ramp one would see many disturbances originate at the location where the vehicles physically enter the freeway, without the corresponding exit from a different lane. When demand is low there is plenty of room for the entering vehicles and all of the information propagates downstream. When the lane drop is limiting flow, the three-lane section is at RCap even though it is free flowing. So a vehicle can only enter the freeway from the ramp by delaying all upstream vehicles in the entered lane (Fig 3f). Upstream of the ramp traffic remains predominantly in state C, but there are frequent drops to state D. In practice it is likely that the upstream moving disturbances eventually consolidate into larger disturbances, but in any event, queuing theory dictates that

---

[6] The figures show the state transitions with straight lines for clarity of presentation; however, the key results would essentially be the same if there were a random walk or some other shape to the interfaces.

[7] Roughly the difference between the two medians at mile 4.8 in Fig 1c

the delays (manifest as lower flows) must propagate to the upstream vehicles. Taken on average, the net result is some average state D' that falls between the two discrete states and some mainline throughput is lost to the on-ramp flow (the difference between the flow at states C and D'). Or in the context of the empirical data of Table 1, station 7 exhibits a higher throughput than stations 5 and 6 in part because of vehicles entering the freeway from the on-ramps consume some of the downstream RCap that would otherwise be available upstream of the on-ramp.

In this scenario, any real capacity drop depends on where you look. In the absence of lane change maneuvers, downstream of the on-ramp the capacity never drops; and upstream of the ramp mainline capacity is merely lost to the on-ramp traffic. If vehicles change lanes (e.g., Fig 3h-i) or can exit at an off-ramp, then some capacity loss will also be seen downstream whenever a downstream moving hole passes (Fig 3g). Many recent publications attribute the capacity drop to lane change maneuvers in the proximity of the APB (e.g., Cassidy and Bertini, 1999; Bertini and Leal, 2005; Laval and Daganzo, 2006; Chung, et al, 2007), but this example illustrates that the capacity drop can also arise due to lane change maneuvers far downstream of the APB.

So far the examples have focused primarily on separable maneuvers. If a second vehicle enters within region D or C of Fig 3d-e, it will either result in a congested state with speeds slower than state D, a longer duration than shown in the diagram, or both (in the end, the traffic must add a delay equivalent to two headways though some combination of these impacts). Each additional vehicle that enters before the delay has been resolved will further compound the situation (e.g., as illustrated in Duret, et al, 2010). The compounding maneuvers may be spatially far apart, they simply have to enter the same upstream moving disturbance. The accommodation time in near free flow traffic could be many seconds if the displacement is only a few mph, in which case the duration of the D bands could be large enough that one might not see the transitions from C to D and back again because the traffic is perpetually in a state of accommodation.

The impact of averaging of traffic states together cannot be overlooked. Almost all hydrodynamic traffic flow models rely on the FD, yet as noted in the Introduction, the shape of this relationship has proven difficult to quantify. All empirical measurements of flow, density and occupancy implicitly average over time, space, or both. While it is commonly accepted that combining data from different traffic states can lead to average states that are not representative of any instantaneous traffic state, this guideline is typically applied at the sampling resolution, usually on the order of 30 sec or longer at conventional loop detectors. Yet the Nyquist sampling criterion dictates that discrete time series data can only resolve features that last at least two samples (see, e.g., Coifman, 1996). While common techniques for reducing noise in time series data filter out features that last only a few samples, e.g., cumulative arrival curves, the resolution is still too coarse. Each of the upstream moving delay waves at state D comprises a second or two of delay (the time it takes to admit one headway at the point of entrance), so each of these drops is either short in duration or small in displacement. In either case, it is impossible to detect their presence individually over the background noise in conventional loop detector data. Similarly, each of the downstream moving holes at state F will only last two headways, far below the resolution of most traffic data collected today.

Although the individual events are below the resolution of commonly available detector data, the aggregate impacts are not. To understand many of the different empirically observed FD shapes, it is beneficial to extend this averaging rule to the microscopic scale, below the resolution of commonly available detector data. Consider average state observed at the three different detector stations, numbered 1-3, in Fig 4a over the single sample period indicated by dashed lines. The state diagram shows the freeway initially at its RCap, state C, from which there is nowhere to go but lower flow. Station 1 sees a transient disturbance to free flow state F from an exiting vehicle, station 3 sees a transient disturbance to congested state D from an entering vehicle, while station 2 sees one of each type of disturbance. Fig 4b shows the resulting averages on the flow density plane, where the number corresponds to the respective station. In all three cases the flow is pulled lower than state C because multiple stationary traffic states are averaged together. Note that in this example the average from station 2 has a lower flow than that from stations 1 and 3 simply because station 2 includes the impacts of two maneuvers while the other stations only include the impact of one maneuver. In practice a sample might include the impact of dozens of maneuvers that occur over several miles. If one then includes region M from Fig 3, it is possible to also see transient states above the triangle, but even then, the speed will not exceed v_{ff}, so there are no forces that can pull the average state to the left of v_{ff}. Note that we deliberately did not specify whether we are observing one lane or all lanes in this example, since the general trend of the impacts would be the same in either case. Even though a lane change maneuver impacts two lanes it will only be observed in one lane at any given location since the two disturbances propagate in different directions.

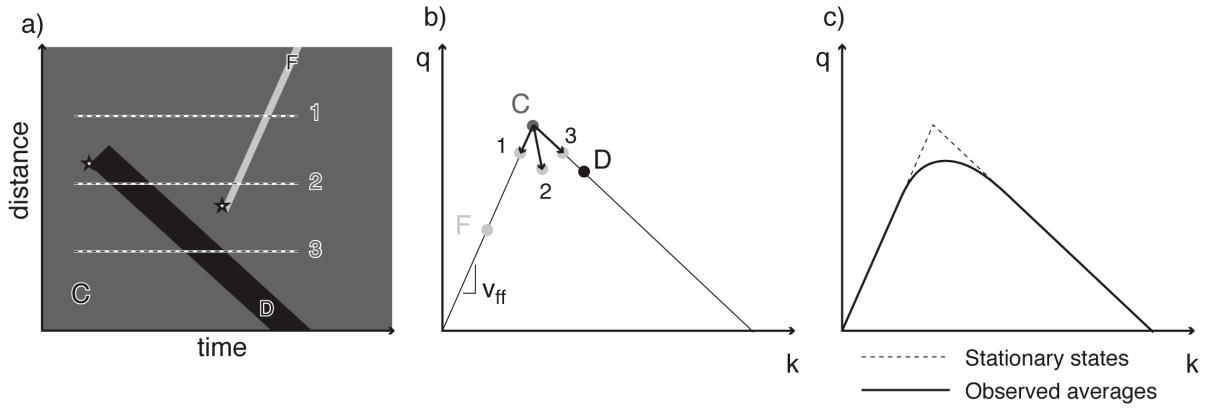

Figure 4, (a) State diagram illustrating how the average state at three detector stations, numbered 1-3, can be displaced in response to disturbances, (b) the net displacement for the examples in the flow-density plane, (c) if the rate of exiting and entering vehicles is correlated with flow, the net result will be that the observed average traffic states will not fall on the underlying, unobserved stationary states.

Although many researchers have stressed the importance of using stationary traffic states (e.g., Breiman and Lawrence, 1973; Cassidy, 1998), if the rate of entering and exiting vehicles is relatively stable over time and space, it is possible for the microscopic events to yield a nearly time stationary system where the traffic state changes slowly over space. The entering and exiting processes could result in a reproducible state diagram that is stable and with aggregate states that are very difficult to differentiate from true stationary, e.g., each of the averaged states in Fig 4b may erroneously be assumed to come from a distinct, stationary traffic state. Thus, making a freeway segment with an underlying triangular FD seemingly exhibit a parabolic FD, e.g., Fig 2b-c. The impacts are most pronounced when the underlying background state is close to the RCap, since the average is more likely to combine states from the free flow and congested regimes; in which case, the observed average traffic state will not fall on the underlying, unobserved set of stationary states. Obviously such distortions as shown in Fig 4b are not limited to state C, any underlying state near capacity could be impacted in a similar manner (e.g., if the detectors observe the disturbances from state A in Fig 3d the resulting average would have a flow greater than q_A and fall somewhere within the ACD triangle in Fig 3c). On the other hand, if the average only contains conditions strictly from the free flow regime or strictly from the congested regime, the average will fall on the triangular relationship (e.g., points 1 and 3 in Fig 4b). If the underlying entrance and exit rates are stable, they can yield a reproducible FD, e.g., Fig 4c. Resorting to the generalized definition of flow and density (e.g., Edie, 1965) would not solve this problem, since the time-space averaging still captures multiple stationary states. When the traffic state is near RCap, the greater the number of entering vehicles that impact the average (i.e., at the lane drop), the further the average will be from the unobserved straight-line free flow regime. As one moves downstream, fewer and fewer entering vehicles impact the average and it moves closer to the straight-line free flow regime. At the same time one moves downstream of more and more exiting vehicles, which pull the average towards the origin, but the resulting average remains along the same straight-line of the free flow regime. This mechanism can account for the change in shape of the fundamental diagram from station 5 to station 7 in the empirical study (Fig 2f). Although quantifying the frequency of lane change maneuvers and modeling the process in detail is left to future research, we do know that the net flow increases from station 4 to station 7 and then drops at station 9 due to the ramp flows (e.g., Table 1). Finally, recall that the maneuvers are on a lane-by-lane basis, so a pair of disturbances (e.g., Fig 3f-g) arises every time a vehicle changes lanes.

4. Discussion and Conclusions

Almost all hydrodynamic traffic flow models rely on the FD between flow and concentration either directly or indirectly, yet the precise relationship remains difficult to capture empirically and debate continues on its shape. This paper examined a lane drop that initially appears to be a single point bottleneck. Superficially all of the downstream detector stations appeared to exhibit free flow conditions while the bottleneck was active. But upon

closer inspection two anomalies became evident in Fig 2 and Table 1, first, close to the lane drop the FD appears to come from a parabolic curve (stations 5 and 6), but further downstream it appears to come from a triangular curve (stations 7 and 9). The simple fact that the shape of the FD curve changes over distance suggests that the bottleneck mechanism does not occur at a discrete point. Second, the observed throughput continued to increase as one travels downstream from station 5 to station 7 even though the number of through lanes remains constant and the RCap is relatively stable. These trends suggest that the flow remains restricted immediately downstream of the lane drop.

The theoretical development showed that the segment downstream of an APB is unique. When the bottleneck is active, downstream conditions will be close to RCap. As a result, any vehicle entering a lane can only do so if all vehicles in the entered lane upstream of the maneuver are delayed by one headway; which in turn triggers a disturbance that propagates upstream through the APB, originating at a point downstream of the APB, something that should not happen if the bottleneck mechanism strictly occurs at a single point in space. Meanwhile, the discharging queue ensures that the traffic is near the true free speed. As a result, the benefits of a large headway behind any exiting vehicle (whether due to a lane change maneuver or an off-ramp) propagate downstream with the surrounding traffic. *In aggregate, the costs of entering vehicles propagate upstream and reduce the throughput of the apparent-point-bottleneck while the benefits of exiting vehicles propagate downstream and are lost*. The microscopic disturbances are too small to be resolved in conventional data aggregation (see, e.g., Coifman and Wang, 2005). But as shown above, their presence provides an explanation for the two anomalies.

Within this theory, the lane drop dominates the bottleneck process but throughput is being limited at many different points. From the empirical data in Table 1, station 7 exhibits a higher throughput than stations 5 and 6 in part because of the vehicles entering the freeway from the on-ramps consume some of the RCap downstream of the lane drop. Similarly, whenever a vehicle changes lanes it too consumes some of the RCap in the entered lane without releasing capacity for the upstream vehicles in the exited lane. In this context one might be tempted to simply take the downstream-most entry point as the singular bottleneck, but since the points of entry are random, such an approach would jump all over. Furthermore, there is no evidence of any standing queues downstream of station 4 in the empirical data. The multiple successive, microscopic flow limiting features offer a potential explanation for the empirically observed anomalies because they can give rise to effects that none of the given flow limiting features could individually.

If the disturbances arise from on-going, stable processes (e.g., inflow from an on-ramp), it may be difficult to identify the original events or their impacts in the first place. The delays upstream of entrances will pull the empirically observed FD away from triangular and towards parabolic (as per averaging microscopic states via Fig 4). The impacts will be greatest close to the lane drop and diminish downstream because the number of downstream entrances progressively decreases as one travels further from the lane drop. This mechanism can account for the empirically observed change in shape of the FD from station 5 to station 7 (Fig 2f). Far enough downstream of the lane drop, we see triangular FD curves with flows higher than the maximum throughput of the APB. The theory developed herein could explain some paradoxes in the literature, e.g., the unexpected scatter in Hurdle and Datta, (1983), the different FD shapes seen by Hall and Gunter (1986) in different lanes, the slow discharge acceleration observed by Persaud and Hurdle, (1988), or the fact that speeds dropped by more than 50% below free speed a half km downstream of the APB when it was active in Bertini and Leal (2005).

Although we studied a lane drop, the present work is applicable to any bottleneck where the RCap remains roughly constant for an extended distance beyond the initial drop in RCap or increase in demand, e.g., a similar near-RCap situation can arise at an on-ramp bottleneck. So the results herein will likely apply to corridors with one or more on-ramp bottlenecks if the number of lanes remains constant.

References

Banks, J.H. (1989) Freeway speed-flow-concentration relationships: more evidence and interpretations, *Transportation Research Record, No 1225*, pp. 53-60.

Banks, J.H., (1990) Flow processes at a freeway bottleneck. *Transportation Research Record, No 1278*, pp. 20-28.

Banks, J.H., (2002) Review of Empirical Research on Congested Freeway Flow, *Transportation Research Record, No 1802*, pp. 225-232.

Bertini, R.L., Leal, M.T., (2005) Empirical study of traffic features at a freeway lane drop. *Journal of Transportation Engineering*, Vol. 131, No. 6, pp. 397-407.

Breiman, L., Lawrence, R.L., (1973) Time scales, fluctuations and constant flow periods in uni-directional traffic, *Transportation Research*, Vol. 7, No. 1, pp. 77-105.

Cassidy, M.J., (1998) Bivariate relations in nearly stationary highway traffic, *Transportation Research Part B: Methodological*, Vol. 32, No. 1, pp. 49-59.

Cassidy, M.J., Bertini, R.L. (1999) Some Traffic Features at Freeway Bottlenecks, *Transportation Research Part B: Methodological*, Vol. 33, No. 1, pp. 25-42.

Chung, K., Rudjanakanoknad, J., Cassidy, M.J., (2007) Relation Between Traffic Density and Capacity Drop at Three Freeway Bottlenecks, *Transportation Research.Part B, Methodological*, Vol. 41, No. 1, pp. 82-95.

Coifman, B. (1996) A New Methodology for Smoothing Freeway Loop Detector Data: an Introduction to Digital Filtering, *Transportation Research Record, No. 1554*, pp. 142-152.

Coifman, B., Krishnamurthy, S., Wang, X., (2003) Lane Change Maneuvers Consuming Freeway Capacity, *Proc. of the Traffic and Granular Flow 2003 Conference*, October 3, Delft, Netherlands, pp 3-14.

Coifman, B., Wang, Y. (2005) Average Velocity of Waves Propagating Through Congested Freeway Traffic, *Proc. of The 16th International Symposium on Transportation and Traffic Theory*, July 19-21, College Park, MD. pp 165-179.

Coifman, B., (2006) *The Columbus Metropolitan Freeway Management System (CMFMS) Effectiveness Study: Part 2 - The After Study*, Ohio Department of Transportation.

Coifman, B., Mishalani, R., Wang, C., Krishnamurthy, S., (2006) Impact of lane-change maneuvers on congested freeway segment delays. *Transportation Research Record, No. 1965*, pp. 152-159.

Del Castillo, J.M. (2010) Two New Models for Flow-Density Relationship, *Proc. of the 89th Annual Meeting of the Transportation Research Board*, paper 10-1093.

Drake, J.S., Schofer, J.L., May, A.D., (1967) A Statistical Analysis of Speed Density Hypotheses, *Highway Research Record, No. 154*, pp 53-87.

Duret, A., Bouffier, J., Buisson, C. (2010) Onset of Congestion due to Low Speed Merging Maneuvers within a Free-Flow Traffic Stream: Analytical Solution, *Proc. of the 89th Annual Meeting of the Transportation Research Board*, paper 10-2070.

Edie, L.C. (1965) Discussion of traffic stream measurements and definitions, *Proc. of the International Symposium On the Theory of Traffic Flow*, Paris, OECD, pp. 139-154.

Goodwin, B.C., Lawrence, R.L. (1972) Investigation of Lane Drops, *Highway Research Record, No. 388*, pp. 45-61.

Greenshields, B.D., (1935) A Study of Traffic Capacity, Proc. of the Highway Research Board, Vol 14, pp. 448-477.

Hall, F.L., Gunter M.A. (1986) Further Analysis of the Flow-Concentration Relationship, *Transportation Research Record, No. 1091*, pp. 1-9.

Hall, F.L., Allen B.L., Gunter M.A. (1986) Empirical Analysis of Freeway Flow-Density Relationships, *Transportation Research Part A: Policy and Practice*, Vol 20, No 3, pp. 197-210.

Hall, F.L., Hurdle, V.F., Banks, J.H. (1992) Synthesis of Recent Work on The Nature of Speed-Flow and Flow-Occupancy (or Density) Relationships on Freeways, *Transportation Research Record, No. 1365*, pp. 12-18.

Hsu, P, Banks, J.H., (1993) Effects of Location on Congested-Regime Flow-Concentration Relationships for Freeways, *Transportation Research Record, No 1398*, pp. 17-23.

Hurdle, V.F., Datta, P.K. (1983) Speeds and Flows on an Urban Freeway: Some Measurements and a Hypothesis, *Transportation Research Record, No. 905*, pp. 127-137.

Jin, J., Ran, B., (2009) Automatic Freeway Incident Detection Based on Fundamental Diagrams of Traffic Flow, *Transportation Research Record, No. 2099*, pp. 65-75.

Koohong, C., Cassidy, M.J. (2004) "Test of theory of driver behavior on homogeneous freeways," *Transportation Research Record, No. 1883*, pp 14-20.

Laval, J.A., Daganzo, C.F. (2006) Lane-Changing in Traffic Streams, *Transportation Research-Part B, Methodological*, Vol. 40, No. 3, pp. 251-264.

Laval, J.A. (2006) Stochastic Processes of Moving Bottlenecks: Approximate Formulas for Highway Capacity, *Transportation Research Record, No. 1988*, pp. 86-91.

Laval, J.A. (2009) Effects of Geometric Design on Freeway Capacity: Impacts of Truck Lane Restrictions, *Transportation Research-Part B, Methodological*, Vol. 43, No. 6, pp. 720-728.

Laval, J.A., Leclercq, L. (2010) Continuum Approximation for Congestion Dynamics Along Freeway Corridors, *Transportation Science*, Vol. 44, No. 1, pp. 87-97.

Lee, H., Coifman, B., (in-review-A) An Algorithm to Identify Chronic Splashover Errors at Freeway Loop Detectors Submitted for publication in *Transportation Research-Part C*.

Lee, H., Coifman, B., (in-review-B) Quantifying Loop Detector Sensitivity and Correcting Detection Problems on Freeways Submitted for publication in *Journal of Transportation Engineering*.

Lee, H., Coifman, B., (in-review-C) Identifying and Correcting Pulse-Breakup Errors from Freeway Loop Detectors Accepted for publication in *Transportation Research Record*.

Lighthill, M., Whitham, G., (1955) On Kinematic Waves II. a Theory Of Traffic Flow on Long Crowded Roads, *Proc. Royal Society of London, Part A*, Vol. 229, No. 1178, pp 317-345.

Munjal, P.K., Hsu, Y.S., Lawrence, R.L. (1971) Analysis and Validation of Lane-Drop Effects on Multi-Lane Freeways, *Transportation Research*, Vol. 5, No. 4, pp. 257-266.

Ogut, K.S., Banks, J.H., (2005) Stability of Freeway Bottleneck Flow Phenomena, *Transportation Research Record, No 1934*, pp. 108-115.

Persaud, B., Hurdle V.F. (1988). Some New Data That Challenge Some Old Ideas About Speed-Flow Relationships. *Transportation Research Record, No. 1194*, pp. 191-198.

Richards, P. (1956) Shock Waves on the Highway, *Operations Research*, Vol. 4, No. 1, pp 42-51.

Transportation Research Board (2000) Chapter 7, Traffic Flow Parameters, *Highway Capacity Manual*, Transportation Research Board.
Wang, C., Coifman, B., (2008) The Effect of Lane Change Maneuvers on a Simplified Car-following Theory, *IEEE Transactions on Intelligent Transportation Systems*, Vol 9, No 3, pp 523-535.
Zhang, H.M., Shen, W., (2009) Numerical Investigation of Stop-and-Go Traffic Patterns Upstream of Freeway Lane Drop, *Transportation Research Record, No. 2124*, pp 3-17.

Appendix A. Repeating the Flow-Occupancy Analysis by Lane

Fig A1 repeats the analysis of Fig 2, only now broken down by lane. This figure shows the curvature in all lanes at stations 5 and 6. Lane 3 at station 6 and lane 1 at station 7 deviate somewhat from the aggregate trend at the respective station, suggesting that the straightening occurs closest to the APB in the outside lane and only later in the inside lane, which is consistent with vehicles entering from the ramps and later merging to the inside lane.

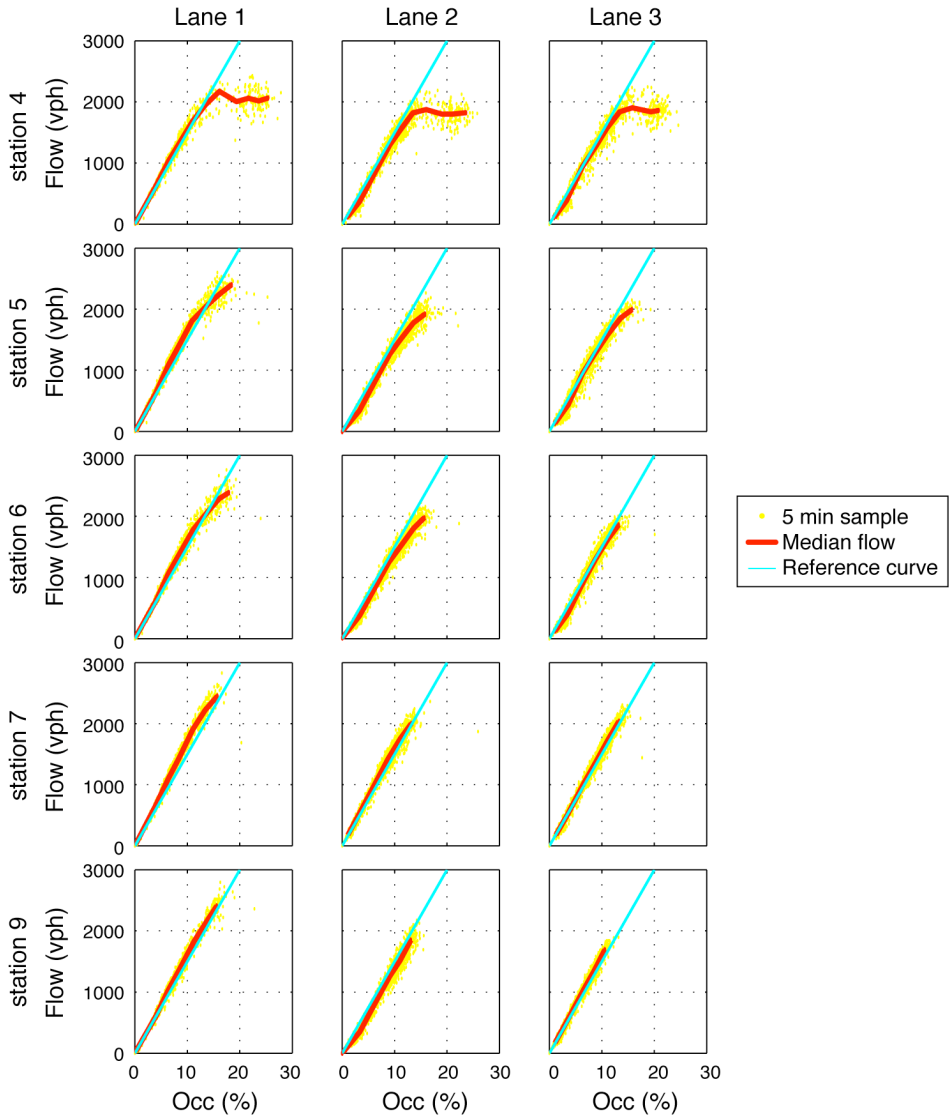

Figure A1, Repeating the analysis of Fig 2 by lane, flow versus occupancy, 5 min samples, over 6 weekdays, one row per station and one column per lane. Each plot includes the median flow curve (by occupancy bin) and for comparison, a common reference curve.