5.3: Queueing and Traffic Flow

Queueing and Traffic Flow is the study of traffic flow behaviors surrounding queueing events. It combines elements from the previous two sections, Queueing and Traffic Flow.

Queueing and Traffic Flow

The first side figure illustrates a traffic bottleneck that drops the roadway from two lanes to one. It allows us to illustrate the changes in capacity, the changes in lane flow, and the stability in the total section flow. Over an extended period of time, by laws of conservation, flow through the bottleneck (q) must equal flow through the upstream section (Q).

\[
\sum_t q = \sum_t Q
\]

A bottleneck causes lane flow (Q_l) to drop, but not section flow (Q), where:

\[
Q = \sum_t Q_l
\]

When researchers observe a “backward-bending” flow-travel time curve, (which we will tackle later in the section) this is occasionally what they are seeing.

\[
\sum_t q \geq \sum_t Q_l
\]

An illustrative plan of a highway bottleneck

Traffic phases in the queueing cumulative input-output (Newell) diagram

The lane flow drops upstream of a bottleneck when demand for a bottleneck exceeds the capacity of the bottleneck. Thus, the same level of flow can be observed at two different speeds on an upstream lane. The first is at freeflow speed.
in uncongested conditions and the second is at a lower speed when the downstream bottleneck is at capacity. In other words, congestion (a queue) forms when Q > \(q_{\text{max}}\) for any period of time. Ultimately, what enters the queue must eventually exit; otherwise the queue will grow to infinity. This is illustrated with the queueing input-output (IO) diagram shown in the second side figure. We can identify four distinct phases in traffic.

- Phase 1 is the uncongested phase when there is no influence of the increasing density on the speeds of the vehicles. The speed does not drop with the introduction of newer vehicles onto the freeway.
- Phase 2 finds the freeway cannot sustain the speed with injection of newer vehicles into the traffic stream. The density increases while speed falls, maintaining the flow.
- Phase 3 shows decreased speed and decreased flows. This is caused by very low speeds cause the queue discharge to drop (slightly) at an active bottleneck, or a queue from a downstream bottleneck may be constraining the flow. As reported by many researchers (Banks, 1991; Hall and Agyemang-Duah, 1991; Persaud and Hurdle, 1991; Cassidy and Bertini, 1999), the reduced bottleneck capacity after breakdown ranges from 0% to 8%. A majority of papers conclude that the section capacity reduction a queue presents is negligible. (As noted above, lane flow upstream of a bottleneck does drop, but this is due to the downstream bottleneck). This is a second source of the “backward-bending” phenomenon.
- Phase 4 is the recovery phase. During this phase, the density of traffic starts decreasing and speed starts increasing.

Traffic phases in a the microscopic fundamental diagram (truncated triangular)

The circled region, ‘A’ (in the third side figure) is typically where we start observing ‘freeway breakdown’. In other words, this region occurs when upstream flow exceeds some critical downstream capacity at a specific point and there is a drop in speed. In the circled region, ‘B’, flow drops are associated with very low speeds. The phenomenon is particularly evidenced by the formation of queues upstream of where the breakdown occurs and a low discharge rate of vehicles due to sustained low speeds.

Traditional queues have “servers”. For example, the check-out line at the grocery store can serve one customer every 5 minutes, or one item every 10 seconds, etc. A conveyor belt can serve so many packages per hour. Capacity is often referred to as belonging to the road. However when we talk about road capacity, it is really a misnomer, as capacity is located in the driver, more precisely in the driver’s willingness and ability to drive behind the driver ahead. If drivers were willing and able to drive behind the vehicle ahead with no gap (spacing between vehicles), and that driver was driving behind the driver ahead of him with no gap, and so on, at high speeds, many more vehicles per hour could use the road. However, while some compression of vehicles occurs in heavy traffic, this situation is unstable because a driver will tap the brakes or even let-up on the accelerator for any number of reasons (to change lanes, to respond to someone else trying to change lanes, because he sees an object in the road, to limit the forces when rounding a corner, etc.). This by definition lowers his speed, which in turn will lower the flow. Risk-averse drivers behind him will slow down even more (the driver has established some unpredictability in behavior, it is reasonable for other drivers to establish an even larger gap to accommodate the unpredictable driver’s behavior, especially given there is a reaction time between the lead driver’s actions, and the following driver’s perception (lead vehicle is slowing), decision (must brake), action (tap the brakes), the vehicle’s response to the action (tighten brakes on wheel)). In this way, maximum flow possible, our capacity \(\langle q_{\text{max}} \rangle\), is a function of the drivers. The road shapes the driver’s willingness to take risks. Drivers will slow down around curves, vehicles may have difficulty accelerating uphill, or even from a slower speed (and even if they don’t, driver’s may provide insufficient fuel before they realize they are going too slow by not giving the vehicle enough gas), merges take time to avoid collision, etc.
Clearly, different drivers and different vehicles (for instance racecars or taxis) could increase the maximum flow through the bottleneck \( (q_{\text{max}}) \). It is better to think of capacity as a maximum sustainable flow (over an extended period of time), given typical drivers’ willingness to follow (subject to highway geometrics and environmental conditions) and their vehicles’ ability to respond to decisions. A series of aggressive drivers may exceed this ‘capacity’ for a short period of time, but eventually more cautious drivers will even out the function.

The IO diagram lets us understand delay in a way that the fundamental diagrams of traffic flow don’t easily allow. The first point to note about the IO diagram is that delay differs for each driver. The average delay can be measured easily (the total area in the triangle is the total delay, the average delay is just that triangle divided by the number of vehicles). The variation (or standard deviation) can also be measured with some more statistics. As the total number of vehicles increases, the average delay increases.

The second point to note about the IO diagram is that the total number of queued vehicles (the length of the queue) can also be easily measured. This also changes continuously; the length of the queue rises and falls with changes in the arrival rate (the tail of the queue edge of the shockwave, where travel speed changes suddenly). Shockwaves indicate a change in state, or speed that suddenly occurs. One such shockwave is found where vehicles reach the back of queues. The arrival rate, coupled with the queue clearance rate, will tell you where the back of the queue is. This is interesting and is where the traveler first suffers delay – but it is not the source of the delay.

Only if the arrival rate exactly equals the departure rate would we expect to see a fixed length of the queue. If the queue results from a management practice (such as ramp metering), we can control the departure rate to match the arrival rate, and ensure that the queue remains on the ramp and doesn’t spill over to neighboring arterials. However that observation suggests that congested “steady state” is likely to be a rare phenomenon, since in general the arrival rate does not equal the service rate.

Queueing analysts often make a simplifying assumption that vehicles stack vertically (queues take place at a point). This is of course wrong, but not too wrong. The resulting travel time is almost the same as if the queue were measured over space. The difference is that the time required to cover distance is included when we make the better assumption, even under freeflow conditions it takes time to travel from the point where a vehicle entered the back of the queue to the point where it exits the front. We can make that correction, but when queueing is taking place, that time is often small compared to the time delayed by the queue. We also assume a first-in, first-out logic, though again this can be relaxed without distracting from the main point. Another assumption we will make for exposition is that this is a deterministic process; vehicles arrive in a regular fashion and depart in a regular fashion. However sometimes vehicles bunch up (drivers are not uniform), which leads to stochastic arrivals and departures. This stochastic queueing can be introduced, and will in general increase the measured delay.

**Hypercongestion**

It has been observed that the same flow can be achieved on many links at two different speeds. Some call this the “backward-bending” phenomenon (Hau 1998, Crozet and Marlot 2001). The queueing analysis framework also has implications for “hypercongestion” or “backward-bending” flow-travel time curves, such as shown in the figure to the right. Recall we identified two sources for “backward-bending” speed-flow relationships. The first has to do with the point of observation. Observing the lane flow upstream of a bottleneck gives the impression of a backward bending
relationship, but this disappears at the bottleneck itself. Under any given demand pattern, flow and speed are a unique pair. When demand is below the downstream active bottleneck’s capacity, a flow on an upstream link can be achieved at high speed. When demand is above the downstream active bottleneck’s capacity, the same flow on the upstream link can only be achieved at a low speed because of queueing. The second has to do with a capacity drop at the bottleneck itself under congested conditions. However, much research reports that this drop is slight to non-existent.

As in the bottleneck, we define lowercase $q$ to be flow (vehicles per hour) departing the front of the bottleneck and uppercase $Q$ to be flow arriving at the back of the bottleneck. We also define $k$ to be density (vehicles per kilometer), $v$ to be speed (kilometers per hour), and $s$ to be service rate (seconds per vehicle). The fundamental diagrams of traffic flow ($q$-$k$-$v$ curves) represent a model of traffic flow as stylized in the traditional textbook representation of the fundamental diagram of traffic flow. The reasons why $q$ should drop as $k$ increases beyond a certain point at an isolated bottleneck are unclear. In other words, why should flow past a point drop just because the number of vehicles behind that point increases? Why should leading traffic be influenced by the behavior of following traffic?

If traffic behaves as a queue through a bottleneck (illustrated above), we should consider reasons why traffic flow departing the queue would not stay at its maximum.

One reason is if the vehicles in the queue could not travel fast enough so that the front of the following car could not reach the point of the front of the leading car in the time allotted the service rate. If a lane serves 1800 vehicles per hour, it serves 1 vehicle every 2 seconds, we say there is a 2 second service rate. If there were a 10 meter spacing between the vehicles (including the vehicle length plus a physical gap), this implies that the service rate would be worse than 1 vehicle every 2 seconds only if it took longer than 2 seconds to travel 10 meters (i.e. speed < 18 km/hour). This is very slow traffic and may properly be called hypercongestion. Why would traffic get that slow if freeflow speed is 120 km/hour? In general traffic departing the front of the queue won’t be traveling that slowly, as the shockwave that reduces speed (the wave of red brake lights) has moved to the back of the queue.

A second reason the bottleneck flow might drop is if the departure flow is affected by external sources, namely a bottleneck further downstream spilling backwards. If the queue at a downstream bottleneck becomes sufficiently long, it will reduce the number of vehicles that can depart the upstream bottleneck. This is because the first bottleneck is no longer controlling, the downstream bottleneck is.

A third reason that bottleneck flow might drop is if the bottleneck is not being fully served (slots for cars are going unfilled). If vehicles are separated by large time headways, then the bottleneck might lose capacity. Thus, if vehicles choose large gaps, the bottleneck flow might drop. However, in congested situations, vehicles tend to follow more closely, not less closely.

For an isolated bottleneck, the departure flow remains (essentially) a constant and arrival flow varies. As more vehicles arrive at the back of the queue, the expected wait at the queue increases. All vehicles will eventually be served. In general, there is no practical constraint on the number of vehicles arriving at the back of the queue, but there is a maximum output flow in vehicles per unit time. Examining traffic upstream of the bottleneck is interesting, but does not get to the root of the problem – the bottleneck itself. This view of hypercongestion is thus not inconsistent with the conclusion drawn by Small and Chu (1997) that the hypercongested region is unsuitable for use as a supply curve in congestion pricing analyses.

BackwardBending.png
Thought Question

Problem

Bottlenecks are shown here to be the main cause of traffic flow congestion and queueing. The examples given are of lane drops, which are not very common (some drivers think they're more common than they actually are). Is this the only way a bottleneck can form?

Solution

Absolutely not. A bottleneck is defined as a constriction of traffic flow, where demand exceeds available capacity. Infrastructure-wise, a bottleneck would be where a lane drop is present, where lane widths decrease, or a sight obstruction lowers the natural free-flow speed. However, bottlenecks can also form because of traffic. Chaotic traffic, primarily where lane changing is occurring, is a prime source of a bottleneck. On-ramps, off-ramps, and weaving areas are the most common examples, but there are many more. Think about how often you have been stuck in standstill traffic where a lane drop is not present.

Additional Questions

1. Why were ideal models of flow vs. density assumed for so long if data did not support it?
2. How do we estimate density of road (in planning or as an after-evaluation?)
3. Explain why the graph of density vs. speed at a bottleneck is not linear using queueing theory.
4. Define time headway and space headway.
5. Why do we need two different speed measures, define each, which is more appropriate? How are these measures related?
6. What are the 3 fundamental principles of traffic flow? What are their relationships?
7. What does the slope on the space time diagram tell you?
8. What is the difference between looking at traffic as a particle rather than a flow?
10. As density increases, what happens to flow?
11. Where are time-space diagrams used?
12. What is space headway on a space-time diagram? What is time headway?
13. What does high density mean?
14. Draw a realistic speed vs. density graph? Compare it with the traditional model.
15. Describe headway speed?
16. How does a bottleneck affect traffic flow upstream of the point?
17. What might cause a falloff of flow as density rises?
18. Why does density reach a maximum when flow is zero?
19. How do these analysis of mathematical relationships between q,k,v help in planning, design, and operations of road?
Variables

- \(q\) - Flow through the bottleneck
- \(Q\) - Flow through the upstream section
- \(Q_l\) - Lane Flow
- \(t\) - time

Key Terms

- Bottleneck
- Queueing
- Traffic Flow
- Lane Flow
- Lane Flow Drops
- Flow-Travel Time Curve
- "Backwards-Bending"
- IO
- Hypercongestion