Measuring traffic congestion

Maxwell G Lay, AM

Peer reviewed paper

This paper has been critically reviewed by at least two recognised experts in the relevant field.

Originally submitted: October 2010.

Abstract

This paper examines traffic congestion, particularly in light of the growing public emphasis on the cost of congestion. Traffic behaviour and management are discussed in the context of congestion and its various causes. The importance of queues is highlighted. The separate roles of intersections and links are explored. It is suggested that many fallacies and misconceptions surround congestion, and doubts are placed on current estimates of the cost of congestion. To overcome this situation, an alternative approach to congestion assessment is proposed with an emphasis on queue formation and catastrophe theory. The new approach provides the basis for a straightforward and implementable way of estimating the occurrence and cost of congestion. The author suggests that some current practices could distort road investment priorities.

INTRODUCTION

Traffic congestion has been a part of daily life since the earliest days of civilised society (IHR 2010; Lay 2010). Readers would all have experienced congestion. Therefore, it might be expected that the phenomenon is well-understood and well-defined. However, this is not the case and there are vastly different definitions in place, leading to a wide range of often questionable estimates of the onset of congestion, of its consequences, and – particularly in the context of the current debate on transport policy – of its financial and social costs. There is a clear need for some common ground.

Members of the general public might well be surprised by the need for a paper such as this. They would easily characterise congestion as waiting in stationary queues, often unpredictable and relatively lengthy travel times, queues of traffic frequently interfering with other traffic, problems entering a traffic stream, local 'traffic jams' stopping movement in all directions in a local area, and finally 'gridlock' when the whole road transport system comes to a standstill. Their one uncertainty might be whether a heavily used but moving system might be deemed to be 'congested.'

SOME DEFINITIONS

The word 'congestion' came to the English language from Latin and has consistently meant an accumulation or heaping. It will be shown that its application to traffic congestion is consistent with this original usage. According to the Oxford English Dictionary, the word congestion began to refer to overcrowded conditions in the 1860s and to traffic conditions in the 1890s. From that time forwards, the definition and measurement of congestion have been poorly understood. For example, Wikipedia (cited here to illustrate the confusion surrounding the term 'congestion', not as an authority) states that:

> *when traffic demand is great enough that the interaction between vehicles slows the speed of the traffic stream, congestion is incurred. As demand approaches the capacity of a road (or of the intersections along the road), extreme traffic congestion sets in.*

The first sentence omits all the congestion motorists experience at intersections and the second sentence gratuitously and unhelpfully offers the reader an undefined new term – 'extreme traffic congestion'. Perhaps even more fanciful and self-fulfilling was the definition offered by the German Minster of Transport to the ECMT (European Conference of Ministers of Transport), as follows:

> *Congestion can be defined as a situation in which transport participants cannot move in a usual or desirable manner. (ECMT 1998)*

At a more professional level, some commentators (e.g. Evans 1992; Boarnet, Kim & Parkany 1998) have defined congestion as a single event based on the ratio of the actual flow to a 'network capacity', implied to be a single defined flow level. With so many different origins and destinations, the idea of a single capacity measure is meaningless and misleading. Even if the terms in the ratio could be defined, it is difficult to see the relevance or usefulness of this ratio-based and very determinant definition, although a measure based on the summation of individual capacities has been shown to have some usefulness (Truong & Hensher 2007). Level of service measures also fall into this numerate but meaningless category (Section 17.4.2 in Lay (2009)).

In a similar vein, a report for the US Federal Highway Administration (FHWA), in a confusion of concepts, states that congestion –

> *relates to an excess of vehicles on a portion of roadway at a particular time resulting in speeds that are slower – sometimes much slower – than normal or 'free-flow' speeds. Congestion often means stopped or stop-and-go traffic. (Cambridge Systematics 2005)*

A common current approach to defining congestion describes it as occurring whenever travel times are greater than the minimum possible travel time. Many commentators have based this minimum time on the time taken by a solitary vehicle to traverse the system without stopping (Port Jackson Partners 2005). This is unrealistic even for a solitary vehicle, as human response times of a second or so, means that intersections with limited sight distance inherently impose real and de facto limits on travelling at typical urban speeds of the order of 20 m/s (Chapter 16 of Lay (2009)). Moreover, in a real city, travellers are obliged to stop at stop signs and red traffic signals, even in the absence of other travellers. The real residents of this real city would not accept that delays caused solely by intersection geometry and traffic control devices represented congestion.

A slightly more realistic approach would be to consider the above delays, but to still use free-flow conditions on the links as the base case. Again, public common sense would not accept that a well-used road is inherently a congested road. The use of such 'nirvana-like' base cases seriously compromises many current estimates of the cost of congestion.

The congestion definitions used by economists are commonly based on using traffic supply and demand curves to find a case that balances benefits and costs (see review in Naudé, Tsolakis and Anson (2005)). It is also common to suggest imposing extra charges to cover externalities and thus produce a rationally-based traffic condition. Although the term 'congestion' is frequently used in these studies, any definition of it is indirectly implied by the results, rather than used to formulate them. For example, Naudé et al. (2005:18) state that:

> *Congestion may be regarded as the point at which an additional road user joins the traffic flow and affects marginal cost in such a way that the marginal social cost of road use exceeds the marginal private cost of road use at the "optimal" level of congestion.*

This definition merely defines an economically optimum traffic flow. That flow need not correspond to any user perception or operator measure of congestion, and will depend on the relative costs assigned to particular activities. Naudé et al. (2005)

also conclude that 'congestion arises when road users are not given the correct signals regarding the true cost of their trips.' However, even if travellers are given correct pricing signals, the outcome might still include the time spent waiting in growing queues.

Given the various misgivings, this paper reviews congestion as an operational aspect of road traffic. It then offers a somewhat new congestion measure that puts aside many past uncertainties and misdirections. The new measure will be seen to depend heavily on the accumulation of vehicles and the subsequent development of queues within the traffic system. It will suggest that many cases labelled as congestion are merely examples of a busy road operating in accordance with design intent and where the traffic behaviour is predictable, repeatable and has an appearance of normality. It will be shown that a doubling of travel time over unimpeded conditions is the norm in an urban street system with traffic control devices, and is not an indication of congestion.

TRAFFIC BEHAVIOUR

To resolve issues with the measurement of congestion requires an understanding of some of the fundamental features of traffic. The road transport system can be perceived as a set of traffic origins and destinations connected together by a seemingly random network of roads.

This paper will use the terminology of 'lanes', 'links' and 'intersections' of links. A 'lane' is a single line of traffic and it is taken that most traffic in a modern road system moves in defined lanes. Furthermore, without loss of generality, the required behaviour and associated definitions can often be based on the simplifying assumption that the traffic in the lane consists of uniform vehicles behaving uniformly and without overtaking other vehicles on their journey from origin to destination. A 'link' is a set of parallel traffic lanes. Links carry traffic between intersections (or nodes) without external interruptions. Links and intersections will usually involve more than one lane of traffic, but it will often simplify the following discussion to concentrate on single lanes of traffic. Sequences of links form routes joining origins and destinations. These routes cross each other at 'intersections'.

The paper is presented assuming that the physical road and traffic system is operating as intended. Any resulting congestion is called 'recurrent', as its occurrence relates to the recurring (mainly diurnal) traffic patterns. 'Non-recurrent' congestion occurs when the system is degraded by an event, such as a road crash or a broken down vehicle. Typically, non-recurrent congestion may account for about a third of urban congestion. In recent years, it has been found by road managers that the effects of non-recurrent congestion can be greatly minimised by the use of dedicated response teams and practised procedures. The conclusions of this paper with respect to congestion definitions, measurement and costing will be seen not to depend on whether the congestion is recurrent or non-recurrent.

Links

The first difference between experts and the public is that many expert analyses of congestion have concentrated solely on the links, whereas the public would see most congestion beginning at intersections. The expert focus on links can perhaps be explained by the fact that link analysis is neater and tidier and produces results that would appear to be intellectually seductive.

It is certainly convenient to firstly examine the link. The behaviour of a line of traffic in a lane is one of car-following, except for the lead vehicle which follows its own, but relatively common, distance–speed–acceleration profile (Section 27.2 of Lay (2009)). Humans have no inherent speed–perception facility (Section 16.4.6 of Lay (2009)) and their key observation in car-following is an estimate of the separation distance between them and the vehicle ahead of them. This driver behaviour is represented in *Figure 1*, where even when stationary at G, drivers leave a gap between vehicles and this gap increases steadily as speed increases and the driver worries about braking in case the vehicle ahead begins to slow down. This regime in which the lead vehicle forces behaviour on the following vehicle is called 'forced flow'. When the vehicle headway[1] exceeds about 4 s, a driver's behaviour is finally unencumbered by any concerns about braking by the vehicle ahead (Chapter 17 of Lay (2009)). This is the 'free-flow' regime.

The form of *Figure 1* can be obtained empirically. It is also easy to produce algebraic forms by making simple assumptions about driver response times and braking performance (Section 17.2.3 of Lay (2009)). The first major attempt at this was by Greenshields (1935), who assumed a linear link between speed and traffic density, where density is the reciprocal of vehicle spacing.

Before becoming too involved in the next algebraic step, it is important to note that *Figure 1* shows a

[1] 'Headway' is the time-based separation of vehicles. 'Spacing' will refer to distance-based separation. Thus, (speed)(headway) = (spacing).

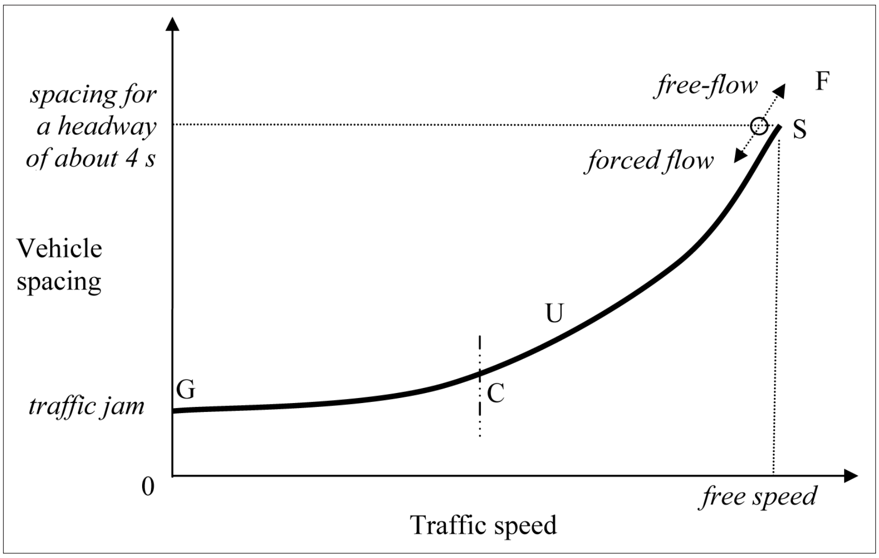


Figure 1
Behaviour of a driver following another vehicle in a traffic lane

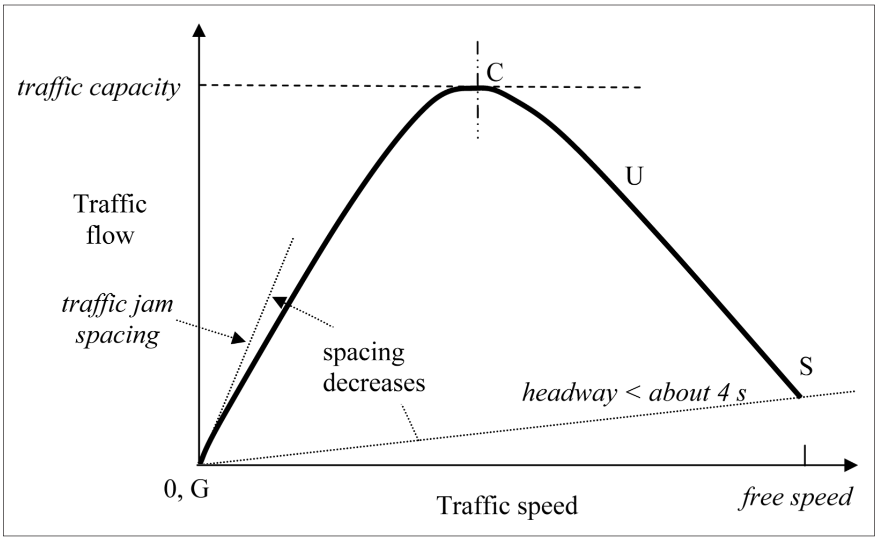


Figure 2
Effect of behaviour in Figure 1 on the flow of a lane of vehicles

deterministic system in the entire range, from stationary at G to free at S. There is a single value of speed for each spacing, and vice versa. The result does not depend on any prior events. As far as drivers on the link are concerned, their speed drops steadily as more drivers occupy the same length of the link – nothing else changes.

However, the traffic flow on the link (measured in vehicles per time unit) is a trade-off between the time disadvantages of going slower and the advantages of fitting more vehicles onto the same length of link at the slower speed. The traffic flow is calculated from the fundamental continuity condition that preserves the number of vehicles and requires that (Section 17.2.1 of Lay (2009)):

$$\text{flow} = \text{speed}/\text{spacing}$$

Using Greenshields' postulate leads to flow as a parabolic function of speed (*Figure 2*). To a close approximation, it predicts a maximum flow at half free speed. This maximum flow is called the capacity of the link, marked by C in *Figures 1* and *2*. The postulate is not realistic and whilst most uninterrupted flow data for motorways and arterials still shows a response of the same form as *Figure 2*, the graph is commonly very skewed and maximum flow occurs at about 80% of free-flow speed (VicRoads 2010, Section 2.2). Nevertheless, the system is still based on *Figure 1* and is determinant. Travellers in a lane would not know that they were at or near the capacity condition, C.

Despite this, many commentators (Tsolakis and Naudé 2006; Button 2004) describe the region CS as 'congestion' and CG as 'hyper-congestion' The 'hyper' term is due to Button, who does not explain its implications. Now, the free-flow speed of a lane depends on the 'traffic environment' of the road and will typically be between 60 and 120 km/h (Section

18.2.5 of Lay (2009)), so the capacity flows from Greenshields' assumption will be at speeds between 30 and 60 km/h. In fact, actual data suggests higher capacity speeds of about 80 km/h, depending on vehicle and driver characteristics (Section 17.4.1 of Lay (2009)). Few travellers would consider travel speeds of 50 km/h or better to represent congested conditions. Indeed, 80 km/h is quite a fast – and arguably adequate – speed in urban areas. Drivers in such traffic would feel they were on an efficient, well-used road and would have no cause to label it as 'congested'. Roads are designed assuming capacity flows (Chapter 17 of Lay (2009)); therefore, all the facilities should be able to accommodate these flows without inappropriate conditions becoming evident.

Whilst there is no justification for labelling the zone UC of *Figure 2* as 'congested', there is a reason why it has also been labelled as a 'stop-start' zone (Tsolakis & Naudé 2006; Button 2004). This arises from an external 'instability' aspect of lane behaviour. If drivers on a link are suddenly forced to increase their estimates of the safe spacing, even if they do no more than touch their brake pedal, their brake lights will send decelerating, and possibly braking, shock waves back through the following traffic. These responses are usually caused by 'unexpected' events, such as merging traffic or a vehicle braking prior to leaving a traffic lane. Thus, a small disturbance can cause a large effect, and such traffic is potentially unstable. This is a micro-example of catastrophe theory in traffic. The effects of such instability will be intermittent congestion. It can be kept to a minimum by avoiding disturbances with measures such as well-designed merging and exit lanes, ramp metering, and good advance signage. Nevertheless, the closer the headways that drivers adopt (and thus, the higher the traffic flow), the more likely it is that these flow instabilities will occur. Empirical data suggests that flows over 1800 veh/h are susceptible to such events (VicRoads 2010, Section 4.2).

When is the CG portion of *Figure 2* encountered? Many commentators (e.g. Tsolakis & Naudé 2006; Button 2004; Evans 1992; Boarnet et al. 1998) see *Figure 2* rotated through 90^o, and disconcertingly describe the region CG as 'backward-bending' and not part of normal traffic (refer also to *Figure 5*). Indeed, Evans (1992:219) wrote that there was 'no logical requirement for the lower branch of the parabola (GC) to exist' in a system. However, traffic moving off after being stopped in a queue (e.g. at a red traffic signal) follows the curve from G to C, and there is nothing in these traffic lines discharging from the queues that is either inherently congested or otherwise unusual. Drivers in such an expanding, accelerating line of traffic heading towards the local free speed or the tail of the next queue ahead would not see themselves as in some new form of congestion. Similarly, a line of drivers coming to a stop follows the curve from C to G, an event which often occurs, even in light traffic.

As the line of traffic on a link reaches C from either flow direction, capacity is reached. What if the flow into the link continues to increase above capacity? There are two main responses. At one extreme, the excess vehicles will not be able to force themselves into the link. Queues will form at the entrance to the link and, if the input flow is unaltered, these queues will increase steadily with time and cause obvious congestion at the entrance. At the other extreme, the excess vehicles will force themselves into the link, spacings will diminish and speeds will drop. More vehicles will be entering the system than are leaving it. Some will be stored in local queues. If the input flow does not alter, the link itself will finally come to a halt as one long queue. The link will be obviously congested.

An important exception to the simple static queue case can occur on a motorway where the ramps at an interchange cannot process all the traffic as it arrives or where some lanes are blocked. The traffic on the motorway may almost come to a standstill, but as some vehicles can still proceed (as opposed to the complete stop caused by traffic control devices), the motorway traffic may still proceed but at a much reduced speed. As capacity drops from C towards G (*Figure 1*), the slow-moving queue will grow in length.

Does such a slow-moving line of traffic fit into the above congestion scenario? As the traffic is moving, some might not consider it to be a queue, but does it represent congestion? The flow in the line will be less than the capacity flow, and it may therefore be a cause of upstream congestion if the upstream flow exceeds the slow-moving flow. Once the downstream constraint is removed, the model in *Figure 1* indicates that the slow-moving line will minimise its travel time by expanding from near G and increasing its speed until inhibited by some forward event or reaching capacity at C.

Traffic flow on a link clearly cannot be seen in isolation. However, Walters (1961) based his ground-breaking paper on congestion pricing on the behaviour in *Figure 2*, and Hills (1993; see also Hills & Gray 2000) pointed out that as a result of this decision 'a generation of economic analysts has been misled into using *traffic flow* as the common measure for relating supply and demand.'

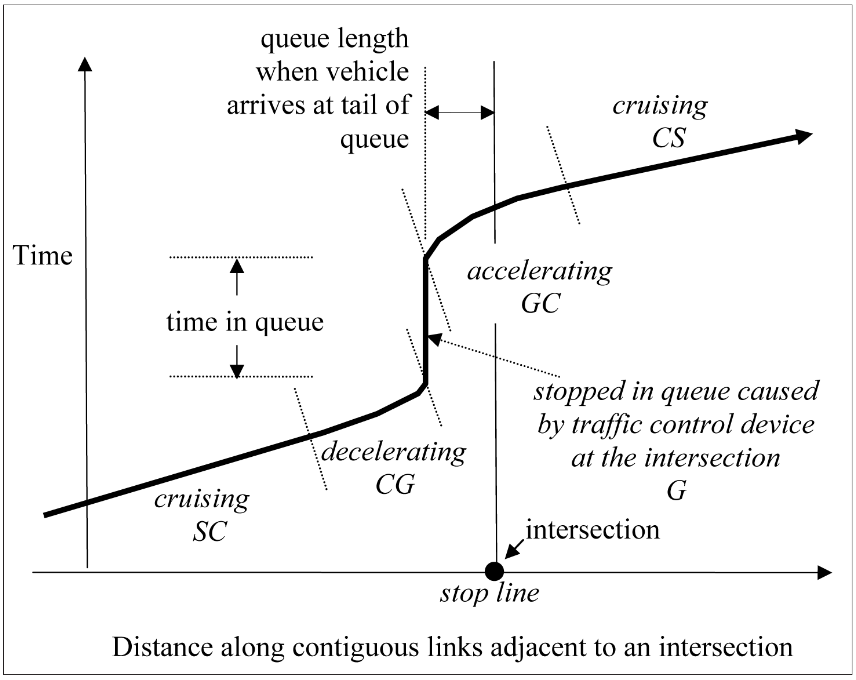


Figure 3
Movement of a vehicle through an intersection

Intersections

Most travellers would see congestion as related to intersections rather than to links, and there is more than a little truth in the view that a freeway is 'the shortest route between two traffic jams' (Meyer & Gomez-Ibanez 1981). This view is also shared by many traffic engineers 'working at the coal-face'. For example, the recent edition of the Australian traffic management guide (Austroads 2008) encourages traffic managers to be more relaxed about reducing the cross-section of arterial roads, stating:

> *Traffic throughput is maintained because, in urban networks, intersections tend to control system capacity. However, care needs to be taken to check for queue lengths developed back from intersections, particularly signals.*

To understand these views, consider a lane of arterial road traffic with a capacity flow of, say, 2000 veh/h arriving at an intersection. The traffic control device at the intersection must share the available time between all links using the intersection. It might, for example, give only half that available time to the arterial lane. In addition, the stopped arterial traffic will have to decelerate to a stop and later accelerate back to speed, and so effectively, the halved-capacity of the intersection lane will be reduced further. It might well be only 40% of the arterial lane capacity, or about 0.4 × 2000 = 800 veh/h. Furthermore, as the traffic flow increases, the intersection will be unable to discharge all of the queued traffic in one cycle of its signals. These residual queues will further diminish the effective capacity of the lane and are a sign of imminent congestion.

As the traffic builds further, these growing intersection queues will also begin influencing the behaviour of other upstream links and intersections. The role of the links in the system is increasingly diminished as the intersections become saturated with traffic. The growing queues are the most common form of urban congestion and lead to local blockages, commonly described as traffic jams. Of course, a traffic jam can also occur on a link, if an event such as a crash blocks that link.

Traffic flow through an intersection is shown in *Figure 3* for the case where the vehicle is only delayed for one cycle of the traffic signals (see Akçelik, Besley & Roper (1999) for a detailed study). Link behaviour is essentially unaltered, but the capacity is directly reduced by the queuing time. Thus, an intersection has four main effects:

- it reduces effective capacity
- it increases journey times
- the tail of the queues may influence the operation of upstream intersections, spreading the local traffic jam, and
- the upstream intersection may supply more vehicles than can be managed by the downstream intersection (e.g. due to inappropriate signal phasings).

Certainly, intersections cannot be excluded from any definition of congestion.

A similar situation occurs with ramp metering, although the queue grows in a more continuous manner. Experience suggests that drivers consider the situation congested if they need to wait more than about four minutes (Minnesota DoT 2010).

William Vickrey was awarded the Nobel Prize in economics in 1996. In 1967, he had used his personal observations of traffic at the end of New York's Lincoln Tunnel to develop new congestion theories. In the process, Vickrey (1967) perceptively called the congestion stage when queues began negatively influencing other intersections as a 'triggerneck', based on the view that a local traffic bottleneck was now triggering other wider events.

At an intersection with only passive right-of-way controls, only the traffic on the links without right-of-way could experience congestion (provided there were no downstream impedances). These vehicles would need to find acceptable crossing gaps in the priority traffic. An acceptable gap would typically be 6 s (Chapter 17 of Lay (2009)), and it was noted above that the headway of vehicles when free-flow begins is about 4 s. Hence, traffic on minor roads may be delayed, even when the overall traffic is not congested.

Travellers passing through a signalised intersection (i.e. one with active controls) would expect to be spasmodically stopped by a red signal – less so if on a route with signals favouring that route and/or with vehicle-actuation – and such a stop and the subsequent queue would not meet any reasonable definition of congestion. Even in light traffic, a vehicle might encounter a delay of about 60 s if it arrives at the end of a green phase, or an average of 30 s (Chapter 23 of Lay (2009)). For a vehicle travelling at a typical upper urban speed of 90 km/h (25 m/s) and a typical urban signal spacing of 800 m, the travel time of 800/25 = 32 s is of the same order as the typical signal delay. The point of these 'back-of-the-envelope' calculations is to suggest that a doubling of travel time (and a halving of average speeds) is the norm in an urban street system with traffic control devices, and is not necessarily an indication of congestion.

As the traffic flow increases, a stage is reached when the entire queue formed during the red phase is not discharged during the subsequent green phase. A residual queue will exist at the beginning of the next red phase. Many would say that such an intersection was congested, particularly if the residual queues lasted for more than one signal cycle. Thus, the beginning of intersection congestion is suggested by the incidence of queues lasting for at least a signal cycle, typically 120 s.

The residual queue condition will occur when the approach flow exceeds flow capacity of the intersection, given by the expression (Lay 2009):

$$[\text{green time}] \times [\text{effective saturation flow}]$$

or when the tail of a downstream or cross-traffic queue prevented vehicles from leaving the intersection smoothly. The latter instance would usually be a case of congestion elsewhere in the system spreading to the intersection being considered.

A key related case is that of a line of traffic moving off from a stop line after a traffic signal has changed from red to green. The lead vehicle will follow its own desired acceleration profile up to its preferred speed. For simplicity, assume it prefers a typical constant acceleration of 0.5 m/s^2 and preferred speed of 20 m/s (72 km/h). The lead vehicle will reach its preferred speed after 40 s and 400 m, and will have moved from G to C. The following vehicles will have spread out from their initial 4 m spacing to a spacing of 40 m when all have reached their preferred speed. Thus, the original queue spreads out as its component vehicles accelerate, and the tenth vehicle in the queue will scarcely have moved after 40 s. Many urban travellers will experience this back-of-queue frustration on a daily basis. This issue is pursued in some detail by Akçelik et al. (1999), who also suggest the existence of a 'saturation speed'.

This behaviour determines the 'saturation flow' of an intersection, which is the traffic flow once the green signal is presented and is independent of any upstream traffic flows or conditions. The total traffic discharged per signal cycle per movement is the product of the relevant saturation flow and green time. The relatively rapid queue build-up if this departure volume is less than the arrival volume has consistently been highlighted in traffic signal studies (e.g. Kimber & Hollis 1979).

The main difference from the link response in segment SUC is that the link flow is assumed to be a constant flow into the system with the input vehicles already at a particular speed, whereas in the departing queue scenario, the vehicles start from zero speed and at the smallest spacing.

For traffic approaching a stop, the lead vehicle will follow its desired deceleration profile and all the following vehicles will do the same, until overruled by the spacing requirement for its current speed. Each vehicle will then be at a different point in the GC curve, until all stop at G.

SYSTEM PERFORMANCE

In studying traffic system performance, it is important to distinguish between the demand to use a facility/system, and the internal performance

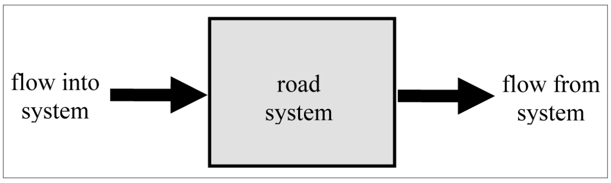


Figure 4
Road system

of the system. This distinction seems to have eluded many previous analysts. At the boundaries of the road network (*Figure 4*), or at trip origins, any growing queues may be observed and reacted to by intending travellers. In response, trip production might be reduced or delayed. Otherwise, vehicles will continue to enter the system and be accommodated within the growing queues. Trip time will increase and the number of unsatisfied vehicles within the system will increase.

A complication arises if only link behaviour is considered. The behaviour in *Figure 2* then applies, and some commentators have struggled with explaining behaviour when the input flows exceed link capacity. Consider the link as the system. The flows into the system may react to what is being observed within the system. Any excess flows either wait at the entrance, causing queues and time delays there, or force their way into the system, taking it from C towards G. This will force more vehicles out of the system and may produce an even worse outcome than in the first case, as there will be both excess new vehicles and a drop in processing capacity.

As traffic flow builds up from its daily low point, *Figure 2* indicates that travel speeds will steadily drop, and *Figure 3* indicates that intersection delay will gradually increase. Thus, travel times will increase steadily as the input flow increases. In most systems, the next stage would occur when some queues have not fully discharged in a traffic signal cycle. This would increase the rate of increase of travel time with input flow. Subsequently, some queues will impede the operation of upstream or cross-traffic intersections, and some links will reach capacity. The vehicles that cannot be accommodated in a link at capacity will either queue in the feeder link or divert to other, less attractive routes from their origin to destination. These events cannot be predicted from the simple link–flow model conventionally used. Furthermore, such a set of interactive events would markedly increase the rate of increase of travel time with input flow. The system would look and behave quite differently. Queues would be a noticeable feature of the system. They would store many of the new vehicles entering the system, but further diminish the service performance of the system.

When a road within a system reaches capacity, some past models have then assigned excess flows above ‘capacity’ to previously less attractive roads. However, in practice, Wardrop’s classic principles of route choice (Chapter 31 in Lay (2009)) have already been applied and traffic has been making individually rational decisions, whilst traffic flows have been increasing on the prime road. Not all travellers have the same characteristics and some will have diverted to alternative routes before capacity is reached on the prime route. Wardrop’s principles ensure that with recurring congestion, under-utilised roads in a busy system rarely, if ever, exist. Thus, surveys of the way pre-trip travellers use advice about the traffic conditions they might expect, is to alter travel times or to alert others of changed arrival times, and not to seek some unused lightly trafficked alternative, knowing that such a search will usually be futile (Karl 2003).

The need to stop thinking in terms of the abstract ‘flow’ on links was made strongly by Hills and Grey in 2000, but their message was largely ignored. Perceptively, they said:

> *Congestion will appear first at specific nodes on a network (usually towards its centre) and will spread out from there, in ways that are influenced not just by the general growth in demand over a peak period but also by those drivers who encounter congestion and adapt their behaviour in response to it. (Hills & Gray 2000)*

The debate over flow curve shape stems mainly from a confusion of purpose. If a supply curve is required, the behaviour in *Figure 2* needs to be recast. First, it is rotated through 90° (*Figure 5a*). Then, keep in mind Hills’ warnings (Hills 1993; Hills & Gray 2000) about how important units and dimensions are in these debates. Speed is replaced by cost/time and then cost/distance[2] to cover one vehicle in the system, and then multiplied by the flow to give total cost/km (*Figure 5b*). The general shape of the curve depends on fleet and driver characteristics (j_1 in Equation 29.18 of Lay (2009)) and the location of G depends on the fleet idling characteristics (j_{t1} in Equation 29.18 of Lay (2009)). The low value of S is due to there being only one or two vehicles in the system compared with about 2000 at C. At G, there are perhaps 2500 idling vehicles. There is nothing in this scenario to support most of the earlier supply curve propositions.

[2] See Equations 29.9 and 29.18 in Lay (2009).

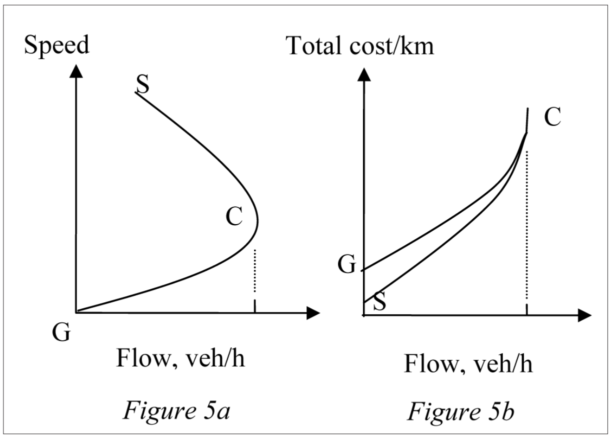


Figure 5
Supply function, derived from recasting Figure 2

In terms of chaos theory, determinant (predictable) behaviour ends when the first vehicle cannot be accommodated within the determinant traffic flow system. This is the onset of chaotic behaviour. Initially, the effect will be barely perceptible and chaos theory suggests that the effects will grow exponentially until the consequences are unpredictable (e.g. when the growing queues begin to block other intersections). This point is called the 'horizon of predictability'.

SOME 'NEW' CONGESTION MEASURES

The performance of a traffic system can be represented by *Figure 6*. It would be inappropriate to argue that congestion can be defined by looking at the shape of the curve in *Figure 6*, as this shape will depend on the units chosen for the two axes. Taking curve gradients merely reproduces the same curve. Nevertheless, a good definition of congestion does lie hidden within the curve but requires a further understanding what the curve represents.

In 1981, the Organisation for Economic Cooperation and Development (OECD) defined congestion as occurring when a small increment in additional traffic caused a major drop in service (OECD 1981). This could be thought to relate to the shape of the curve in *Figure 6*, but we will see below that it reflects a deeper understanding of congestion and is linked to the earlier discussion of how congestion actually occurs as a result of a marked changed in traffic behaviour. Regrettably, few have subsequently adopted the OECD's perceptive definition.

In recent times, catastrophe theory – first formalised by Lorenz in the 1960s – has provided new insights into system behaviour (Strogatz 1994). Consider a system in internal equilibrium, such as a line of balls running quickly down a dished channel. All is predictable, unless a small lateral disturbance causes a ball to curve sideways and over the side of the channel. This point at which there could be such a dramatic change of behaviour is known as a bifurcation point, with one path leading to a 'catastrophe'. It has already been noted that a disturbance within a traffic link could invoke a response explained by catastrophe theory. At the system level, any flows above capacity will grow into a bifurcation point. Indeed, for a link carrying traffic, *Figure 2* could be seen as a 'fold catastrophe'.

The link to catastrophe theory is even stronger than this, as the graphs in *Figure 2* and *Figure 5a* are forms of a logistic curve (see, for example, Equation 33.1 in Lay (2009)) used generally to explain population growth and later applied widely in catastrophe theory. In these 'population' terms, the traffic flow is proportional to the traffic density and to how close that density is to jam density. This leads to the 'classic' parabolic link between flow and speed.

What should be covered by a congestion measure in order that it would meet reasonable user

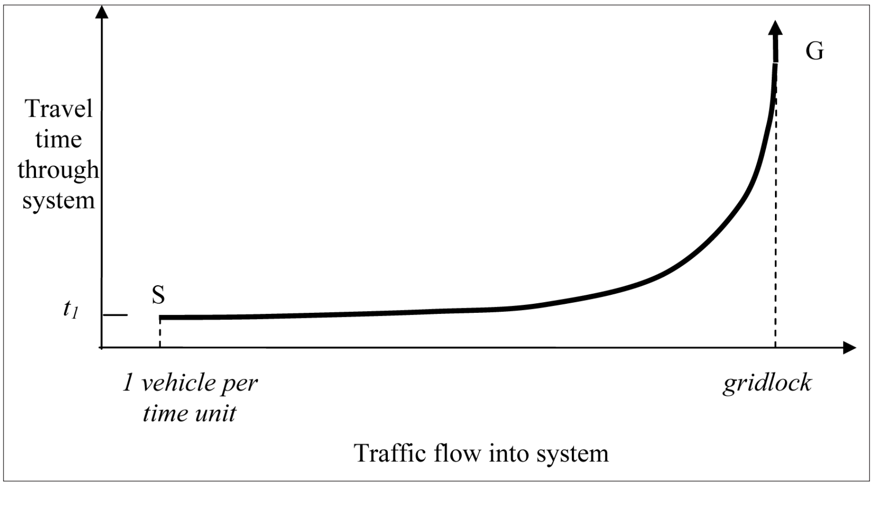


Figure 6
Basic performance parameters (the time unit should be larger than the travel time).

expectations and be acceptable technically? A useful analogy is with a river experiencing flood flows. As the flow increases, it will reach capacity for the river flowing between its normal banks. Any excess flow then overtops the banks at seemingly random points and is determined by factors that had no influence on the flow and capacity calculations for the preceding normal conditions. The subsequent flow in the flood plane also follows a new set of rules and directions. As with traffic congestion, the subsequent events are neither determinant, nor reversible or predictable from the factors used to estimate the initial flow conditions.

From the preceding sections, it is proposed that traffic congestion in a system occurs in five stages, as follows:

1. When the flows into the system are low, the travel times through the system are constant and unaffected by the flow level.
2. The flows increase and travel times now begin to gradually increase as the flow increases[3]. The behaviour is predictable, repeatable and has an appearance of normality.
3. As flows continue to increase, traffic begins to accumulate within the transport system and at an increasing rate. A small further increase in traffic flow causes a disproportionately large increase in travel times. Local traffic jams begin to occur.
4. In the fourth stage, the effects of the accumulating traffic are no longer localised, but begin to interact negatively with other parts of the system. In a traffic jam, there is still the possibility for some vehicles to leave the jam and move onto unjammed links. Thus, queues in traffic jams may be moving, albeit slowly.
5. Eventually the spreading effects of stage 4 bring the system to gridlock, and traffic is stationary.

The travel time increases in the third and fourth stages cannot be predicted solely from knowledge of the properties of the internal system. In technical terms, the behaviour has become unstable and chaotic. System response is not predictable solely from the original traffic flow considerations.

The accumulation of traffic in stages three and four occurs predominantly through the growth of stationary queues. The queues are the cause of both the reality and the perception of congestion. If there are no queues, there is no congestion. In urban areas, most congestion will be the result of queues at intersection stop lines. The link congestion, which has dominated most technical and economic discussions to date, will largely be confined to inter-city roads and some long radial city freeways, where the traffic comes mainly from smoothly operating entry ramps. In all these cases, queuing is easily observed and measured.

Thus, congestion can be easily measured at each intersection via routinely recorded queue lengths and times, and then aggregated across a system to give meaningful and timely cost estimates. These prime indicators of congestion can be taken directly from most traffic signal systems[4] or readily measured by aerial photography. Thus, it is conceptually easy to envisage a current traffic control system providing a real-time congestion measure. Timely measures of congestion would help to prioritise actions to help alleviate that congestion.

It must be emphasised that the queues in question are not the ‘normal’ queues that occur inevitably at a traffic signal red phase and are cleared during the next green phase. Congestion queues are ones that exceed these signal phase queues and accumulate with time. Measuring queues also avoids being misled by the spatial and temporal variability of congestion, which was one of Hills’ (1993) more trenchant criticisms of conventional processes.

Total daily queue lengths may reflect a growth in traffic. To manage this, a better measure might be to divide the queue length summation by the total number of trips made or the total trip time accumulated during the measurement period. Thus, an indicative measure might be queuing time as a proportion of trip time.

The cost of traffic congestion cannot be reliably measured or predicted by the first-order closed models used for most current estimates of congestion cost as their underlying assumptions are well-removed from reality. It is difficult to see the usefulness of most current estimates of congestion costs.[5] However, a relevant and useful measure of the cost of congestion flows directly from the queue length data. Once the number of hours of vehicle stoppage in queues (or delay in slow-moving motorway queues) is known, then the product of this total and the value of their time

[3] Such a slow-moving line of traffic could be classed as a congestion-related queue as long as the vehicle spacings remained at jam levels and were not increasing.

[4] This is certainly possible with the SCATS-based systems with which the author is most familiar. For data on SCATS, see Lay (2009, Chapter 23).

[5] The time spent by vehicles decelerating before and accelerating after stop lines can be conservatively neglected as that waste time is only avoided by vehicles that are not required to slow down at a traffic control device.

gives an estimate of the cost of congestion, which would be far more accessible, accurate, reliable and relevant than the estimates currently in use (Hills 1993; BTRE 2007; BTCE 1996).

Of course, there are many uncertainties about the value of time (Section 31.2.3 of Lay (2009)) and it is not feasible to resolve them at this stage of the current paper. It should be said that the two Federal Government reports, BTRE (2007) and BTCE (1996), represent an advance on most earlier work in that they do attempt to account for intersection delays and recognise their importance.

At a more general level, the suggested congestion measure and its associated costing would mean that road investments might be directed more to operating existing roads at capacity (e.g. with ramp metering), to ensuring the quick detection of and rapid response to 'incidents' which cause unexpected and unpredictable deterioration in traffic conditions, to the completion of effective traffic networks, and to avoiding short-term fixes that merely transfer congestion from one location to another.[6]

CONCLUSIONS

The paper has reviewed traffic congestion and the various measures and definitions associated with it. It has been shown that many of these measures contain major flaws that cast serious doubt on their application to such tasks a estimating the cost of congestion. By considering how a heavily used traffic system behaves, and with some insights from catastrophe theory, a 'new' measure of congestion is proposed based on the accumulation of traffic in a system, particularly as evidenced by growing queues. This measure is technically sound and leads to a readily observed and measurable outcome. The new measure can be directly and easily applied to calculating the cost of congestion.[7]

REFERENCES

Akçelik, R, Besley, M & Roper, R 1999, *Fundamental Relationships for Traffic Flows at Signalised Intersections*, ARRB Research Report ARR 340, ARRB Transport Research Ltd., Vermont South, Vic.

Austroads 2008, *Guide to Traffic Management Part 7: Traffic Management in Activity Centres*, by Brindle, RE, Marks, D & Croft, P, Report AGTM07/09, Austroads, Sydney, NSW.

Boarnet, K, Kim, E & Parkany, E 1998, *Measuring Traffic Congestion*, ITS Working Paper UCI-ITS-WP- 98-3, Institute of Transportation Studies, University of California, Irvine CA, USA.

BTCE 1996, *Traffic Congestion and Road User Charges in Australian Capital Cities*, BTCE Report 92, Bureau of Transport and Communications Economics, Canberra, ACT.

BTRE 2007, *Estimating Urban Traffic and Congestion Cost Trends for Australian Cities*, Working Paper 71, Bureau of Transport and Regional Economics, Canberra, ACT.

Button, KJ 2004, *Road Pricing*, Final Report of ITS Center Project, Center for ITS Implementation Research, George Mason University, Fairfax VA, USA.

European Conference of Ministers of Transport (ECMT) 1998, *Traffic congestion in Europe*, ECMT Roundtable 110, European Conference of Ministers of Transport, Paris.

Evans, A 1992, 'Road congestion pricing: When is it good policy?' *Journal of Transport Economics & Policy*, 26, pp.213–44.

Cambridge Systematics 2005, *Traffic Congestion and Reliability: Trends and Advanced Strategies for Congestion Mitigation*, report by Cambridge Systematics Inc. with Texas Transportation Institute, for U.S. Department of Transportation, Federal Highway Administration (FHWA), Washington DC.

Greenshields, B 1935. 'A study in highway capacity', *Highway Research Board Procedings*, 14, p.458.

Hills, P 1993, 'Road congestion pricing; when is it good policy? Comment and rejoinder', *Journal of Transport Economics & Policy*, 27(1), pp.91–105.

Hills, P & Gray, P 2000, 'Characterisation of congestion on an urban road network subject to road-use pricing – a fundamental review', paper presented at *9th International Conference on Travel Behaviour Research*, Gold Coast, Queensland, July 2–5.

Institute of Historical Research (IHR) 2010, *Conference on Blocked Arteries: Circulation and Congestion in History*, Institute of Historical Research, University of London, November 2010, CD-ROM.

Karl, CA 2003, 'The Learning Driver: Meeting Traveller Information Needs', Doctor of Business Administration thesis, Australian Graduate School of Entrepreneurship, Swinburne University, Melbourne, published by VDM Verlag, Germany, 2009.

Kimber, R & Hollis, E 1979, *Traffic Queues and Delays at Road Junctions*, TRL Report 909, Transport Research Laboratory, Crowthorne.

Lay, MG 2009, *Handbook of Road Technology*, 4th Edn., Spon, London.

Lay, MG 2010, 'An operational history of congestion', *Conference on Blocked Arteries: Circulation and Congestion in History*, Institute of Historical Research, University of London, November 2010, viewed 20 February 2011 <www.history.ac.uk/events/event/1160>.

[6] Some of these are explored in Lay (2010).

[7] Regrettably, the (retired) author no longer has access to the data needed to illustrate the points in the paper, but would be delighted to participate in any study directed at applying 'automatically' collected queue data to real-time estimates of congestion costs.

Meyer, J & Gomez-Ibanez, J 1981, *Autos, Transit and Cities*, Harvard UP, Cambridge Mass.

Minnesota DoT 2010, *Twin Cities Metro Area Ramp Meter Study*, Minnesota Department of Transportation, St Paul, MN.

Naudé, C, Tsolakis, D & Anson, G 2005, *Defining Transport Congestion*, ARRB contract report VC71854, Victorian Competition and Efficiency Commission, Melbourne, Victoria, Australia.

OECD 1981, *Traffic Control in Saturated Conditions*, OECD Road Research Programme, Organisation for Economic Cooperation and Development, Paris.

Port Jackson Partners 2005, *Reforming and Restoring Australia's Infrastructure*, Business Council of Australia, Sydney.

Strogatz, S 1994, *Non-Linear Dynamics and Chaos*, Perseus, Cambridge, MA.

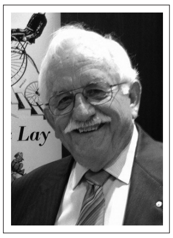
Truong, T & Hensher, D 2007, 'A reassessment of the characterisation of congestion on an urban road network – some theoretical suggestions and illustrative experiments', Ch. 1 in Coto-Millán, P & Inglada, V (Eds.) *Essays on Transport Economics*, Springer, Heidelberg.

Tsolakis, D & Naudé, C 2006, 'Road pricing option for addressing congestion: lessons and possibilities for Australia', *Road & Transport Research* 15(4), pp.64–78.

Vickrey, W 1967, 'Optimisation of traffic and facilities', *Journal of Transport Economics & Policy*, 1(2), p.124.

VicRoads 2010, *Freeway Ramp Signals Handbook*, VicRoads, Kew.

Walters, A 1961, 'The theory and measurement of private and social cost of highway congestion', *Econometrica*, 29, pp.676–699.



Maxwell G Lay

Dr Lay was a Director of ConnectEast. He is an Advisor to Roads Australia and a Professorial Fellow at Melbourne University. In 2000, ARRB named its library the 'M.G. Lay Library'. In 2003, he was awarded a Centenary Medal, followed by an Order of Australia in 2006, particularly for 'the development of new contract management processes, and as an educator and historian'. In 2009, he was awarded the Roads Australia Gold Medal and in 2010, the RACV entered his name in its Golden Book. Also in 2010, ITS Australia gave him its first lifetime achievement award and named the 'Dr Max Lay Award'. From 1986 to 2008 he was a Director of RACV, and President from 1999 to 2002. He was President of the Australian Automobile Association from 2000 to 2002. He has authored over 760 documents, most recently a history of traffic congestion.

Contact

Dr Max Lay

Email: mmlay@bigpond.net.au