Overview of the Causes of Congestion

Artificial Intelligence Research Laboratory, ETRI *October 27, 2020, Sung-Soo Kim*

Abstract

Traffic congestion is a condition in transport that is characterised by slower speeds, longer trip times, and increased vehicular queueing. Traffic congestion on urban road networks has increased substantially, since the 1950s. This technical memo describes an overview of the causes of traffic congestion and the concept of three-phase traffic theory.

1 Introduction

Traffic congestion is a condition in transport that is characterized by slower speeds, longer trip times, and increased vehicular queueing. Traffic congestion on urban road networks has increased substantially, since the 1950s [1]. When traffic demand is great enough that the interaction between vehicles slows the speed of the traffic stream, this results in some congestion. While congestion is a possibility for any mode of transportation, this article will focus on automobile congestion on public roads.

As demand approaches the capacity of a road (or of the intersections along the road), extreme traffic congestion sets in. When vehicles are fully stopped for periods of time, this is known as a *traffic jam* or (informally) a *traffic snarl-up*. Traffic congestion can lead to drivers becoming frustrated and engaging in road rage.

Traffic congestion results from the *imbalance* between the supply of and the demand for transportation facilities [2]:

- The *supply* is constrained by history and geography, by transportation management and operating practices, and by the level of investment on streets and highways
- The *demand* results from the concentration of travel in space and in time

Congestion can be classified in two categories: *recurring* and *nonrecurring*

- *Recurring Congestion* is the delay travelers regularly experience/expect during known travel times—such as the morning and evening rush hours
- *Nonrecurring Congestion* delay is caused by non-predictable (random) events that disrupt traffic flow. These include incidents such as vehicle breakdowns or crashes; road repair and inclement weather; special events that create sudden surges in demand such as the end of a sports event; and natural or man-made disasters. Nonrecurring congestion can either create *new congestion* (in the off-peak periods), or can increase the delay experienced during periods of recurring congestion.

2 Summary of Causes

Traffic congestion occurs when a volume of traffic or modal split generates demand for space greater than the available street capacity; this point is commonly termed saturation. There are a number of specific circumstances which cause or aggravate congestion; most of them reduce the capacity of a road at a given point or over a certain length, or increase the number of vehicles required for a given volume of people or goods. About half of U.S. traffic congestion is *recurring*, and is attributed to sheer weight of traffic; most of the rest is attributed to traffic incidents, road work

and *weather events* [3]. In terms of traffic operation, rainfall reduces traffic capacity and operating speeds, thereby resulting in greater congestion and road network productivity loss.

Traffic research still cannot fully predict under which conditions a *traffic jam* (as opposed to heavy, but smoothly flowing traffic) may suddenly occur. It has been found that individual incidents (such as accidents or even a single car braking heavily in a previously smooth flow) may cause *ripple effects* (a cascading failure) which then spread out and create a sustained traffic jam when, otherwise, normal flow might have continued for some time longer.

Nearly a century ago Miller McClintock [4] stated that congestion is due to three general causes: (1) the inability of the streets to hold a sufficient number of vehicles and to process them at an adequate speed, (2) the inclusion of elements in the traffic stream which hamper its free flow, and (3) the improper or inadequate direction and control of traffic.

Today the causes of traffic congestion are more specifically known and include (1) large concentrations of demand in time and space—including temporal surges in travel demand on roadways of generally constant capacity physical, operational, and design deficiencies that create bottlenecks, (2) traffic demand that exceeds roadway capacity, and (3) physical and operational bottlenecks.

Congestion generally increases with city size. This happens because activity concentrations are larger, and travel distances are longer as cities grow. Economists view chronic congestion as a pricing-induced problem. They argue that the absence of marginal cost pricing contributes to congestion because average cost pricing makes road use more attractive than it would be if prices would rise with congestion.

2.1 Concentration of Trips in Space and Time

If all travel demand were evenly distributed among the various sections of the urban area, the traffic congestion problem would be a rare event. Similarly if all travel were evenly distributed to each hour of the day there would be little, if any, congestion.

But travel demand patterns reflect the concentration in time and space of daily activities: where and when people work, shop, recreate, move goods and provide services. It is the peaking of these spatial and temporal travel patterns that con- tributes to the recurring traffic congestion problem.

2.2 Growth in Population, Employment, Car Use and Insufficient Capacity

Growth in population, employment, and car use (*vehicle miles of travel*—VMT) increase congestion on streets and highways where capacity growth has not kept pace with growth in VMT.

There are several factors that contribute to and shape the growth in population, employment, and vehicle miles of travel (VMT) in urban areas.

2.3 Bottlenecks

Bottlenecks are perhaps the most common cause of congestion. They result from the convergence of a greater number of lanes in the upstream roadways than are available in the downstream roadways. Bottlenecks delay is typically found in hours of peak flow where the number of lanes converging on a roadway, bridge or a tunnel exceeds the number of lanes these facilities have. An early example of a 1940 bottleneck at the Holland Tunnel in New York City is shown in Fig. 1, where traffic from 27 lanes is merging into two Tunnel lanes.

Bottlenecks are also created by roadway incidents that reduce block travel lanes and restrict traffic flow, or they are created by bad weather conditions (e.g., ice on a bridge), a work zone, poorly timed traffic signals, or driver behavior.

Figure 1: Holland Tunnel Bottleneck (1940)—27 lanes trying to get into 2 lanes., Source Reference [5], p 110

3 Classification

Qualitative classification of traffic is often done in the form of a six letter A-F *level of service* (LOS) scale defined in the Highway Capacity Manual, a US document used (or used as a basis for national guidelines) worldwide. These levels are used by transportation engineers as a shorthand and to describe traffic levels to the lay public. While this system generally uses delay as the basis for its measurements, the particular measurements and statistical methods vary depending on the facility being described. For instance, while the percent time spent following a slower-moving vehicle figures into the LOS for a rural two-lane road, the LOS at an urban intersection incorporates such measurements as the number of drivers forced to wait through more than one signal cycle.[20]

Traffic congestion occurs in time and space, i.e., it is a *spatio-temporal process*. Therefore, another classification schema of traffic congestion is associated with some common spatiotemporal features of traffic congestion found in measured traffic data. Common spatiotemporal empirical features of traffic congestion are those features, which are qualitatively the same for different highways in different countries measured during years of traffic observations. Common features of traffic congestion are independent on weather, road conditions and road infrastructure, vehicular technology, driver characteristics, day time, etc. Examples of common features of traffic congestion are the features [J] and [S] for, respectively, the wide moving jam and synchronized flow traffic phases found in Kerner's *three-phase traffic theory*. The common features of traffic congestion can be reconstructed in space and time with the use of the ASDA and FOTO models.

3.1 Three-phase Traffic Theory

The *fundamental diagram* of traffic flow is a diagram that gives a relation between the *traffic flux* (vehicles/hour) and the *traffic density* (vehicles/km). A *macroscopic* traffic model involving traffic flux, traffic density and velocity forms the basis of the fundamental diagram. It can be used to predict the capability of a road system, or its behaviour when applying inflow regulation or speed limits.

Three-phase traffic theory is a theory of traffic flow developed by Boris Kerner between 1996 and 2002. It focuses mainly on the explanation of the physics of traffic breakdown and resulting congested traffic on highways. Kerner describes three phases of traffic, while the classical theories based on the fundamental diagram of traffic flow have two phases: *free flow* and *congested traffic*.

Kerner's theory divides congested traffic into two distinct phases, *synchronized flow* and *wide moving jam*, bringing the total number of phases to three:

1. Free flow (F)
2. Synchronized flow (S)
3. Wide moving jam (J)

The word "*wide*" is used even though it is the length of the traffic jam that is being referred to.

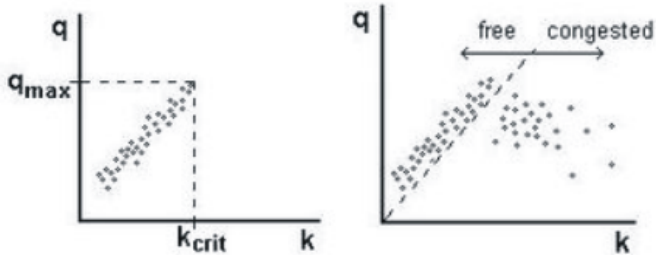

Figure 2: Measured flow rate versus vehicle density in free flow, Flow rate versus vehicle density in free flow and congested traffic (fictitious data)

A phase is defined as a state in space and time.

Free flow (F): In free traffic flow, empirical data show a positive correlation between the flow rate q (in vehicles per unit time) and vehicle density k (in vehicles per unit distance). This relationship stops at the maximum free flow $q_{\max}$ with a corresponding critical density k_{crit}. (See Figure 2.)

Synchronized flow (S): In this phase, the downstream front, where the vehicles accelerate to free flow, does not show this characteristic feature of the wide moving jam. Specifically, the downstream front of synchronized flow is often fixed at a bottleneck.

Wide moving jam (J): A so-called "wide moving jam" moves upstream through any highway bottlenecks. While doing so, the mean velocity of the downstream front v_g is maintained. This is the characteristic feature of the wide moving jam that defines the phase J.

References

[1] R. W. Caves, *Encyclopedia of the City*, 2005.

[2] J. C. Falcocchio and H. S. Levinson, *Road Traffic Congestion: A Concise Guide*. Springer, 2015.

[3] A. Essien, I. Petrounias, P. Sampaio, and S. Sampaio, "The impact of rainfall and temperature on peak and off-peak urban traffic," in *Database and Expert Systems Applications*, ser. Lecture Notes in Computer Science. Springer Nature, 2018, pp. 399–407.

[4] M. M, *Street Traffic Control S.l.* McGraw-Hill Book Company, New York, 1925.

[5] G. NB, *Magic Motorways*. Random House, New York, 1940.