A. Chang

Synchronization of Traffic Signals in Grid Networks

Abstract: A method of synchronizing traffic signals interconnected in an arbitrary network is presented. The procedure consists of using a simplified mathematical model for traffic to relate the vehicular delay within the network to the signal parameters and then searching over these parameters to minimize the delay. The technique has been used to synchronize traffic signals in San Jose, California and has yielded a ten percent reduction in the average delay per car in comparison with the signal settings determined by the city traffic department with conventional engineering methods.

Introduction

This paper is concerned with synchronizing two-phase traffic signals in an arbitrary network. The signals are assumed to have a common period, and therefore, the problem is to specify the relative phasing and green duration of the signals to satisfy some desired goal or criterion. In the special case when the signals are on a single arterial street, they can be set so that a car can go from one end to the other without stopping, provided the driver maintains the speed used in setting them. The portion of the cycle for which this is possible is called the bandwidth for that direction. For arterials, traffic engineers traditionally have considered that the problem is to maximize the bandwidth for one direction while maintaining some specified bandwidth in the other, Recently, Little has shown how this problem can be defined and solved as a mixed-integer linear program.[1]

In an arbitrary network of intersecting streets it may not be possible to obtain concurrently a nonzero bandwidth for every street. Thus, in general, maximum bandwidth may not be a suitable design objective for networks. The work of Helly and Baker suggests further that even for an arterial, large bandwidth probably is of little value when traffic is heavy.[2] One reason for this is that the presence of queues, which inevitably form in heavy traffic, is not taken into consideration in designing for maximum bandwidth.

In this paper, the design criterion is the total vehicular delay in the system. An idealized mathematical model is used to relate the movement of traffic to the signal settings, and the total delay in the network is computed and used as the criterion for judging the effectiveness of the settings. The objective is to find signal parameters that minimize the total delay. While this approach is straightforward in principle, the complexity of traffic flow and the combinatorial aspects of the problem preclude any chance of obtaining a complete solution. Both the modelling problem, which is to obtain an accurate, yet computationally efficient, model of traffic flow, and the problem of minimizing the objective function are extremely difficult and are resolved here only to a limited extent.

A model for traffic flow

In this section, a model of traffic flow will be given which is suitable for the numerical computation needed in the synchronization problem. Typically, problems of practical interest will have about ten intersections when the signals are on a single arterial, and on the order of fifty when they are in a network. As traffic flow over such areas is a complex phenomenon involving a large number of vehicles, it will be necessary to limit consideration to a few facets of the flow.[3] In the model used here, the discrete nature of cars is disregarded, and traffic is thought of as continuous flow. The main physical variables considered are vehicular flow rates and queues. Both of these variables are defined at the intersections, and not on the streets that connect them. In this sense, the model is discrete in space but continuous in time.

Let the signalized intersections in the network of interest be numbered in some way, and let i and j be any two adjacent intersections such that cars can go from i to j. We make the following assumption concerning all such adjacent pairs (i, j):

A.1 *All cars travelling from i to j move at the same speed,* v_{ij}.

As a consequence of A.1, the flows at points on streets connecting intersections are equal to flows at intersections except for a time shift. More explicitly, let $f_i(t)$, expressed in units of cars/sec, be the rate of flow leaving intersection

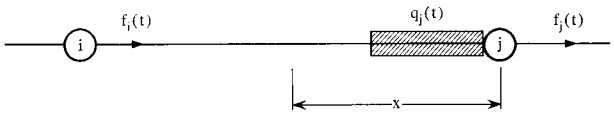

Figure 1 Schematic representation of one direction of traffic at a pair of intersections.

i towards j at time t (Fig. 1). Then the flow at a point a distance x from j, in the absence of a queue at j extending as far back as x, is given by the equation

$$f_x(t) = f_i\left(t - \frac{d_{ij} - x}{v_{ij}}\right), \tag{1}$$

where d_{ij} is the distance from i to j.

In traffic, of course, A.1 is not observed experimentally. It would be more realistic, for example, to allow the speed of each car to be a random variable with some distribution. However, it would then be necessary to trace the motion of each vehicle throughout the system or, equivalently, to determine the flow at points between intersections, and the simplicity expressed by (1) is lost. As we intend to use the model for setting traffic signals, the speeds v_{ij} may be considered the ideal or design speeds. The actual dispersion in speeds generally becomes important only when the system covers an area that requires a relatively long time to traverse.

The number of cars in a queue at an intersection can be readily expressed in terms of its initial value and the flows from the intersection and the previous adjacent intersection. For notational simplicity, only the equation for one direction of flow through j will be given, the others being identical in form. We assume the length of a queue is proportional to the number of cars it contains. Accordingly, let ρ_j, which depends on the number of lanes on the street connecting i and j, denote the length per queued car at j. Then using (1), the queue at j at time t is given by the equation

$$q_j(t) = q_j(0) + \int_0^t \left[f_i\left(\tau - \frac{d_{ij} - \rho_j q_j(\tau)}{v_{ij}}\right) - f_j(\tau) \right] d\tau. \tag{2}$$

In deriving (2), the length of the queue, $\rho_j q_j(\tau)$, has been assumed to be less than d_{ij} in the interval $(0, t)$.

In assuming ρ_j is a constant, we have also neglected the time lag between when the first car in the queue starts and when the last car begins moving. For, differentiating (2) and putting $f_i = 0$, the length of the queue, l_j, in the model satisfies the differential equation,

$$\frac{dl_j}{dt} = -\rho_j f_j(t).$$

In reality, the length of the queue is determined by the position of the last car which does not begin moving until the "starting wave" emanating from the front of the queue propagates to the rear. For this reason, the model tends to underestimate the number of cars which are stopped by other stationary cars. A more realistic equation for l_j would be of the type

$$\frac{dl_j}{dt} = \rho_j \left\{ f_i\left(t - \frac{d_{ij} - l_j(t)}{v_{ij}}\right) - f_j\left(t - \frac{l_j(t)}{v_p}\right) \right\},$$

where v_p is the velocity of propagation of the starting wave. However, in view of the other simplifying assumptions made, this refinement has not been considered essential.

We now give a model for the flow variables. To accomplish this, it is convenient to think of a flow as arising from two components. As before, only the equation for one direction of flow at j will be given. First, if the signal at j is green for the direction (i, j), and there is no queue at j, the flow leaving j is just the flow from i delayed by the travel time d_{ij}/v_{ij} (Fig. 1). On the other hand, should a queue form at j, it will give rise to a component of f_j once the signal turns green. We will make the following assumption concerning the departure of cars from queues:

A.2 *Cars leave queues at a constant rate, accelerating to their desired velocity in a negligible amount of time.*

With A.2, the second component of flow is equal to a constant, r_j, whenever $q_j > 0$ and signal j is green. Note that the two components of flow are mutually exclusive, because a queue interrupts the free flow of vehicles from the previous intersection. Let $I_{q_j}(t)$, the indicator function of q_j, be defined as

$$I_{q_j}(t) = \begin{cases} 0 & \text{if } q_j(t) = 0 \\ 1 & \text{if } q_j(t) > 0. \end{cases}$$

Then, combining the two components, the flow leaving j at time t is given by

$$f_j(t) = \begin{cases} 0 & \text{if signal is red} \\ r_j I_{q_j}(t) + [1 - I_{q_j}(t)] f_i\left(t - \dfrac{d_{ij}}{v_{ij}}\right), & \\ \text{otherwise.} & \end{cases} \tag{3}$$

In (3), the amber phase is ignored; it may be considered as part of either the red or green phase.

Equations (2) and (3) were derived for intersections j in the "interior" of the network. For an intersection on the boundary, the flow from i in (2) and (3) is replaced by a source: $f_i(t)$ is set equal to some prescribed function. In principle, the source waveforms should be chosen to match as closely as possible the flows observed empirically. However, except for the average number of cars which

the source should supply per cycle, the exact shape of the waveform to use is generally difficult to determine. In lieu of a solution to this difficulty, we will make the following simplifying assumption:

A.3 *Cars arrive at the boundary intersections in platoons. Within each platoon, they arrive separated by the same time interval. Furthermore, the temporal arrival pattern repeats itself from cycle to cycle.*

Corresponding to A.3, the source waveforms are periodic piecewise constant functions. Note that A.3 includes the case where cars arrive at a constant rate.

Assuming that each flow f_i continues on to a single intersection (i.e., there are no turns), the model is complete.[4] For if the initial conditions in the network, the timing of the signals, and the sources are specified, (2) and (3) can be solved, provided there is no queue that overflows into its adjacent intersection.

The numerical solution of (2) and (3) is greatly simplified by A.3, since the solutions to (3) are now piecewise constant and, consequently, the solutions to (2) are piecewise linear. It therefore suffices to compute the values of the solutions at times where their slopes change.

A second important consequence of A.3 is that the solution of (2) and (3) is periodic for sufficiently large t, provided that the inputs are low enough to preclude congestion. Furthermore, the periodic solution is independent of the initial conditions in the network. These two facts can be proven by considering a single intersection fed by a periodic source and then showing that the flow leaving it is eventually periodic. The periodic solution will be called the steady-state solution.

Synchronization problem

One signal in the network may be chosen as a reference. Then, the phasing of any other signal is determined by specifying the interval between the beginning of "main street" green at the two intersections. This inteval, called the *offset*, may range from zero up to the period of the signals. The other signal parameter, called the *split*, is the duration of main street green. Its value is restricted by such factors as pedestrian crossing times. The synchronization problem is to specify the offset and split of each signal, within their permissible range, to accomplish some objective.

The criterion which we use to judge the performance of a set of signal parameters is the total steady state delay in the network which we define as

$$D = \sum_i \int_0^T q_i(t)\, dt. \tag{4}$$

In (4), T is the period of the signals, and the $q_i(t)$ are the steady-state solutions of (2) and (3) for some fixed set of sources. The sum is taken over all intersections and directions in the network. D is a function of the signal parameters alone, and the objective is to find a set of these variables which minimizes D.

Properties of D

In this section, some properties of the function D are given, but the associated proofs are only outlined. It will be clear from these properties that D does not have the mathematical structure needed to apply directly any of the standard mathematical programming methods.

P.1 *In general, D is not a continuous function of its arguments.*

To see this, consider an intersection with a queue being serviced at a rate r_1 toward which cars are flowing in a platoon at a rate $r_2 > r_1$. Let the signal at the intersection be set so that the queue is just dissipated when the next platoon arrives at the intersection. Then, if the timing of the signal is changed so that the queue is not dissipated in time, there will be a discontinuity in the delay, because the platoon now will be stopped, and the queue will increase.[5]

It can be shown, however, that D has the following property:

P.2 *D is continuous if the queue service rate at each intersection is at least as great as the possible flow rates into the intersection.*

The conditions required in P.2 are satisfied, for example, if the cars from each source pass through intersections that have the same queue service rates, and if the intensity of each source is sufficiently low.

The q-functions are piecewise linear functions, so from (4) it is clear that D is a quadratic function of all the times t_k where the slopes of the q-functions change. Let p_i denote either the offset or the split of signal i. It can be shown that the times t_k are piecewise linear functions of p_i. Moreover, the possible slopes of the q-functions depend only on the source flow rates, the velocities v_{ij}, and the intersection constants ρ_j and r_j, and not on the signal parameters. Combining the last two statements, it follows that the graph of D as a function of p_i with all other signal parameters held constant is a piecewise quadratic curve. Unfortunately, the "corner points" of such curves are not simply related to the signal parameters, and this property of D is not very useful for minimizing the function.

An individual signal can influence the total network delay by affecting the delay at the signal itself. In addition, by governing the time at which cars leave the intersection, the signal affects the delay these cars experience at succeeding intersections. Within our model, these are the only ways that a signal can affect the total delay. Furthermore, because turns are excluded in the model, the effect is felt

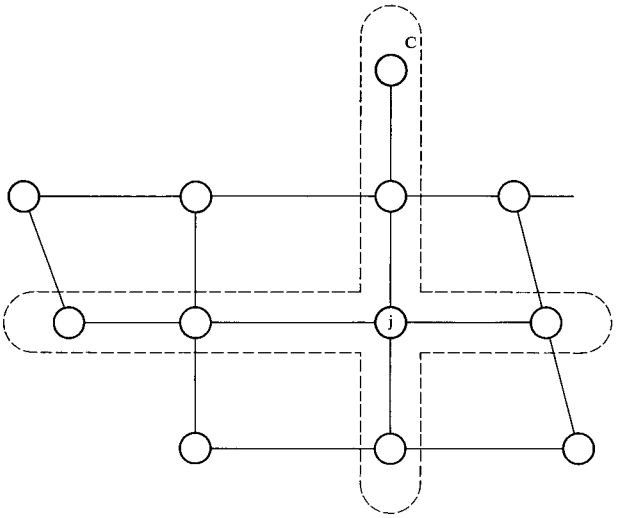

Figure 2 The signal at j affects only the intersections in the contour C.

only in a subset of the network. In Fig. 2, the influence of signal j is confined to the area enclosed by the contour C. Partitioning the network in this way is quite useful. For example, to compute the partial derivative of D, only the intersections in contours like C need be considered at each step.

Procedure for minimizing D

In view of the structure of D, there is probably no foolproof algorithm for minimizing D short of evaluating it for "all possible" signal settings. As the latter is not computationally feasible even for moderate-size networks, it has been necessary to resort to a heuristic search procedure. The procedure, which has been used with some success, normally consists of two stages. Usually a coarse search, which may involve dividing the given network into subsystems, is carried out first. The result obtained is then used as a starting point in a finer search for a local minimum of the function. The procedure will be described briefly in this section, and two examples will be given in the following section.

The search procedure for the case when all the signals are on a single arterial will be described first. An arterial is said to be symmetric if $v_{ij} = v_{ji}$ for all pairs (i, j). It has a half-cycle synchronization if the time between the midpoints of the red phases of every pair of signals is an integral multiple of a half cycle. Morgan and Little showed that among the half-cycle synchronizations is a solution to the maximum equal bandwidth problem.[6] The starting point for the refined search is found by evaluating D for all possible half-cycle synchronizations and choosing one which gives the lowest delay.[7] If the number of signals is n, then 2^{n-1} evaluations of D are required. The setting obtained is refined by a variable step-size gradient procedure or a sequence of searches over regions defined by allowing subsets of the parameters to vary.[8] In the latter case, the region is usually divided with a uniform grid spacing, although a Fibonacci-type search has also been used for single variable searches.[9] Some care is taken in choosing the step size in the gradient procedure because it is more time consuming to compute ∇D than D itself, and because ∇D may be discontinuous.

The procedure for networks is similar to the method used for arterials, except for the determination of a starting point. In a typical network, it is usually possible to single out a few streets which are more heavily travelled than the others. These streets are treated preferentially by synchronizing them as arterials in the first stage of the search.[10] The settings obtained are then combined and used as a starting point for a gradient or sequential search procedure as described above.

The determination of a good starting point is crucial to the minimization procedure. In a real situation, one may also use the settings that are currently on the street. It may also be possible in networks, especially those with several one-way streets, to calculate by hand a reasonably good starting point.

Examples

The first example was studied by Morgan and Little and is a stretch of Euclid Avenue in Cleveland.[6] Figure 3 shows the space-time diagram of a maximum equal bandwidth synchronization for the arterial given in Ref. 6. The parallel sloping lines define the time during which a car may enter the system and go at 50 feet/second from one end to the other without stopping. It is clear that this synchronization

Figure 3 Space-time diagram for part of Euclid Avenue. Maximum equal bandwidth, speed = 50 ft/second, cycle length = 65 seconds, red times are 0.47, 0.40, 0.40, 0.47, 0.48, 0.42, 0.40, 0.40, 0.42. Horizontal lines indicate duration of red lights.

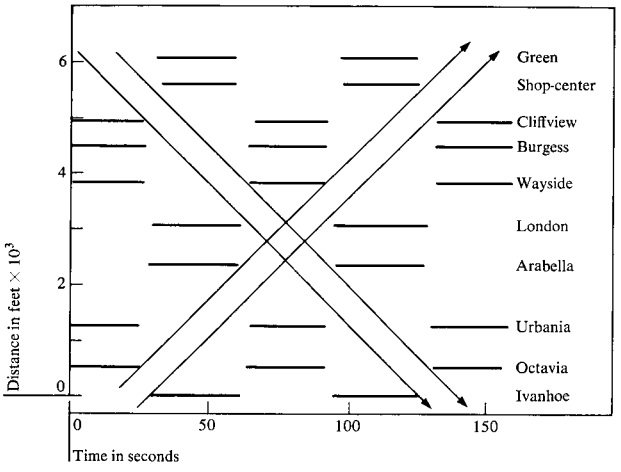

Table 1 Search parameters used for Euclid Avenue.

Iteration	N_1	N_2	Δ (*Cycles*)	D
0				0.89
1	2	0	0.25	0.101
2	2	2	0.25	0.101
3	2	0	0.05	0.091
4	2	0	0.02	0.088
5	2	2	0.02	0.088

Table 2 Offsets for Euclid Avenue. Maximum equal bandwidth solution taken from Ref. 6.

Intersection	*Offset for Maximum Equal Bandwidth (Seconds)*	*Offset Determined by Method of Paper (Seconds)*
Ivanhoe	0	0
Octavia	30.2	31.2
Urbania	30.2	27.9
Arabella	0	1.9
London	0.2	0.6
Wayside	29.5	31.2
Burgess	30.2	31.2
Cliffview	30.2	31.2
Shopping center	64.2	60.5
Green	64.2	63.7

gives zero delay in our model if the platoons supplied by the sources at each end of the arterial are properly phased and shorter than the bandwidth, which is 0.237 cycles. To test our method, the sources were set to supply platoons 0.25 cycles long and phased so that the midpoints of the green phase at boundary intersections and the platoons coincided.

Only the offsets were varied in the search, and initially all offsets were set equal to zero. The intersections were numbered consecutively, and each step in the procedure consisted of a search over the offsets of signal i and $i+1$, $1 \leq i \leq 9$, with the other offsets held constant. Let $o_i(k)$ denote the offset of signal i after the k^{th} step. Then the next search over the offsets of signal i and $i+1$ was in the rectangular grid defined by the equations

$$o_i = o_i(k) + (n - N_1)\Delta, \quad n = 0, 1, 2, \cdots, 2N_1$$

$$o_{i+1} = o_{i+1}(k) + (m - N_2)\Delta,$$

$$m = 0, 1, 2, \cdots, 2N_2.$$

The grid is centered at the previously computed offsets for the signals, Δ is the mesh size, and the parameters N_1 and N_2 determine the number of points in the grid. Considering a search over each of the nine pairs of signals as one

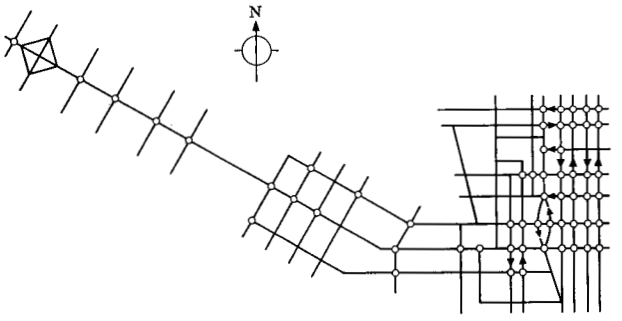

Figure 4 San Jose network.

iteration, the search converged after five iterations. The set of offsets so determined has the property that varying any two adjacent offsets cannot reduce the computed delay. Table 1 gives the value of the parameters Δ, N_1, N_2 and the delay after each iteration. The offsets of the maximum equal bandwidth solution and the computed solution are approximately the same and are given in Table 2 (see also Fig. 3).

The second example is a 59-intersection area in San Jose which is shown schematically in Fig. 4. There are several one-way streets in the system; these are indicated by arrows in Fig. 4. The network is actually two separate systems, because eight intersections in the western part of the system are three-phase signals with eighty-second cycle lengths, and the others are two-phase with 55-second cycle lengths. Consequently, the former was treated as an arterial (with the left-turn green considered as part of the cross-street green) while the latter was considered a single network.

The initial starting point for the search procedure in both systems was zero offset with splits approximately equal to those that had actually been in use in the city. Several signal parameters which resulted from the preliminary coarse search procedure were adjusted by hand. This was easy to do because of the one-way streets. Otherwise, the search procedure used was as described in the preceding section.

All of the signals in the area are currently controlled by an IBM 1710 Control System. Vehicle detectors have been placed thoughout the system in sufficient numbers so that almost every lane on each street is covered. With the present programming system, the number of vehicles that pass over each detector every five seconds is recorded. By knowing when cars pass over the detectors and the phase of the signals, the computer determines approximately the number of cars which stop and the vehicular delay at each intersection. A useful figure of merit for a synchronization is the delay/car for the total system, which is defined as the ratio of the total delay in the system to the total number of cars served by the intersections in the system. In recent tests, over a two-week period, the delay/car was

about ten percent less for the signal settings determined by our procedure in comparison to those determined by the city traffic engineering department using conventional engineering methods. Similar improvements also were measured by a floating car trip-time method.

References and footnotes

1. J. D. Little, "The Synchronization of Traffic Signals by Mixed-Integer Programming," Working Paper 129-65, Sloan School of Mangement, MIT, August 1965.
2. W. Helly and P. G. Baker, "Accelerated Noise in a Congested Signalized Environment," Report R & D 65-2, Engineering Department of the Port of New York Authority, June 1965.
3. Traffic simulation computer programs which attempt to model the detailed structure in traffic have been written but are too slow for actually determining signal settings. See A. M. Blum: "A General Purpose Digital Simulator and Examples of its Application," *IBM Systems Journal* **3**, No. 1, 41-50 (1964).
4. Turns can be modelled by introducing sources and sinks within the network without substantially changing the theory given.
5. The platoon can pass through the intersection without stopping if $q(t) = 0$ even though $r_2 > r_1$. See Eq. (3).
6. J. T. Morgan and J. D. Little, "Synchronizing Traffic Signals for Maximal Bandwidth," *Operations Research* **12**, 896-912 (1964).
7. Computing time on a 7094 for a ten-intersection problem requires approximately fifty-five seconds.
8. The number of variables in each search normally ranges from one to three.
9. D. J. Wilde, *Optimum Seeking Methods*, Prentice Hall, Englewood Cliffs, New Jersey, 1964.
10. A signal common to two arterials is assigned to only one of them.

Received September 14, 1966.