2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

Estimating Daily Start Times of Periodic Traffic Light Plans from Traffic Trajectories

Ori Rottenstreich, Tom Kalvari, Nitzan Tur, Eliav Buchnik, Shai Ferster, Dan Karliner, Omer Litov, Danny Veikherman, Avishai Zagoury, Jack Haddad, Dotan Emanuel and Avinatan Hassidim
Google Research, Israel

***Abstract*— In recent years, the wealth of available vehicle location data from connected vehicles, cell phones, and navigation systems has been introduced. This data can be used to improve the existing transportation network in various ways. Among the most promising approaches is traffic light optimization. Traffic light optimization has the potential to reduce traffic congestion, air pollution and GHG emissions. The first step in such optimization is the understanding of the existing traffic light plans. Such plans are periodic but, in practice, often start every day at arbitrary times, making it hard to align traffic trajectories from various days toward the analysis of the plan. We provide an estimation model for estimating the daily start time of periodic plans of traffic lights. The study is inspired by real-world data provided, for instance, by navigation applications. We analyze the accuracy of such computations as a function of the characteristics of the sampled traffic and the length of the evaluated time period.**

I. Introduction

Road transportation is responsible for over 10% of the world Greenhouse gases (GHGs) [1] and in several countries a person spends on average over half an hour per day in traffic delays. Studies showed that emissions and travel time can be reduced by a careful design of traffic light plans that match traffic trends. Designing efficient traffic light plans requires inputs related to the intersection where the traffic light is located and its traffic [2], [3], [4]. These include intersection properties such as the structure of the intersection, the allowed movements crossing the intersection, the periodicity of traffic light plans in the intersection and the currently operating plans.

Such input can be learned from traffic data expressed as vehicle trajectories [5], [6], [7]. A trajectory is a series of pairs of timestamps and GPS locations of vehicles. A potential common source for such trajectories is navigation applications, often adopted by a subset of the vehicles and accordingly represent a sampled part of the traffic.

The Google Green Light project [8] helps to reduce emissions in cities by analyzing Google Maps driving trends to build intelligent recommendations that optimize the timing and coordination of traffic lights. The project is already deployed in over 12 cities such as Rio de Janeiro, Seattle, Bangalore, Hamburg, Haifa, Jakarta and Budapest. It currently affects more than 25 million drivers every month. Initial deployments at intersections show a reduction of up to 30% in stopping and 10% in GHG emissions.

Authors contributed equally to this work.

As traffic light plans do not change frequently, analyzing them based on several-day data allows higher accuracy. While this is challenging since a periodic plan often starts at arbitrary times each day, *computing the daily start times allows aligning trajectories from different days that match identical parts of the plan to better estimate the plan.*

Contributions. In this paper, we study the estimation based on traffic trajectories of *Day Shifts* - the times a periodic plan of a traffic light starts on each day. We present a graph-based method with three steps to compute them based on trajectories and analyze the accuracy of the computations based on the amount of available information.

Terminology. We detail the basic terminology of this study.

Movement - A movement refers to intersection traffic sharing the same pair of incoming and outgoing directions.

Traffic plan - An intersection is associated with a periodic traffic plan. In each cycle, the plan has a sequence of various phases, each allowing traffic of some movements. Two movements that cross each other cannot be allowed at the same phase. While there can be several traffic plans during the day, we refer to a period of hours with a single plan that repeats itself. The length of the periodic plan is called the *cycle time*.

Crossing time - The time a vehicle enters an intersection (crossing a stop line for its movement). The time can be computed from the vehicle trajectories.

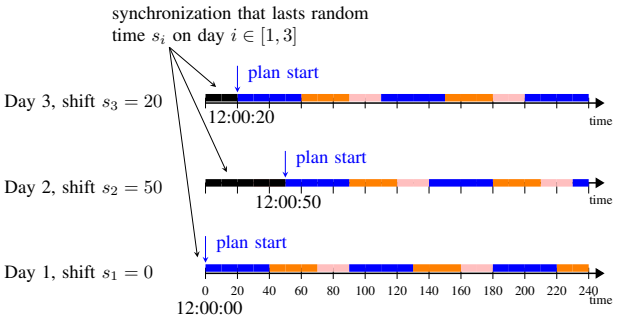


Fig. 1. Day Shifts Estimation: The same traffic plan starts every day at a different time following a synchronization process that lasts every time some random time (shown in black). The plan is an ordered sequence of three phases (shown in blue, orange and pink). On day $i \in [1,3]$ the plan starts at time s_i and repeats itself every cycle time of $C = 90$ seconds.

TABLE I

SUMMARY OF THE MAIN NOTATIONS

Symbol	Meaning
C	plan cycle time
s_i	plan time shift for day i
G	graph representing days and their mutual shift differences
M	bound on the maximal error of mutual shift differences
m	number of days

II. The Day Shifts Estimation Problem

Consider a signalized intersection with a periodic traffic light plan, that repeats itself based on its cycle time. Typically, the cycle time is in the range of 1-4 minutes. Over the days, the plan starts at different times following a short daily synchronization process of the traffic light that takes some random time every day. To understand the phases of a plan and analyze its performance, it is important to align the traffic samples collected between days based on the phases of the plan they match. Computing the daily start times of the plans is crucial to allow such an alignment. We refer to the differences in the time a plan starts each day as daily shift values. This paper studies how shift values can be estimated based on traffic trajectories.

Problem Statement. Consider a plan that applies for several days with a known cycle time C. For vehicles in a single movement, modeling crossing times (the times vehicles cross their stop line) on day i as independent random variables that distribute modulo C as $s_i + \lambda$, where s_i is the shift of day i and $\lambda \sim \Lambda$ is an unknown but fixed crossing time distribution Λ. We wish to estimate all s_i. The number of crossing times in each day is distributed like some Poisson distribution, independent between different days.

Example 1. *Consider a plan with a cycle time of $C = 90$ seconds that operates in three days. On day 1 the plan starts at 12:00:00 (namely at noon). The plan repeats itself every $C = 90$ seconds and starts again at times 12:01:30, 12:03:00, etc. On days 2 and 3 the plan starts at 12:00:50 and 12:00:20 (respectively), and repeats every $C = 90$ seconds. These start times translate to day shift values of $0, 50, 20$ seconds respectively. Fig. 1 illustrates the plan for the three days with different colors for various parts of the plan.*

We aim to compute the shift values $(s_1, s_2, s_3) = (0, 50, 20)$.

Table I summarizes the main notations of the study.

III. The Proposed Three-Step Approach

A. Overview of the Approach.

We propose a three-step approach to detect the daily shift values. The intuition behind the approach is as follows. First, we correlate distributions of pairs of days to estimate modular differences. We then use small cycles in the modular difference graph to estimate non-modular differences. Finally, we use non-modular differences to estimate the individual day shifts by solving a least-squares problem.

We suggest the three following steps.

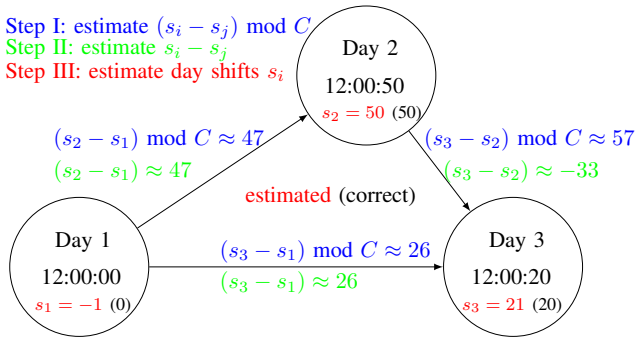


Fig. 2. Illustration of Steps I-III for estimating day shift values. There are three days with a cycle time of $C = 90$ seconds of the periodic plan. Each node represents a day and indicates the (unknown) start time of the plan.

- **Step I** - Estimate the mutual modular difference of shift values between pairs of days modulo the cycle time, up to some additive error. The output of this step is a weighted graph $G = (V, E)$ with nodes representing days. The weight D_{ij} of a directed edge $e = (i, j)$ between two days i, j indicates the estimated mutual shift from day j to day i (i.e. $s_i - s_j$) modulo the cycle time C. The graph is not necessarily complete based on the accuracy of estimations.
- **Step II** - Use estimates of modular differences $(s_i - s_j)$ modulo C to estimate non-modular differences $s_i - s_j$.
- **Step III** - Use all estimates of differences $s_i - s_j$ to estimate the values of the day shifts s_i for all days in V.

In some cases, the result of Step I could be used directly to estimate the day shift values by arbitrarily setting the shift of one day and considering its mutual shift from others. Steps II and III are important as the graph is not necessarily complete. Moreover, improved accuracy can be achieved with the additional steps that take advantage of the complete information from Step I to overcome the inherent potential error in estimating the mutual modular difference of shift values.

The following example illustrates the steps.

Example 2. *To illustrate the approach, we refer again to shift day values from Example 1 and Fig. 1 where the cycle time is $C = 90$ and we aim to compute shift values such as $(s_1, s_2, s_3) = (0, 50, 20)$. We illustrate steps I-III in Fig. 2. First, in Step I, we use traffic information to estimate differences in pairs of shift values.*

Assume that shift values between pairs of days computed at Step I yield approximate values of $(s_2 - s_1)$ mod $90 \approx 47$, $(s_3 - s_1)$ mod $90 \approx 26$, and $(s_3 - s_2)$ mod $90 \approx 57$.

In Step II, we wish to find the approximate values of the differences (without the modular restriction), so we would get $s_2 - s_1 \approx 47$, $s_3 - s_1 \approx 26$ and $s_3 - s_2 \approx -33$.

In step III, we use the approximate differences without modulo to estimate the values of the day shifts, deriving for instance $(s_1, s_2, s_3) \approx (-1, 50, 21)$.

Next, we detail the three steps of the proposed approach.

B. Step I - estimating mutual modular differences in day shifts

The first step in our algorithm is to estimate the modular differences of day shifts, i.e. estimating $(s_i - s_j) \bmod C$ for all i, j. The core idea is that for estimating $(s_i - s_j) \bmod C$, we correlate crossing time distributions in days i and j mod C, to find a shift between the days which best aligns them.

Since Λ is in practice a continuous (and relatively well-behaved) distribution, it might be best to use a continuous similarity test (such as the Kolmogorov-Smirnov test) [9].

To derive provable bounds, we view the distribution Λ in a discrete setting. We partition the range $[0, C)$ into B bins of equal size. The bins correspond to $[0, \frac{C}{B})$, $[\frac{C}{B}, \frac{2C}{B})$, ..., and we index them as bins number $0, 1, ..., B-1$ respectively. We approximate the day shifts as multiples of $\frac{C}{B}$[1]. By making this translation, we may view Λ as a distribution on the bins, i.e. $\Lambda_0, \Lambda_1, ..., \Lambda_{B-1}$ where $\Lambda_i := \Pr_{\lambda \sim \Lambda}(\lambda = i)$, and we approximate s_i as integral translations $\hat{s_i} := \left[\frac{B}{C} \cdot s_i\right]$ on the bins, i.e. assume that on day i, each sample distributes like $(\hat{s_i} + \lambda) \mod B$ for $\lambda \sim \Lambda$.

We wish to use sampled crossing times of two days to estimate their modular difference. For this estimation, we provide a method and analyze its correctness. For each possible difference, we consider the inner product of the sample distributions of the two days. Say that for each bin t, on day i we have X_t crossing times at this bin and on day j we have Y_t crossing times at this bin (we think of the bins as modular, so $X_{t+B} := X_t$, $Y_{t+B} := Y_t$). We consider the following *score* of a potential modular difference D:

$$\text{Score}_D := \sum_{t=0}^{B-1} X_t Y_{t+D}. \tag{1}$$

Let $\mu_D := \mathbb{E}[\text{Score}_D]$ be the mean score. First, we can see that it is maximal when D is the correct offset $\hat{s_j} - \hat{s_i}$:

Lemma 1. *The maximum $\max_D(\mu_D)$ is achieved for $D = \hat{s_j} - \hat{s_i}$, and it is unique if Λ has no smaller period than B.*

We prove this lemma in Appendix A. To generate a confidence interval for the correct offset, we must also analyze the tail distribution of Score_D. Each such score is an inner product between two vectors of Poisson variables. We use the following lemma to bound its tail distribution.

Lemma 2. *Let B be a positive integer, let ϵ be a positive real number, and let $\theta_0, \theta_1, ..., \theta_{B-1}$ and $\nu_0, \nu_1, ..., \nu_{B-1}$ be nonnegative real numbers. Take the following independent random variables:*

$$X_i \sim \textit{Pois}(\theta_i), Y_i \sim \textit{Pois}(\nu_i) \quad (\forall i \in \{0, ..., B-1\}) \tag{2}$$

Define the following random variable: $Z := \sum_{i=0}^{B-1}(X_i \cdot Y_i)$. Its mean is $\mu := \sum_{i=0}^{B-1}(\theta_i \cdot \nu_i)$. Denote $L := 2\log\left(\frac{2B}{\epsilon}\right)$. For any positive t, it holds that

$$\textit{Pr}(|Z - \mu| \geq t) < \epsilon + e^{-\frac{2t^2}{5BL^4 + 5L(\sum_i \theta_i^3 + \sum_i \nu_i^3)}}. \tag{3}$$

The proof for this lemma is given in Appendix B. We may bound θ_t, ν_t with high probability using the results of the Poisson random variables X_i, Y_i, and we may use that to gain an upper bound on $\left(\sum_i \theta_i^3 + \sum_i \nu_i^3\right)$ (with high probability). Since μ_D distributes like Z in the lemma, we may use it to bound the probability that some t maximizes μ (so by Lemma 1 is the correct difference), for any distribution Λ. We join these possible t-s into a confidence interval for the difference. Note that the distribution Λ affects the distribution of the size of the confidence interval. If Λ is uniform, all scores distribute identically, so the confidence interval typically contains the entire range $[0, B)$. On the other extreme, if Λ were supported only on one value, then we would need very few samples for the confidence interval to contain only the correct difference.

For every pair of days, we generate some estimate for their modular day shift difference, with a confidence interval attached. There is redundancy in the differences between all pairs of day shifts, so we will only use confidence intervals under some threshold size for estimating individual day shifts. We discuss this tradeoff more in Subsection III-E.

C. Step II - from modular to non-modular differences

In this algorithm step, we use the modular difference estimates of day shift differences $s_i - s_j$ to derive estimates for the non-modular differences. To do so, we assume that the modular estimates computed in Step I have an error smaller than M (which depends on the distribution Λ and the number of sample points). We can write this as

$$s_i - s_j = D_{i,j} + k_{i,j} \cdot C - \beta_{i,j}, \tag{4}$$

where $D_{i,j}$ are the modular estimates, $k_{i,j}$ are integral, and $\beta_{i,j}$ satisfy $|\beta_{i,j}| < M$. Denote $D_{j,i} = -D_{i,j}$, $k_{j,i} = -k_{i,j}$ and $\beta_{j,i} = -\beta_{i,j}$.

To translate modular difference estimates to non-modular ones, we must compute the $k_{i,j}$ values.

Note a degree of freedom in these values: For any day i and integer a, if we increase s_i by $a \cdot C$ and decrease $k_{i,j}$ by a for all j (and maintain the other values), all equations hold.

To find the $k_{i,j}$ values, we leverage cycles in G. The cycle length can be bounded based on the graph diameter.

Definition 1 (Graph diameter)**.** *For a graph $G = (V, E)$, define the distance between two vertices as the number of edges in a shortest path between them. The graph diameter is the maximal distance between any pair of vertices.*

Assumption 1 Our algorithm works under the assumption that the diameter of the graph G (indicating the availability of modular differences) is at most $\frac{C}{4M} - \frac{1}{2}$.

The algorithm works as follows. First, take a BFS tree F around some node v. For any node, its distance to v in the tree is equal to its distance to v in the graph, which is at

[1]We can see $\frac{C}{B}$ as a minimal time resolution such as seconds. Such a minimal resolution implies an error with an order of $\frac{C}{B}$, which we do not discuss.

most the graph diameter. Leveraging the discussed degree of freedom in the values of the day shifts s_i, we assume without loss of generality that all $k_{i,j}$ of edges of F are 0. We now need to determine $k_{i,j}$ for edges of G not in F. For each such edge between i, j, close it to a cycle using the unique path between them on the tree, say $i = u_0$–u_1–...–$u_r = j$, and sum Equation (4) over that cycle. Denote $u_{r+1} := i$, we derive

$$\sum_{p=0}^{r} \left(s_{u_p} - s_{u_{p+1}}\right) = \sum_{p=0}^{r} \left(D_{u_p,u_{p+1}} + k_{u_p,u_{p+1}} \cdot C - \beta_{u_p,u_{p+1}}\right)$$

The left-hand side is 0 as a sum of differences over a cycle. On the right-hand side all $k_{u_p,u_{p+1}}$ are 0 except for $k_{j,i}$, so $0 = \sum_{p=0}^{r} D_{u_p,u_{p+1}} - \sum_{p=0}^{r} \beta_{u_p,u_{p+1}} + k_{j,i} \cdot C$ and accordingly

$$k_{i,j} = \frac{1}{C}\left(\sum_{p=0}^{r} D_{u_p,u_{p+1}} - \sum_{p=0}^{r} \beta_{u_p,u_{p+1}}\right). \quad (5)$$

The value r, as the path length over the tree F between two nodes in G, equals at most twice the graph diameter, namely $r \leq \frac{C}{2M} - 1$. Accordingly, the cycle length $r+1$ is at most $\frac{C}{2M}$.

The sum $\sum_{i=0}^{r} \beta_{u_p,u_{p+1}}$ is smaller in absolute value than $(r+1) \cdot M$. The last inequalities imply together that

$$\frac{1}{C} \cdot |\sum_{i=0}^{r} \beta_{u_p,u_{p+1}}| < \frac{1}{C} \cdot (r+1) \cdot M \leq \frac{1}{C} \cdot \frac{C}{2M} \cdot M = \frac{1}{2}.$$

As the left-hand side in Equation 5 has an integer value, we can derive all $k_{i,j}$ values from $D_{i,j}$ values as $k_{i,j} = \left[\frac{1}{C} \cdot \sum_{i=0}^{r} D_{u_p,u_{p+1}}\right]$ where [] denotes the rounding operation.

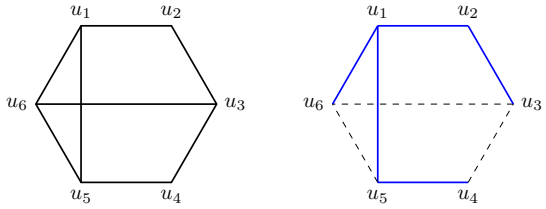
Example 3. *Fig. 3(a) shows an example graph G with 6 nodes $u_1, u_2, \ldots, u_6$ that refer to 6 days. The graph has a diameter of 2. Fig. 3(b) shows in blue a potential BFS tree F with 5 edges. Accordingly, we assume that the five $k_{i,j}$ values for the tree edges equal 0, namely $k_{1,2} = k_{2,3} = k_{4,5} = k_{1,5} = k_{1,6}$. The values for other edges can be computed based on cycles in the tree connecting them. For instance, based on the cycle u_4–u_5–u_1–u_2–u_3–u_4, we can compute the value of $k_{3,4}$. The length of such a path is at most $r+1$ such that r is at most the value of the graph diameter.*

D. Step III - using day shift differences to determine day shifts

Step III uses the estimates for differences of day shifts computed at Step II to estimate the day shifts themselves. We explain the proposed approach for this estimation. Denote

$$d_{i,j} := D_{i,j} + k_{i,j} \cdot C.$$

For each edge (i, j), we have that $s_i - s_j = d_{i,j} - \beta_{i,j}$, where $d_{i,j}$ is known and $\beta_{i,j}$ is small. We can find a solution that best approximates the differences (in terms of least mean squared error) by solving a simple least-squares problem. We know that $\beta_{i,j}$ are the estimation error we got in Subsection III-B. By assuming asymptotic normality of the estimators (as defined in [10]) and given enough data, $\beta_{i,j}$ are roughly normal.



(a) a graph G (b) a BFS tree F in G

Fig. 3. Example of a graph G with 6 nodes and diameter 2 (in (a)). A node refers to a day and an edge to the availability of estimated shift value differences. A BFS tree F in G with 5 edges is shown in blue in (b).

Denote by A a matrix with a column for each node and a row for each edge, where in each row there is a value 1 on the column matching one of the edge endpoints and a value -1 on the column for the other edge endpoint. In this row, the value is 0 on all other columns. This gives us $As = d + \beta$ where s, d, β indicate vectors of the corresponding values.

Consider a solution to the instance of the least-squares problem, namely a vector s' that minimizes $As' - d$. The vector s' is our estimate for s. We wish to investigate the error of this estimate, $|s - s'|$. We must first mention another degree of freedom - if we add a constant to all s_i, the differences do not change. This means that we can, at most, hope to estimate the day shifts up to a constant. This expresses a degree of freedom in the problem formulation - we may denote any point in the traffic light plan as the "start" of the plan, shifting all s_i by a constant (modulo C). To resolve this degree of freedom, we add a constraint that $\sum_i s_i = 0$, which uniquely determines the day shifts.

Denote by $U := \mathbb{R}_0^V$ the vector space of all vectors indexed by vertices which sum up to zero, and $W := \mathbb{R}^E$ the vector space of all vectors indexed by edges. We think of A as a linear transformation from U to W.

Denote by m the number of days for which shift values are estimated that m is the number of vertices in G. We show that up to a constant shift, the solution for the least-squares instance estimates the day shifts up to a small error when each edge in the graph appears with at least some probability p that is not very small. To do so, we rely on the following assumption:

Assumption 2 An edge in the graph G of the differences of day shift values (for m days) appears with at least some probability p such that $p > H \cdot \frac{log(m)}{m}$, for a constant H to be expressed later.

We express the accuracy of the solution to the least-squares problem as a solution for the shift values based on the following. [2]

[2]In this analysis, we assume for simplicity that estimated shift differences are available with some probability independently for the various pairs of days. In practice, this might not be fully accurate due to some days with unusual traffic patterns. Note that the result also holds if the probability for the availability of estimations varies among pairs while referring to the minimal probability over all pairs.

Theorem 1. *For any positive real $\epsilon > 0$ there exists a real constant $H > 0$ and integer $m_0 > 0$, such that for any positive integer m with $m > m_0$ and real $p \in [0, 1]$ which satisfies Assumption 2, the following holds. For a random Erdös-Rényi graph $G = G(m, p)$ with a vertex set V, for any real $\sigma > 0$ and any assignment of values to its vertices $s : V \to \mathbb{R}$ with $\sum_{v \in V} s_v = 0$, if we take "noisy differences" along the edges*

$$d_{i,j} := s_i - s_j + \beta_{i,j}$$

where $\beta_{i,j}$ are i.i.d. normal random variables with mean 0 and standard deviation σ, taking s' to be the ordinary-least-squares solution to the above problem gives an approximation to s with root mean squared error at most

$$(H \cdot \sigma)/(m \cdot p) \tag{6}$$

with probability at least $1 - \epsilon$.

The proof of Theorem 1 can be found in Appendix C.

E. Explicit Parameter Dependence

In Subsection III-B, we had a configurable tradeoff between the error of the day shift differences and the number of the differences that we are able to produce. Are fewer, more accurate day shifts better? Or do we prefer more, even though they are off by more? The analysis of Step II gives us an explicit dependence that we must reach in order for the algorithm to work - particularly Assumption 1. Equation (6) in Step III determines how accurate the final result will be.

Note that in Subsection III-D we claimed the "differences have a normal noise with standard deviation σ", and in Subsection III-C we needed to assume a "guaranteed bound M on difference error". We may use $M := 2\sqrt{\log(m)}\sigma$, to get a global bound with high probability. Such a value of the bound on the difference error is satisfied with high probability following the distribution with standard deviation σ based on Hoeffding's inequality.

To translate the diameter in Assumption 1 to a demand on p, we rely on a result by Klee and Larman in [11] on the diameter of random graphs; For a random graph $G(m, p)$, the diameter almost surely satisfies $\text{diameter}(G) = \left\lfloor \frac{1}{1+\frac{\log(p)}{\log(m)}} + o(1) \right\rfloor$.

To satisfy Assumption 1, we need

$$\text{diameter}(G) + \frac{1}{2} \leq \frac{C}{4M} = \frac{C}{8\sqrt{\log(m)}\sigma}$$

$$8\sqrt{\log(m)}\sigma \left(\left\lfloor \frac{1}{1 + \frac{\log(p)}{\log(m)}} + o(1) \right\rfloor + \frac{1}{2} \right) \leq C, \tag{7}$$

which gives us an explicit dependence between p and σ. Note that for the case where the diameter is constant, it holds that $p > m^{-1+\epsilon}$ and Assumption 2 is necessarily satisfied.

Given that these assumptions hold, we know from Equation (6) that the final root mean squared error is at most $\frac{\sigma}{\sqrt{mp}}$. This means that in Step I when we decide on the tradeoff between σ and p, we want to minimize $\frac{\sigma}{\sqrt{mp}}$ under the assumption that the Inequality (7) holds.

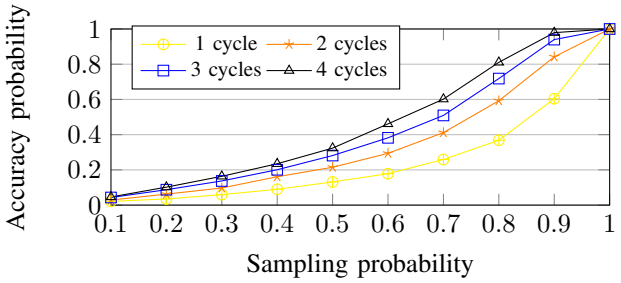


Fig. 4. Accuracy in mutual difference for two days - Impact of the sampling probability for data with various duration values.

IV. Experimental Evaluation

In this section, we conduct experiments to illustrate the proposed approach and examine its accuracy while focusing on the mutual difference for pairs of days. We examine the accuracy of estimating the modular difference of the day shifts of two days. This task was described in detail in Section III-B under the name of *estimating mutual modular differences*.

We refer to a plan with a cycle time of $C = 90$ seconds with three phases, similar to the plan illustrated in Fig. 1. Let (s_1, s_2) be the daily shift values for the two days. We aim to find $(s_2 - s_1)$ mod 90, the modular difference of the day shift of the two days. We refer to the estimation based on the first of the three phases. Recall that the estimation method finds the modular difference as the difference D that maximizes the score Score_D from Eq. 1.

The range $[0, C)$ is partitioned to $B = 30$ bins, each of 3 seconds: $[0, \frac{C}{B})$, $[\frac{C}{B}, \frac{2C}{B})$, ... Assume that the first phase lasts 30 seconds among the $C = 90$ seconds in each cycle and thus refers to 10 bins. The distribution Λ for each bin implies some probability for a vehicle's arrival within the bin's time. As the bin refers to a relatively short period, we assume the number of arrivals in the bin is at most one. We refer to that probability as the sampling probability.

Fig. 4 shows the probability of computing the exact mutual difference vs. the sampling probability for various duration values of the data. The sampling probability has a significant impact on the accuracy probability. For instance, with data of the duration of a single cycle ($C = 90$ seconds), the accuracy probability is 0.021, 0.131 and 0.604 for sampling probabilities of 0.1, 0.5 and 0.9, respectively. Increasing the duration of the data allows higher accuracy. For instance, for the mentioned sampling probability of 0.9, when the duration is set to 2 cycles ($C = 90$ seconds), the accuracy increases to 0.842. Similarly, with the same sampling probability of 0.9, for longer data periods of 3 and 4 cycles, the accuracy probabilities reach even higher values of 0.940 and 0.979.

We also consider scenarios of low sampling probability. We examine the duration of the data required for achieving some required accuracy probability in estimating the modular difference of the day shift or alternatively to satisfy bounds on the error of the estimated mutual difference. Fig. 5 shows the minimal duration in units of cycles (each of $C =$

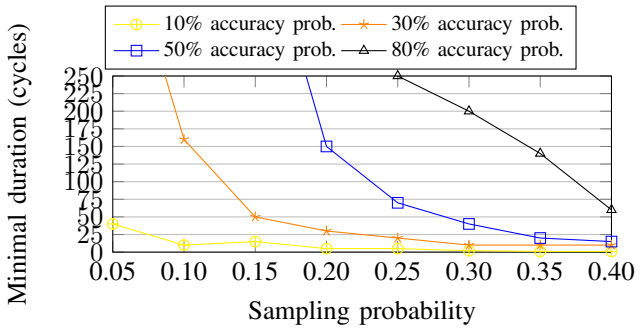


Fig. 5. Accuracy in mutual difference for two days - Data minimal duration for low sampling probabilities and various accuracy probabilities.

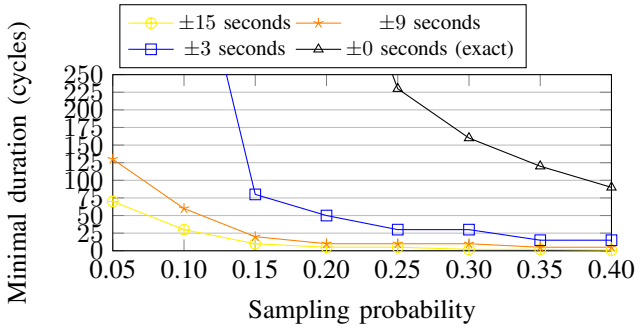


Fig. 6. Accuracy in mutual difference for two days - Data minimal duration for 80% success in having an estimation error within the particular error bound.

90 seconds) for sampling probabilities within $[0.05, 0.40]$. For space limits, duration values of more than 250 cycles (roughly 6.25 hours) are not shown for low sampling probabilities and high accuracy probabilities. As we can see, for an accuracy probability of 50% (blue curve), an average duration of 150 cycles (3.75 hours) is required for a sampling probability of 0.20. The average required duration reduces to only 40 cycles (a single hour) when the sampling probability is 0.30.

Similarly, Fig. 6 shows the duration of the data required to estimate (with an accuracy probability of 80%) the correct mutual shift with some bounded error. For computing a shift with a distance of at most 3 seconds from the correct value, a duration of 80 cycles (2 hours) is necessary when the sampling probability is 0.15. The duration drops to 30 cycles (45 minutes) for a sampling probability of 0.25. Allowing inaccuracies in estimating the mutual shift, such as bounded by 9 or 15 seconds allows low minimal duration values. In such cases, the duration is no larger than 20 cycles when the sampling probability is 0.15 or larger.

V. Conclusions

This paper studies a basic estimation problem of the daily shift values in the start times of periodic plans in traffic lights. For the problem, we showed a graph-based approach that includes three steps: In the first step we estimate mutual modular differences of day shifts modulo the cycle time. In the second step, we explain how to derive the global mutual differences among day shifts. In the third and last step, we compute the shift value for each of the days. We studied the impact of the number of days and traffic arrival distributions on the accuracy of the estimations. As a future work, we aim to extend this study towards the estimation of additional features that refer to the traffic light plans such as the exact start and end time in each of the phases of the periodic plan.

References

[1] F. D. Albuquerque, M. A. Maraqa, R. Chowdhury, T. Mauga, and M. Alzard, "Greenhouse gas emissions associated with road transport projects: Current status, benchmarking, and assessment tools," *Transportation Research Procedia*, vol. 48, pp. 2018–2030, 2020.

[2] D. I. Robertson, "Transyt: a traffic network study tool," 1969.

[3] B. L. Smith, W. T. Scherer, T. A. Hauser, and B. B. Park, "Data–driven methodology for signal timing plan development: A computational approach," *Computer-Aided Civil and Infrastructure Engineering*, vol. 17, no. 6, pp. 387–395, 2002.

[4] B. Park and A. Kamarajugadda, "Development and evaluation of a stochastic traffic signal optimization method," *International journal of sustainable transportation*, vol. 1, no. 3, pp. 193–207, 2007.

[5] V. Protschky, C. Ruhhammer, and S. Feit, "Learning traffic light parameters with floating car data," in *IEEE International Conference on Intelligent Transportation Systems*, 2015.

[6] X. Zhan, R. Li, and S. V. Ukkusuri, "Link-based traffic state estimation and prediction for arterial networks using license-plate recognition data," *Transportation Research Part C: Emerging Technologies*, vol. 117, p. 102660, 2020.

[7] O. Rottenstreich, E. Buchnik, S. Ferster, T. Kalvari, D. Karliner, O. Litov, N. Tur, D. Veikherman, A. Zagoury, J. Haddad, D. Emanuel, and A. Hassidim, "Probe-based study of traffic variability for the design of traffic light plans," in *International Conference on Communication Systems & Networks (COMSNETS)*, 2024.

[8] Google, "Green Light - Using Google AI to reduce traffic emissions," 2023. [Online]. Available: https://sites.research.google/greenlight/

[9] N. V. Smirnov, "Approximate laws of distribution of random variables from empirical data," *Uspekhi Matematicheskikh Nauk*, no. 10, pp. 179–206, 1944.

[10] "Random variable theory," in *Markov Processes*. San Diego: Academic Press, 1992, pp. 1–58.

[11] V. Klee and D. Larman, "Diameters of random graphs," *Canadian Journal of Mathematics*, vol. 33, no. 3, p. 618–640, 1981.

[12] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak Mathematical Journal*, vol. 23, no. 2, pp. 298–305, 1973. [Online]. Available: http://eudml.org/doc/12723

[13] T. Kolokolnikov, B. Osting, and J. Von Brecht, "Algebraic connectivity of erdös-rényi graphs near the connectivity threshold," *Manuscript in preparation*, 2014.

Appendix

A. Proof of Lemma 1

Proof: We assumed that the number of samples in each day distributes like a Poisson random variable, independently between different days. Say that on each day p, the expected number of samples on day p is γ_p. This means that the number of crossing times at bucket t on day p distributes like a Poisson random variable with expectation $\gamma_p \Lambda_{t+s_p}$ (where the sum is taken modulo B). Recall that we denoted the number of crossing times at bucket t on day i as X_t and the number of crossing times at bucket t on day j as Y_t. This

means that we may compute the expected score directly:

$$\begin{aligned}\mu_D &= \mathbb{E}\left[\text{Score}_D\right] = \mathbb{E}\left[\sum_{t=0}^{B-1} X_t Y_{t+D}\right] \\ &= \sum_{t=0}^{B-1} \mathbb{E}\left[X_t\right]\mathbb{E}\left[Y_{t+D}\right] = \sum_{t=0}^{B-1} \gamma_i \Lambda_{t+\hat{s}_i}\gamma_j\Lambda_{t+\hat{s}_j+D} \\ &= \gamma_i\gamma_j \sum_{t=0}^{B-1} \Lambda_{t+\hat{s}_i}\Lambda_{t+\hat{s}_j+D}.\end{aligned}$$

By the Cauchy–Schwarz inequality the above equals at most

$$\gamma_i\gamma_j \sqrt{\sum_{t=0}^{B-1}\Lambda_{t+\hat{s}_i}^2}\sqrt{\sum_{t=0}^{B-1}\Lambda_{t+\hat{s}_j+D}^2} = \gamma_i\gamma_j\sum_{t=0}^{B-1}\Lambda_t^2.$$

Furthermore, equality is attained exactly when the two series are proportional, i.e. there is some α such that $\Lambda_{t+s_i} = \alpha\Lambda_{t+s_j+D}$ for all t. However, the sum of the Λ-values is 1 (on both sides), so α must be 1. This means that equality is achieved exactly when

$$\Lambda_{t+\hat{s}_i} = \Lambda_{t+\hat{s}_j+D} \quad \forall t \in \{0,1,...,B-1\}.$$

This holds when $D = \hat{s}_i - \hat{s}_j$, therefore it maximizes μ_D, as we wished to show. Moreover, if μ_D achieves this equality for any other value of D, it must hold that

$$\Lambda_t = \Lambda_{t+\hat{s}_j+D-\hat{s}_i} \quad \forall t \in \{0,1,...,B-1\}. \tag{8}$$

Since $\hat{s}_j + D - \hat{s}_i \neq 0 \bmod B$, it means that Λ is periodic.

B. Proof of Lemma 2

The correctness of the lemma follows as a result of common statistical inequalities, but it takes quite a bit of computation to see that. We first need to prove two other lemmas:

Lemma .1. *For any real θ, ϵ with $\theta \geq 0$ and $\epsilon > 0$, for a random variable $X \sim$ Pois(θ), it holds that*

$$Pr\left(|X-\theta| > \sqrt{2\theta\log\left(\frac{1}{\epsilon}\right)} + 2\log\left(\frac{1}{\epsilon}\right)\right) < \epsilon.$$

Lemma .2. *For any real $\theta_1, \theta_2, \epsilon$ with $\theta_1, \theta_2 \geq 0$ and $\epsilon > 0$, for random variables $X_1 \sim$ Pois(θ_1) and $X_2 \sim$ Pois(θ_2), if we denote*

$$L := 2\log\left(\frac{2}{\epsilon}\right),$$

it holds that

$$Pr\left(|X_1X_2 - \theta_1\theta_2| \geq 3\sqrt{L}\left(\theta_1^{\frac{3}{2}} + \theta_2^{\frac{3}{2}}\right) + 3L^2\right) < \epsilon.$$

We now present the proofs of all three lemmas.

Proof: [Proof of Lemma .1] We use the following standard tail inequality for Poisson variable, which holds for all positive α:

$$\Pr\left(|X-\theta| > \alpha\right) < e^{-\frac{\alpha^2}{2(\theta+\alpha)}}. \tag{9}$$

Let $\alpha := \sqrt{2\theta\log\left(\frac{1}{\epsilon}\right)} + 2\log\left(\frac{1}{\epsilon}\right)$.

It holds that

$$\begin{aligned}\alpha\left(\alpha - 2\log\left(\frac{1}{\epsilon}\right)\right) &> \left(\alpha - 2\log\left(\frac{1}{\epsilon}\right)\right)^2 \\ &= \sqrt{2\theta\log\left(\frac{1}{\epsilon}\right)}^2 = 2\theta\log\left(\frac{1}{\epsilon}\right) \\ \alpha^2 &> 2\theta\log\left(\frac{1}{\epsilon}\right) + 2\alpha\log\left(\frac{1}{\epsilon}\right) \\ \frac{\alpha^2}{2(\alpha+\theta)} &> \log\left(\frac{1}{\epsilon}\right).\end{aligned}$$

Plugging that into Equation (9), we derive

$$\Pr\left(|X-\theta| > \alpha\right) < e^{-\frac{\alpha^2}{2(\theta+\alpha)}} < e^{-\log\left(\frac{1}{\epsilon}\right)} = \epsilon,$$

as needed.

Proof: [Proof of Lemma .2] For $i \in \{1,2\}$, denote $\epsilon_i := \sqrt{\theta L} + L$. From Lemma .1, we know that for each $i \in \{1,2\}$, with probability greater than $1 - \frac{\epsilon}{2}$, it holds that $|X_i - \theta_i| < \epsilon_i$. This means that both inequalities hold with probability greater than $1-\epsilon$. In that case,

$$\begin{aligned}&|X_1X_2 - \theta_1\theta_2| \\ &= \left|\theta_1\left(X_2-\theta_2\right) + \theta_2\left(X_1-\theta_1\right) + \left(X_1-\theta_1\right)\left(X_2-\theta_2\right)\right| \\ &\leq \theta_1\epsilon_2 + \theta_2\epsilon_1 + \epsilon_1\epsilon_2.\end{aligned}$$

We bound this quantity using the inequality of arithmetic and geometric means:

$$\begin{aligned}&\theta_1\epsilon_2 + \theta_2\epsilon_1 + \epsilon_1\epsilon_2 \\ &\quad= \theta_1\sqrt{L}\left(\sqrt{\theta_2}+\sqrt{L}\right) + \theta_2\sqrt{L}\left(\sqrt{\theta_1}+\sqrt{L}\right) \\ &\qquad + L\left(\sqrt{\theta_1}+\sqrt{L}\right)\left(\sqrt{\theta_2}+\sqrt{L}\right) \\ &\quad= \sqrt{L}\left(\theta_1\sqrt{\theta_2} + \theta_2\sqrt{\theta_1}\right) + L\left(\theta_1+\theta_2\right) \\ &\qquad + L(\sqrt{\theta_1\theta_2} + \sqrt{\theta_1 L} + \sqrt{\theta_2 L} + L) \\ &\quad\leq \sqrt{L}\left(\theta_1^{\frac{3}{2}} + \theta_2^{\frac{3}{2}}\right) + L\left(\theta_1+\theta_2\right) + \\ &\qquad L\cdot\frac{\theta_1+\theta_2+\theta_1+L+\theta_2+L+L}{2} \\ &\quad= \sqrt{L}\left(\theta_1^{\frac{3}{2}} + \theta_2^{\frac{3}{2}}\right) + 2L\theta_1 + 2L\theta_2 + \frac{3}{2}L^2.\end{aligned} \tag{10}$$

We can leverage another inequality of arithmetic means, using the fact that

$$\begin{aligned}L\theta_i = \sqrt{L}\left(L^{\frac{3}{2}}\theta_i^{\frac{3}{2}}\theta_i^{\frac{3}{2}}\right)^{\frac{1}{3}} &\leq \sqrt{L}\frac{L^{\frac{3}{2}} + \theta_i^{\frac{3}{2}} + \theta_i^{\frac{3}{2}}}{3} \\ &= \frac{L}{3} + \frac{2\sqrt{L}\theta_i^{\frac{3}{2}}}{3}.\end{aligned} \tag{11}$$

Plugging that into Equation (10), we get that with probability greater than $1-\epsilon$, we have as needed the following

$$\begin{aligned}&|X_1X_2 - \theta_1\theta_2| \\ &< \sqrt{L}\left(\theta_1^{\frac{3}{2}} + \theta_2^{\frac{3}{2}}\right) + \frac{4\sqrt{L}\theta_i^{\frac{3}{2}}}{3} + \frac{4\sqrt{L}\theta_i^{\frac{3}{2}}}{3} + \frac{4L}{3} + \frac{3}{2}L^2 \\ &< 3\sqrt{L}\left(\theta_1^{\frac{3}{2}} + \theta_2^{\frac{3}{2}}\right) + 3L^2.\end{aligned}$$

We are now ready to present the proof for Lemma 2 based on Lemma .1 and Lemma .2. *Proof:* [Proof of Lemma 2] Use Lemma .2 on all pairs X_i, Y_i, with $\epsilon' := \frac{\epsilon}{B}$. We have $L = 2\log\left(\frac{2B}{\epsilon}\right) = 2\log\left(\frac{1}{\epsilon'}\right)$ and $M_i := 3\sqrt{L}\left(\theta_i^{\frac{3}{2}} + \nu_i^{\frac{3}{2}}\right) + 3L^2$. By that Lemma, we know that for each $i \in \{1, 2, ..., B-1\}$, with probability greater than $1-\epsilon'$,

$$|X_i Y_i - \theta_i \nu_i| \geq M.$$

This means that with probability greater $1 - \frac{\epsilon}{2}$, all of these inequalities hold. We consider the random variable $\hat{Z}$ which is Z conditioned on this event and use Hoeffding's inequality, to get that

$$\Pr\left(|\hat{Z} - \mu| > t\right) \leq \exp - \frac{2t^2}{\sum_i M_i^2}.$$

Looking closer at the denominator of the exponent, we get:

$$\begin{aligned}\sum_i M_i^2 &= \sum_i \left(3\sqrt{L}\left(\theta_i^{\frac{3}{2}} + \nu_i^{\frac{3}{2}}\right) + 3L^2\right)^2 \\ &\leq \sum_i \left(9L\left(\theta_i^3 + \nu_i^3\right) + 9L^4\right) \\ &= 9BL^4 + 9L\left(\sum_i \theta_i^3 + \sum_i \nu_i^3\right)\end{aligned}$$

so the probability that $|\hat{Z} - \mu| > t$ is at most

$$\epsilon + e^{-\frac{2t^2}{9BL^4 + 9L(\sum_i \theta_i^3 + \sum_i \nu_i^3)}} \leq \epsilon + e^{-\frac{2t^2}{5BL^4 + 5L(\sum_i \theta_i^3 + \sum_i \nu_i^3)}},$$

as needed.

C. Proof of Theorem 1

Proof: The estimate s' we use for s is the Ordinary Least Squares solution $s' := (A^T A)^{-1} A^T d$ ($A^T A$ is an invertible linear transformation from U to itself since the graph is connected), so we get an error of

$$\begin{aligned}s - s' &= (s - (A^T A)^{-1} A^T (As - \beta)) \\ &= (A^T A)^{-1} A^T \beta.\end{aligned}$$

Denote $Q := (A^T A)^{-1} A^T$. The error is $Q\beta$, so we wish to investigate the typical size of $Q\beta$ for normal β. The expected squared error is therefore

$$\mathbb{E}\left[< Q\beta, Q\beta >\right] = \mathbb{E}\left[< Q^T Q\beta, \beta >\right].$$

Since $Q^T Q$ is symmetric, we may diagonalize it with an orthonormal basis of eigenvectors v_i corresponding to eigenvalues λ_i, and write β in that basis, which gives that the expected inner product above is equal to

$$\sum_i \lambda_i \mathbb{E}\left[< \beta, v_i >^2\right] = \sigma^2 \sum_i \lambda_i \tag{12}$$

where λ_i are the eigenvalues corresponding to v_i, i.e. the eigenvalues of $Q^T Q$. The expected root mean squared error is therefore

$$\mathbb{E}\left[\sqrt{\frac{||s - s'||^2}{m}}\right] \leq \sqrt{\frac{\mathbb{E}\left[< Q\beta, Q\beta >\right]}{m}} \tag{13}$$

$$= \frac{1}{\sqrt{m}}\sigma\sqrt{\sum_i \lambda_i}, \tag{14}$$

where the first inequality is Jensen's inequality applied on the concave function $x \to \sqrt{x}$. Note that the nonzero eigenvalues of $Q^T Q$ are the same as those of QQ^T, which may be rewritten as

$$QQ^T = (A^T A)^{-1} A^T A (A^T A)^{-1} = (A^T A)^{-1}.$$

If we considered the transformation $A^T A$ on R^V (rather than R_0^V), we would see that the matrix $A^T A$ is a matrix of special interest - it contains, on its diagonal, the degrees of the nodes of the graph, and for every pair of days i, j, it contains either -1 on location (i, j) an edge connects them and 0 otherwise. This matrix is often called the *Laplacian* of the graph, and its spectral decomposition is a much-studied object in Spectral Graph Theory. Note that the smallest eigenvalue of the Laplacian is 0 with eigenvector $(1, 1, ..., 1)$ (see for instance [12]), but we restrict it to $\mathbb{R}_0^V$, which is the perpendicular space to its 0-eigenspace, which means that its smallest eigenvalue on this subspace is the second smallest eigenvalue of the original Laplacian, $\hat{\lambda}_2 = \lambda_2(A^T A)$. This quantity is often referred to as the *algebraic connectivity* of the graph G. Particularly, from [13, Theorem 1.4], we know that for a random graph with m vertices where each pair of vertices has an edge with probability p independently, if $p = \omega\left(\frac{log(m)}{m}\right)$, then we get that $\hat{\lambda}_2 = mp + O(\sqrt{mp\log(m)})$ with high probability. In other words, we can choose H, m_0 to be large enough so that we get that if $p > H\frac{\log(m)}{m}$, then with probability at least $1 - \frac{\epsilon}{2}$,

$$\begin{aligned}\hat{\lambda}_2 &\geq mp - \sqrt{H}\sqrt{mp\log(m)} \\ &\geq \left(1 - \frac{1}{\sqrt{H}}\right) mp \geq \frac{mp}{1 + \frac{2}{\sqrt{H}}}.\end{aligned}$$

From Equation (13), we learn that the overall root mean squared error in our estimate will be in expectation at most $\sigma\sqrt{\sum_i \lambda_i}$ for λ_i the eigenvalues of $(A^T A)^{-1}$. These are the inverses of the eigenvalues of the Laplacian $A^T A$, which we know are all at least $(1 + o(1))mp$. So the expected root mean squared error is at most

$$\left(1 + \frac{2}{\sqrt{H}}\right)\frac{1}{\sqrt{m}}\sigma\sqrt{\sum_i \frac{1}{mp}}$$

$$= \left(1 + \frac{2}{\sqrt{H}}\right)\sqrt{H}\frac{1}{\sqrt{m}}\sigma\sqrt{\frac{1}{p}} = \left(1 + \frac{2}{\sqrt{H}}\right)\frac{\sigma}{\sqrt{mp}}.$$

This is a bound for the expectation of the root mean squared error. From Markov's inequality, we know that with probability at least $1 - \frac{\epsilon}{2}$, we have that the root mean squared error is at most

$$\frac{2}{\epsilon}\left(1 + \frac{2}{\sqrt{H}}\right)\frac{\sigma}{\sqrt{mp}}.$$